EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

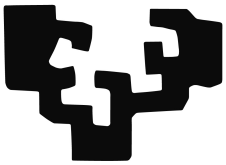
Doktoretza tesia

Integrating Outside Knowledge and Spatial Reasoning in Vision-and-language Models

Ander Salaberria Saizar

2024

eman ta zabal zazu



EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Integrating Outside Knowledge and Spatial Reasoning in Vision-and-language Models

Ander Salaberria Saizarrek Eneko Agirre eta Gorka Azkuneren zuzendaritzapean eginiko tesi txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia, 2024ko Iraila.

The difference between us and a computer
is that the computer is blindingly stupid,
but is capable of being stupid many,
many million times a second.

— Douglas Adams

Eskerrak

Lehenik eta behin, nire eskerrik beroenak tesi honen zuzendariei, Eneko eta Gorkari, etapa honetako pauso guztietan gidatu izanagatik, zuen ekarpen guztiengatik eta nirekin pazientzia izateagatik. Momentu gozo eta latzenetan hor egon zarete eta zuengandik ikasi dudanak ez dauka preziorik. Zuzendariekin batera, mila esker Oier eta Aitorri astero lan honen garapenari tarte bat eskaintzeagatik. Paregabeak izan zarete!

Ixakide guztiei, lan orduak arinagoak bihurtzeagatik eta eskua emateko beti prest egoteagatik. Lagun ederrak topatu ditut talde honetan, behin eta berriro topatu ditudanak eguneroko hamaiketako eta bazkal ordutan, (gehienbat Matia kaleko) tabernetan, sagardotegietan, asteburuko eskapadetan, rokodromoan (ez nuen espero bizitzan bat zapalduko nuenik...) etab. Mila esker esperientzi guzti horiengatik, ziur nago gehiagotan parte hartzeko irrika ez dudala galduko.

318ko tropari, nire eguneroko lana arinagoa bihurtzeagatik. Bulegoa berotu ez ezik, asko lagundu duzue taldean ni integratua sentiaraztearekin. 3D inprimagailua berriz martxan jartzea zaila iruditzen zait, baina alternatibaren bat bilatu beharko dugu... 314koek krispeta-festa klandestinoak zituzten eta 313ko etorri berriek baratze bat dute jada... ezin gara atzean gelditu!

Olia, Anar eta Iñigori, tesiak dakartzan alde ilunak argiztatzeagatik eta kanpoan ondo gosaltzeari beste esanahi bat emateagatik.

I am also quite grateful to Frank Keller. Thank you for allowing me to collaborate and hosting me at the University of Edinburgh. My stay there widened my view of how research can be carried out in different groups and it was an invaluable experience.

Eranskin dedikatu bat sortzen ez badut arazoak izango ditudanez familiarteko guztiak izendatzeko, agurtutzat eman zure burua (badakizu, familiako bazkari gehienetan bezala). Mila esker denoi zuen animo eta hitz goxoengatik. Batez ere, eskerrak nire gurasoei edozein gauzetarako beti laguntzeko prest egoteagatik,

baita ordenagailuaren aurrean zer demontre egiten ari naizen ulertzeko interesa erakustegatik ere. Tesia bukatu ondoren ez dut aitzakiarik izango Goienetxetik gehiagotan ez pasatzeko!

Ezagutzen nauzuenek badakizue beti eskertzen ditudala mahai jokoez inguratutako solasaldiak eta behar baino aldagai ezezagun gehiago dituzten bidaiak. Eskerrik asko horietan parte hartu duzuen guztiei.

Azkenik, eskerrak eman nahi dizkiet bereziki: Eneritzeri, tesiaren gora-beheretan lehenengo filan egoteagatik; Albari, mundua ikusteko eta bizitzeko era ezberdinak daudela erakusteagatik; eta Maiteri, bai Edinburgon eta baita Errenterin ere gida gisa jokatzeagatik... ez ahaztu toaila!

Esker instituzionalak

Eusko Jaurlaritzako Hezkuntza Sailari, ikerketa-lan hau egiteko emandako ikertzaileak prestatzeko bekarengatik, baita egonaldi internazionala burutzeko dirulaguntzarengatik ere.

Abstract

The fields of natural language processing (NLP) and computer vision (CV) have lately emerged thanks to recent advancements in computational power, data quantity, and an evergrowing research community. The bridge between NLP and CV has also advanced, particularly in tasks requiring the grounding of textual and visual modalities, such as visual question-answering (VQA) and text-to-image generation. This paves the way for more sophisticated systems and applications across various domains. Nevertheless, these systems still face weaknesses that have no trivial solution.

The goal of this thesis is to explore two limitations of current Vision-and-language models (VLMs): world knowledge integration and spatial reasoning. This dissertation can be divided into two main parts, one for each limitation that we tackle. In the first part, we verbalize images to better leverage world knowledge that is implicitly encoded in language models. In contrast, in the second, we exploit the generation of synthetic data from object annotations to aid the spatial reasoning of both language models and text-to-image generators.

More in-depth, visio-linguistic tasks, such as VQA, usually need to reason over an image by integrating world knowledge. As previous work has shown that pre-trained language models encode this knowledge, we propose an unimodal (text-only) approach by generating captions from images automatically and discarding the image from the rest of the inference. We show that using only textual representations to encode the language model's input is especially effective for VQA tasks requiring external knowledge. In addition, we show that our unimodal approach outperforms VLMs of a comparable number of parameters, while we also observe that both approaches are complementary regardless of the need for world knowledge. Our qualitative analysis reveals that automatic captions often fail to capture the information needed to answer the prompted question, which seems to be balanced by the better inference ability of our unimodal model.

Entering the field of spatial reasoning, we show that text-only language models can learn to ground spatial relations (*left of* or *below*) if they are provided with explicit object locations and they are properly trained to leverage them. We feed this spatial knowledge by using location tokens that represent bounding box information, which are extracted using an off-the-shelf object detector. In order to learn how to link each spatial relation to different sets of location tokens, we define simple heuristics that specify whether a given relation is fulfilled or not, and we use that signal to build a synthetic dataset and fine-tune language models. By doing so, we set the new state-of-the-art for the VSR dataset, even improving the performance of VLMs. Our analysis shows that our text-only language models can generalize beyond the relations seen during training to some extent, learning also more useful information than that encoded in the heuristics mentioned earlier.

We also tackle the task of text-to-image generation by following a similar approach. We hypothesize that current systems do not accurately depict spatial relations in generated images due to the lack of them in the training data. Therefore, we introduce the Spatial Relation for Generation (SR4G) dataset, which contains: synthetic captions composed of 14 different explicit spatial relations, 9 million image-caption pairs for training, and more than 60 thousand captions for evaluation. We also provide an unseen split in order to test generalization, with different sets of objects used during training, development and testing. We show that fine-tuning two different Stable Diffusion models (denoted as SD_{SR4G}) yields significant improvements in the VISOR metric, an evaluation metric specifically designed to check whether an image contains a specific spatial relation or not. The improvement holds in the unseen split, showing that SD_{SR4G} is able to generalize to unseen objects. This way, we improve the state-of-the-art with fewer parameters and avoid complex architectures involving layout generation and large language models.

Laburpena

Hizkuntza naturalaren prozesamendua (NLP) eta konputagailu bidezko ikusmenaren (CV) alorrak asko hazi dira azkenaldian. Bultzada hau ordenagailuen kalkuluahalmen eta eskuragarri dagoen datu kopuruaren hazkundeari esker lortu da, baita etengabe hazten ari den ikerketa-komunitateari esker ere. NLP eta CV-ren arteko zubian aurrerapenak lortu dira ere bai, batez ere testu eta ikusmen modalitateen oinarritzea eskatzen duten zereginetan, hala nola, ikusizko galdera-erantzute (VQA) eta testuan baldintzatutako irudi sorkuntza. Horrek sistema eta aplikazio sofistikatuagoetarako bidea zabaltzen du hainbat domeinutan. Dena den, sistema hauek konponbide errazik ez dituzten ahuleziak dituzte oraindik.

Tesi honen helburua egungo ikusizko hizkuntza-ereduen (VLM) bi ahulezi aztertzea da: munduko ezagutzaren integrazioa eta arrazoinamendu espaziala. Tesi hau bi zati nagusitan bana daiteke, jorratzen dugun ahulezi bakoitzeko bana alegia. Lehenengo zatian, irudietatik goiburukoak sortzen ditugu hizkuntza-ereduetan inplizituki kodetuta dagoen munduko ezagutza hobeto aprobetxatzeko. Bigarrenean, aldiz, objektu anotazioetatik datu sintetikoak sortzen zentratu gara arrazoinamendu espazialaren ikasketari laguntzeko, bai hizkuntza-ereduetan eta baita testu bidezko irudi sortzaileetan ere.

Gehiago sakonduz, VQA bezalako ikusmen-testu atazetan ohikoa da irudi baten gaineko arrazoinamendua burutzea munduko ezagutza integratuz. Hizkuntza-eredu aurrentrenatuek ezagutza hau kodetzen dutela erakutsi denez, modalitate bakarra (testua soilik) erabiltzea proposatzen dugu, irudietatik goiburukoak automatikoki sortuz eta irudi bera gainerako inferentzietatik baztertuz. Hizkuntza-ereduaren sarrera kodetzeko testua soilik erabiltzea bereziki eraginkorra dela erakusten dugu munduko ezagutza eskatzen duten VQA atazetarako. Horrez gain, gure hurbilpen unimodalak pareko parametro kopuruak dituzten VLM-ak gainditzen dituela erakusten dugu. Bi aldaera hauek osagarriak direla antzeman dugu, munduko ezagutza beharrak dituzten VQA atazekin eta ezagutza behar hori gabe-

koekin ere bai. Gure analisi kualitatiboak goiburuko automatikoek galdera erantzuteko behar den informazioa sarritan ez dutela jasotzen agerrarazten du. Hala ere, gabezi hau inferentzia hobeagoak egiteko kapazitatearekin orekatzen dela dirudi.

Arrazoinamendu espazialaren alorrean sartuz, testua soilik jasotzen duten hizkuntza-ereduek erlazio espazialak (*ezkerrean* edo *azpian*) oinarritzen ikas ditzaketela erakutsi dugu. Ikasketa hau burutzeko ezinbestekoa da objektuen kokapen esplizituak ereduari ematea eta behar bezala prozesatzen ikasteko atazak erabiltzea. Gure kasuan, ezagutza espazial hori objektuen kaxa inguratzaileen informazioa kodetuz lortzen dugu token berezi batzuk erabiliz, hots, kokapen-tokenak. Kokapen-token hauek publikoki eskuragarri dagoen objektu detektore bat erabiliz eskuratzen ditugu. Erlazio espazial bakoitza kokapen-token multzoekin lotzen ikasteko, erlazio jakin bat betetzen den ala ez zehazten duten erregela sinpleak definitzen ditugu. Erregela hauekin datu-multzo sintetiko bat eraiki dezakegu eta hizkuntza-ereduak doitu. Horrela, VSR datu-multzoaren artearen egoera ezarri dugu, VLM-en errendimendua hobetuz. Gure analisiak testua soilik erabiltzen duten hizkuntza-ereduak entrenamenduan zehar ikusitako erlazioetatik haratago orokortu dezaketela erakusten du hein batean, lehen aipatutako erregeletan kodetutakoa baino informazio baliagarriagoa ere ikasiz.

Testu bidezko irudi sorkuntza atazari ere aurre egiten diogu antzeko hurbilpen bat jarraituz. Artearen egoerak ez ditu erlazio espazial esplizituak ondo irudikatzen eta, gure ustez, entrenamenduan erabiltzen diren datu-multzoetan hauen agerpena urria delako. Hori dela eta, *Spatial Relations for Generation* edo SR4G datu-multzoa aurkezten dugu. SR4G-ek 14 erlazio espazial esplizitu ezberdinez osatutako goiburuko sintetikoak ditu, 9 milioi irudi-goiburuko pare definituz entrenamendurako eta 60K goiburuko baino gehiago ebaluatzeko. Gainera, datu-multzoaren *unseen* bertsio bat definitu dugu, goiburukoetan objektu ezberdinak zehazten direlarik entrenamendu, garapen eta ebaluazio azpimultzoetan. Stable Diffusion ereduak SR4G datu-multzoan doitzeak (SD_{SR4G}) hobekuntza nabarmenak ematen ditu VISOR metrikan, erlazio espazialak irudietan betetzen diren ala ez neurtzen duen ebaluazio metrika automatikoa dena. *Unseen* bertsioan hobekuntzak mantentzen dira, SD_{SR4G} eredu doituak ikusten ez dituen objektuetara orokortzeko gai dela erakutsiz. Horrela, artearen egoera hobetzen dugu parametro gutxiago erabiliz eta arkitektura konplexuak saihestuz.

Gaien aurkibidea

Taulen zerrenda

Irudien zerrenda

1. CHAPTER

Introduction

This thesis belongs to the intersection of two academic fields: natural language processing (NLP) and computer vision (CV). Both fields aim to endow machines with human capabilities. In other words, while NLP enables language processing and generation, CV allows machines to mimic sight and understand their surroundings. This way, their intersection becomes intuitive, as people often reason and talk about what they see. Nevertheless, the effective integration between vision and language by machines has historically been a challenge, as grounding two different modalities, such as images and text, is still an open problem that has mainly been mitigated thanks to brute force (e.g. massive amounts of data and computational power).

This dissertation has been undertaken in the Ixa group inside the HiTZ research centre at the University of the Basque Country. HiTZ is considered the reference NLP research team in Spain and one of Europe's top NLP research centres. Since their beginnings, Ixa and HiTZ have been pioneers in developing NLP tools, paying special attention to the development of language tools for Basque. Furthermore, the group participates in worldwide-level research projects, contributing not only to Basque but also to many more languages.

Research in the Ixa group has mostly focused on text-only tasks until recently, as current trends in multimodality have sparked new ideas to mix different modalities in the field of NLP. That is the case of this dissertation, where we dive into the limitations that current state-of-the-art vision-and-language models have, and explore different approaches to solve them.

1.1 Motivation

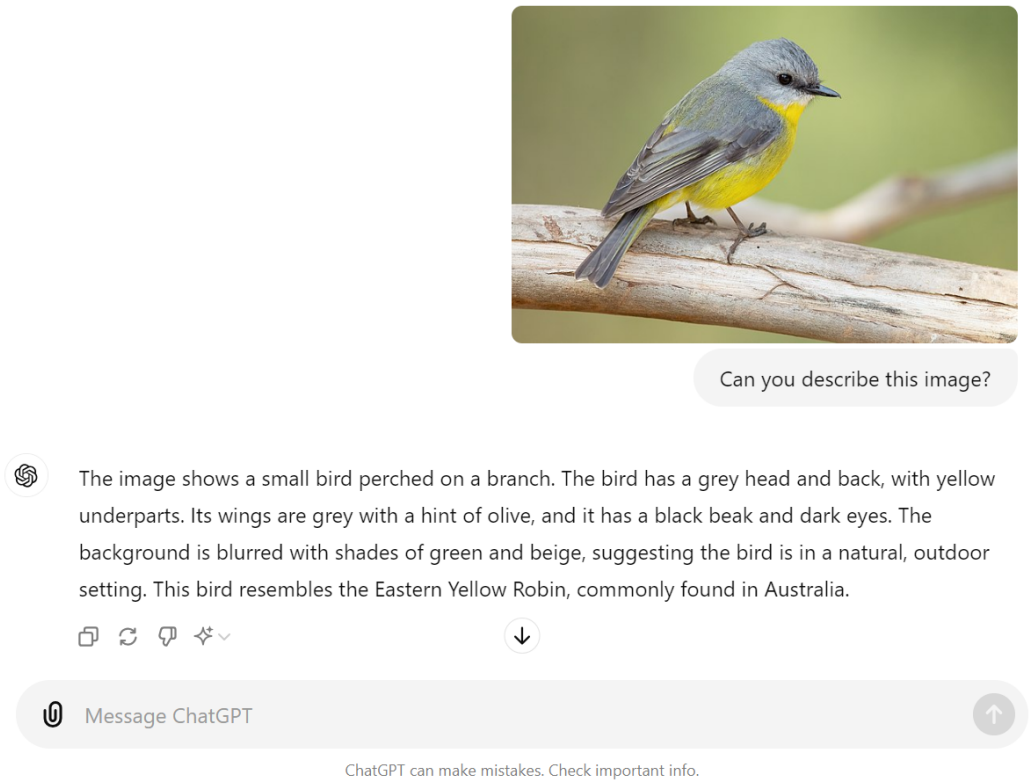
The recent uproar in NLP has been achieved thanks to the emergence of language models (LMs) and later on large language models (LLMs), gargantuan statistical models that, due to their capacity, can achieve general-purpose language generation. Their capabilities of generating grammatically accurate and verbose text with an apparent understanding of its semantics have convinced many that natural language understanding (NLU) and even artificial general intelligence (AGI) are near. This overreaction is fueled by the tendency of humans to add meaning and intent where there is none (Bender *et al.* 2021), misleading NLP researchers and the general public when interacting with LLM-based applications such as ChatGPT and Claude.

One key aspect of achieving NLU or AGI is to link language with sensory data. As previously mentioned, humans communicate what they perceive in their surroundings, which shapes our language. Multimodality is key to grounding language in the real world, as it takes into account additional context beyond just text. Without this grounding, language is not directly linked to the physical world, disabling the model's ability to understand or convey the meaning of what is being said. The latest LLMs have started integrating different modalities, and visual signals are the most common due to the abundance of image-text pairs available on the internet (e.g. captions of photographs). Therefore, these vision-and-language models (VLMs) learn to solve visual reasoning, image captioning, and visual dialogue, among other tasks.

Figure 1.1 shows what state-of-the-art large VLMs like GPT-4o can do. In this case, the user asks for information about an image with a bird, and the model answers with a detailed description. This answer also includes world knowledge, as it correctly identifies the species of the bird (an Eastern Yellow Robin) and gives more information about its main habitat (Australia). In short, there is no inaccuracy in its response and the question has been successfully answered.

Even though GPT-4o's architecture, training, and inference procedures are unknown to the public, we can expect that it has seen the prompted image alongside text describing this bird during training, as this image was extracted from the Wikipedia page of Eastern Yellow Robins.[1] By ingesting massive amounts of crawled data from the Internet, these large models have memorized plenty of it. This makes their evaluation difficult, as discerning whether they memorized the generated text or are generalizing becomes harder. Nevertheless, they also tend

[1] https://en.wikipedia.org/wiki/Eastern_yellow_robin. (accessed on 01/07/2024)



1.1 Figure – Example of ChatGPT's visual dialog capabilities. This version of ChatGPT uses the multimodal GPT-4o model, accessed on 30/06/2024.

to hallucinate by generating plausible facts without linking them to evidence, a major limitation of current language models.

Large VLMs show good zero-shot and few-shot capabilities, like unimodal LLMs (Brown *et al.* 2020). However, both share limitations, such as the previously mentioned hallucinations. Even though VLMs show strong performance in several tasks, such as object detection and image captioning, they struggle with many others involving reasoning, knowledge retrieval, or compositional understanding. Research on these topics has recently emerged. For example, reasoning is being tackled with Chain-of-Thought approaches (Wei *et al.* 2022c), and Retrieval Augmented Generation has become a standard for knowledge retrieval with LLMs (Lewis *et al.* 2020; Gao *et al.* 2023).

It is worth mentioning that text-to-image generators also fall into the category of VLMs. In other words, VLMs do not only cover models that generate just text

1.2 Figure – Two examples of visual question-answering, where the goal is to answer a question about a given image. A VLM able to solve the question to the left needs to access knowledge about goldfinches. In contrast, locating and assessing the relative positions between the beak and the flower is necessary for the other.

but also images. As their name implies, they generate images conditioned on textual signals, which enables an intuitive way of generating visual representations by giving a short description. Dall-E and Midjourney are two examples of popular products that give this service. Therefore, grounding both modalities is needed for the correct functionality of text-to-image models. They also share similar weaknesses compared to other VLMs, including spatial reasoning and compositional understanding.

1.2 Goals and Research Lines

The main goal of this thesis is to explore the limitations of current VLMs and develop new approaches to tackle them. We focus on two limitations: i) *world knowledge integration*, where we explore different ways to better leverage the implicit knowledge found in language models, and ii) *spatial reasoning*, where the lack of grounded spatial relations in text corpora and vision-and-language datasets is fought by generating synthetic data from object annotations. Figure 1.2 shows visual question-answering (VQA) examples where these limitations are key to solving them correctly. More specifically, our research lines are the following:

[**RL1**] : **The leverage of implicit knowledge found in language models with different modalities.** In this research line we have analyzed the use of visual features and/or textual representations to represent and reason about an image on a VQA task with world knowledge needs. Existing state-of-the-art systems focus on mixing different modalities including text, images, graphs... to retrieve specific knowledge and give an answer.[2] Our approach has focused on better leveraging implicit knowledge encoded in the pre-trained weights of LMs, instead of retrieving this knowledge from text corpora or knowledge graphs.

[**RL2**] : **Creation of synthetic datasets to enhance spatial reasoning capabilities.** The lack of explicit spatial relations in text corpora used to pre-train VLMs hurts their capabilities to understand and perform several tasks correctly. We have aimed to use object annotations (e.g. object labels, attributes, and bounding boxes) and hand-crafted heuristics to generate synthetic data containing these relations. From this data, we have fine-tuned LMs and text-to-image generators to enhance their ability to reason with explicit spatial relations.

[**RL2.1**] : **Development of text-only language models that reason better with spatial relations.** In this research line we have improved spatial reasoning on text-only language models. By definition, unimodal language models learn statistical language patterns without grounding the text in the real world. We tackle this issue by verbalizing layout information of images via location tokens, pairing them with their respective spatial relations, and defining a training task to learn the bridge between spatial relations and location tokens.

[**RL2.2**] : **Development of text-to-image models that generate correct spatial relations more consistently.** Even if text-to-image generators use image-text pairs to learn the task properly, they struggle to consistently depict the spatial relations mentioned in the input caption. Assuming that the lack of image-text pairs containing these relations during their training process has a big role in this, we define a data generation procedure to fill the gap in data and fine-tune these models to show an improvement in the generation of spatial relations.

[2] We are referring to the state-of-the-art of the first half of this dissertation (2020-2022), during the development of this research line, as the advent of LLMs has changed this paradigm.

1.3 Structure of the Thesis

This dissertation is written in two different languages: English and Basque. The English written block is composed of Chapters 1, 2 and 6, whereas Chapters 3, 4 and 5 complete the Basque block.

English block: In Chapter 1, we begin by introducing the topic of this work and the motivation behind it (Section 1.1), following it with its goals and the explored research lines (Section 1.2). After discussing the structure of this document (Section 1.3), we list all the scientific contributions made during the duration of this thesis (Section 1.4). Subsequently, Chapter 2 dives into the related work and building block in which this thesis has been established (Section 2.1). We introduce VLMs and Diffusion Models (Section 2.2), and describe their current limitations regarding spatial reasoning and world knowledge integration (Section 2.3). Finally, Chapter 6 presents our main contributions, conclusions, and new lines of work that this thesis provides.

Basque block: This block contains the main work of this thesis, divided into 3 chapters. Chapter 3 provides our contributions regarding world knowledge integration by better leveraging the implicit knowledge found in language models on Visual Question-answering tasks. Then, in Chapters 4 and 5, we shift the topic to spatial reasoning, and utilize synthetic data to improve: i) spatial reasoning capabilities of text-only language models, and ii) the generation of explicit spatial relations in text-to-image generation. All these chapters follow the same structure. We start by elaborating on each work's motivation and contributions. Then, we focus on the methodology used to set up our experiments and carry them out with the corresponding analysis. We conclude each chapter by summarizing its main conclusions.

Note for non-Basque-speaking readers: The reader can check Appendix A to read the original papers/preprints containing all chapters of the Basque block in their recommended reading order.

1.4 List of Scientific Contributions

In this section, we present the scientific contributions developed during my PhD student years. This section is split into two parts. Firstly, we present the publications and preprints that build this manuscript. After that, we list the ones that are not part of it.

1.4.1 Contributions that are part of the thesis

[**A.1**] Salaberria et al. (ESWA 2023) presented in Chapter **3**

Salaberria A., Azkune G., Lopez de Lacalle O., Soroa A., and Agirre E. (2023). Image Captioning for Effective Use of Language Models in Knowledge Based Visual Question Answering. In *Expert Systems with Applications*.

Abstract: Integrating outside knowledge for reasoning in visio-linguistic tasks such as visual question-answering (VQA) is an open problem.[3] Given that pre-trained language models have been shown to include world knowledge, we propose to use an unimodal (text-only) training and inference procedure based on automatic off-the-shelf captioning of images and trained language models. More specifically, we verbalize the image contents and allow language models to better leverage their implicit knowledge to solve knowledge-intensive tasks. Focusing on a visual question-answering task which requires external knowledge (OK-VQA), our contributions are (i) a text-only model that outperforms pre-trained multimodal (image-text) models of a comparable number of parameters; (ii) confirmation that our text-only method is especially effective for tasks requiring external knowledge, as it is less effective in standard a VQA task (VQA 2.0); and (iii) our method attains results in the state-of-the-art when increasing the size of the language model. We also significantly outperform current multimodal systems, even though augmented with external knowledge. Our qualitative analysis of OK-VQA reveals that automatic captions often fail to capture relevant information in the images, which seems to be balanced by the better inference ability of the text-only language models. Our work opens up possibilities to further improve inference in visio-linguistic tasks.

[3] In this document we use outside/external/world knowledge interchangeably.

[**A.2**] Azkune et al. (NN 2024) presented in Chapter **4**

Azkune G., **Salaberria A.**, and Agirre E. (2024). Grounding Spatial Relations in Text-only Language Models. In *Neural Networks*.

Abstract: This paper shows that text-only Language Models (LM) can learn to ground spatial relations like *left of* or *below* if they are provided with explicit location information of objects and they are properly trained to leverage those locations. We perform experiments on a verbalized version of the Visual Spatial Reasoning (VSR) dataset, where images are coupled with textual statements which contain real or fake spatial relations between two objects of the image. We verbalize the images using an off-the-shelf object detector, adding location tokens to every object label to represent their bounding boxes in textual form. Given the small size of VSR, we do not observe any improvement when using locations, but pretraining the LM over a synthetic dataset automatically derived by us improves results significantly when using location tokens. We thus show that locations allow LMs to ground spatial relations, with our text-only LMs outperforming Vision-and-language Models and setting the new state-of-the-art for the VSR dataset. Our analysis shows that our text-only LMs can generalize beyond the relations seen in the synthetic dataset to some extent, learning also more useful information than that encoded in the spatial rules we used to create the synthetic dataset itself.

[**A.3**] Salaberria et al. (*Preprint* 2024) presented in Chapter **5**

Salaberria A., Azkune G., Lopez de Lacalle O., Soroa A., Agirre E., and Keller F. (2024). Improving Explicit Spatial Relationships in T2I through an Automatically Derived Dataset. *arXiv preprint arXiv:2403.00587*.

Abstract: Existing work has observed that current text-to-image systems do not accurately reflect explicit spatial relations between objects such as *left of* or *below*. We hypothesize that this is because explicit spatial relations rarely appear in the image captions used to train these models. We propose an automatic method that, given existing images, generates synthetic captions that contain 14 explicit spatial relations. We introduce the Spatial Relation for Generation (SR4G) dataset, which contains 9.9 million image-caption pairs for training, and more than 60 thousand captions for evaluation. In order to test generalization we also provide an *unseen* split, where the set of objects in the train and test captions are disjoint. SR4G is

the first dataset that can be used to spatially fine-tune text-to-image systems. We show that fine-tuning two different Stable Diffusion models (denoted as SD_{SR4G}) yields up to 9 points improvements in the VISOR metric. The improvement holds in the *unseen* split, showing that SD_{SR4G} is able to generalize to unseen objects. SD_{SR4G} improves the state-of-the-art with fewer parameters and avoids complex architectures. Our analysis shows that improvement is consistent for all relations.

1.4.2 Contributions outside the thesis

Lopez de Lacalle et al. (ECAI 2020)

Lopez de Lacalle O., **Salaberria A.**, Soroa A., Azkune G., and Agirre E. (2020). Evaluating multimodal representations on visual semantic textual similarity. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*.

Salaberria et al. (IkerGazte 2021)

Salaberria A., Campos J. A., García I., and Fernandez de Landa J. (2021). Itzulpen automatikoko sistemen analisia: Genero alborapenaren kasua. In *Fourth Conference for Basque Researchers (IkerGazte 2021)*.

Fernandez de Landa et al. (IkerGazte 2021)

Fernandez de Landa J., García I., **Salaberria A.**, and Campos J. A. (2021). Twitterreko Euskal Komunitatearen Eduki Azterketa Pandemia Garaian. In *Fourth Conference for Basque Researchers (IkerGazte 2021)*.

García-Ferrero et al. (SemEval 2023)

García-Ferrero I., Campos J. A., Sainz O., **Salaberria A.**, and Roth D. (2023). IXA/Cogcomp at SemEval-2023 Task 2: Context-enriched Multilingual Named Entity Recognition Using Knowledge Bases. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval 2023)*.

Agerri et al. (*Book chapter* 2023)

Agerri R., Agirre E., Aldabe I., Aranberri N., Arriola J. M., ..., **Salaberria A.**, ... and Soroa A.(2023). State-of-the-Art in Language Technology and Language-centric Artificial Intelligence.. In *European Language Equality: A Strategic Agenda for Digital Language Equality*.

Fernandez de Landa et al. (SIGUL 2024)

Fernandez de Landa J., García-Ferrero I., **Salaberria A.**, and Campos J. A. (2024). Uncovering Social Changes of the Basque Speaking Twitter Community during COVID-19 Pandemic. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*.

Miranda et al. (Under Review - NeurIPS 2024)

Miranda I., **Salaberria A.**, Agirre E., and Azkune G. (2024). BiVLC: Extending Vision-Language Compositionality Evaluation with Text-to-Image Retrieval. *arXiv preprint arXiv:2406.09952*.

2. CHAPTER

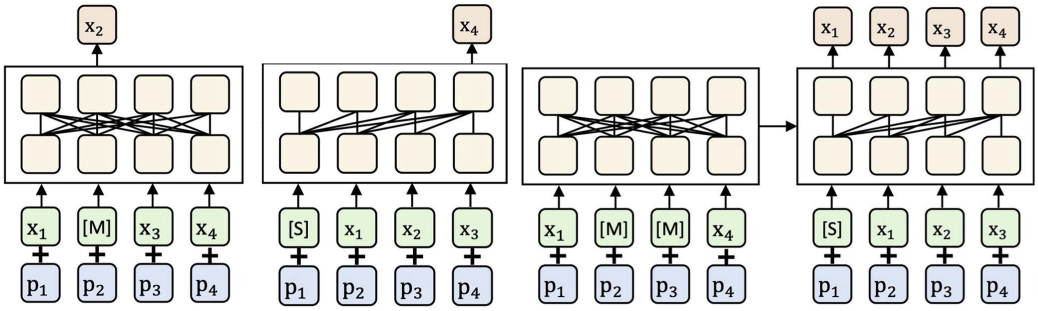
Background

This thesis explores two limitations of contemporary vision-and-language models (VLMs). Therefore, in this chapter we explain the research landscape in which this work is located, exploring relevant datasets and state-of-the-art systems for our work. Before entering the domain of VLMs, we briefly introduce their building blocks (Section 2.1) to discuss then the state-of-the-art of two different kinds of VLMs (Section 2.2): language models with visual components and text-to-image generators. Finally, we explore the recent advances in world knowledge integration and spatial reasoning (Section 2.3).

Before starting with the related work, it is noteworthy to mention that the state-of-the-art has changed significantly since the start of this thesis. During this chapter, we have distinguished between the state-of-the-art at the time of development of this work and the research landscape at the time of writing this manuscript.

2.1 Unimodal Systems

The term vision-and-language model (VLM) has been extensively used in the literature for transformer-based models adapted to process visual and textual data. Even though some VLMs have been trained to process both modalities from scratch (Wang *et al.* 2022a; Driess *et al.* 2023), most have adapted unimodal systems and use them as building blocks for VLMs (Li *et al.* 2019: 2020; Ramesh *et al.* 2022; Li *et al.* 2023a), which include language models and visual encoders.



(a) Encoder-only **(b)** Decoder-only **(c)** Encoder-decoder

2.1 Figure – Transformer architectures are divided into three main categories: encoder-only, decoder-only and encoder-decoder. Source figures from Wang *et al.* (2023).

Language Models

Language modelling is a well-established research topic that centres on creating probabilistic models of natural language. The landscape has drastically changed since the initial approach of Shannon (1951) with n-gram language models. Neural language models emerged as robust alternatives to n-gram models (Bengio *et al.* 2000) and task-specific variants of recurrent neural networks (RNNs) were popularly used for many NLP tasks over the last decade (Mikolov *et al.* 2010; Cho *et al.* 2014; Sutskever *et al.* 2014).

The adoption of two key concepts defined the starting landscape of this thesis. On the one hand, the pre-training and fine-tuning paradigm allowed language models to learn the underlying patterns and semantic knowledge present in unlabelled text corpora. This allows us to fine-tune these pre-trained task-agnostic models to downstream tasks (Peters *et al.* 2018). On the other hand, Vaswani *et al.* (2017) introduced the transformer architecture. It employs self-attention mechanisms to calculate an attention score for every token (or text unit) in text sequences, effectively modelling the influence each word has on the others in parallel. This parallelization capability greatly surpasses RNNs, enabling the efficient pre-training of language models on massive datasets using several GPUs.

The transformer architecture was originally defined as an encoder-decoder architecture, as its initial goal was to tackle machine translation, a sequence-to-sequence downstream task. Even though the following transformer models used the pre-train and fine-tune paradigm, they employ different neural architectures depending on the downstream tasks they are meant to solve.

- Encoder-only: These models only use the encoder part of Vaswani *et al.* (2017) the transformer architecture. This way, attention layers can access all tokens of the initial sentence (see Figure 2.1a). Pre-training typically involves corrupting a sentence (e.g., by masking random words), challenging the model to find or reconstruct the original sentence, as well as predicting whether two given sentences come one after the other or not. A special classification token prepended to the input called $[CLS]$ with a classification layer on top is used for this prediction task. Encoder models excel in tasks requiring a comprehensive understanding of the entire sequence, such as sentence classification, named entity recognition, and question-answering. Devlin *et al.* (2019) proposed BERT, which established the norm of using encoder-only models for language understanding tasks (Liu *et al.* 2019; Yang *et al.* 2019; He *et al.* 2020).

- Decoder-only: For any given token, the attention layers of these models can only access the words positioned before it in the sentence (see Figure 2.1b). These auto-regressive models are typically pre-trained by predicting the next token in the sequence. Decoder-only models like GPT (Brown *et al.* 2020) and LLAMA (Touvron *et al.* 2023) are particularly well-suited for text generation tasks.

- Encoder-decoder: The original transformer architecture follows this encoder-decoder approach. In this approach, the decoder attends to all the previously generated tokens to generate the next token, as well as the final representations of the context previously fed to the encoder (see Figure 2.1c). Raffel *et al.* (2020) showed with their T5 models that almost all NLP tasks can be cast as a sequence-to-sequence generation task. Thus, an encoder-decoder language model is a unified model that can perform every natural language understanding and generation task.

Recent advancements with language models pre-trained on extensive textual corpora have significantly enhanced performance in downstream NLP tasks. In addition to acquiring linguistic knowledge, these models also store world knowledge embedded in the training data (Petroni *et al.* 2019), which we call implicit knowledge. Not only that, the size of these models is also correlated with their capacity to store knowledge and adapt to new tasks (Kaplan *et al.* 2020). This fact pushed researchers to increase the capacity of these models, increasing their number of trainable parameters up to four orders of magnitude compared to BERT's base model (Achiam *et al.* 2023).

Large language models (LLMs) showcase emerging abilities that smaller models do not have (Wei *et al.* 2022b), including in-context learning capabilities that remove the necessity to do task-specific fine-tunings (Brown *et al.* 2020). Examples of general-purpose LLM families include GPT (Brown *et al.* 2020), LLAMA (Touvron *et al.* 2023) and PALM (Chowdhery *et al.* 2023).

Visual Encoders

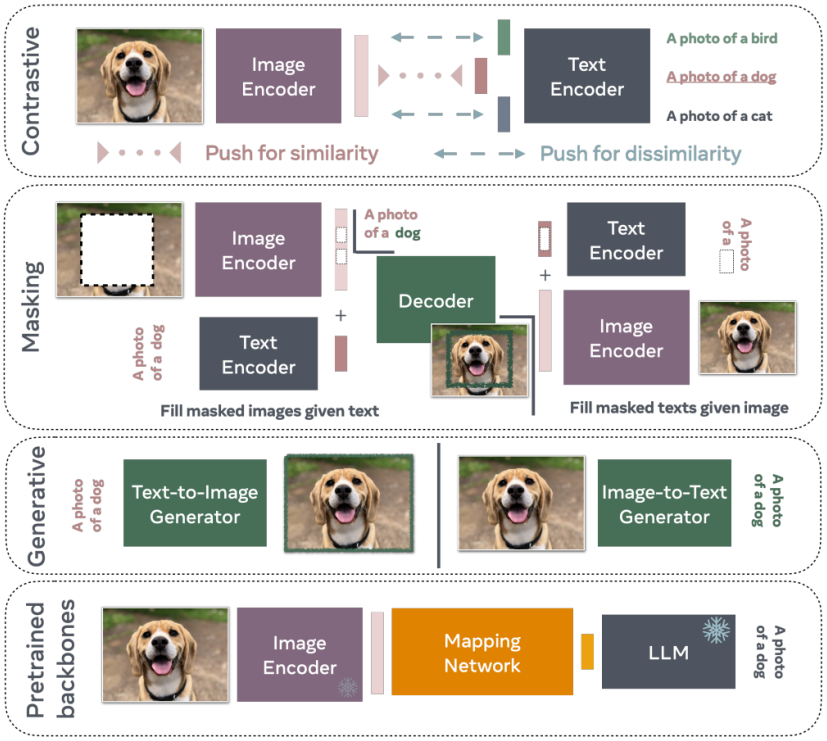
When RNNs started gaining traction in the field of NLP, a similar phenomenon occurred with convolutional neural networks (CNNs) within the field of computer vision. With the advent of deep learning, Krizhevsky *et al.* (2012) popularized the use of deep CNNs for image recognition tasks by significantly outperforming previous state-of-the-art, and Simonyan and Zisserman (2015) showed the relevance of the capacity and depth of these CNNs. Their initial vanishing gradient issues were addressed by adding skip connections between convolutional layers (He *et al.* 2016; Xie *et al.* 2017; Szegedy *et al.* 2017), settling CNNs as the defacto choice for encoding visual representations until recently.

The transformer architecture is not tied to process just text. Dosovitskiy *et al.* (2020) adapted an encoder-only transformer for image recognition tasks, by dividing images into non-overlapping patches, reshaping them into one-dimensional embeddings that the encoder can work with, and adding a classifier head on top of the encoder. CNNs and transformer-based models show an overall similar performance in image recognition tasks, and, nowadays, both approaches are still being used. Nevertheless, we are interested in the visual representations these classifiers learn before feeding them to their classification heads, as these one-dimensional embeddings encode relevant visual information that can be fed to VLMs.

Finally, these visual encoders are also used to build the backbones of object detectors, which we will employ when working on spatial reasoning, as they detect and, therefore, locate objects in a given image. Faster R-CNN is an example of a CNN-based object detector (Ren *et al.* 2015), whereas DETR is a transformer-based alternative (Carion *et al.* 2020).

2.2 Vision-and-language Models

Vision-and-language models encompass a variety of models that leverage visual and textual data. These models can be divided into 4 groups (see Figure 2.2). In this dissertation, we use or compare models belonging to all groups, which are:



2.2 Figure – Different families of VLMs: contrastive, masked, generative, and VLMs with pre-trained backbones. The first two families have inputs on both sides with outputs in the middle, while in the other cases, inputs are on the left and outputs on the right. Source figure from Bordes *et al.* (2024).

- Contrastive VLMs: They are dual encoders that learn to project visual and textual representations into the same multimodal space.
- VLMs with masking objectives: Similar to encoder-only language models, they use self-supervision by learning to recover masked image regions and text tokens with multimodal context (e.g. image-caption pairs).
- Generative VLMs: These models generate images, captions, or both. They can condition their output on different input modalities as well.
- VLMs with pre-trained backbones: They adapt LLMs to vision by learning to map visual features encoded by pre-trained image encoders to the LLM. As LLMs are kept frozen during adaptation, they maintain their emerging abilities, such as in-context learning.

Many approaches can be considered in more than one of these groups. For example, VLMs with pre-trained backbones are generative by definition, as their LLM backbones generate texts. It is also true that text-to-image generators are rarely called VLMs in the literature, even though they leverage both vision and language and rely on grounding both modalities.

Considering the work done in this dissertation, we divide this section into three subsections. First, we focus on vision-and-language encoders. These encoders can be divided into two groups mentioned in Figure 2.2: i) VLMs with masking objectives, which are the state-of-the-art systems contemporary to our work presented in Chapters 3 and 4, and ii) contrastive VLMs, which are used as building blocks in some of the models introduced in Chapter 5. Then, we deepen into the current state-of-the-art composed of generative models that are either native large VLMs (trained from scratch) or VLMs based on large pre-trained backbones. Finally, we specifically discuss text-to-image generators, as we work with their spatial reasoning capabilities in Chapter 5.

2.2.1 Vision-and-language Encoders

Following the trend of transformer-based language models, the first transformer-based VLMs had primarily encoder-only architectures (Li *et al.* 2019; Lu *et al.* 2019; Radford *et al.* 2021). Their language understanding capabilities initially outshined the text generation abilities of early transformer decoders, which pushed the use of encoder-only language models and VLMs. These vision-and-language encoders use image-caption pairs to learn grounded latent spaces of both modalities by training with one of the following approaches: masking objectives or contrastive learning.

Use of masking objectives

This training paradigm is a natural progression of encoder-only language models. As mentioned in Section 2.1, language encoders learn by reconstructing input sentences with random masked words, and predicting if two sentences come one after the other or not. These objectives can be tweaked for VLMs that work with image-caption pairs. Thus, we can learn to align regions of images with the caption by partially masking the caption (Masked Language Modeling or MaskLM) and predicting whether the image corresponds to the caption (Image-Text Matching or ITM).

For example, VisualBERT (Li *et al.* 2019) employs a pre-trained object detection model, Faster R-CNN (Ren *et al.* 2015) to identify and represent objects of an image. Then, it feeds those representations to a BERT-like language model and defines MaskLM and ITM objectives to implicitly align components of a caption with corresponding regions of an image using the self-attention layers of the language model encoder. More recent approaches propose alternative masking objectives, such as PrefixLM (Wang *et al.* 2021) and Masked Image Modeling (Assran *et al.* 2023), an analogous objective of MaskLM whose goal is to reconstruct masked region features of images.

In early encoder-only VLMs, there was no consensus on several architectural decisions. For instance, Li *et al.* (2019) and Su *et al.* (2020) opt for single-stream approaches, that is, they encode both modalities on the same module and self-attention layers are applied to all modalities at the same time. Meanwhile, Lu *et al.* (2019) and Tan and Bansal (2019) establish dual-stream encoders, where each modality is encoded separately and the mapping between modalities is mainly computed in cross-attention layers. Comparing contemporary systems it seems that using single-stream or dual-stream backbones does not affect their performance much (Li *et al.* 2019; Su *et al.* 2020; Lu *et al.* 2019; Tan and Bansal 2019). Therefore, single-stream transformer encoders became more popular due to their simpler architecture and the tendency to create VLMs by adapting text-only language models, which are single-stream by definition. Another relevant decision was to use grid representations of images instead of object regions (Jiang *et al.* 2020; Kim *et al.* 2021), as grid representations offer similar performance and remove dependencies with pre-trained object detectors.

Instead, architectural choices including model capacity, pre-training objectives, and datasets used are key factors for their good performance. Popular pre-training datasets used to train VLMs with masking objectives include: Conceptual Captions (Sharma *et al.* 2018), MS-COCO (Lin *et al.* 2014), SBU (Ordonez *et al.* 2011), and Visual Genome (Krishna *et al.* 2017).

VLMs with masking objectives constitute the state-of-the-art contemporary to the research presented in Chapters 3 and 4, as, at the time of development, they offered the strongest performance on visual question-answering and spatial reasoning, respectively.

Contrastive learning

Contrastive learning consists of mapping input images and texts into a multimodal feature space using two unimodal encoders. Given a batch of N image-text pair

representations $\{(\mathbf{v}_i, \mathbf{t}_i) : \text{where } i \in [1, ..., N]\}$, this alignment follows the InfoNCE loss (Oord *et al.* 2018), minimizing the distance between embeddings of matching image-text pairs and maximizing the rest (see Equation 2.1).

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{i=1}^{N} \log \frac{e^{(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}}{\sum_{j=1}^{N} e^{(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)}} \tag{2.1}$$

Models like CLIP (Radford *et al.* 2021) and ALIGN (Jia *et al.* 2021) use the cosine distance as the similarity metric between text and image embeddings defined in Equation 2.1, where τ is a learnable temperature parameter. Another approach, LiT (Zhai *et al.* 2022b), proposes a method to fine-tune the text encoder using the same training loss while maintaining the image encoder frozen. This technique aims to enhance the text encoder's ability to interpret image embeddings from the image encoder. Other methods, such as FLAVA (Singh *et al.* 2022), combine contrastive learning with additional pre-training strategies (including masking objectives) to align vision and language embeddings effectively.

These contrastive models need millions of image-caption pairs to learn rich representations. In the literature, contrastive models usually use larger training datasets than VLMs with masking objectives of comparable size, at least an order of magnitude larger (Li *et al.* 2019; Radford *et al.* 2021). Moreover, the batch size used during training also conditions the quality of the learned multimodal space. A high batch size implies more negative image-caption pairs in each training step, which allows for more diverse and difficult negative samples. The need for higher batch sizes has pushed the use of transformer-based visual encoders (e.g. ViT models) instead of CNN-based encoders (e.g. ResNet models), as the former ones are more efficient for training and offer similar performance (Radford *et al.* 2021).

Even though early transformer-based models like CLIP and ALIGN were trained on private vision-and-language datasets with millions of instances, researchers can now train their models using public open-source datasets like PMD (Singh *et al.* 2022) and LAION (Schuhmann *et al.* 2022), which were used to train FLAVA and the open-source version of CLIP (Cherti *et al.* 2023), respectively.

In our work, we use CLIP models in two use cases described in Chapter 5. On the one hand, its text encoder is used in text-to-image generation to obtain meaningful representations of the input caption and use them to condition diffusion models while generating the image. On the other hand, CLIP's zero-shot classification capabilities enable it to be adapted into an open-vocabulary object detector, OWL-ViT (Minderer *et al.* 2022), which will come in handy when automatically evaluating the spatial reasoning capabilities of image generators.

2.2.2 More Recent VLMs

The most recent VLMs are orders of magnitude larger than the previously mentioned encoders. With the emergence of general-purpose LLMs, alternative vision-and-language models appeared, showcasing similar capabilities. Since transformer-based language models and VLMs emerged, both kinds of systems have followed the same architectural trends. Therefore, it is easy to find vision-and-language alternatives to text-only language models. For instance, OFA (Wang *et al.* 2022a) follows the same principles as the text-only T5 model (Raffel *et al.* 2020). Both are generative encoder-decoders that build general-purpose models, with the distinction that OFA also integrates vision-only and vision-and-language tasks during its pre-training.

Currently, the VLM families described in Figure 2.2 do not share the same popularity, as the use of encoder-only models has declined over the last few years, following a similar fashion to text-only encoders. Contrastive models are used to obtain rich multimodal representations, but the performance and versatility of large generative VLMs in downstream tasks make VLMs with masking objectives a less appealing alternative. However, Zeng *et al.* (2022b) proposed mixing both masking and contrastive objectives to build more robust encoders, showing that this research line is still active and gives competitive results in some vision-and-language tasks, e.g. visual question-answering (Zeng *et al.* 2024; Luo *et al.* 2024).

As seen before, generative LLMs can be adapted to leverage visual representations, but there are different ways to build large VLMs. Some are pre-trained from scratch, like OFA and CM3LEON (Yu *et al.* 2023). In the case of PALM-E (Driess *et al.* 2023), they fine-tune an existing LLM alongside a visual encoder to incorporate other modalities. Many approaches use frozen LLMs and visual encoders, enabling the LLM to leverage the new modality by mapping visual representations. These approaches include naive adaptations that tweak LLMs for specific vision-and-language tasks by learning simple linear projections (Koh *et al.* 2023), as well as models that learn more complex mappings, such as BLIP-2 (Li *et al.* 2023a) that uses a Q-former to map visual representations and Flamingo (Alayrac *et al.* 2022) which employs a Perceiver model (Jaegle *et al.* 2021).

Another paradigm that has been applied to VLMs is instruction tuning. This simple method fine-tunes a model on a labelled dataset of instructional prompts and corresponding outputs to adapt LLMs to interact with users and follow their commands (Wei *et al.* 2022a). By creating a vision-and-language instruction dataset and fine-tuning a text-only LLAMA-2 model with the dataset, Liu *et al.* (2024) builds an instruction-tuned model that leverages vision and language.

2.2.3 Text-to-image Generators

Many text-to-image systems have been proposed in the last few years. In general, we can distinguish between those based on auto-regressive transformer architectures, such as the original Dall-E (Ramesh *et al.* 2021), the multi-task system OFA (Wang *et al.* 2022a) or CogView2 (Ding *et al.* 2022); and those based on diffusion models (DMs), pioneered by GLIDE (Nichol *et al.* 2022), which evolved into latent diffusion models (LDMs) and are becoming the de-facto architecture for the latest text-to-image generators, such as Stable Diffusion (Rombach *et al.* 2022) and Attend-and-Excite (Chefer *et al.* 2023).

In our work, we fine-tune open-source LDMs with the strongest spatial reasoning capabilities (Gokhale *et al.* 2023), that is, Stable Diffusion models. The rest of this section describes the basic concept of DMs, the architecture of text-to-image LDMs, training objectives, and common evaluation metrics.

Basic concepts of DMs: Diffusion models learn to recover noisy representations of images iteratively. During training, noise is added to real images in the forward pass, and the model is optimized to estimate that same noise to reverse the process.

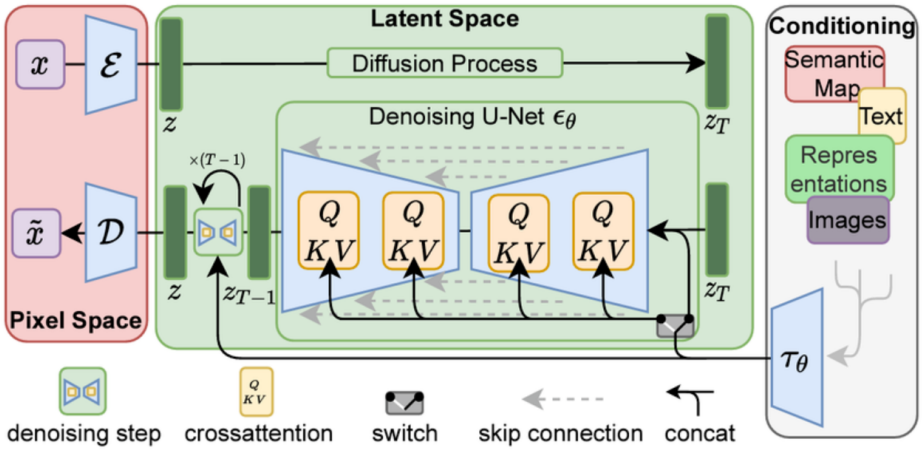
- Forward pass: The addition of noise is a Markov chain where Gaussian noise is added until traces of the original image are removed. Given an image sampled from the data distribution $x_0 \sim q(x_0)$, the forward pass generates x_t iteratively with $q(x_t|x_{t-1})$ in T diffusion steps. There are no learnable parameters in the forward pass, as the rate at which Gaussian noise $\mathcal{N}$ is added to x_{t-1} is fixed before training.

- Reverse pass: DMs learn to recover noisy images during training. Starting from an image sampled from $p_\theta(x_t)$, we generate $x_0 \sim p_\theta(x_0)$ images trying to mimic the true data distribution $q(x_0)$.

$$\mathcal{L}_{DM} := \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\Big] \tag{2.2}$$

  The optimized function is defined in Equation 2.2, where ϵ is the added noise during the forward pass and $\epsilon_\theta(x_t, t)$ is the predicted noise of the denoising module given the noisy input and the number of t steps in which the noise has been applied to the input image.

  At inference time, the reverse process is repeated T times, where, on each step t, the predicted noise $\epsilon_\theta(x_t, t)$ is sampled and removed from x_t to obtain x_{t-1}.



2.3 Figure – Architecture of latent diffusion models, which can condition its output on different modalities. Source figure from Rombach *et al.* (2022).

In summary, DMs are probabilistic models designed to learn a data distribution $q(x_0)$ by iterative denoising a normally distributed variable, which corresponds to learning the reverse process of a fixed Markov Chain of length T.

Architecture. DMs responsible for image generation are composed of a diffusion and denoising module, but conditional LDMs used for text-to-image generation can be divided into three different modules (see Figure 2.3):

- Perceptual image compressor: This module encodes and decodes images into a low-dimensional latent space using $\mathcal{E}$ and $\mathcal{D}$ respectively. Compared to the high-dimensional pixel space, this space is more suitable as LDMs can focus on the relevant semantic bits of $q(x_0)$ and train in a computationally more efficient space.

- Diffusion and denoising module: The module responsible for the forward and reverse passes. Compared to DMs that work in pixel space (Nichol *et al.* 2022), LDMs use latent representations (Rombach *et al.* 2022).

- Conditioning mechanism: Both DMs and LDMs can be conditioned on captions, or even other modalities. This conditioning information denoted as y can be fed to a domain-specific encoder τ_θ that projects y to an intermediate representation $\tau_\theta(y)$. These representations condition the generation by employing cross-attention layers in the denoising module, which enable tasks such as text-to-image generation.

Stable Diffusion models (Rombach *et al.* 2022; Wallace *et al.* 2024) utilize VQ-GAN as their perceptual image compressor (Esser *et al.* 2021), a UNET architecture for denoising (Ronneberger *et al.* 2015) and CLIP's text encoder to encode meaningful conditioning representations (Radford *et al.* 2021).

Regarding data needed to learn the denoising objectives, DMs without any conditioning mechanism only need images, learning image generation in an unsupervised manner. For text-to-image generation, DMs need labelled data consisting of image-caption pairs. The literature uses datasets with millions of instances to learn these denoising objects. That is the case of Stable Diffusion models, that use the public LAION dataset for their training process (Schuhmann *et al.* 2022).

Evaluation. Evaluating text-to-image generation is not a trivial task. Many aspects of the image can be evaluated, so different metrics have been proposed over the last few years. Fréchet Inception Distance (FID) is a widely used metric for quantitatively assessing image quality and photorealism (Heusel *et al.* 2017). More specifically, FID compares the mean and standard deviation of the deepest layer in Inception v3 of synthetic and real images. A high FID value means the generated images are far from real-world images. Inception Score is a related Inception-based metric that assesses the photorealism of generated images (Salimans *et al.* 2016). If we want to analyze how well synthetic images are aligned with their corresponding captions, R-Precision (Xu *et al.* 2018) and CLIPScore (Hessel *et al.* 2021) are two popular options.

In our work, we use FID during the fine-tuning of diffusion models to analyze that image quality does not decay during the fine-tuning process. Apart from that, we are interested in evaluating the alignment between images and captions containing spatial relations. As R-Precision and CLIPScore are not specifically designed to evaluate spatial reasoning abilities and might not encode the meaning of spatial relations, we will resort to more suitable metrics mentioned in Section 2.3.2.

2.3 Limitations of VLMs

As the field of VLMs continues to expand with more robust approaches, their abilities to generate captions, answer questions... about images have drastically increased. However, they still struggle with tasks that need external knowledge[1] or spatial reasoning, and thus methods to alleviate this need have been explored.

[1]We use external, outside, or world knowledge interchangeably, referring to knowledge that is not present in the training data but can be found in external knowledge sources.

2.3.1 World Knowledge Integration

Visual question-answering (VQA) tasks are commonly used to analyze world knowledge integration in VLMs. Given an image and a question about it, the goal is to answer that question correctly. To solve these tasks, VLMs *only* require mapping visual information with the given question and, if needed, retrieving relevant knowledge to answer the question, which makes them suitable for building methods that integrate world knowledge. Many VQA tasks in the literature can be solved with just the visual information of the image (Antol *et al.* 2015; Goyal *et al.* 2017; Johnson *et al.* 2017). However, others demand leveraging external knowledge to infer the answer, that is, knowledge-based VQA tasks.

Good examples of knowledge-based VQA tasks are KB-VQA (Wang *et al.* 2017a), KVQA (Shah *et al.*, 2019), FVQA (Wang *et al.* 2017b) and OK-VQA (Marino *et al.* 2019). KVQA requires knowledge about named entities (e.g. Barack Obama, White House, United Nations), which is provided as a graph. FVQA annotates questions by selecting a fact from a fixed knowledge base but its size is small. KB-VQA is even smaller, presenting template-based questions whose answers can be obtained reasoning over commonsense resources or Wikipedia. In contrast, OK-VQA requires knowledge from unspecified external resources and, although smaller than KVQA in terms of the number of images and question-answer pairs, it is considerably bigger than the other knowledge-based VQA datasets and requires more varied knowledge sources. Therefore, we have chosen OK-VQA for our experiments, which is evaluated using the standard VQA score defined by Goyal *et al.* (2017).

As a side note, after the development of our work in world knowledge integration, a new knowledge-based VQA task was released, A-OKVQA (Schwenk *et al.* 2022). It is twice as big as OK-VQA, but it mainly focuses on questions that require commonsense reasoning, which makes finding the needed information in external knowledge sources more difficult than in OK-VQA.

Implicit vs. Explicit (Symbolic) Knowledge

Knowledge is encoded in different shapes and forms. Text, graphs, or tabular data are common modalities in which knowledge is stored. On the one hand, text corpora, such as ThePile (Gao *et al.* 2020), can be scrapped from the Web and contain knowledge from different domains. On the other, knowledge graphs (Speer *et al.* 2017; Ilievski *et al.* 2021) and tabular data (Parikh *et al.* 2020) can be used to find knowledge sparse in text (e.g. commonsense).

System	Year	Implicit Knowledge	Symbolic Knowledge	VQA Score
VLM Encoders				
ConceptBERT	2020	VilBERT	ConceptNet	33.7
KRISP	2021	BERT	ConceptNet, hasPart KB DBPedia, Visual Genome	38.9
RVL	2021	LXMERT	ConceptNet, Wikidata	39.0
MAVEx	2022	VilBERT	ConceptNet, Wikipedia Google Images	41.4
LLMs / Large VLMs				
PICa	2022	GPT-3	None	48.0
KAT	2022	GPT-3	Wikidata	54.4
REVIVE	2022	GPT-3	Wikidata	58.0
PromptCap	2023	GPT-3	None	60.4
Prophet	2023	GPT-3	None	61.1
PaLM-E	2023	PaLM	None	66.1

2.1 Table – Summary of OK-VQA systems, their knowledge sources, and performance measured with VQA score. See text for references.

As Marino *et al.* (2021) specifies, knowledge can be represented in two types. Explicit or symbolic knowledge encompasses knowledge that can be explicitly found in different modalities (text, graphs,...), whereas implicit knowledge is embedded into some non-symbolic form (e.g. weights of a language model, as they implicitly capture language-based knowledge during pre-training). In our work, we analyze several approaches to better leverage the implicit knowledge embedded in language models and VLMs. Therefore, we deeply analyze the state-of-the-art VLMs evaluated in OK-VQA, emphasizing their use of different knowledge sources.

Integrating Knowledge in VLMs for OK-VQA

Table 2.1 summarizes the state-of-the-art for OK-VQA during the development of this thesis. We distinguish their knowledge sources and split them into two groups depending on their backbone models: i) VLM encoders with masking objectives and ii) LLMs or large VLMs.

In short, this division establishes a change in the paradigm of state-of-the-art VLMs to solve OK-VQA. Until 2022, OK-VQA systems focused on adding symbolic knowledge of different modalities to the VLMs. However, the advent of LLMs showed that models with higher capacity can leverage their implicit knowledge better, and the efforts of researchers started to be spent on better leveraging this embedded knowledge. We now describe the approaches mentioned in Table 2.1 with more detail.

- ConceptBert (Gardères *et al.* 2020) was the first system to use multimodal transformers and symbolic knowledge for OK-VQA. It is based on a combination of a pre-trained BERT to encode questions, a graph convolutional neural network to encode triplets extracted from the ConceptNet knowledge graph (Speer *et al.*, 2017), and a multimodal transformer (VilBERT) to jointly represent and reason over image features and encoded question tokens.

- KRISP follows a similar approach (Marino *et al.* 2021), combining a VLM with symbolic knowledge. In this case, the backbone model of KRISP is MM_{BERT}, based on VisualBert (Li *et al.* 2019), and initialized with the weights of a pre-trained BERT. Additionally, authors built a knowledge graph fusing DBPedia (Auer *et al.* 2007), ConceptNet (Speer *et al.* 2017), VisualGenome (Krishna *et al.* 2017) and hasPart KB (Bhakthavatsalam *et al.* 2020). They used different image feature encoders and question tokens to obtain a subset of the full graph relevant to the target question and image. Finally, using a graph convolutional neural network, they combined the symbolic and implicit knowledge to predict the final answer.

- MAVEx (Wu *et al.* 2022) and RVL (Shevchenko *et al.* 2021) showed different ways to combine implicit and symbolic knowledge. MAVEx used a pre-trained VilBERT to generate various candidate answers which were later reranked using answer-specific knowledge retrieval. They also used both textual and visual knowledge resources, including images searched using Google, sentences from Wikipedia articles, and concepts from ConceptNet. On the other hand, RVL trained the two-stream multimodal transformer LXMERT (Tan and Bansal 2019) with an auxiliary objective that aligned its representations with knowledge graph embeddings retrieved from ConceptNet and Wikidata.

These models (Gardères *et al.* 2020; Marino *et al.* 2021; Wu *et al.* 2022; Shevchenko *et al.* 2021) employ different symbolic knowledge sources. Nevertheless, we have noticed that the improvement obtained by adding symbolic knowledge is minor. MAVEx is the only one with a significant improvement. However, due to its design, the model is limited to answering a set of answer candidates generated by only accessing implicit knowledge. This shows the dependency of current systems on the encoded knowledge found in VLMs. So, we focus on implicit knowledge (as opposed to explicitly encoded knowledge) which we exploit by first verbalizing images and then feeding these captions to a pre-trained language model.

Contemporary to our work, Yang *et al.* (2022) proposed PICa, which prompts GPT-3 via the use of image captions and object tags. By feeding in-context examples of OK-VQA to the LLM, they set the new state-of-the-art for OK-VQA. They also show that the careful selection of these in-context examples and an ensemble of GPT-3 models boost PICa's performance further. These findings fueled the use of LLMs, especially GPT-3, for knowledge-based VQA. Some of these approaches tried to retrieve relevant knowledge from Wikidata via CLIP models, which include KAT (Gui *et al.* 2022) and REVIVE (Lin *et al.* 2022). However, OK-VQA's leaderboard[2] is currently led by Prophet (Shao *et al.* 2023) and PROMPTCap (Hu *et al.* 2022), which do not rely on symbolic knowledge.

Finally, it is worth mentioning that, to the best of our knowledge, PALM-E sets the current state-of-the-art in OK-VQA (Driess *et al.* 2023) with a VQA score of 66.1 points (doubling ConceptBert's score), although these results are not reflected in OK-VQA's leaderboard. PALM-E adapts a 540B parameter transformer-based decoder-only LLM (Chowdhery *et al.* 2023) by injecting multimodal observations into the language embedding space as if they were language tokens. As OK-VQA is one of the many vision-and-language datasets used in the training phase of PaLM-E, its good performance is not unexpected.

2.3.2 Spatial Reasoning

Spatial reasoning (SR) involves understanding and processing the spatial relations between objects within visual scenes or textual descriptions. This capability is crucial for tasks that require understanding how objects are positioned relative to each other and how they interact within a given space. We explore SR in two settings. On the one hand, we explore how language models can learn to interpret

[2] https://okvqa.allenai.org/leaderboard.html. (accessed on 06/07/2024)

spatial relations correctly. On the other, we dive into text-to-image generation to evaluate and improve the generation of correct spatial relations. Note that, in this dissertation, we focus on explicit spatial relations (e.g. the man above the horse), not on implicit ones (e.g. the man riding the horse, where the same relative position between objects is defined implicitly).

SR in Language Models

Datasets. The spatial commonsense knowledge of current LMs and VLMs is evaluated from different angles thanks to many available datasets. For example, Bagherinezhad *et al.* (2016) and Elazar *et al.* (2019) focus on the acquired commonsense knowledge of models about object scales, e.g. do they know that a person is bigger than an ant? In that sense, they do not provide a specific scene context, but rather ask about generic object scale relations, so the dataset they provide is not useful for our work. Collell *et al.* (2018) and Elu *et al.* (2021) propose datasets and methods to generate bounding boxes from textual descriptions. Although the evaluation approach is suitable for testing spatial grounding, these methods focus on implicit spatial relations, which we are not interested in.

Intending to evaluate both object scales and spatial relations, Liu *et al.* (2022b) and Zhang *et al.* (2022) provide new unified datasets. As the objective of these works is to evaluate whether VLMs learn more spatial commonsense than language models, the datasets are purely textual, so they do not provide any means to ground spatial relations (they assume the grounding occurs in a previous training process), and, hence, they are not useful for our work. Interestingly, authors find that VLMs, and more concretely text-to-image systems, perform much better than text-only LMs.

There are other ways to test the spatial inference and reasoning capabilities of VLMs. CLEVR was one of the pioneering works on testing compositional language and elementary visual reasoning (Johnson *et al.* 2017). Using 3D-rendered images of simple objects such as spheres, cones, and cubes, different questions are generated automatically. A model needs to process the image and the question to provide an answer. Although CLEVR can be used to test spatial grounding, it has two major drawbacks: i) questions not only cover spatial grounding but some other concepts such as compositional language and attribute identification, and ii) spatial relations are limited to four, i.e. *left*, *right*, *behind* and *in front*. The natural extension of CLEVR is GQA (Hudson and Manning 2019), which shares similar ideas but is built on natural images. Although spatial grounding is essential for this task, compositional language is also evaluated. As both dimensions appear

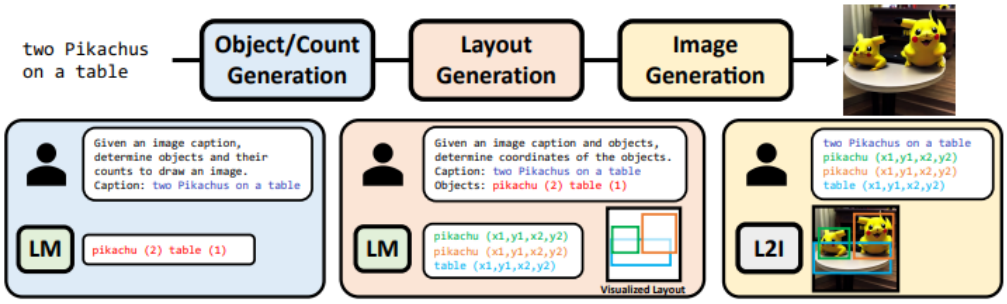
together, we believe this dataset is not the best option for our purposes.

In the text-only scenario, SpartQA provides another synthetic dataset (with a small subset annotated by humans). Given a spatial description of a scene with explicit relations, a model has to answer some spatial questions about that scene. The task is specially focused on spatial reasoning capabilities, such as transitivity, and does not provide any means to ground spatial relations, as its target is the reasoning process. Moreover, similar datasets have been recently proposed as an extension and improvement of SpartQA (Mirzaee and Kordjamshidi 2022).

In our work, we use the recent Visual Spatial Reasoning (VSR) dataset (Liu *et al.*, 2023) to evaluate the spatial grounding capabilities of text-only language models. VSR has been designed to test spatial grounding capabilities, covering 65 different spatial relations over natural images collected from COCO (Lin *et al.*, 2014) and annotated by humans. Given an image, they provide a caption that describes a spatial relation between two objects that appear in the image. That relation can be real or fake, and the model has to infer precisely that, i.e. whether the caption is aligned with the given image. We believe VSR is a good candidate to evaluate the grounding of explicit spatial relations for language models and VLMs. Nevertheless, as text-only language models cannot process images, we propose a way to verbalize those images and run meaningful experiments.

Encoding layout information. Although VLM architectures may vary, the basic idea is to input the models with textual tokens and visual features. As transformers are feed-forward networks they do not consider the input order, thus, positional encodings represent word order (Vaswani *et al.* 2017). A similar idea is used also for visual features. LXMERT (Tan and Bansal 2019), for instance, uses the x_0, y_0, x_1, x_2, W, H coordinates of a bounding box for a given visual feature, projects them linearly, and sums it to the visual feature itself before inputting it to the transformer. Alternatively, ViLT (Kim *et al.* 2021) does not use any object detector, but works directly on image patches. They use positional embeddings to represent the order of those patches in the image, very similar to the positional embeddings of textual tokens.

Regarding text-only language models, to the best of our knowledge, Patel and Pavlick (2022) represent scenes with textual tokens on which spatial grounding and reasoning can be performed. More concretely, they propose to create grid-like structures with textual tokens inside the vocabulary of the language model. Their proposal is interesting, but it is limited to toy experiments since they can only represent *small* scenes and six spatial relations: *left*, *right*, *up*, *down*, *top* and *bottom*. In contrast, our approach described in Chapter 4 covers complex scenes depicted in natural images and 23 spatial relations.



2.4 Figure – VPGen's Pipeline. It decomposes text-to-image generation in three steps: i) lists and counts the objects described in the input caption, ii) generates the layout conditioned on the amount of counted objects, and iii) generates the image conditioned on the initial caption and the generated layout. Source image from Cho *et al.* (2023b).

SR in Image Generation

In the last few years, text-to-image systems have improved in photorealism and efficiency, but recent work has shown that their performance for explicit spatial relations is not good (Gokhale *et al.* 2023; Cho *et al.* 2023b). These models struggle to correctly draw textual descriptions like *a cat on top of a table*. To overcome these limitations, VPGen (Cho *et al.* 2023b) and LayoutGPT (Feng *et al.* 2023) propose pipeline systems, combining Large Language Models to generate layouts from textual prompts and layout-to-image generators (Li *et al.* 2023b). The difference between both systems is that VPGen fine-tunes Vicuna-13B (Chiang *et al.* 2023) to generate layouts from textual descriptions, whereas LayoutGPT relies on Llama-2-7B (Touvron *et al.* 2023) and in-context learning for the same purpose.[3] See Figure 2.4 for more details on VPGen's pipeline.

To avoid complex and large pipeline systems, (Yang *et al.* 2023) propose ReCo, an end-to-end system that uses layout descriptions in the input. In our approach described in Chapter 5, we also focus on end-to-end systems. Nevertheless, we avoid inserting layout information into the input, as this imposes a substantial burden on users compared to simple text inputs.

Evaluation. To evaluate the performance of text-to-image generators for explicit spatial relations, dedicated datasets have been created, since commonly used datasets like COCO (Lin *et al.* 2014), CC12M (Changpinyo *et al.* 2021) or LAION

[3]Originally they used LLMs from the OpenAI GPT family, but they have released a publicly available Llama2-based variant of LayoutGPT, which we use in this work.

(Schuhmann *et al.* 2022), contain very few examples of explicit spatial relations. For example, Gokhale *et al.* (2023) proposes the SR_{2D} dataset, composed of synthetic captions created by combining two objects in the COCO object vocabulary and four explicit spatial relations. SR_{2D} only contains captions and can not be employed for training. Similarly, Feng *et al.* (2023) published the Numerical and Spatial Reasoning dataset (NSR-1K) which does include caption-image pairs. The spatial part contains only 1021 image-caption pairs (738 for train and 283 for test, with no development split) for 4 relations, insufficient for accurate evaluation and too small for training.

There are many approaches for the automatic evaluation of explicit spatial relations in generated images (Gokhale *et al.* 2023; Cho *et al.* 2023b; Feng *et al.* 2023. These evaluations rely on: i) object detectors to locate objects in the image, and ii) heuristic rules that use bounding box coordinates to decide whether a given relation is fulfilled. This implies the need for labeled evaluation data (Feng *et al.* 2023; Gokhale *et al.* 2023), that is, each caption must have a spatial triplet containing two objects and their spatial relation. On the contrary, VPEval (Feng *et al.* 2023) uses GPT-3 to generate evaluation programs conditioned on the caption, and runs these programs with the aid of visual tools (e.g. object detector, OCR). Instead of more noisy evaluations, VPEval does not need labeled spatial triplets.

3. KAPITULUA

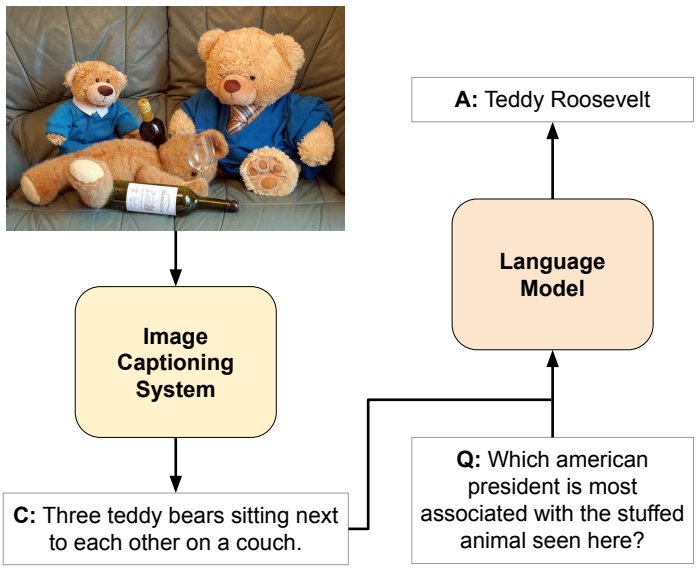
Ezagutza Inplizituaren Erabilera VQA Sistemetan

3.1 Motibazioa eta Ekarpenak

Ikusmen-testu ataza gehienak definitzerakoan, ataza ebazteko adierazgarria den informazioa datu-multzoko irudi eta testuetan zehazten da. Horiek dira, adibidez, galdera-erantzute bisuala, edo VQA, (Antol *et al.*, 2015) eta ikusizko inferentziaren (Xie *et al.*, 2019) kasuak. Hala ere, ataza batzuk ebazteko datu hauetatik at dagoen ezagutza eskura eduki behar da. Kapitulu honetan, kanpo ezagutzan oinarritutako VQA ataza batean murgildu gara, OK-VQA atazan hain zuzen ere (Marino *et al.*, 2019). Bertan, galderak ondo erantzuteko irudiaren edukia izatea ez da nahikoa. Irudi eta testu pareekin ebatz daitezkeen VQA atazak ez bezala, ataza honek kanpo ezagutza txertatu, prozesatu eta ezagutza horretatik erantzunak inferitzeko ahalmena duten ereduen beharra dauka.

OK-VQA atazan baliagarria den kanpo ezagutza bi azpimultzotan banatu daiteke (Marino *et al.*, 2021): (i) ezagutza sinbolikoa, grafo edota beste datu egitura batzuen bidez adierazi daitekeena, ConceptNet (Speer *et al.*, 2017) ezagutza grafoa adibidez; eta (ii) ezagutza inplizitua, testu corpus erraldoietan entrenatutako neurona sareen pisuetan kodetua azaltzen dena. Azkeneko esaldia indartuz, transformer arkitekturan oinarritutako hizkuntza-eredu aurrentrenatuak (Devlin *et al.*, 2019; Liu *et al.*, 2019; He *et al.*, 2020) arrakasta handiarekin erabili dira ezagutza iturri inplizitu gisa (Petroni *et al.*, 2019).

Hau horrela izanik, kapitulu honetan aurrentrenatutako hizkuntza-ereduak ezagutza inplizitu iturri gisa erabili ditugu. OK-VQA atazan hizkuntza-ereduen era-



3.1 Irudia – Galdera eta irudi bat emanda, irudiaren edukia hitzez adierazten dugu goiburuko baten bidez, inferentzia unean aurrentrenatutako hizkuntza-eredu bat erabiliz. Gaur egungo ezagutzan oinarritutako hizkuntza-ereduak eredu multimodalak baino hobeak dira, bai beraien orokortze ahalmenean eta baita inferentzia unean ere.

bilera ohikoa bada ere, hainbat eratan integratzen dira modalitate anitzekin lan egiterakoan. Izan ere, eredu hauek ikusizko eta testuzko sarrera datuez elikatu behar dira ataza behar bezala ebazteko. Hizkuntza-ereduak testua bakarrik prozesatzeko diseinatuta daude, testu corpus erraldoietan estentsiboki entrenaturik. Hori dela eta, testuan soilik oinarritutako sistema batek ezagutza inplizitu hau hobeto aprobetxatuko duenaren hipotesia jarri dugu mahai gainean.

OK-VQA ikusmen-testu ataza denez, irudiak automatikoki berbalizatzea proposatzen dugu irudiari buruzko informazioa hitzez adierazteko modu gisa. Gure kasuan, irudiak berbalizatzen ditugu goiburukoen bitartez, hau da, irudia deskribatzen duen esaldi baten bitartez. Behin goiburuko hauek sortzen direnean, planteatzen dugun metodoak testuzko ereduak bakarrik erabiltzen ditu. Irudia berbalizatzeko garaian informazio galera bat dagoela badakigu eta testua bakarrik erabiltzeak hasierako galera hori konpentsa dezakeen egiaztatu nahi dugu. Goiburukoetan oinarritutako eredu hau (*Caption based Model* edo CBM) 3.1. Irudian antzeman daiteke.

Gure hipotesia balioztatzeko, OK-VQA atazaren gaineko esperimentazio zabala aurkezten dugu. Modalitate anitzeko transformer-ra ikusmen-testu atazetan irudi eta testu pareak prozesatzeko erabiltzen den eredu estandarra da, eta guk proposatutako CBM ereduarekin konparatu dugu. Hizkuntza-ereduen tamainaren eragina aztertu dugu ere bai, OK-VQA atazan ereduaren gaitasun eta kapazitateak nola eragiten duen ikusteko.

Gure ikerketaren ekarpenak ondorengoak dira:

- Goiburukoak irudiak baino eraginkorragoak dira OK-VQA atazan tamaina berdineko ereduak erabiltzen direnean. Gainera, pareko emaitzak lortzen dituzte beste VQA datu-multzo osagarriekin doitzen badira.

- Hizkuntza-ereduen tamaina eta kapazitatea handitzeak artearen egoerako emaitzak lortzeko ahalmena ematen du, gaur egungo modalitate anitzeko transformer-ak hein handi batean gaindituz. Hobekuntza hau oraindik egonkortu ez dela antzeman dugu.

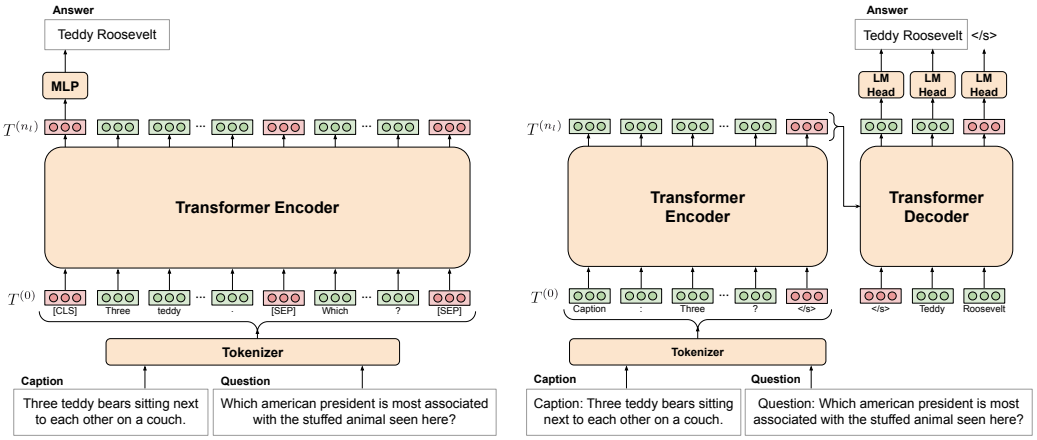
- PICa ereduaren (Yang *et al.*, 2022) testuinguru bidezko ikasketaren erabilera konplexuak ez du gure eredu txikiagoaren doikuntza gainditzen, hau da, T5 batean (Raffel *et al.*, 2020) oinarritutako gure sistemak emaitza konparagarriak lortzen ditu GPT-3 ereduaren bost inferentzien bateratzearekin konparatuz. Kontuan izan behar dugu GPT-3 ereduaren parametro kopurua 15 aldiz handiagoa dela.

- OK-VQAn goiburukoen ekarpena VQA ataza estandar batean (Goyal *et al.*, 2017) baino dezente handiagoa da. Honek testu modalitateko ereduak kanpo ezagutza behar den kasuetan bereziki eraginkorrak direla adierazten du.

Gure VQA eredua bizitza errealeko kasu ezberdinetara egokitu daiteke itsuei edota ikusmen urria duten pertsonei laguntzetik (Gurari *et al.*, 2018) gaur egungo asistente birtualak hobetzera (Tulshan and Dhage, 2018). Gure eredua bereziki onuragarria da munduko ezagutza behar duten aplikazioetarako. Beraz, bai hezkuntza eta baita aisira begira ere aplikatu ahalko litzateke.

Lan honetan garatutako kodea publikoki atzigarri dago.[1]

[1]URL: https://github.com/salanueva/CBM



(a) CBM_{BERT} **(b)** CBM_{T5}

3.2 Irudia – Proposatu ditugun CBM ereduen eskemak.

3.2 Metodologia

Atal honetan inplementatu ditugun ereduak, erabilitako atazak eta ikasketa algoritmoak deskribatu ditugu. Inplementazio lanerako *Pytorch* (Paszke *et al.*, 2019), *Pytorch Lightning* eta *Transformers* (Wolf *et al.*, 2020) liburutegiak erabili ditugu.

3.2.1 Inplementatutako Ereduak

Goiburukoetan oinarritutako eredua (CBM)

Gure goiburukoetan oinarritutako eredua, CBM deritzoguna (*Caption Based Model*), bi pausotan banatzen da: (i) goiburuko sortzaile sistema batek irudi baten deskribapen motz bat sortzen du eta (ii) hizkuntza-eredu batek irudi horri buruzko galdera bat erantzuten du, irudiaren informazio iturri gisa goiburukoa bakarrik erabiliz.

Lehenengo pausoa burutzeko OSCAR eredua (Li *et al.*, 2020) erabili dugu. Hainbat modalitate anitzeko atazetan artearen egoera zehazten duen transformer kodetzailea da, irudi goiburuko ataza hauetako bat izanik. Modalitate anitzeko transformer-etan ohikoa den bezala, OSCAR ereduak Faster R-CNN (Ren *et al.*, 2015) deitzen den aurrentrenatutako objektu detektore bat erabiltzen du irudi eskualdeen ezaugarriak eta eskualde horiei dagozkien etiketak lortzeko. OSCAR-en aurrentrenamenduan zehar ezaugarri eta etiketa hauek eskuz anotatutako goiburu-

koekin batera erabiltzen dira, (Anderson *et al.*, 2018) lanean bezala. OSCAR-en goiburuko sormen gaitasuna antzekoa da bere bi tamaina ezberdineko bertsioetan. Horregatik, oinarrizko eredua erabili dugu gure esperimentu guztietarako, hots, bietatik txikiena.

OSCAR eredua goiburuko sorkuntzan doitzeko pausoan, OK-VQA atazaren ebaluazioan dauden irudi eta goiburukoen azpimultzo bat erabili zen. Esperimentuen zuzentasuna bermatzeko eta kutsadura arazoak saihesteko aurrentrenatutako OSCAR eredua goiburuko sorkuntza atazan doitu dugu. Horrela, ebaluazioko instantzia hauek entrenamendu prozesutik kanpo utzi ditugu.

Bigarren pausorako bi hizkuntza-eredu ezberdin erabili ditugu: BERT eredua, modalitate anitzeko ereduekin konparaketa zuzenak egiteko, eta T5 familiako ereduak, handituz doazen hizkuntza-ereduen gaitasuna aztertzeko.

CBM$_{\text{BERT}}$. Lehen hurbilpen honetan, BERT-base transformer kodetzaile aurrentrenatua erabili dugu hizkuntza-eredu gisa. Goiburuko eta galdera tokenizatuen sekuentziak elikatzen dizkiogu BERT ereduari $T^{(0)} = \{\mathbf{t}_i^{(0)} | i = 1, \ldots, n_t\}$ eta, ondoren, $[CLS]$ edo sarrera sekuentziaren lehen tokenaren irteera jasotzen dugu: $\mathbf{t}_1^{(n_l)}$, non n_t sekuentziaren token kopurua eta n_l transformer ereduaren geruza kopurua diren (ikus 3.2a. Irudia).

Hizkuntza-eredua VQA atazetan doitzeko, sailkapen burua gehitu diogu $[CLS]$ tokenaren irteera bektoreari. Nahiz eta VQA (Antol *et al.*, 2015; Goyal *et al.*, 2017) eta OK-VQA (Marino *et al.*, 2019) atazek domeinu irekiko erantzunak izan, lan honen garapenean zehar, artearen egoerako ereduek sailkapen arazo bezala planteatu zuten ataza (Zhang *et al.*, 2021; Marino *et al.*, 2021), entrenamenduko datuetatik eratorritako hiztegi itxiak definituz. Joera hau jarraituz, gure sailkapen burua geruza ezkutu bat duen geruza anitzeko pertzeptroi edo MLP batekin eraiki dugu, bere sarrera bezala $\mathbf{t}_1^{(n_l)}$ jasotzen duena. MLP hau 3.1. Ekuazioan definitzen dugu.

$$\begin{aligned} \mathbf{h} &= LayerNorm(GELU(\mathbf{W}_h \mathbf{t}_1^{(n_l)} + \mathbf{b}_h)) \\ \hat{\mathbf{y}} &= Softmax(\mathbf{W}_{\hat{y}} \mathbf{h} + \mathbf{b}_{\hat{y}}) \end{aligned} \tag{3.1}$$

MLP honen geruza ezkutuan GELU aktibazio funtzioa erabiltzen dugu normalizazio geruza bat (Ba *et al.*, 2016) aplikatu aurretik. MLP honen parametro ikasgarriak $\mathbf{W}_h \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_h \in \mathbb{R}^{d_h}$, $\mathbf{W}_{\hat{y}} \in \mathbb{R}^{d_h \times n_{label}}$ eta $\mathbf{b}_{\hat{y}} \in \mathbb{R}^{n_{label}}$ dira, non n_{label} sailkapen atazan definitutako erantzun kopurua eta d_h geruza ezkutuaren dimentsionalitatea den, gure kasuan $d_h = 768$ izanik.

CBM_{T5}. Gure bigarren hurbilpenean T5 transformer kodetzaile deskodetzaile aurrentrenatuak erabili ditugu (Raffel *et al.*, 2020). Lan hau garatu zenean eredu hauek artearen egoera definitzen zuten galdera-erantzute atazetan. Tamaina ezberdineko bost T5 eredu daude publikoki eskuragarri, txikienak 60M parametro eta handienak 11B izanik. CBM_{BERT} jarraituz, T5 ereduei goiburuko eta galdera sekuentzia tokenizatuak elikatzen dizkiegu ere bai, $T^{(0)} = \{\mathbf{t}_i^{(0)} | i = 1, \ldots, n_t\}$. Hala ere, esaldi bakoitzaren hasieran *"caption:"* eta *"question:"* bezalako aurrizkiak gehitu ditugu. Honako hau T5-en aurrentrenamenduan erabilitako testuzko baldintzak ahal den heinean imitatzeko burutu da, hizkuntza-ereduak aurrentrenamenduan ikasitakoa hobeto aprobetxatzen lagunduz. BERT ez bezala, T5 hizkuntza-eredu sortzailea da. Beraz, erantzun bat sailkatu beharrean, T5ek modu irekian sortzen du erantzuna, hurbilpen honetan sailkapen burua alde batera utziz (ikus 3.2b. Irudia).

Modalitate anitzeko transformer-a (MM_{BERT})

Gure CBM_{BERT} eredua transformer arkitekturan oinarritutako modalitate anitzeko MM_{BERT} ereduarekin (Marino *et al.*, 2021) konparatu dugu. BERTen aldaera honek eskualde bisualen eta galderen adierazpen bektorialak erabiltzen ditu sarrera gisa. Beste hitzetan, BERT testuzko sarrerak soilik prozesatzeko diseinatuta dagoen bitartean, MM_{BERT}ek bere bektore geruza egokitzen du ezaugarri bisualak jaso ahal izateko ere bai.

Faster R-CNN eredua erabili dugu n_v eskualde bisualen ezaugarriak $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{n_v}\}$ kalkulatzeko. $v_i \in \mathbb{R}^{d_v}$ bakoitzak irudiko objektu bat adierazten du, non gure kasuan $d_v = 2048$ den. $\mathbf{V}$ ezaugarri bisualek ez dute objektuen posizioa irudian kodetzen, arazo hau ezaugarri bakoitzari dagokion kaxa inguratzailearen koordenatuak konkatenatuz konpondu daitekeelarik. Hasierako esperimentu batzuk burutu ondoren informazio gehigarri honek VQA eta OK-VQA atazei ez diela laguntzen ondorioztatu genuen.

Erabili dugun Faster R-CNN-aren bertsioak ResNeXt-152 (Xie *et al.*, 2017) erabiltzen du bizkarrezur eredu gisa, eta MMF liburutegiko inplementazioa erabili dugu irudi eskualdeen ezaugarriak kalkulatzeko.

CBM eta MM_{BERT} ereduen arteko konparaketa errazteko $[CLS]$ tokenaren irteera bektorea erabiltzen dugu MLP sailkatzailea elikatzeko. Kontuan izan hau ez dela jatorrizko MM_{BERT} (Marino *et al.*, 2021) ereduaren parekoa, jatorrizko inplementazioak azken transformer geruzako irteera bektore guztien batezbestekoa erabiltzen baitu sailkatzailea elikatzeko.

VQA: What color is the bear? brown
OK-VQA: What species of bear is this? grizzly

VQA: What is the weather like? cloudy
OK-VQA: Why would one suspect that this is not chicago? sign

VQA: Are the animals in captivity? yes
OK-VQA: Which valuable material grows on this animal's face? ivory

3.3 Irudia – VQA 2.0 eta OK-VQA datu-multzoen adibide batzuk. VQA ataza-ko galderak irudiaren edukiari buruzkoak dira, eta OK-VQA atazako galderek, berriz, kanpo ezagutza behar dute.

Galderatan oinarritutako eredua (Q_{BERT})

Goiburukoen ekarpena neurtu nahi dugunez, sarrera gisa galdera bakarrik jasotzen duen eredu bat entrenatu dugu ere bai, irudiari buruzko informaziorik jasotzen ez duena. Eredu horri Q_{BERT} deritzogu eta CBM_{BERT} ereduaren ablazio gisa ikus daiteke.

3.2.2 VQA Datu-multzoak

Gure esperimentuetarako datu-multzo nagusia OK-VQA (Marino *et al.*, 2019) da, modalitate anitzeko ataza honek hizkuntza-ereduen ezagutza inplizituaren erabilera ebaluatzea ahalbidetzen baitu. Era berean, VQA 2.0 (Goyal *et al.*, 2017) datu-multzoaren gainean esperimentuak egin ditugu, bi arrazoik motibatuta: (i) ereduen doitze prozesuan datu osagarri gisa erabiltzeko (OK-VQA atazan doitu aurretik); eta (ii) ereduen arteko errendimendu ezberdintasunak bi eszenario ezberdinetan aztertzeko: VQA ataza estandar batean, eta baita ezagutza behar handiko VQA ataza batean ere. 3.3. Irudian bi datu-multzoen adibideak azaltzen dira.

VQA 2.0

Datu-multzo honek irudiei buruzko galdera irekiak ditu. Galdera hauek objektuak eta hauen atributuak irudian identifikatzen, erlazioak detektatzen eta objektuak kontatzen zentratzen dira. Datu-multzo hau COCO datu-multzotik (Lin *et al.*, 2014) erauzita dauden 204K irudiz eta 1.1M galderaz osatuta dago, galdera bakoitzak 10 giza anotazio dituelarik balizko erantzun gisa. Ataza hau ebazteko uneko artearen egoera ez zen testu sorkuntzan oinarritzen. Izan ere, hauen paradigma

nagusia aldez aurretik definitutako erantzun sorta baten gaineko sailkapena burutzea da. Hiztegi hau entrenamendu instantzien erantzun ohikoenak batuz sortu da, eta VQA 2.0ren kasuan 3.129 erantzun posible definitu dira.

VQA 2.0 datu-multzoa hiru azpimultzotan banatzen da: entrenamendua (*train*), garapena (*val*) eta ebaluazioa (*test*). VQA 2.0-ko garapeneko irudi batzuk OK-VQA atazako ebaluazio zatian berrerabiltzen dira. Beraz, edozein kutsadura saihesteko, ez dugu VQA 2.0-ko garapen zatia erabiltzen ez entrenamendurako eta ezta hiperparametroak aukeratzeko ere.

VQA atazetarako ebaluazio metrika estandar bat proposatu zen (Antol *et al.*, 2015). Metrika horretan sistemaren erantzun bat erabat zuzena dela zehazten da hamar giza anotazioetatik gutxienez hirutan erantzun hori agertzen bada. Anotazio horietan erantzun jakin bat x aldiz agertzen bada, asmatze-tasa metrika hau 3.2. Ekuazioan definitzen da.

$$acc = \min\left(\frac{x}{3}, 1\right) \tag{3.2}$$

Horrez gain, metrika honekin ereduen ebaluazioa anotatzaileen arteko adostasunarekin konparatu nahi zen (Antol *et al.*, 2015). Giza anotazioetan 10 instantzia bakarrik daudenez, asmatze-tasa kalkulatzeko 10 anotazioetatik 9 aukeratzeko konbinazio guztiak hartzen dira kontuan. Horrela, 9 anotazioko konbinazio bakoitzarekin 3.2 kalkulatzen da, metrika honen amaierako balioa konbinazio guztien arteko batezbestekoa izanik.

OK-VQA

OK-VQA datu-multzoa COCO datu-multzoko 14.031 irudiren eta eskuz anotatutako 14.055 galderaren gainean eraiki da. VQA 2.0 atazan bezala, galdera bakoitzak 10 giza anotazio ditu eta metrika bera erabiliz ebaluatzen da. Ezagutza behar handiko VQA datu-multzo gisa, galdera bakoitza erantzuteko iruditik kanpoko ezagutza eskuratu behar da. Hala ere, kanpo ezagutza hori ez dago ez hornitua ezta identifikatua ere, hots, ez dago zeregin horretarako eskura dauden ezagutza iturrien zerrendarik, ataza ebaztea zailduz.

Datu-multzo honen bi bertsio daude publikoki atzigarri. Bertsio bakoitzean erantzunak nola normalizatzen diren aldatzen da, erantzunen erro bilaketa algoritmoa aldatuz. OK-VQA v1.0 bertsioan erabiltzen den erro bilaketak existitzen ez diren hitzak itzultzen ditu kasu batzuetan (adibidez, “poni tail”, “pony tail” beharrean). OK-VQA v1.1-ean erro bilaketa ezberdin bat aplikatzen da, erantzunen hiztegi koherenteagoa lortuz. Esperimentu guztietan OK-VQA v1.1 erabili dugu.

3.2.3 Ikasketa Algoritmoa

VQA landu duten aurreko lanetan ez bezala, ez dugu entropia gurutzatu bitarra erabiltzen gure sailkapen ereduetan. Izan ere, hasiera batean egindako esperimentuek klase leuneko entropia gurutzatuak (*soft cross entropy* edo SCE) azkarrago konbergitzen duela erakutsi zuten. SCE galera funtzioa 3.3. Ekuazioan definitu da. Bertan, $\mathbf{y}$ VQA atazan erabiltzen den ebaluazio metrikak (ikus 3.2. Ekuazioa) erantzun posible guztiei esleitzen dien balioez osatutako probabilitate bektorea da.

$$\mathcal{L}_{SCE}(\mathbf{y}, \hat{\mathbf{y}}) = -\mathbf{y} \cdot \log \hat{\mathbf{y}} \tag{3.3}$$

CBM_{T5}en inguruan, eredu sortzaile hauek doitzeko entropia gurutzatua erabili dugu. Ondorioz, ereduak sarrera sekuentzia bakoitza dagokion irteera sekuentziari mapatzen ikasten du. Hala ere, honela doitzeak VQA atazetan aurki ditzakegun giza anotazioekin talka egiten du, galdera bakoitzak hainbat balizko erantzun ditu eta. Hori konpontzeko, entrenamendu aro bakoitzean balizko erantzun bat aukeratzen dugu irteera sekuentzia gisa.

Aurretiazko esperimentuek jada erakutsi zuten nola erantzun posible guztietatik ausaz bat aukeratzea kaltegarria zela. Azken finean, hainbat erantzunek akats ortografikoak dituzte, karaktere kate hutsak dira edota ez daukate zentzurik. Hori dela eta, doikuntza fasean zehar VQA ebaluazio metrikarekin puntuazio osoa lortzen ez dituzten erantzunak baztertzen ditugu. Beste hitzetan, gutxienez bi anotatzailek emandako erantzunak bakarrik hartzen ditugu kontuan [2].

3.3 Esperimentuak

Atal honek gure esperimentuen ezarpenen xehetasunak ematen ditu, eta 3.2.1. Atalean definitutako ereduen errendimendua erakusten du artearen egoerarekin alderatuz ere bai. Gainera, analisi sakon bat egiten da antzemandako hobekuntzak nondik datozen ulertzeko.

[2]Galdera batek ez baditu arau hauek betetzen dituen erantzunik, galdera hau ez da erabiltzen entrenamendurako. OK-VQA atazako entrenamendu azpimultzoan 112 instantzien kasua da.

3.3.1 Esperimentazio Ezarpenak

Erabilitako hiperparametroak (Marino *et al.*, 2021) lanetik hartu ditugu CBM_{BERT}, MM_{BERT} eta Q_{BERT} ereduen esperimentu guztietarako. Doikuntza bakoitza 88K pausotan burutzen dugu AdamW optimizatzailea (Loshchilov and Hutter, 2017) eta 56ko sorta tamaina erabiliz. Ikasketa-tasaren aldetik $5 \cdot 10^{-5}$eko balio maximoa definitzen dugu, kosinu planifikatzailea aplikatuz entrenamenduan zehar 2K pausoko beroketa linealarekin.

CBM_{T5} ereduei dagokienez, tamaina ezberdineko bost T5 eredu daude eskura, 60M eta 11B parametro artekoak. Eredu guztiek OK-VQA atazan duten errendimendua erakusteko, aurretik aipatu ditugun hiperparametroak erabili ditugu ondorengo aldaketekin:

- Tamaina ezberdinetako ereduek konbergitzeko entrenamendu pauso kopuru ezberdinak behar dituzte. Kopuru hau zehazteko ondorengo metodologia proposatzen dugu. Entrenamendu instantzien %20a erabili dugu garapenerako, gelditzen den %80a doikuntzarako erabiliz 20K pausotan zehar. Ondoren, garapenean VQA ebaluazio metrika hoberena duen pausoarekin gelditzen gara. Prozesu hau hiru aldiz egiten dugu garapen instantzia berdinak erabiliz. Horren ondoren, hiru entrenamenduen batez besteko pauso kopurua kalkulatzen dugu hiperparametro honen balio finala zehazteko.

- Entrenamendu pauso kopurua ereduaren tamainaren arabera aldatzen denez, doikuntza prozesuan zehar $5 \cdot 10^{-5}$eko ikasketa-tasa konstante bat erabiltzea erabaki dugu. Beraz, ez dugu beroketa edota entrenamendu pausoetan baldintzatzen diren ikasketa-tasa planifikatzailerik erabiltzen.

Sailkapen ereduekin egindako esperimentu guztiak 12GB-eko vRAM memoria duen GPU bakar batean burutu dira, gehienez 12 orduko iraupena izan dutelarik. CBM_{T5} eredu handiagoekin egindako esperimentuetan, ordea, 4 NVIDIA A100 GPU erabiltzera iritsi gara, bakoitzak 80GB vRAM dituelarik. Ereduak handitzen doazen heinean, GPU kopurua eta hainbat hiperparametro aldatzen joan gara sorta tamaina efektiboa berdin mantentzeko helburuarekin. Gainera, 11B parametroko ereduarekin DeepSpeed-en *ZeRO Stage 2* optimizazio algoritmoa erabili behar izan dugu (Rajbhandari *et al.*, 2020), CPU-aren memoria GPU-arekin konpartitzea ahalbidetuz. Hala ere, entrenamendu iraupena gehienez 4 ordukoa izan da kasu honetan, entrenamendu pauso gutxiago behar baitira beste ereduekin konparatuz CBM_{T5} handiena doitzeko.

Emaitza kontsistenteak lortzeko esperimentu bakoitza hiru aldiz burutu dugu, ebaluazio metrikaren batez besteko balioak eta desbideratze estandarrak emanik.

Eredua	Asmatze-tasa	+ VQA doikuntza	Parametroak
Q_{BERT}	21,2 ±0,2	23,0 ±0,2	112M
MM_{BERT}	29,6 ±0,6	35,7 ±0,3	114M
CBM_{BERT} (ours)	**32,5** ±0,4	**36,0** ±0,4	112M

3.1 Taula – Gure hiru sailkapen ereduen errendimendua OK-VQA atazan (hurrenez hurren, galderetan bakarrik oinarritutako eredua, irudietan oinarritutakoa eta goiburukoetan oinarritutakoa). Emaitzak VQA atazan aurretik doitu gabe eta doituta daude zatituta zutabeetan. VQA-ren batez besteko puntuazioa eta desbideratze estandarra erakusten dugu 3 entrenamenduetan zehar.

3.3.2 Irudi eta Goiburukoen Erabilera

3.2.1. Atalean aurkeztutako hiru ereduen emaitzak ageri dira 3.1. Taulan. Bertan OK-VQA atazan doitutako ereduen emaitzak aurki daitezke, baita aurretik VQA atazan doitu diren ereduen emaitzak ere. Kontuan izan eredu hauek arkitektura, tamaina eta gure entrenamendu aurreko pisuak partekatzen dituztela.

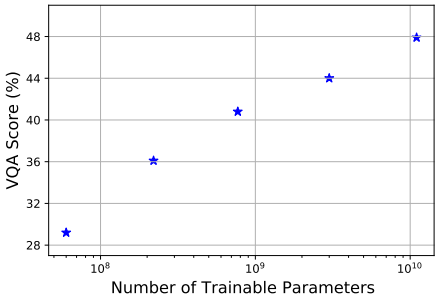
Ikus dezakegunez, ereduak galderekin bakarrik elikatzeak (Q_{BERT}) errendimendu oso baxua ematen du beste bi sistemekin alderatuta, 13 puntu gutxiago arte lortuz. Honek irudiaren edozein adierazpen erabiltzea (bai testuala eta baita bisuala ere) funtsezkoa dela erakusten du galderei zuzen erantzuteko. Gainera, VQA-ren aurrentrenamenduak dakarren hobekuntza eredu ezberdinen artean konparatzeak gehiago bermatzen du esandakoa. Izan ere, Q_{BERT} ereduak 2 puntu baino gutxiago hobetzen du aurrentrenamendu honen ondorioz, eta beste biek, berriz, 4-6 puntu gehiago lortzen dituzte.

Goiburukoen ekarpena. CBM_{BERT} eta MM_{BERT} ereduen errendimendua konparatzen dugunean, ikusmen-testu aurrentrenamendurik ez dagoenean CBM_{BERT} hobeto dabilela ikus dezakegu OK-VQA atazan. Hala ere, eredu horiek antzeko ataza multimodal batean entrenatzen ditugunean (kasu honetan, VQA 2.0) bien asmatze-tasak 4-6 puntu handitzen dira, antzeko emaitzak lortuz.

OK-VQA atazaren entrenamendu instantzia kopurua oso txikia da VQA-rekin konparatzen badugu (9K vs. 410K instantzia). Gure ustez, OK-VQA atazan MM_{BERT}-en doikuntza ez da nahikoa sarrera modalitate berrira egokitzeko. Hala ere, CBM_{BERT}-ek testu hutsa erabiltzen duenez, entrenamendu txikiarekin doikuntza burutzea eraginkorragoa dela ikusi dugu.

Ereduak	Asmatze-tasa	Param.
$\text{CBM}_{\text{T5-Small}}$	29,2 ±0,2	60M
$\text{CBM}_{\text{T5-Base}}$	36,1 ±0,5	220M
$\text{CBM}_{\text{T5-Large}}$	40,8 ±0,4	770M
$\text{CBM}_{\text{T5-3B}}$	44,0 ±0,7	3B
$\text{CBM}_{\text{T5-11B}}$	**47,9** ±0,2	11B

3.2 Taula – CBM_{T5} eredu sortzaileen errendimendua OK-VQA atazan.



3.4 Irudia – CBM_{T5} ereduen tamaina eta hauen errendimenduen arteko korrelazioa. Ardatz horizontala eskala logaritmikoan dago.

3.3.3 Hizkuntza-ereduen Tamaina

T5 ereduak aldez aurretik galderei erantzuteko hainbat atazetan doitu direnez, zuzenean OK-VQA atazan burutzen dugu doikuntza, hots, ez dugu beste ikusmentestu ataza batean doikuntzarik burutzen aldez aurretik.

3.2. Taulan, OK-VQAn doitu eta ebaluatutako tamaina ezberdineko CBM_{T5} ereduen emaitzak ageri dira. Kontuan hartu $\text{CBM}_{\text{T5-Base}}$ eta VQA-n doitutako CBM_{BERT} ereduek emaitza konparagarriak lortzen dituztela. Esperotako emaitzak dira; izan ere, bi ereduak galdera-erantzute atazetan entrenatu dira parametro kopuru konparagarriak izanik. Azken finean, $\text{CBM}_{\text{T5-Base}}$ bi BERT-base ereduren pareko diren kodetzaile eta deskodetzailez dago osatuta.

3.2. Taulako emaitzak 3.4. Irudian agertzen dira, gure ereduen tamaina OK-VQA atazan duten errendimenduarekiko logaritmikoki proportzionala dela erakutsiz. Izan ere, emaitza hauek Kaplan *et al.*, 2020 lanean aipatutako eskalatze legeak betetzen dituzte. Gure eredurik handienak ere joera hori jarraitzen du, eta ez dirudi oraindik hobekuntza hau moteltzen ari denik. Emaitza hauek ereduaren tamainak bere gaitasunean duen garrantzia erakusten dute. Eredu guztiak corpus berarekin aurrentrenatu eta ataza berdinetan doitu dira. Hala ere, tamaina ezberdintasunak eredu handienei laguntzen die, corpus horretatik ateratako informazioa hobeto aprobetxatuz eta OK-VQA ebazteko behar den kanpoko ezagutza barneratuz. Hau horrela izanik, kontuan hartzekoa da gure eredu handienak modalitate anitzeko ereduak baino askoz errendimendu altuagoa erakusten duela.

Eredua	Asmatze Tasa		Param.
ConceptBERT (Gardères *et al.* 2020) *	31,4	(+sym. 33,7)	348M
MAVEx (Wu *et al.* 2022)	35,2	(+sym. 41,4)	353M
KRISP (Marino *et al.* 2021)	37,1	(+sym. 38,9)	116M
RVL (Shevchenko *et al.* 2021) *†	37,3	(+sym. 39,0)	208M
PICa-Base (Yang *et al.* 2022)	42,0	(+tags 43,3)	175B
PICa-Full (Yang *et al.* 2022) (Ensemble)	46,9	(+tags **48,0**)	175B
CBM_{BERT} (ours)	36,0		112M
$CBM_{T5\text{-}11B}$ (ours)	**47,9**		11B

3.3 Taula – OK-VQA atazaren artearen egoera. +sym. ereduari ezagutza sinbolikoa txertatu zaiola adierazten du. +tags etiketak, berriz, objektu etiketen erabilera adierazten du (goiburukoekin batera). * duten ereduen emaitzak OK-VQA v1.0 erabiliz daude kalkulatuta, eta † ikonoak kutsadura arazoak adierazten du (ikusi testuan).

Izan ere, ez dago argi modalitate anitzeko eredu handiagoek gure CBM_{T5} handienaren emaitzetara iristeko gaitasuna lortuko dutenik. Lan hau garatzean ezin izan genuen hipotesi hori frogatu, une horretan ez baitzegoen tamaina bereko eredu konparagarririk publikoki eskuragarri. Hala ere, ezagutza behar handiak dituzten atazetan, hala nola OK-VQA atazaren kasuan, egungo modalitate anitzeko ereduek (Lu *et al.*, 2019; Li *et al.*, 2019; Tan and Bansal, 2019) gure CBM ereduak baino okerrago ibiliko direla uste dugu. Izatez, eredu hauek aurrentrenatzerakoan erabilitako testuak goiburukoez edota irudiei lotutako deskribapen txikiz osatuta daude. Corpus mugatu batetik hiztegi eta ezagutza mugatu bat ikasteko aukera dute soilik, T5 bezalako ereduak eraikitzeko erabiltzen diren corpus aberatsago eta askoz handiagoen kasuan ez bezala.

3.3.4 Artearen Egoerarekin Konparaketa

3.3. Taulan artearen egoerako ereduren emaitzak ageri dira hiru multzotan banatuta: i) modalitate anitzeko transformer-etan oinarritutako sailkapen ereduak, ezagutza sinbolikoaren erabilera gehitzen dituztenak; ii) GPT-3 eta testuinguru bidezko ikasketan oinarritutako eredu sortzaileak; iii) gure goiburukoetan oinarritutako ereduak.

Sailkapen ereduen artean KRISP (Marino *et al.*, 2021), MAVEx (Wu *et al.*, 2022) eta RVL (Shevchenko *et al.*, 2021) ereduek antzeko errendimendua erakusten dute ezagutza inplizitua bakarrik erabiltzen dituzten aldaeretan, aurrentrenamenduko ataza eta modalitate anitzeko eredu ezberdinetan oinarrituta badaude ere. Kontuan izan behar dugu RVL-ek kontaminazio arazo bat duela, OK-VQA atazako ebaluazio irudiak bere aurrentrenamenduan erabili baitira. Orokorrean ezagutza sinbolikoa gehitzeak 2 puntuko hobekuntza ekartzen duela antzeman dugu, salbuespena MAVEx izanik. Izan ere, MAVEx-en kasuan kanpo ezagutza hainbat iturrietatik eskuratzen du, ConceptNet (Speer *et al.*, 2017), Wikipedia and Google Images-eko ezagutza elkartuz.[3]

PICa (Yang *et al.*, 2022) GPT3 ereduaz baliatzen da (Brown *et al.*, 2020) artearen egoera berri bat ezartzeko. Aurreko ereduek ez bezala, testua sortzen du sailkapena burutu beharrean, eta testuinguru bidezko ikasketaz baliatzen da. Bere oinarrizko ereduaren emaitzak (PICa-Base) jada ikusitakoak baino hobeak dira ezagutza sinbolikoaren gehigarririk gabe. Bi teknika ezberdin aplikatuz are gehiago hobetzen dituzte emaitzak (PICa-Full): i) 5 GPT-3 ereduren bateratzea burutzea, eta ii) testuinguruan erabilitako adibideen hautaketa egiteko heuristiko batzuk erabiltzea.

3.3. Taulan bi emaitza azaltzen dira PICa eredu bakoitzeko: i) automatikoki sortutako goiburukoak bakarrik erabiliz lortutako emaitzak (gure kasuan bezala), eta ii) automatikoki lortutako objektu etiketak gehitzen dituztenenak, hobekuntza txikiak erakusten dituztenak.

Gure CBM_{BERT} sistemak modalitate anitzeko transformer-en pareko errendimendua erakusten du. Aipagarria iruditzen zaigu, gure ereduetan ez baitugu zuzenean inolako irudi ezaugarririk erabiltzen, goiburukoak baino ez. Kontuan izan eredu guzti hauen tamainak konparagarriak direla. Gainera, gure eredu generatiboaren tamaina handitzen badugu, gaur egungo modalitate anitzeko ereduak gainditzen ditugu, PICa-Full ereduaren pareko emaitzak lortuz. Izan ere, $\text{CBM}_{\text{T5-11B}}$-ek goiburukoak bakarrik erabiltzen dituen PICa-Full baino emaitza hobeak lortzen ditu, gure eredua 15 aldiz txikiagoa bada ere.

3.3.5 Analisia

Atal honetan hainbat esperimentu gehiago burutzen ditugu. Lehenik eta behin, OK-VQA atazan lortutako emaitzak VQA 2.0 atazakoarenekin konparatzen ditu-

[3]Emaitza hau 3 MAVEx ereduen bateratzearekin lortu da, hirurek modalitate anitzeko transformer bera partekatzen dutelarik. MAVEx eredu batek 40,3ko asmatze-tasa lortzen du.

Eredua	Asmatze-tasa
MM_{BERT}	**65,8**
PICa-Full	56,1
CBM_{BERT} (ours)	59,6

3.4 Taula – MM_{BERT} eta testuzko sarrera bakarrik jasotzen duten bi ereduren (PICa-Full eta CBM_{BERT}) errendimendua VQA 2.0 atazako garapen azpimultzoan.

gu, ezberdintasunak arrazoituz. Ondoren, CBM_{BERT} eta MM_{BERT} batzen ditugu beraien arteko osagarritasuna aztertzeko. Jarraian, CBM_{BERT} gizakiek idatzitako goiburukoekin doitzen dugu, OSCAR (Li *et al.*, 2020) ereduarekin lortutakoekin konparatuz. Azkenik, analisi kualitatibo bat egiten dugu sortutako erantzunen gainean.

Emaitzak VQA 2.0 atazan

OK-VQA atazan modalitate bakar eta anitzeko ereduak antzeko emaitzak lortzen dituzte. VQA 2.0-n, berriz, beste joera bat ikusten dugu. 3.4. Taularen arabera CBM_{BERT}-ek 59,6ko asmatze-tasa lortzen du, eta MM_{BERT}-ek, berriz, 6 puntu gehiago. Gure ustez, irudi bat goiburuko bihurtzean informazioa galtzen delako gertatzen da hori, galderari erantzuteko behar den informazioa berbalizazio prozesuan gal daiteke eta. Hori bereziki garrantzitsua da VQA 2.0-rako, galdera gehienak irudiaren edukiari, erlazio espazialei eta objektuen atributuei buruzkoak baitira (ikus 3.3. Irudia). PICa-ren kasuan antzeko jokaera antzeman dezakegu. Eredu honek objektu etiketak erabiltzen ditu ere bai, berbalizazioan daukagun informazio galera minimizatzeko. Hala ere, ez du gure sistemak bezain ondo funtzionatzen. Are gehiago, ereduaren parametro kopurua 1.000 aldiz txikiagoa bada ere, gure CBM_{BERT}-ek PICa gainditzen du. Honek entrenamendu kopuru handiak eskura daudenean doikuntza egiteak daukan garrantzia erakusten du, testuinguru bidezko ikasketarekin alderatuz behintzat, VQA 2.0-n bezala.

VQA eta OK-VQA atazen arteko errendimendu ezberdintasunak adierazgarriak dira. Izan ere, OK-VQA atazan goiburukoek informazio nahikoa ematen digutela iradokitzen du ezberdintasun horrek, hau da, ezagutza behar handiko modalitate anitzeko atazetan behar adina informazio ematen digutela. Hala ere, erantzuna irudian aurki daitekeen galdera-erantzute sistemetan modalitate anitzeko transformer ereduak egokiagoak direla dirudite.

Eredua	Asmatze tasa	+ VQA Doikuntza
Fusio Goiztiarra	32,5 ±0,4	38,2 ±0,8
Fusio Berantiarra	34,0 ±0,4	**38,6** ±0,2

3.5 Taula – Fusio goiztiar eta berantiar ereduen errendimendua OK-VQA atazan.

Informazio bisuala eta goiburukoak nahasten

Modalitate ezberdinek kodetzen duten informazioa ezberdina dela eta, CBM_{BERT} eta MM_{BERT} osagarriak diren aztertu nahi izan dugu. Gure hipotesia CBM_{BERT}-ek hizkuntza-ereduak bereganatutako ezagutza inplizitua aprobetxatu dezakeela da. MM_{BERT}-ek, aldiz, irudi eskualdeetan aurki daitezkeen xehetasunak adierazi eta erabiltzeko aukera du. Beraz, osagarritasuna aztertzeko bi fusio definitu ditugu.

Fusio goiztiarra. Galdera bakoitzeko irudiaren goiburukoa eta ezaugarriak elikatzen dizkiogu hizkuntza-ereduari galderarekin batera. Sistema hau galdera batek (testua), goiburuko batek (testua) eta irudiaren eskualdeko ezaugarriek osatutako sarrera multimodal bat prozesatzen duen MM_{BERT} eredua bezala ikus daiteke. Eredu hau BERT-base baten pisuekin hasieratzen dugu eta doikuntza ohiko entrenamendu instantziekin burutzen dugu.

Fusio berantiarra. Kasu honetan, CBM_{BERT} eta MM_{BERT} ereduak bakoitza bere aldetik entrenatzen ditugu, 3.2.1. Atalean zehaztutako sarrerekin. Horrela, irteerak inferentzia denboran konbinatzen ditugu azken erantzuna lortzeko. Konbinazioa egiteko, bi ereduen irteera probabilitateak biderkatzen dira klase bakoitzeko, irteerako probabilitate handienarekin geldituz.

3.5. Taulan bi fusio eredu hauen errendimendua agertzen da. CBM_{BERT} eta MM_{BERT} ereduen errendimendua bizpahiru puntutan igotzen dute ia kasu guztietan. CBM_{BERT}-ekin alderatuta hobekuntzarik ez dagoen kasu bakarra VQA aurre entrenamendurik gabeko fusio goiztiarrarena da. OK-VQA atazaren entrenamendu instantzia kopuru txikiak eragina izan duela uste dugu, zailtasunak sortuz modalitate testual eta bisualen arteko zubia ikasteko. Hala ere, VQA aurrentrenamendua gehitzean ereduek ikusitako datu kopurua izugarri handitzen da, eta antzeko portaera erakusten dute bai fusio goiztiarreko ereduak eta baita berantiarrak ere.

Bi modalitateen osagarritasuna VQA atazan aztertu da ere bai. Fusio goiztiarrak 67,8 puntuko asmatze-tasa lortzen du VQA 2.0 atazako garapen azpimultzoan, eta fusio berantiarrak, berriz, 67,7 puntu lortzen ditu MM_{BERT}-en errendimendua 2 puntutan hobetuz. Emaitzek gure hipotesia balioztatzen dute, agertoki honetan irudi eskualdeko ezaugarriak eta goiburukoak osagarriak direla erakutsiz.

Giza goiburukoak

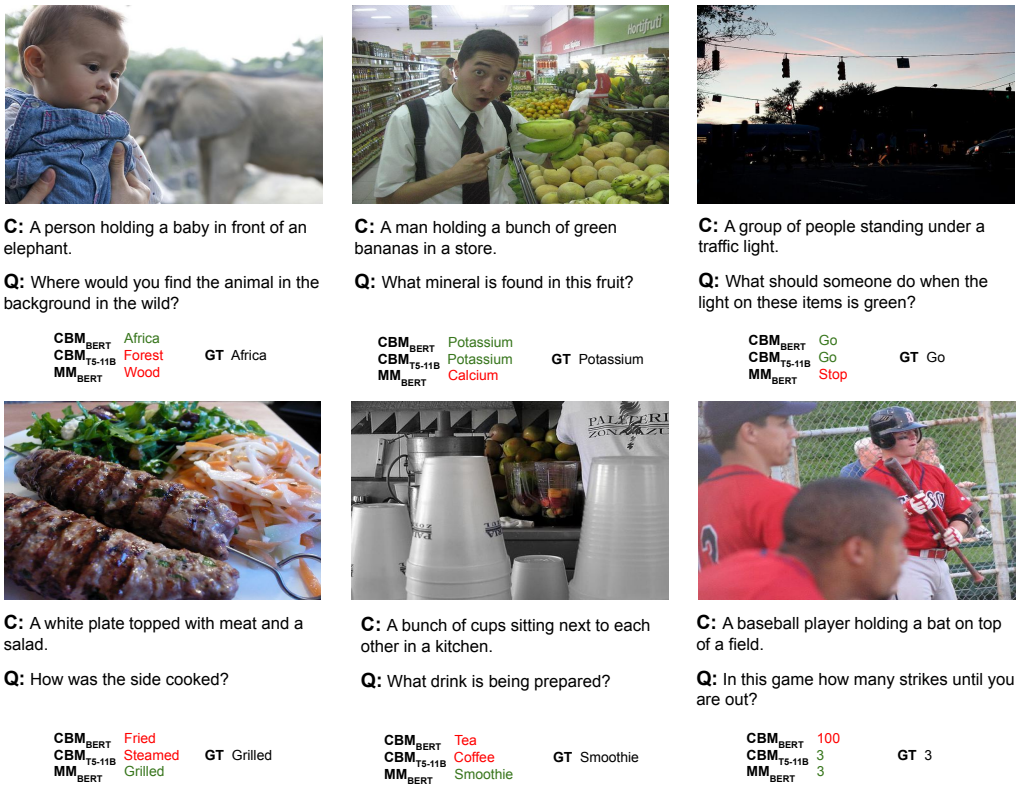
Proposatutako CBM ereduan goiburuko sortzaileak dituen eraginak neurtzeko, gizakiek idatzitako eta OSCAR bidez sortutako goiburukoen arteko aldea erakusten dugu. OK-VQA ataza COCO (Lin *et al.*, 2014) datu-multzoko irudien gainean eraikitzen denez, irudi bakoitzak bost goiburuko ezberdin ditu. CBM_{BERT} doitzeko goiburuko hauek erabiltzen ditugu OK-VQA atazan, VQA aurrentrenamendurik gabe eta aurreko esperimentuen ezarpen eta hiperparametro berdinak jarraituz. Esperimentu bakoitza hiru aldiz errepikatzen dugunez, goiburuko ezberdin bat aukeratzen dugu errepikapen bakoitzeko, goiburuko bera entrenamendu osoan zehar erabiliz. OK-VQA atazako ebaluazio instantzien irudi bakoitzeko bost goiburuko ditugunez, eredu bakoitza hiru aldiz ebaluatzen dugu goiburuko aukeraketa prozesu bera jarraituz.

3.1. taulak dagoeneko erakusten du $32{,}5 \pm 0{,}4$ko asmatze-tasa eta desbideratze estandarra lortzen dugula, OSCAR-ekin sortutako goiburukoak erabiliz. Hala ere, giza goiburukoak erabiltzen ditugunean batezbeste $32{,}3 \pm 0{,}3$ko asmatze-tasa lortzen dugu. Bi kasuetan antzeko emaitzak lortzen ditugu, eta OSCAR bidez sortutako goiburukoek CBM_{BERT} ereduak behar adina informazio dutela erakusten dute.

Analisi Kualitatiboa

Bai CBM_{BERT}-ek bai MM_{BERT}-ek antzeko emaitzak lortzen dituzte VQA aurrentrenamenduarekin (ikus 3.1. Taula), baina ebaluazio azpimultzoko instantzien $\%38{,}7$an bi ereduen erantzuna ezberdina da, horietako bakarra zuzena izanik. 3.5. Irudian kasu hauetako adibide batzuk ipini ditugu. $\text{CBM}_{\text{T5-11B}}$-ren erantzunak ere gehitu ditugu aurreko emaitzekin konparatzeko.

Goi ezkerreko adibidetik hasita, CBM_{BERT}-ek elefanteak Afrikakoak direla ondoriozta dezakeela ikus dezakegu; MM_{BERT}-ek, aldiz, ez. Hain zuzen, sortutako goiburukoak irudian aurkitutako animalia elefante bat dela adierazten du, eta galderari erantzuteko behar den lehen urratsa egiten du. Horrela, hizkuntza-eredua bere ezagutza inplizitua erabiltzera bideratu dezakegu nahi dugun erantzuna lortzeko. CBM_{T5}-ek *forest* sortzen du erantzun gisa. Erantzuna guretzat baliagarria izan daitekeen arren, erantzuna ez dago balizko erantzunen zerrendan, eta ez dugu ontzat hartzen. Goiko errenkadako beste bi adibideek antzera jokatzen dute. Goiburukoak galderaren eta irudiaren arteko lotura errazten du. Galdera bat irudiari buruzkoa denean (“this fruit” eta “these items", hurrenez hurren) eta goiburukoak objektu horiek aipatzen baditu (“bananas"eta “traffic light”), hizkuntza-ereduak



3.5 Irudia – OK-VQAko instantziak non CBM_{BERT} eta MM_{BERT} ereduen arteko batek bakarrik ondo erantzuten duen. $\text{CBM}_{\text{T5-11B}}$ek itzulitako erantzunak konparazio gisa jarri ditugu. **C** OSCARrek sortutako goiburukoak dira. Erantzun zuzena berdez dago, okerra, berriz, gorriz.

hobeto aprobetxatzen du bere ezagutza inplizitua eta arrazoitzeko duen gaitasuna galderari erantzuteko. Puntu honekin jarraituz, goiko eskuineko adibidea interesgarria da. Izan ere, irudiak semaforo gorriak erakusten dituen bitartean, argi berdeen eraginei buruz galdetzen da. Horrek MM_{BERT} engainatu dezake argi gorriek (eta ez berdeek) eragiten duten efektuari erantzuteko baldintzatuz.

3.5. Irudiko beheko errenkadak goiburukoak informazio nahikoa ematen ez dituen bi adibide erakusten ditu. Lehenengo kasuan, CBM ereduek ez dakite zehazten ea haragia frijituta, parrilan eginda ala lurrunean egosita dagoen, goiburukoak ez baitie galdera horri erantzuteko informazio nahikoa ematen. Hala ere, MM_{BERT}-ek irudiaren seinale bisualak atzitu ditzake, haragia parrilan egin dela antzeman dezakeelarik. Bigarren adibidean ere antzeko gauza bat gertatzen

da. Goiburukoak edariaren osagairik zehazten duen bitartean, irudian frutak ikusten ditugu. Eskuineko adibideko goiburukoak inferentziarako beharrezkoa den informazioa dauka, baina CBM_{BERT}-ek erantzun okerra itzultzen du. Goiburuko horrekin, "this game" hitzekin beisbolari buruz hitz egiten ari dela deduzitu dezakegu. Hala ere, CBM_{BERT}-ek ezin izan du erantzun hiru *strike* nahikoak direnik jokalari bat kanporatzeko; $CBM_{T5\text{-}11B}$ and MM_{BERT}-ek, berriz, erantzun zuzena ematen dute.

Oro har, adibide horiek ezaugarri bisualak eta goiburukoak osagarriak direnaren hipotesiari eusten diote. Halaber, gure ereduak abantailak erakusten ditu ereduaren interpretagarritasunari dagokionez, bereziki gure metodoa oker dagoen kasuetan antzeman daitekeena. Kasu batzuetan, 3.5. Irudiko behe ezkerreko bi adibideetan bezala, galderari erantzuteko behar den objektua edo ezaugarria ez dago goiburukoan. Beste kasu batzuetan, eskatutako informazioa goiburukoan dago, baina inferentzia okerra da.

3.4 Ondorioak

Kapitulu honetan VQA sistema bat aurkezten dugu, goiburuko baten bidez irudiak deskribatzen dituena eta gero testu datuekin bakarrik lan egiten duena. Eredu honek OK-VQA atazan emaitza oso onak lortzen dituela ikusi dugu, iruditik kanpo dagoen ezagutza eskuratzeko beharra duten galderetan abantaila erakusten duelarik. Gure analisiak irudiak berbalizatzean dagoen informazio galera testua bakarrik prozesatzen duten hizkuntza-ereduen inferentzia gaitasun handiagoarekin orekatzen dela erakusten du. Era berean, hizkuntza-eredu batek duen gaitasunaren garrantzia ere erakusten dugu, bertan dagoen ezagutza inplizitua aprobetxatuz, artearen egoeraren emaitzak lortuz, gaur egungo modalitate anitzeko ereduen errendimendua alde handiarekin gaindituz eta 15 aldiz handiagoko GPT-3 ereduaren emaitzak berdinduz. Modalitate anitzeko ereduekin alderatuta, publikoki atzigarri dauden hizkuntza-ereduen tamaina edota gaitasuna askoz handiagoa da, ezagutza behar handiko atazetan onuragarria dela erakusten dugularik.

Etorkizunean, irudien deskribapen aberatsagoek emaitzak are gehiago hobetu ditzaketen aztertu nahiko genuke. Gainera, ezagutza grafo sinbolikoak testuarekin bakarrik lan egiten duten hizkuntza-eredu handietan txertatzea ikertu nahi dugu. Izan ere, OK-VQA atazako sailkapen ofizialean[4] dauden lehenengo postuetako ereduek ez dute kanpo ezagutza iturririk erabiltzen, eredu hauek hizkuntza-eredu handietako ezagutza inplizitua erabiltzen baitute ezagutza iturri gisa (Shao *et al.* 2023; Hu *et al.* 2022).

[4] https://okvqa.allenai.org/leaderboard.html. (2024/07/06ean atzituta)

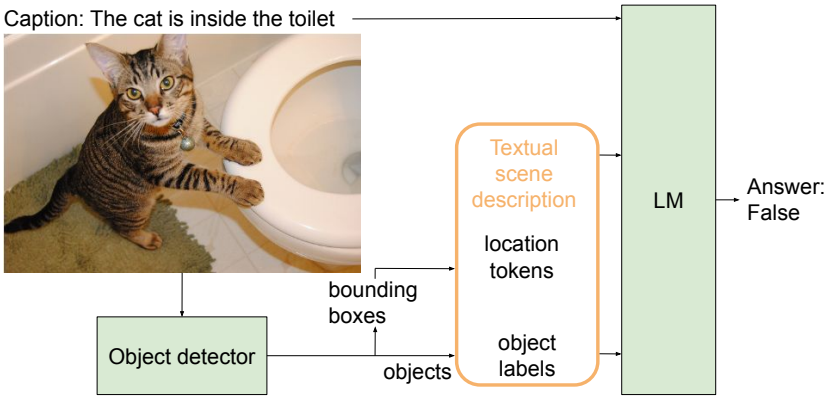
4. KAPITULUA

Arrazoinamendu Espaziala Ikasten Hizkuntza-ereduetan

4.1 Motibazioa eta Ekarpenak

Ezker eta *azpian* bezalako erlazio espazialak era naturalean oinarritu daitezke irudietan. Hau horrela, ikusizko hizkuntza-ereduak (*vision-and-language models* edo VLM) testua mundu errealeko kontzeptuetan oinarritzeko aukerarik egokiena dirudite. Hala eta guztiz ere, CLIP (Radford *et al.*, 2021), VisualBERT (Li *et al.*, 2019), LXMERT (Tan and Bansal, 2019) edo ViLT (Kim *et al.*, 2021) bezalako xede orokorreko VLM-ek erlazio espazialen gainean arrazoitzeko arazoak dituztela antzeman da (Liu *et al.*, 2022b: 2023). Egoera are okerragoa da testua soilik prozesatzen duten hizkuntza-ereduentzat, non erlazio espazialen arrazoinamenduan VLM-en atzetik geratzen diren (Liu *et al.*, 2022b).

Oinarritze eta arrazoinamendu espazialak oso interesgarriak dira testua soilik erabiltzen den atazetan (Liu *et al.*, 2022b; Mirzaee *et al.*, 2021; Mirzaee and Kordjamshidi, 2022). Testu hutsezko ataza horiek ebazteko alternatiba bat VLM-ak testu sarrerarekin soilik elikatzea izango litzateke. Hala ere, hurbilpen hau jada landu da eta eraginkorra ez dela ikusi da (Tan and Bansal, 2020). Azken finean, VLM hauek ez dira testu hutsezko atazetan azaltzen diren testu aberats eta anitzetan entrenatu, eta horrek VLM-en ahalmena oztopatzen du testu hutsezko atazak ebazterako garaian.



4.1 Irudia – Erlazio espaziala duen goiburuko bat eta dagokion irudi bat emanda, VSR atazan goiburukoa irudiari dagokion ala ez zehaztu behar da. Datu multzoaren testu hutsezko alternatiba bat proposatzen dugu, non objektu detektore batek irudian azaltzen diren objektuen etiketak eta kokapenak itzultzen dituen (kaxa inguratzaileetatik eratorritakoak). Informazio hau irudietan azaltzen diren eszenen testuzko deskribapen gisa erabiltzen dira. Deskribapen hau eta goiburukoa hizkuntza-eredu batera elikatuz, hizkuntza-ereduen oinarritze espazialeko gaitasunak aztertu ditzakegu.

Kapitulu honetan, beste bide bat aztertzen dugu eta testu hutsezko hizkuntza-ereduen oinarritze espazialean zentratzen gara. Ikusizko informazioa testura itzultzeko egungo joera jarraituz (Yang *et al.*, 2022; Zeng *et al.*, 2022a; Wang *et al.*, 2022b; Liu *et al.*, 2022a), testu tokenak era berri batean erabiltzea proposatzen dugu, mundu errealeko eszenak irudikatzeko eta aurrentrenatutako hizkuntza-ereduak hobeto aprobetxatzeko. Xehetasunetan sartuz, kokapen tokenak erabiltzea proposatzen dugu eszena bateko objektuen posizioak eta tamainak zehazteko.

Erabilitako kokapen tokenak dagoeneko hizkuntza-ereduaren hiztegian dauden zenbakien tokenak erabiliz definitzen dira, hots, tokenizatzailean zehaztutako hiztegian agertzen diren tokenak erabiliz. Horrela, objektu baten posizio eta tamaina zehazteko lau kokapen token eta objektuaren izena erabiltzen ditugu (objektuaren hainbat atributu gehitu ditzakegularik ere bai). Irudiaren testuzko adierazpen honi esker, hizkuntza-ereduek *ezker*[1] bezalako erlazio espazialak dagozkien kokapen token sekuentziekin erlaziona ditzakete, erlazio horiek oinarritzeko modua eskainiz.

[1]Hemendik aurrera erabilitako erlazio espazial esplizituak ingelesez azaltzen dira, VSR-ren goiburukoekin bat egiteko.

Gure hipotesia kokapen token horiek hizkuntza-ereduen erlazio espazialak oinarritzeko modu eraginkor bat ahalbidetzen dutela da. Hau balioztatzeko, *Visual Spatial Reasoning* (VSR) datu-multzoaren bertsio berbalizatu batean egin ditugu esperimentuak (Liu *et al.*, 2023). Datu multzoak irudi eta goiburuko pareak ditu, non goiburukoak irudian agertzen diren bi objekturen arteko erlazio espazial bat zehazten duen. Horiekin batera, etiketa boolear bat dator, goiburukoa irudian betetzen den ala ez zehazten duena.

4.1. Irudiak erakusten duen moduan, VSR ataza honela lantzen dugu: (i) objektu detektore bat erabiliz eszenaren testu deskribapenak lortzen ditugu, (ii) deskribapen horietan kokapen tokenak gehitzen ditugu, objektu detektoreak antzemandako kaxa inguratzaileen informazioaz baliatuz, (iii) goiburukoa eta eszenaren deskribapen testuala kateatzen ditugu eta lortutako token sekuentzia hizkuntza-ereduari elikatzen diogu, (iv) hizkuntza-eredua sailkapen bitarra burutzeko doitzen dugu. Gainera, aldez aurretik hizkuntza-eredua trebatzeko aukera eskaintzen dugu guk garatutako datu-multzo espazial sintetikoan (*Synthetic Spatial Training Dataset* edo SSTD).

Gure esperimentuen ondorioz, ondorengo ekarpenak egiten ditugu:

1. Kokapen tokenak eraginkorrak dira erlazio espazialak oinarritzeko, gure ereduaren emaitza positiboek erakusten duten bezala.

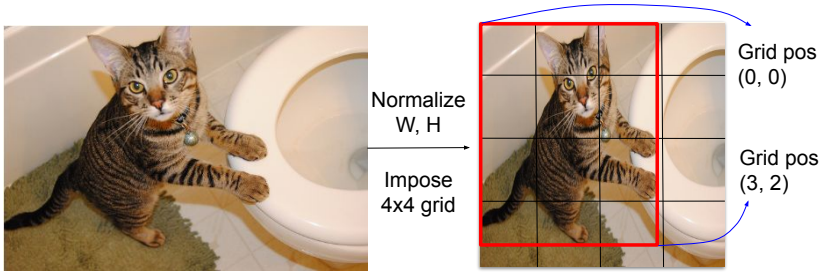
2. VSR datu-multzoaren entrenamendu azpimultzoa txikiegia da erlazio espazial eta objektuen kokapenen arteko lotura ikasteko. Hala ere, automatikoki sortutako SSTD datu-multzoak horretarako aukera ematen du. Bitartean, hizkuntza-ereduak kokapen informaziorik gabe erabiltzeak huts egiten duela erakutsi da.

3. Datu multzo sintetikoan trebatutako hizkuntza-ereduak hein batean orokortu daitezke datu sintetikoetan aztertu ez diren erlazio espazialetara. Bereziki aipagarria da sakonera informazioa eskatzen duten erlazioen errendimenduan bultzada hau antzematea.

4. Testuaz bakarrik baliatzen diren gure hizkuntza-ereduek artearen egoerako emaitzak lortzen dituzten VLM-ak gainditzen dituzte VSR atazan.

Gure kodea, ereduak eta datu-multzoak edozeinen eskura utzi ditugu[2] .

[2]URL: https://github.com/gazkune/SpatialLM



4.2 Irudia – BB koordenatuak kokapen token bihurtzeko adibide bat. Kasu honetan, 4×4ko sareta tamaina zehazten dugunez, katuaren (kaxa gorria) kokapen tokenak $(0, 0, 3, 2)$ dira.

4.2 Metodologia

Atal honetan, *kokapen tokenen* kontzeptua definitzen dugu hizkuntza-ereduetan erlazio espazialen oinarritzea burutzeko. Gainera, erabili ditugun datu-multzoak zehazten ditugu, baita maneiatutako ereduak eta hauek doitzeko ezarpenak ere.

4.2.1 Testuzko Deskribapen Espazialak

VSR irudi-testu datu-multzo bat dela kontuan hartuta, testuz deskribatu nahi dugun eszena irudi batek definitzen du. Eszena hori testu deskribapen batean adierazten dugu artearen egoera den objektu detektore bat erabiliz, VinVL (Zhang *et al.*, 2021). Irudi bat emanik, VinVL-ek objektuen zerrenda bat sortzen du hauen izena, atributu eta kaxa inguratzailearen koordenatuekin. Gehiago zehaztuz, VinVL-ek detektatutako objektu bakoitza $O = \{name, attr_1, \ldots, attr_n, BB\}$ gisa adierazten da, non $BB \in \mathbb{R}^4$ kaxa inguratzailearen koordenatuak diren. $BB = \{x_0, y_0, W, H\}$ kaxaren goi ezker erpineko koordenatuak, zabalera eta alturaren balio normalizatuak dira.

Ondorengo prozedura jarraitu dugu BB horiek kokapen token bihurtzeko (ikus 4.2. Irudia): i) irudiaren zabalera eta altuera normalizatu $[0, 1]$ tartean, ii) irudia zatitu $(G \times G)$ tamainako sareta erregular batean, eta iii) (x_0, y_0, x_1, y_1) balioak saretan dauden gelaxkaren indizeak erabiliz diskretizatu $(\hat{x}_0, \hat{y}_0, \hat{x}_1, \hat{y}_1)$ koordenatu diskretuak lortzeko. Koordenatu diskretu horiek tokenizatzailetik pasatzean bihurtzen dira kokapen tokenetan. Ondorioz, detektatutako objektu bakoitzeko lau kokapen token edo koordenatu diskretuen sekuentzia lortzen dugu. Beraz, gure irudiaren testu deskribapena $Descr(S) = \{O_0, O_1, \ldots O_N\}$ objektuen ize-

Caption: The person is ahead of the cow.
Label: True.

Caption: The cat is inside the toilet.
Label: False.

4.3 Irudia – VSR datu-multzoko bi instantzia.

nak, posizioak eta tamainak definitzen dituen sekuentzia bat da. Sekuentzia horretan objektu bakoitzaren deskripzioak honako formatua jarraitzen du: $O_i = \{\hat{x}_0^i, \hat{y}_0^i, \hat{x}_1^i, \hat{y}_1^i, name_i\}$. Kontuan izan VinVL-ek objektu bakoitzaren atributuen zerrenda ere itzultzen duela. Kontrakoa adierazi ezean, atributu horiek irudiaren testu deskribapenean baztertzen ditugu.

VSR atazarako irudi guztien testu deskribapenak sortzen ditugu. Horrela, datu-multzoan emandako goiburukoekin kateatzen ditugu eta hizkuntza-ereduari elikatzen dizkiogu. Gaur egungo hizkuntza-ereduek defektuz dauzkaten posizio bektoreei esker, hizkuntza-ereduek kokapen tokenen ordena eta objektuen izenekin duten korrespondentzia behar bezala interpretatzen ikasi dezakete. Adibidez, 4.2. Irudiko katuaren irudian, *cat* objektuaren deskribapen testuala hau izango litzateke: *0 0 3 2 cat*. Adibide honetan gure sareta tamaina $G = 4$ dela jakinda, irudiaren ezkerreko aldea estaltzen duen katu bat dela interpretatzen da. Irudiaren objektu guztiekin antzera egingo genuke gure eszenaren testu deskribapena eraikitzeko.

Kontuan izan VSR-ren kasuan eszena irudi batekin definitzen dela. Baina, orokorrean, testu edota grafoak bezalako beste modalitate batzuetatik erator genitzake. Esate baterako, eszena baten testu deskribapen naturala emanda (adibidez, "katu bat mahai baten gainean dago"), eszenaren testu deskribapenak lor litezke kokapen tokenak gehituz. Hala ere, lan honen irismenetik at utzi dugu, ez baitugu datu-multzo egokirik aurkitu honetarako. Ikus 4.4. Atala etorkizunean aurreikusita dugun ikerketaren ingurukoak jakiteko.

Bertsioa	Entrenamendua	Garapena	Ebaluazioa	Guztira
random	7.083	1.012	2.024	10.119
zero-shot	5.440	259	731	6.430

4.1 Taula – VSR datu-multzoko instantzia kopuruak.

4.2.2 Erlazio Espazialen Datu Multzoak

VSR Datu Multzoa

VSR datu-multzoak irudi eta testu pare naturalak ditu, ikusmen-testu ereduen oinarritze gaitasunak aztertzea ahalbidetzen duena. 4.3. Irudian ikus daitekeenez, irudi baten testuzko deskribapena daukagu eskura, non irudiko bi objekturen arteko erlazio espaziala esplizituki zehazten den. Erlazio espazial hori egia ala gezurra izan daiteke irudi horretan. Ataza behar bezala ebazteko, ereduek 65 erlazio espazial ezberdin ondo oinarritzen ikasi behar dituzte, 7 kategoriatan multzokatzen direnak erlazioaren arabera: albokotasuna (*adjacency*), norabidea (*directional*), orientazioa (*orientation*), proiektiboa (*projective*), gertutasuna (*proximity*), topologikoa (*topological*) eta bestelakoa (*unallocated*).

Datu multzoak bi bertsio ezberdin ditu: *random* eta *zero-shot*. Lehenengoan datu-multzorako anotatu diren instantzia guztiak entrenamendu, garapen eta ebaluazio azpimultzoetan zatituta daude, guztira 10.119 instantzia edukirik. Bigarrenean, berriz, instantzia kopurua txikiagoa da, bertsio honen azpimultzo ezberdinetan objektu bera agertzea ekidin baita. Horrela, ataza ebazteko erabili den ereduari ez zaio uzten objektuen arteko agerpenen estatistikak memorizatzen. Murrizketa hau dela eta, *zero-shot* azpimultzoak 6.430 instantzia ditu guztira. 4.1. Taulan bi bertsioen entrenamendu, garapen eta ebaluazio instantzia kopuruak zehazten dira.

VSR datu-multzoaren gainean egindako esperimentuen arabera (Liu *et al.*, 2023), VLM onenak gizakien errendimendutik oso urrun daude. Gizakiek bi bertsioetan 95,4 puntuko asmatze-tasa lortzen duten bitartean, *random* bertsioko eredurik onenak, hau da, LXMERT-ek (Tan and Bansal, 2019), 70,1 ingurukoa lortzen du, *zero-shot* bertsioan oraindik gehiago okertuz (63,0). Gizakien eta VLM-en arteko errendimendu ezberdintasun honek erlazio espazialak hobetzeko oraindik lan asko egiteko dagoela erakusten du. Kontuan izan ausaz erantzuten duen eredu batek 50,0 puntuko asmatze-tasa lortuko lukeela.

Kategoria	Erlazio Espazialak
Objektu baten posizioa	*top left*, *bottom left*, *left*, *top right*, *bottom right*, *right*, *top*, *bottom*, *center*
Bi objekturen arteko tamaina konparaketa	*wider*, *narrower*, *taller*, *shorter*, *larger*, *smaller*
Bi objekturen arteko posizio konparaketa	*surrounding*, *inside*, *left of*, *above*, *right of*, *below*, *overlapping*, *separated*

4.2 Taula – Gure SSTD datu-multzoko 23 erlazioak hiru kategoriatan sailkatuta.

SSTD Datu Multzoa

Erlazio espazial esplizituak dituzten modalitate anitzeko entrenamendu datuak, irudi eta goiburuko pareez osatuta daudenak, urrikak izan ohi dira. Gure hurbilpenaren bigarren osagai gisa erlazio espazialak dituen datu-multzo sintetiko bat eraiki dugu, *Synthetic Spatial Training Dataset* (SSTD). SSTD hizkuntza-ereduei kokapen token eta erlazio espazialen arteko loturak erakusteko dago pentsatuta. Irudi bat emanik, objektu detektore bat erabiltzen dugu testuzko deskribapen bat sortzeko. Deskribapen hau objektu zerrenda eta kokapen tokenen sekuentzia batez osatzen da, objektu bakoitzaren kokapen informazioa berbalizatuta emanik. Bi objektu eta hauen kaxa inguratzaileak kontuan hartuta, arau eta txantiloi sinple batzuk erabili ditugu bi objektuen arteko erlazio espazialei buruzko galdera bitarrak sortzeko, bai positiboak eta baita negatiboak ere. Alternatiboki, objektu bakar baten irudiko posizio absolutuari buruz ere galderak sortzen ditugu. 4.4. Irudiak aurreko pausoak jarraituz sortu dugun adibide bat erakusten du. SSTDren abantaila garrantzitsuenak hauek dira: i) milaka adibide sor ditzakegu, ii) eskulan arina eskatzen du, erlazio bakoitzeko arau eta txantiloiak bakarrik zehaztu behar baitira[3], iii) erraz heda daiteke beste erlazio espazialetara, eta iv) datu-multzo hau testuarekin bakarrik lan egiteko edota ikusmen-testu ataza gisa erabil daiteke.

SSTD eraikitzeko COCO datu-multzoko 2014 bertsioa erabili dugu (Lin *et al.*, 2014). SSTD-n COCO datu-multzoko entrenamendu eta garapen azpimultzoak erabili ditugu, COCO-ko instantzien banaketa errespetatuz. Gizakiek anotatutako kaxa inguratzaileak erabili beharrean, VinVL ereduarekin lortutakoak erabili ditugu, VinVL-ek eskaintzen duen hiztegiaren tamaina COCO-rena baino askoz handiagoa baita (1.848 klase 80 beharrean). VinVL-eko klase kopurua handia-

[3] $\sim$5 orduko lana behar izan dugu lan honetan zehaztutako arauak eta txantiloiak zehazteko.

Q: Is man right of horse?
Descr: 0 3 16 29 horse 14 7 26 31 man 22 6 31 31 baby 21 5 28 10 tree 0 5 23 31 building…
A: Yes.

4.4 Irudia – Irudi batetik SSTD garapenerako sortu dugun instantzia, ondorengoak batzen dituena: galdera (Q), deskribapena (Descr) eta erantzuna (A). Irudia baztertu egiten dugu dena sortu ondoren. Deskribapen partziala erakusten dugu, kasu honetan 44 objektuz osatuta baitago. Kokapen tokenak 32×32ko sareta diskretu koordinatuak dira, objektu bakoitzaren kaxa inguratzailearen koordinatuak zehazten dituztenak. Adibide gisa, zaldiaren kasurako $(0, 3)$ eta $(16, 29)$.

goa denez, COCO-ren klaseen azpiklase kontsidera daitezke. Adibidez, COCO-n *person* klasea aurkitzen dugun bitartean, VinVL-eko detekzioetan *woman*, *man*, *child* edota *girl* aldaerak ditugu. Honek dibertsitate handiagoa ematen dio SSTD-ri. Egia da VinVL-ek errore batzuk gehitzen dituela egindako detekzioetan, bai klase edota kaxa inguratzaileetan. Hala ere, testuarekin bakarrik lanean ibiliko garenez, detekzioen eta irudien lerrokatzeak guztiz zuzenak ez izateak ez digu eragiten. Azken finean, detekzioetako klase eta kaxa inguratzaileekin lerrokatuta dauden erlazio espazial zuzenak sortu nahi ditugu.

SSTD-ko instantziak sortzeko erabili ditugun erlazio espazialen anbiguetaterik gabeko zerrenda bat definitu dugu (Johnson *et al.*, 2018) lanean erabilitakoan oinarrituz. Adibide bat jartzegatik, bi kaxa inguratzaile izanik objektu bat bestearen ezkerretara (*left of*) dagoen erabakitzea ez da anbiguoa. Bestalde, kaxa hauek erabiliz ezin dugu jakin bi objektu hauen arteko gertutasuna, *close to* erlazioaren

kasua izango litzatekeena. Izan ere, bi kaxak gertu badaude ere irudiaren planoan, ez dute zertan sakonera antzekoa izan behar. Honek erlazio batzuetarako irudiak ematen duen testuinguruaren beharra dagoela erakusten du, kaxa inguratzaileetatik at lortu behar dena. Zentzu horretan, ez gara saiatu SSTD-rako erabili ditugun erlazioak VSR-koekin bat etortzea, anbiguetaterik ez edukitzearen helburua hartu dugu aintzat bakarrik eta. Hori dela eta, SSTD oinarritze espaziala behar duten beste atazetarako erabilgarria izan beharko luke. 4.2. Taulan inplementatutako erlazio guztiak eta hauen kategoriak aurki daitezke, erregela sinple batzuen bitartez zehazten direnak (iruzkin gehiago B.1. Eranskinean). Prozesu hau jarraitzen dugu SSTD-ko instantzia bat sortzeko:

1. Irudi bat hartu eta detektatu ditugun objektu kopurua begiratzen dugu. Objektu bat edo biren arteko erlazioak zehaztu ditugunez, 4.2. Taulako hiru kategorien artean ausaz bat aukeratzen dugu detektatutako objektu kopuruaren arabera. Objektu bat bakarrik detektatu ezkero, zuzenean "Objektu baten posizioa" kategoria aukeratzen dugu. Bi objektu edo gehiagoren kasuan, bi objekturen arteko erlazioei pisu gehiago ematen diegu, hots, $\%70$ko probabilitatea esleitzen diegu bi objekturen arteko erlazioei eta $\%30$ objektu batekoei. Kategoria aukeratuta, behar ditugun objektuak ausaz aukeratzen ditugu (kategoriaren arabera objektu bat edo bi).

2. Sortuko dugun galderaren erantzuna baiezkoa ala ezezkoa izango den aukeratzen dugu ausaz, SSTD-n *yes* eta *no* erantzunak orekatuta sortzen ditugula bermatuz. Eskuz zehaztutako berbalizazio txantiloiak erabiliz, galdera sortzen dugu aurreko pausoan aukeratutako kategorian dagoen erlazio espazial bat ausaz aukeratuz. B.1. Eranskinean aurki daitezke txantiloi hauek.

3. Galdera sortu ondoren irudia berbalizatzen dugu, bi berbalizazio ezberdin eraikiz: i) VinVL-ek detektatutako objektu guztien konkatenazioa, objektu bakoitzaren izena bere kokapen tokenekin lagunduz; eta ii) kokapen tokenik gabeko bertsioa, hots, objektu izenez bakarrik osatutako zerrenda. Beste berbalizazio batzuk erraz erabili genitzake, goiburukoak adibidez. Hala ere, alternatiba hauek ez dira interesgarriak gure esperimentuetarako, kokapen token eta erlazio espazial esplizituen arteko oinarritzea aztertu nahi dugu gure esperimentazioan.

4. Hortaz, SSTDko instantzia bakoitza galdera, eszenaren testuzko deskribapena eta erantzun batez dago osatuta, testuzko ereduetan irudia alde batera uzten delarik.

Prozedura hau jarraituz, irudi bakoitzeko instantzia anitz sortu ditzakegu. Zentzu horretan, SSTD-k ez dauka instantzia kopuru zehatz bat: erabiltzaileak zehaztu dezake zenbat instantzia erauzi nahi dituen irudi bakoitzeko. Gure ikasketa espazialean zehar, aro bakoitzean ausazko instantzia bat sortzen dugu COCO datumultzoko entrenamendu irudi bakoitzeko. Honen ondorioz, ereduek aro kopurua bider 80K adibide ezberdin ikusten dituzte entrenamenduan zehar, non 80K COCO-ko entrenamendu azpimultzoko irudi kopurua den. Azkenik, VSR COCOn oinarrituta dagoenez, kontaminazio arazoak ekiditeko ez ditugu VSR-ko garapen eta ebaluazio azpimultzoetan agertzen diren irudiak erabiltzen SSTD-ko entrenamendu irudi gisa.

4.2.3 Ikasketa Algoritmoa

Lan honetan ondorengo bi faktoreren garrantzia aztertu nahi dugu: i) kokapen tokenen erabilera hizkuntza-ereduetan, eta ii) token hauek erabiltzen ikasteko proposatutako SSTD datu-multzoaren erabileraren onurak. Horretarako, BERT-base (Devlin *et al.*, 2019) eredua erabili dugu gure hizkuntza-eredu gisa eta era ezberdinetan doitu dugu, bai kokapen tokenekin entrenatuz (ala ez) eta baita eredua SSTD-n aurrentrenatuz ere (ala ez). Sailkapen buru bat gehitu dugu $[CLS]$ bektorearen gainean ($\mathbf{t}_1^{(n_l)}$, non n_l transformer kodetzailearen azkeneko geruzaren indizea den) sailkapen bitarra burutzeko. Sailkapen buru hau geruza anitzeko pertzeptroi (MLP) gisa definitu dugu, geruza ezkutu bat gehitu diogularik. Erabilitako MLP-a 4.1. Ekuazioan definitzen dugu.

$$\begin{aligned} \mathbf{h} &= LayerNorm(GELU(\mathbf{W}_h\mathbf{t}_1^{(n_l)} + \mathbf{b}_h)) \\ \hat{\mathbf{y}} &= Sigmoid(\mathbf{W}_{\hat{y}}\mathbf{h} + \mathbf{b}_{\hat{y}}) \end{aligned} \tag{4.1}$$

Esperimentuetan zehar tamaina ezberdineko ereduak erabiliko ditugu. Horretarako, BERT-Large eredua erabili dugu (sailkapen burua gehituz 4.1. Ekuazioa jarraituz), baita T5 familiako ereduak ere (Raffel *et al.*, 2020). Bide batez, kodetzaileak bakarrik ez diren ereduen azterketa burutu dugu ere bai, T5 familiako ereduak kodetzaile eta deskodetzaile banaz osatuta daudelarik. T5 ereduei sarrerako testua elikatzeko txantiloi bat erabili dugu, esaldi baten aurretik aurrizki batzuk ezarriz. Adibidez, goiburuko baten aurretik *"caption:"* zehazten dugu eta eszenaren deskribapenerako, aldiz, *"context:"*. Honen helburua T5 ereduen aurrentrenamenduan erabilitako sarrera imitatzea da, hizkuntza-ereduak aurrentrenamenduan ikasitakoa hobeto aprobetxatzeko pentsatua. Azkenik, T5 eredu

generatiboa denez, testu irekia sortzen du eta, beraz, *"yes"* eta *"no"* tokenak sortzeko probabilitateak konparatzen ditugu, sailkapen bururik gehitu gabe.

SSTD-ko entrenamenduaren garapena balidatzeko SSTD-ko garapen azpimultzo estatiko bat definitu dugu, azpimultzo honen instantziak ausaz sortuz eta berdinak erabiliz entrenamendu guztietan zehar. COCO-ko garapen azpimultzoko irudiak erabili ditugu eta, irudi bakoitzeko instantzia bat sortu dugunez, gure garapen azpimultzoa 40.504 instantziez osatuta dago. SSTD-ko entrenamendu aro bakoitzaren ondoren, garapenean ebaluatzen dugu eredua. Entrenamendua bukatu ostean, garapenean emaitza onenak lortu dituen ereduarekin geratzen gara, VSR atazan doitzeko erabiliko dena.

4.3 Esperimentuak

Atal honek gure esperimentuen ezarpenen xehetasunak ematen ditu. Ondoren, ikasketa espaziala eta kokapen tokenak erabiltzeak dakartzan onurak aztertu ditugu eta hizkuntza-ereduen tamainekin jolastu dugu, artearen egoerarekin konparaketak burutuz ere bai. Gainera, analisi sakon bat egiten da antzemandako hobekuntzak nondik datozen ulertzeko.

4.3.1 Esperimentazio Ezarpenak

Esperimentuetan zehar VSR atazako *random* bertsioa erabili dugu, bere tamaina handiagoa baita. Doikuntza prozesuan zehar, ereduak entrenamendu azpimultzoa erabiliz ikasten du eta garapen azpimultzoan hoberen dabilen eredua aukeratzen dugu ebaluazioa burutzeko. VSR-ko egileen gomendioak jarraituz, hiru entrenamendu ezberdinen batezbestekoak erakusten ditugu esperimentu guztietan, hauen desbiderapen estandarrarekin batera.

Aukeratutako hiperparametroak 3. Kapituluan erabilitakoetan oinarritzen dira, hau da, ez dugu hiperparametro bilaketa berririk burutu esperimentazio honetan. Egin diren aldaketa bakarrak ereduak gure makinen espezifikazioetara egokitzeko beharrezkoak direnak izan dira. Ondoren, doitu dugun eredu bakoitzarekin erabilitako hiperparametroak zerrendatzen ditugu.

BERT-base ereduko esperimentuetan 20K pauso zehaztu ditugu SSTD eta VSR atazetako entrenamenduetarako, AdamW optimizatzailea erabiliz. 56ko sorta tamaina eta 5×10^{-5}eko ikasketa tasa maximoa erabili ditugu. Ikasketa tasa aldakorra zehaztu dugu kosinu planifikatzailea erabiliz 2K pausoko beroketa linealarekin. NVIDIA A30 GPU (24GB VRAM) bakarra erabili dugu esperimentu

guztietarako, entrenamendu bakoitzak gehienez 5 ordu iraun duelarik.

BERT-large ereduaren kasuan entrenamendu pauso kopuru bera erabili dugu, baina 32ko sorta tamaina eta 10^{-5} ikasketa tasa maximoa zehaztuz. Ez ditugu optimizatzaile edota beste hiperparametroetan aldaketa gehiagorik egin, BERT-base ereduearen hiperparametroekin jarraituz. Hori bai, kasu honetan GPU handiago bat erabili behar izan dugu, NVIDIA A100 GPU (80GB VRAM) bat hain zuzen ere. Hala eta guztiz ere, eredu honekin entrenamendu bakoitza burutzeko gehienez ere 5 ordu behar izan ditugu.

T5 ereduak 88K pausotan zehar entrenatzen ditugu espazialki, T5-3B kenduta, 20K pausoz entrenatu dugularik bere tamaina dela eta. Ikasketa espazialaren ostean, 20K pausoz doitzen ditugu VSR atazan. 32ko sorta tamaina eta 5×10^{-5}eko ikasketa tasa maximoa zehaztu dugu hauen entrenamendurako. Optimizatzailea eta ikasketa tasa planifikatzaileak BERT ereduekin erabilitakoak izan dira. T5 ereduen entrenamendu denborak eta entrenatzeko erabili diren GPU-ak honakoak izan dira:

- T5-Base: NVIDIA A30 GPU bat erabili da, ikasketa espazialak $\sim$20 ordu behar izan duelarik eta VSR-ko doitzeak, aldiz, $\sim$4 ordu.
- T5-Large: NVIDIA A100 GPU bat erabili da, ikasketa espazialak $\sim$28 ordu behar izan duelarik eta VSR-ko doitzeak, berriz, $\sim$10 ordu.
- T5-3B: NVIDIA A100 GPU bat erabili da, ikasketa espazialak $\sim$20 ordu behar izan duelarik eta VSR-ko doitzeak, berriz, $\sim$15 ordu.

Azkenik, esperimentu guztietan zehar $G = 32$ sareta tamaina zehaztu dugu.

4.3.2 Ikasketa Espazialaren Eragina

4.3. Taulak VSR-ko ebaluazioan lortzen ditugun emaitzak erakusten ditu ikasketa espaziala edota kokapen tokenak erabiltzearen arabera. Taulako lehenengo blokeak VSR-en doitutako BERT-base ereduaren errendimendua erakusten du. Bertan kokapen tokenen erabilerak ezberdintasunik ez dakarrela ikus daiteke. Hala ere, bigarren blokean antzemangarria da ezberdintasuna. Bloke hauetako ereduak SSTD datu-multzoan doitu dira lehenago, non kokapen tokenak erabiltzen dituen eredua den hobekuntzak lortzen dituen bakarra. Hobekuntza hau nabarmena da beste ereduekiko, $\sim$12 puntu absolutuko aldea erakutsiz. Emaitza hauek kokapen tokenak informazio espaziala kodetzeko eta hizkuntzaren oinarritzea burutzeko egokiak direla erakusten dute. Kokapen token hauek ikasteko erdibideko entrenamendu bat beharrezkoa dela ikusi dugu, kasu honetan SSTD bidez burutu dena.

	Eredua	Kokapen Tokenak	Asmatze Tasa
Hizkuntza-ereduak	BERT-base	Ez	$62{,}11_{\pm 0{,}88}$
		Bai	$61{,}60_{\pm 0{,}92}$
Hizkuntza-ereduak Ikasketa Espazialarekin	BERT-base	Ez	$61{,}83_{\pm 0{,}28}$
		Bai	$\mathbf{73{,}69}_{\pm \mathbf{0{,}88}}$

4.3 Taula – BERT-base eredua harturik, kokapen token edota ikasketa espaziala erabiliz lortutako emaitzak VSR-ko ebaluazio azpimultzoan, batezbesteko asmatze-tasa eta desbiderapen estandarra adieraziz.

Eredua	Kokapen Tokenak	Asmatze Tasa
BERT-base	Ez	$76{,}96$
	Bai	$94{,}49$

4.4 Taula – SSTD-ko garapen azpimultzoko asmatze-tasak kokapen tokenen erabileraren arabera.

Beste aldetik, 4.4. Taulak SSTD datu-multzoko garapenean lortutako emaitzak ikus ditzakegu. Emaitza hauek ez dira hain garrantzitsuak, baina emaitza horien gaineko azterketak datuen eta doitutako ereduen joerak interpretatzen laguntzen du. Ataza honetan, ausaz erantzuten duen eredu batek 50 puntuko asmatze-tasa lortuko luke, sailkapen bitarra burutu behar baitu. Interesgarria da kokapen tokenik gabeko ereduak $76{,}96$ puntu lortzea. Honek kokapen tokenik gabe datuetan dauden hainbat alborapen ikasi ditzakeela erakusten digu, baina ataza ondo ebazten ikasteko nahikoa ez dela antzematen da ere bai. Kokapen tokenak gehitzean ~ 17 puntu inguru gehiago lortzen ditu ereduak, $94{,}49$ko asmatze-tasa lortuz eta ataza ebazteko token hauen garrantzia erakutsiz.

4.3.3 Artearen Egoerarekin Konparaketa

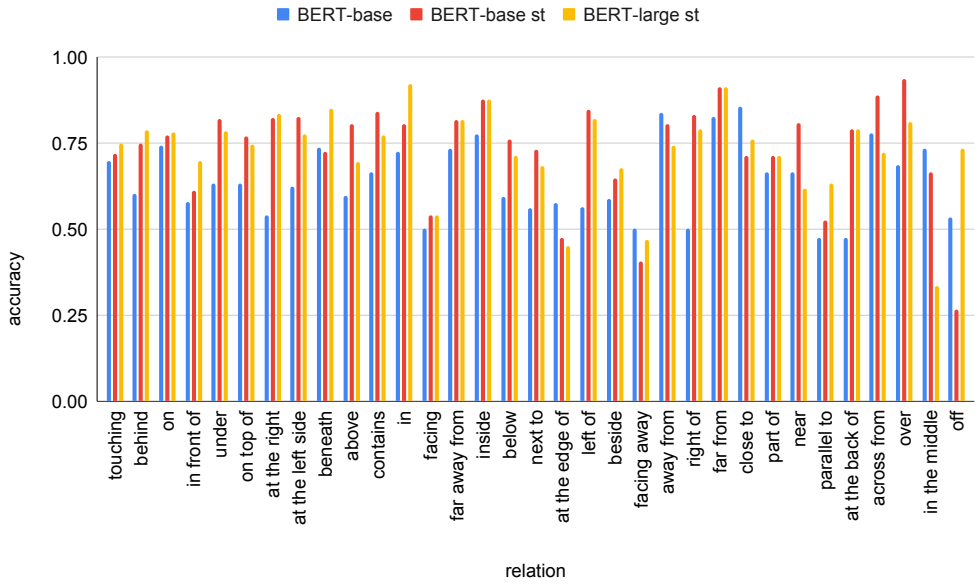
Atal honetan gure emaitzak VSR-ko artearen egoerarekin konparatzen ditugu. Gainera, hizkuntza-ereduen tamainarekin jolasten dugu lortu ditzakegun errendimendu hobekuntzak aztertzeko. Horretarako, BERT-Large eredua eta T5 familiako ereduak erabili ditugu.

	Eredua	Parametroak	Asmatze-tasa
Ikusizko Hizkuntza-ereduak	CLIP$_{prompting}$	632M	$55{,}2_{\pm 1{,}4}$
	VisualBert †	110M	$57{,}4_{\pm 0{,}9}$
	ViLT	87M	$69{,}3_{\pm 0,9}$
	LXMERT	240M	$70{,}1_{\pm 0,9}$
Hizkuntza-ereduak Ikasketa Espazialarekin	BERT-base	110M	$73{,}69_{\pm 0{,}88}$
	BERT-large	336M	$\mathbf{74{,}44_{\pm 0{,}73}}$
	T5-base	220M	$73,09_{\pm 0{,}59}$
	T5-large	770M	$\mathbf{74{,}49_{\pm 0{,}36}}$
	T5-3B	3B	$\mathbf{74{,}52_{\pm 0{,}25}}$

4.5 Taula – VSR-ko ebaluazio azpimultzoan lortutako emaitzak, batezbesteko asmatze-tasa eta desbiderapen estandarra adieraziz. Lehenengo blokean artearen egoera definitzen duten ikusizko hizkuntza-ereduak aurki daitezke, erreferentziak testuan agertuz. Informazio espaziala erabiltzen ez dituzten ereduek † bat dute. Bigarren blokean guk doitutako hizkuntza-ereduak aurkitzen dira.

4.5. Taulan artearen egoera definitzen duten VLM eta gure hizkuntza-ereduekin lortutako emaitzak daude. VLM ereduen emaitzak (Liu *et al.*, 2023) lanetik erauzi ditugu. VLM onenak, LXMERT-ek, $70{,}1$eko asmatze-tasa lortzen du, eta gure eredu guztiek nabarmenki gainditzen dute. Hala ere, guk informazio garrantzitsua galtzen dugu irudiaren deskribapena erabiltzean, objektuen izenak eta hauen kaxa inguratzaileak bakarrik erabiltzen ditugu eta. Gure ereduen artean onenak hiru hizkuntza-eredu handienak dira (letra lodiz daudenak 4.5. Taulan), 74 puntuko asmatze-tasa baino gehiago lortuz, LXMERT-ek baino 4 gehiago.

Emaitza hauetatik kokapen token eta ikasketa espaziala erabiltzeak estrategia ona direla ondorioztatu dezakegu. Gainera, hizkuntza-ereduek informazio espaziala maneiatzeko gaitasuna dutela erakutsi dugu, arrazoinamendu espazial testuala edota dokumentuen egiturari buruzko atazak lantzeko ate berriak irekitzen dituena. Hala ere, hizkuntza-ereduen tamaina handitzeak dakartzan etekinak oso txikiak direla ikusi dugu VSR atazarako. Egia da gure eredu onena 3B parametroko T5 eredua dela, baina T5-large edota BERT-large batekiko diferentzia oso txikia da, hau da, estatistikoki ez dira esanguratsuak. Gainera, azkeneko bi eredu hauek 4 eta 10 aldiz txikiagoak dira T5-3B ereduarekiko, hurrenez hurren. Hiperparametro bilaketarik burutu ez denez, baliteke hizkuntza-eredu handienen hobekuntza nabarmena bihurtzea bilaketa sakonago bat egin ezkero.



4.5 Irudia – Hiru BERT ereduren arteko konparaketa, asmatze-tasak erlazio espazialka zehaztuz. Erlazioak (ardatz horizontalean) VSR-ko agerpen kopuruaren arabera daude ordenatuta altuenetik baxuenera. Irakurketa errazteko, VSR-ko ebaluazio azpimultzoan 15 aldiz baino gehiagotan azaltzen diren erlazioak bakarrik ipini ditugu. Hiru ereduek kokapen tokenak erabiltzen dituzte, eta "st" akronimoak eredu horrek VSR-en doitu aurretik ikasketa espaziala jaso duela adierazten du.

4.3.4 Analisia

Atal honetan erlazio espazial bakoitzarekin lortutako emaitzak banaka aztertzen ditugu. Bide batez, gure sistema erregeletan oinarritutako algoritmo batekin konparatzen dugu, baita VLM batekin ere. Azkenik, objektu atributuak erabiltzeak dakartzan ondorioen analisi bat burutzen dugu.

Erlazio Espazial Bakoitzeko Emaitzak

VSR-ko 65 erlazioetan lortutako emaitzak banaka konparatu nahi izan ditugu, ikasketa espaziala egin duten hizkuntza-ereduen errendimendua erlazioka aztertzeko. Helburua, beraz, ereduaren tamainak eta ikasketa honek nola eragiten duten aztertzea da. 4.5. Irudiak hiru hizkuntza-eredu ezberdinen emaitzak erakusten

ditu erlazioka. Aukeratu ditugun ereduak hauexek dira: BERT-base ikasketa espazialik gabe, BERT-base berdina ikasketa espazialarekin eta BERT-large ikasketa espazialarekin ere bai.

Orokorrean, ikasketa espazialak erlazio guztietan laguntzen du, salbuespen batzuekin. Adibidez, orientazioa lantzen duen *facing away* edota albokotasuna adierazten duen *at the edge of*. Esperotako emaitzak dira, izan ere, SSTDk ez ditu erlazio horiek kontuan hartzen, orientazioa ezin baita kaxa inguratzaileetatik inferitu, ezta albokotasuna ere, ez baitugu sakonera informaziorik. VLM hauentzat orientazioa identifikatzea ataza zaila dela dirudi (Liu *et al.* 2023). Beraz, ildo honetan ikertzea jarraitu beharreko bidea dela iruditzen zaigu.

Dena den, hizkuntza-ereduen orokortze ahalmenean efektu positiboak ikusi ditugu. Kaxa inguratzaileek 3D informazioa kodetzen ez badute ere, ikasketa espaziala egin ondoren hobekuntzak antzeman ditugu sakonera lantzen duten erlazioetan: *behind* eta *in front of*, besteak beste. Gure hipotesietako bat SSTD-k tamainari buruzko erlazioak dituela da (*wider*, *smaller*...) eta, hortaz, ikasketa espazialak sakonera inferitzeko eskura duen informazioa erabili dezake. Beste hitz batzuetan, ohikoak diren tamainari buruzko erlazioak eta kaxa inguratzaileen informazioa konbinatzen ikasi dezakete. Adibidez, pertsonak katuak baino handiagoak direnez, pertsonaren kaxa inguratzailea katuarena baino txikiagoa bada, urrunago egon beharko luke irudian. Kasu hauek gehiago ikertzeko asmoa dugu, erregela aritmetikoen bitartez bakarrik deskribatu ezin daitezkeen erlazioetan orokortze ahalmena antzeman baitugu. Atariko analisi kualitatibo bat burutu dugu eta B.2. Eranskinean aurki daiteke.

4.5. Irudian VSR eta SSTD datu-multzoetan agertzen diren erlazioak ikus daitezke, non beraien errendimendua asko hobetzen den ikasketa espaziala burutu ondoren. Bi entrenamenduen artean transferentzia bidezko ikasketa burutzea espero genuen, semantikoki oso antzekoak baitira erlazio horiek. Hala ere, *beneath* erlazioaren kasuan, SSTD-n dagoen *below* erlazioarekin lotuta badago ere, ikasketa espaziala duen BERT-base ereduak ez du ikasketa gabekoa gainditzen. BERT-large ereduak, ordea, bai gainditzen duela, $+12$ puntu lortuz.

Analisira testuinguru gehiago gehitzeko, 4.6. Taulak VSR datu-multzoko erlazio kopuruak kategorietan erakusten ditu multzokatuta. SSTD-ko erlazioekin berdina burutzen dugu eta kategoria bakoitzean SSTD-rekin ikasketa espaziala burutzeak dakartzan hobekuntzak ipintzen ditugu. VSR-ko 65 erlazioetatik SSTD-k 17 erlazio estaltzen ditu. Gainerako erlazioetako batzuk antzeko edo aurkako esanahiak dituzte eta horietan entrenatzea lagungarria izan daiteke. Adibide gisa, VSR-ko *detached to* erlazioa SSTD-ko *overlapping* erlazioari lotuta dago. Oro har, *overlapping* erlazioa betetzen ez dituzten objektuen kaxa inguratzaileek *de-*

VSR kategoria	VSR Erlazioak	SSTD Erlazioak	Δ Asmatze-tasa
Adjacency	10	2	$+4{,}7$
Directional	11	2	$+2{,}9$
Orientation	4	0	$+9{,}1$
Projective	12	8	$+14{,}4$
Proximity	5	0	$+1{,}1$
Topological	18	5	$-1{,}2$
Unallocated	5	0	$+56{,}8$

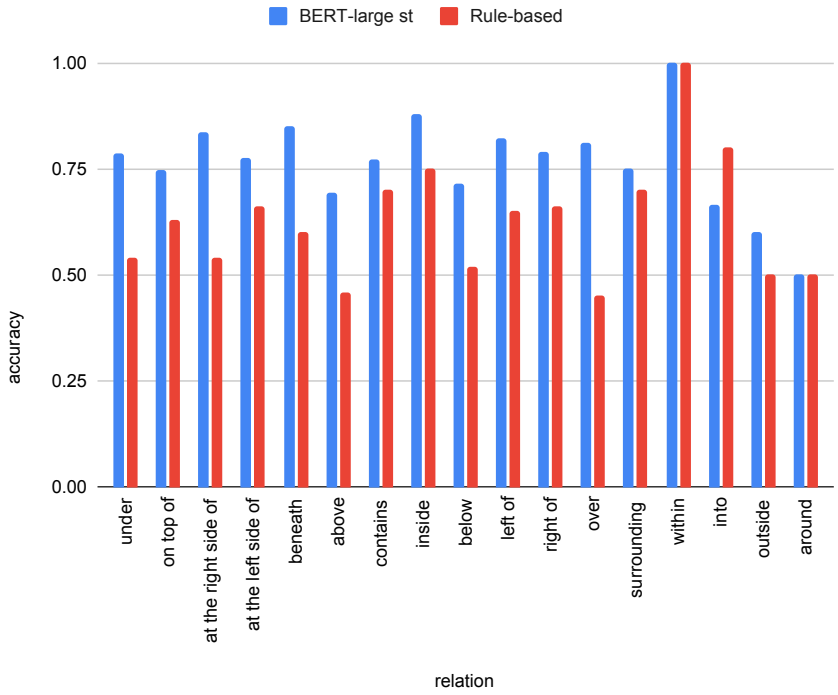
4.6 Taula – VSR-ko kategoria bakoitzeko erlazio kopuruak zehazten ditugu lehenengo zutabean. Bigarrenean, SSTD-rekin berdina erakusten dugu. Azkeneko zutabean, berriz, bi BERT-base ereduren arteko diferentzia erakusten dugu VSR-ko ebaluazio azpimultzoan, ikasketa espaziala egiteak dakartzan hobekuntzak zehaztuz.

tached to erlazioa beteko dute. Errendimendu aldaketaren zutabeari erreparatuz (4.6. Taulako azkeneko zutabeari, hain zuzen), ikasketa espaziala kategoria guztientzat onuragarria dela erakusten dugu, *topological* kategorian izan ezik, non diferentzia oso txikia den. *Unallocated* kategoriak hobekuntza oso handia erakusten du, $+56{,}8$ puntu absolutukoa, baina ez da oso esanguratsua, 51 kasu bakarrik baitaude ebaluzio azpimultzoan. Orokorrean, SSTD-n ondo ordezkatuta dauden kategoriak kontsistenteki hobetzen dira VSR atazan. Horiek dira *projective* ($+14{,}4$), *adjacency* ($+4{,}7$) eta *directional* ($+2{,}9$) kategorien kasuak. Zentzu horretan, *orientation* kategoriaren hobekuntza harritzekoa da ($+9{,}1$).

Azkenik, hizkuntza-ereduen tamainari begira, BERT-base eta BERT-large ereduen arteko diferentziak ez dira kontsistenteak erlazioka. Ez dugu portaera nabarmenik antzeman beraien artean.

Erregela bidezko sistemarekin konparaketa

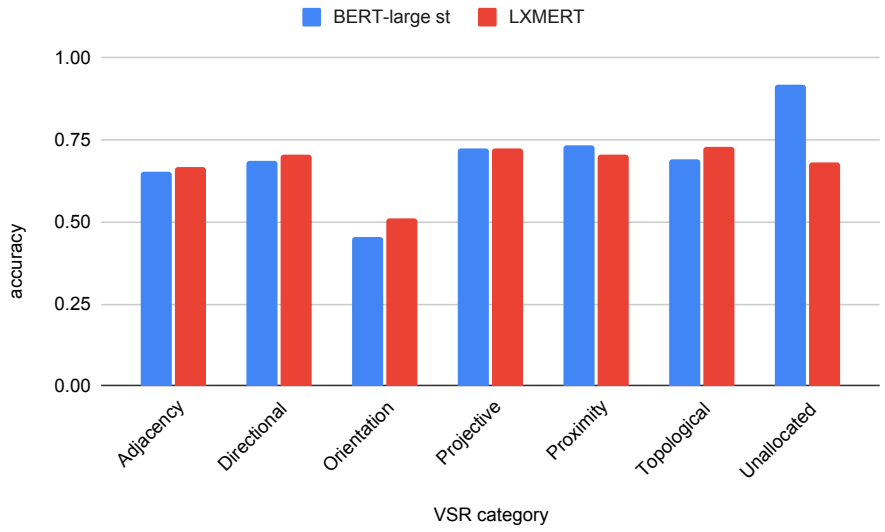
Gure emaitzetatik ikasketa espaziala burutu duten hizkuntza-ereduen gaineko galdera interesgarri bat sortzen da: eredu hauek zerbait ikasten al dute definitu ditugun erregela espazialetatik haratago? Galdera hau erantzuteko erregeletan oinarritutako algoritmo sinple bat zehaztu dugu, SSTD sortzeko erabili ditugun erregela espazial berdinak erabiliz. Algoritmo honen inplementazioari buruzkoak B.3. Eranskinean daude eskura. Erregela hauek erabiliz VSR-ko ebaluazio azpimultzoko $\%38$a ebatzi daitezkeela ikusi dugu. Hala ere, goiburuko eta deskribapen



4.6 Irudia – VSR-ko ebaluazio azpimultzoan BERT-large eta erregeletan oinarritutako algoritmoaren arteko konparaketa erlazioka. Kaxa ingurutzaileen informazioa erabiliz ebatz daitezkeen erlazioak bakarrik hartu ditugu kontuan.

espazialen arteko objektuen lerrokatze desegokia dela eta, instantzia guztien $\%25$ bakarrik ebatzi daitezke erregelak erabiliz. Instantzia hauetarako lortzen dugun asmatze-tasa $60{,}7$ puntukoa da, ikasketa espaziala duten hizkuntza-ereduen errendimendutik urruti gelditzen delarik. Gainerako instantziak ausaz ebazten baditugu (ebaluazio azpimultzoko $\%75$ dena gutxi gorabehera), asmatze-tasa $52{,}4$ puntura jaisten da. Gure eredu hoberenak $74{,}5$ puntu lortzen ditu, $22{,}1$ puntuko hobekuntza erakutsiz erregeletan oinarritutako algoritmoarekiko.

4.6. Irudian erregela bidezko sistema eta gure BERT-large ereduen arteko konparaketa aurki daiteke, emaitzak erlazioka azalduz. Ikus daitekeenez, erregelak erabiliz ebatz daitezkeen erlazioetan ikasketa espaziala egin duen BERT-large eredua da garaile, hiru erlaziotan izan ezik. *within* eta *around* kasuetan emaitza berdinak lortzen ditugu, eta *into* erlazioan, berriz, erregeletan oinarritutako algorit-



4.7 Irudia – Ikasketa espaziala egin duen BERT-large eta LXMERT ereduen arteko konparaketa VSR-ko ebaluazio kategorietan banatuta.

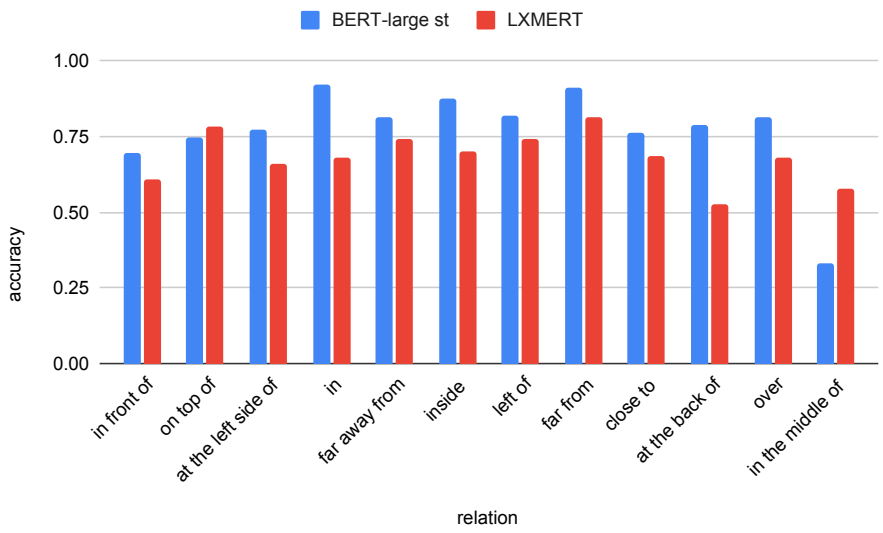
moak lortzen ditu emaitza hoberenak.[4] Emaitza hauetatik gure hizkuntza-ereduek erregela espazialetan kodetuta dagoen informazioa baino gehiago ikas dezaketela ondorioztatu dezakegu.

VLM batekin Konparaketa

Kapitulu honen ardatz nagusia ez bada ere, interesgarria da ikasketa espaziala duen hizkuntza-eredu bat VLM batekin konparatzea. Analisi horretarako BERT-large eta LXMERT ereduak erlazioka ebaluatzen ditugu VSR-ko ebaluazio azpimultzoan.

4.7. Irudian bi ereduekin lortutako asmatze-tasak agertzen dira, kategorietan banatuta. Antzeman daitekeenez, ez daude ezberdintasun handirik bi ereduen artean, *unallocated* kategorian kenduta, non BERT-large ereduak LXMERT-eri nabarmenki irabazten dion (92 vs. 68). Dena den, erlazioka egiten badugu konparaketa, ondorio interesgarriagoetara iritsi gaitezke. 4.8. Irudian hori egiten dugu, 4 puntu absolutu baino handiagoko diferentzia duten erlazioak bistaratuz bakarrik.

[4]Kontuan izan VSR-ko ebaluazio azpimultzoan 6 instantzia bakarrik daudela *into* erlazioarekin, emaitzak esanguratsuak ez izanik.



4.8 Irudia – Ikasketa espaziala egin duen BERT-large eta LXMERT ereduen arteko konparaketa VSReko erlazioetan banatuta. Bi ereduen arteko diferentzia 4 puntu absolutukoa baino handiagoak diren erlazioak azaltzen dira bakarrik.

4 puntuko diferentzia nabarmena dela diogu VSR-ko ebaluazio azpimultzoan bi ereduen arteko batezbesteko diferentzia 4 puntukoa delako.

Ikus daitekeenez, BERT-large ereduak LXMERT gainditzen du *in front of, at the left side of, in, far away from, inside, left of, far from, close to, at the back of* eta *over* erlazioetan. Alde batetik, zentzua du BERT-large bi dimentsioko informazioa bakarrik behar duten erlazioak hobeto ebaztea (*at the left side of, left of* eta *over*). Kontua da BERT-large ereduak informazio gehiago behar dituzten beste erlazio batzuetan hobeto dabilela ere bai (*in front of, in, far away from, inside, far from, close to* eta *at the back of*). Erlazio hauetarako gainerako ikusizko informazioa erabilgarria izan beharko luke, baina badirudi LXMERT-ek ez dakiela informazio hori ondo maneiatzen. Beste aldetik, LXMERT-ek bi erlazioetan bakarrik gainditzen du nabarmenki BERT-large (*on top of* eta *in the middle of*). *On top of* erlazioaren kasuan ez dugu hau gertatzeko arrazoi garbirik ikusten. *In the middle of* kasuan, ordea, BERT-large ereduaren errendimendua oso txarra da, BERT-base ereduarena baino askoz okerragoa. Erlazio honetan BERT-base eta LXMERT-ek pareko emaitzak lortzen dituzte. Beraz, joera hau erlazio honen instantzia kopuru urriarekin lotuta dagoela uste dugu (15 bakarrik baitaude).

Objektu Atributuen Erabilera

VinVL-ek, hau da, erabili dugun objektu detektoreak, objektuen izena eta kaxa inguratzaileaz kanpo, hauen atributuak itzultzen ditu. Atributu hauek koloreak, egoerak (*open hand*, *standing boy*), tamainak, texturak (*striped jacket*), materialak (*brick wall*), dira besteak beste. Ikasketa espazialean aldaketak egin ditugu atributu zerrendak deskribapen espazialari gehitzeko eta, ondoren, BERT-base bat entrenatu dugu 4.3.1. Ataleko hiperparametroak erabiliz. Ondoren, SSTD-ko garapenean emaitza onenak dituen eredua VSR-en doitzen dugu, berriz ere atributuak gehituz deskribapenetara. Horrela, VSR-ko ebaluazioan lortzen dugun asmatzetasa $74{,}1$ puntukoa da, 4.3. Taulan erakusten dugun BERT-base ereduaren desbiderapen estandarraren barruan dagoena. Hortaz, ataza honetarako VinVL bidez erauzitako atributuak erabiltzea ez dela onuragarria ondorioztatzen dugu. Hori bai, kontuan hartzekoa da orientazioa edota sakonera bezalako atributuak VSR atazarako baliagarriak izan daitezkeela, aurreko azpiatalean aipatu den bezala.

4.4 Ondorioak

Kapitulu honetan erlazio espazialak hizkuntza-ereduetan oinarritzeko modu berri bat aurkeztu dugu, kokapen tokenen bidez ahalbidetzen dena. Kokapen token eta erlazio espazialen arteko oinarritzea ikasi ahal izateko, SSTD datu-multzoa proposatzen dugu. Testuzko datu-multzo hau anbiguetaterik gabeko erlazio espazialak lantzen ditu, irudi errealetatik erauzitako objektuen arteko erlazioak hain zuzen ere. *Visual Spatial Reasoning* datu-multzoko bertsio berbalizatu bat erabiltzen dugu esperimentuak burutzeko, hizkuntza-ereduen oinarritze eraginkorra burutzen dugula erakutsiz. Gainera, VLM-kin konparatzean emaitza hobeak lortzen ditugula erakusten dugu, gure hurbilpenak funtzionatzen duela adieraziz.

Gainera, hizkuntza-ereduen tamaina handitzean artearen egoera berria definitu dugu VSR atazan. Hala ere, hobekuntza oso txikiak antzeman ditugu hizkuntza-ereduen tamaina handitu dugun heinean. Honek erlazio espazialen oinarritzean tamaina ez dela garrantzitsua adieraz dezake, beste ikerketa lerroei atea irekiz.

Etorkizunean ikasketa espaziala sakondu nahi dugu, orientazioa eta sakonera bezalako kategoria berriak gehituz, adibidez. Arrazoinamendu espaziala lantzen duten testuzko atazak landu nahi ditugu ere bai, SpartQA (Mirzaee *et al.*, 2021) eta RESQ (Mirzaee and Kordjamshidi, 2022), adibidez. Bertan, lengoaia naturaleko irudien deskribapenak eraldatu nahi ditugu kokapen tokenak txertatzeko eta hauen onurak ataza horietan aztertzeko.

5. KAPITULUA

Erlazio Espazialek Baldintzatutako Irudien Sorrera

5.1 Motibazioa eta Ekarpenak

Stable Diffusion (Rombach *et al.* 2022) eta Dall-E 3 (Betker *et al.* 2023) bezalako testu bidezko irudi sortzaileak arreta handia jaso dute azken aldian, beraien errendimendua asko hobetu baita. Hala ere, sistema hauek perfektuak izatetik urruti daude, hainbat ahulezia erakutsiz. Adibidez, artearen egoeran erlazio espazial esplizituak ondo adierazten ez dituztela antzeman da (Gokhale *et al.* 2023; Cho *et al.* 2023b). Gabezi hau oztopo handia bihurtzen da testu bidezko irudi edizioan edota beste aplikazio batzuetan (Kawar *et al.* 2023).

Esan bezala, erlazio espazial esplizituak irudikatzeko artearen egoerak errendimendu eskasa erakusten du. Gure ustetan, honen zergatia irudi sortzaileen entrenamenduan erabiltzen diren goiburukoek erlazio espazial gutxiegi dituztela da. Hipotesiari eusteko analisi bat burutu dugu ingelesezko LAION-2B datu multzoaren gainean (Schuhmann *et al.* 2022), Stable Diffusion familiako eredu irekiak entrenatzeko erabili baita. LAION-2B datu multzoko goiburukoak sarean eskuragarri dauden irudien *alt-text* eremutik erauzi dira. Goiburuko horien guztien gainean erlazio esplizituak automatikoki bilatu ditugu (*left*, *below* etab.) eta goiburukoen $\%0{,}72$etan bakarrik aurkitu ditugu. Gainera, goiburuko hauen $\%64{,}1$ek *left* eta *right* erlazioak dituzte, gaur egungo irudi sortzaileek ikasi ezin dituztenak. Izan ere, entrenamendu irudiei ausazko iraulketa horizontala aplikatzen zaie irudi kopurua artifizialki handitzeko, baina goiburukoei ez zaie pareko transformaziorik aplikatzen, bien arteko lerrokatze espaziala galduz.

Erlazio espazialak dituzten goiburukoen faltagatik motibatuta, entrenamendu datuak sortzen zentratu gara irudi sortzaileen artearen egoera hobetzeko. Kontuan izan behar dugu hurbilpen hau irudi sortzaileen arkitekturan aldaketak burutzearen osagarria dela (Cho *et al.* 2023b; Feng *et al.* 2023). Zehazki, erlazio espazial esplizituak dituzten goiburuko sintetikoak automatikoki sortzen ditugu, irudi errealekin parekatuz. 4. Kapituluan datu sintetikoak sortzeko landutako hurbilpena jarraituz, COCO datu multzoko (Lin *et al.* 2014) objektu anotazioak erabiltzen ditugu bi kaxa inguratzaileren arteko erlazio espazialak inferitzeko. Horietatik, goiburuko sintetikoak sortzen ditugu, irudi errealekin lerrokatuta daudenak, eta pare horiekin datu multzo bat osatzen dugu, *Spatial Relations for Generation* edo SR4G deritzoguna.

SR4G datu multzoa Stable Diffusion (SD) familiako bi eredu doitzeko erabili dugu, erlazio espazial esplizituak dituzten irudi eta goiburuko pareekin ikasteak ereduen gaitasunak hobetuko dituela aurreikusiz. Doitutako ereduak ebaluatzeko eta jatorrizko SD ereduekin konparatzeko proposatu berri den VISOR metrika erabili dugu (Gokhale *et al.* 2023), erlazio espazial gehiagotara hedatzen duguna.

Lan honen ekarpenak ondorengoak izan dira:

1. SR4G datu multzoa sortu dugu, testu bidezko irudi sortzaileak arrazoinamendu espazialean doitzeko, garatzeko eta ebaluatzeko lehen baliabidea, 14 erlazio esplizituekin osatuta dagoena.

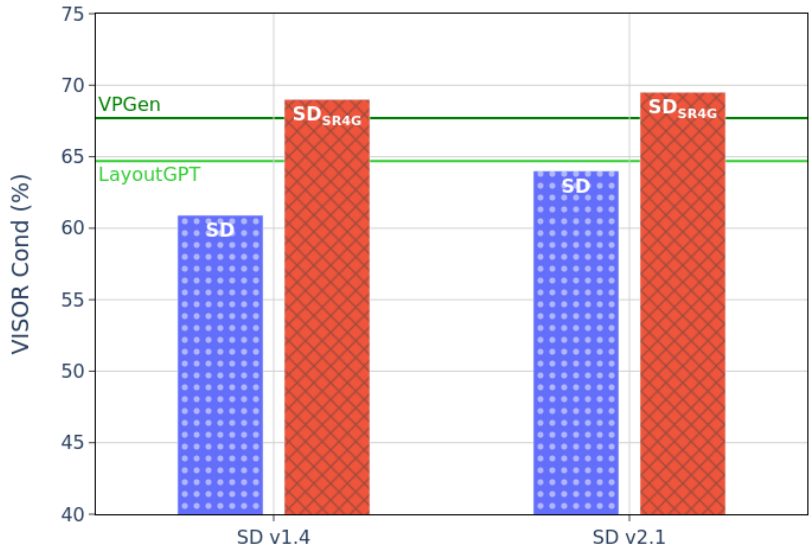
2. Gure esperimentuek erakutsi dute, SR4G-ko datuekin doitzeak erlazio espazialen ulermena hobetzen duela.

3. Hobekuntza hau entrenamenduan zehar ikusi ez dituen objektuetan antzeman da baita ere, doitutako ereduak erlazio hauek ikasteko gaitasuna eta objektu berrietan erlazioak orokortzeko ahalmena dutela erakutsiz.

4. Gure emaitzak artearen egoera gainditzen dute irudi sortzaileen ulermen espazialaren alorrean (Cho *et al.* 2023b; Feng *et al.* 2023) tamaina txikiagoko ereduekin eta arkitektura konplexu edota hizkuntza-eredu handien erabilera ekiditen (5.1. Irudia).

Gure kodea, ereduak eta datu multzoak edozeinen eskura utzi ditugu.[1]

[1] URL: https://github.com/salanueva/SR4G



5.1 Irudia – Stable Diffusion ereduak (v1.4 eta v2.1) gure SR4G datu multzoan doitzeak dakarren errendimenduaren hobekuntza nabarmena da. Artearen egoera gainditzen du ezagutza espaziala jorratzen duten irudi sortzaileetan, lerro horizontalekin adierazten direnak (ikus 5.3.3. Atala xehetasun gehiago jakiteko).

5.2 Metodologia

Atal honetan, SR4G datu multzoa nola sortu dugun zehaztu dugu, goiburuko sintetikoen sorrera zehaztuz. Gainera, SD ereduak nola doitu ditugun zehaztu dugu, baita ereduen ebaluazio automatikoa nola gauzatu dugun ere.

5.2.1 SR4G: Erlazio Espazial Esplizituen Sorkuntzarako Datumultzo Sintetikoa

Irudi eta goiburuko pareez osatutako datu multzoen gabeziak ikusita, goiburuko sintetiko eta irudi erreal pareak sortzea proposatzen dugu, pare hauek SR4G datu multzoa sortzeko erabiliz (*Spatial Relations for Generation*). Aurreko lanetan erabilitako erlazio espazial kopurua handitu dugu (Gokhale *et al.* 2023; Cho *et al.* 2023b; Feng *et al.* 2023), proiektiboak (*projective*) eta tamaina (*scale*) erlazioetatik at, erlazio topologikoak (*topological*) gehituz baita ere. Erabilitako erlazio espazial ez anbiguoen zerrenda ondorengoa da:

- **Proiektiboak:** *left of*, *right of*, *above* eta *below*.
- **Topologikoak:** *overlapping*, *separated*, *surrounding* eta *inside*.
- **Tamaina erlazioak:** *taller*, *shorter*, *wider*, *narrower*, *larger* eta *smaller*.

Gure helburua entrenamendu, garapen eta ebaluaziorako datu multzo bat sortzea da. Entrenamendurako irudi eta goiburuko pareak behar ditugu, baina ereduen errendimendua ebaluatzeko, berriz, goiburukoekin soilik nahikoa dugu. Izan ere, aurreko lanetan egindakoa jarraituz (Gokhale *et al.* 2023; Cho *et al.* 2023b), irudi sortzailearen irteerak ez dira irudi errealekin konparatzen ebaluazio garaian. Ebaluazioaren inguruko xehetasunak 5.2.2. Atalean deskribatzen dira.

Ebaluaziorako goiburukoak

Goiburuko sorkuntza burutzeko, $\langle$*subject, relation, object*$\rangle$ hirukote espazialen zerrenda bat definitzen dugu. Hirukote espazial batek erlazio espazial bat (*relation*) zehazten du, goiburuko batean azalduko den izena (*subject*) eta objektuaren (*object*) arteko ezaugarri espazial bat zehaztuz. Gure hasierako hirukote espazialen zerrenda osatzeko COCO (Lin *et al.* 2014) datu multzoan definituta dauden 80 objektuen pare guztiak eraikitzen ditugu (objektu errepikatuak dituztenak kenduta), 3.160 objektu pare eraikiz. Pare guztiak 14 erlazio espazialekin konbinatuz, 88.480 hirukote espazialekin gelditzen gara.

Hauetako hirukote espazial batzuk ez dira *arruntak*. Adibidez, oso zaila da $\langle$*skis, above, toothbrush*$\rangle$ edo $\langle$*truck, inside, cat*$\rangle$ hirukoteak irudi errealetan aurkitzea, ez direlako ez ohikoak ezta zentzudunak ere. Arruntak ez diren hirukoteak kendu nahi ditugu sortzen ari garen datu multzotik. Hortaz, COCO-ko entrenamenduan erabilitako irudietan gutxienez behin agertzen diren hirukoteak identifikatu ditugu. Gutxienez agerpen bat duten hirukote hauek erabili ditugu amaierako zerrenda osatzeko, hirukote guztien $\%68{,}8$ erabiliz ebaluazioko goiburukoak zehazteko (60.836 hirukote guztira).

Hirukoteak goiburuko bihurtzeko eskuz definitutako txantiloiak erabili ditugu. Txantiloi hauek ahalik eta sinpleenak izaten saiatu gara (ikus C.1.1. Eranskina). Goiburuko hauek bi objekturen arteko erlazio espaziala zehazten dute soilik, beste xehetasunik gehitzea ekiditen dugularik, ebaluazioa erlazio espazialetan bakarrik zentratu nahi dugu eta.

Ikasketarako irudi eta goiburuko pareak

Ikasketarako irudi errealekin lotuta dauden erlazio espazialak dituzten goiburukoak behar ditugu. COCO 2017 bertsioko entramendu azpimultzoa erabili dugu irudi errealak eta hauen objektu anotazioak lortzeko. Horietatik goiburuko sintetikoak lortzeko metodologia bat definitzen dugu bi pausotan zatitzen dena: i) objektu anotazioetatik hirukote espazialak sortzen ditugu eta ii) hirukote hauetatik goiburukoak eraikitzen ditugu.

I irudi bat eta irudi horri dagokion n objektuz osatutako $O_I = \{o_1, o_2, \ldots, o_n\}$ zerrenda izanik, gure helburua O_I zerrendan dauden bi objektuz eta hauen arteko balizko erlazio espazial batez osatutako hirukote espazial bat sortzea da $\langle o_s, r, o_o \rangle$, non $s, o \in \{1, \ldots, n\}$. Objektu bakoitza bere l_i klasea eta $bb_i = \{x_i^0, y_i^0, x_i^1, y_i^1\}$ kaxa inguratzailearekin daukagu etiketatuta, hau da, objektu bakoitzaren izena eta bere posizioa/tamaina irudian ezagutzen ditugu.

$T_I = \{t_1, \ldots, t_m\}$ zerrenda I irudian agertzen diren balizko hirukoteez osatuta dago, non m irudi horretan azaltzen den hirukote kopurua den. Beraz, $t_j = \langle l_s, r, l_o \rangle$ hirukotea SR4G datu multzoko zerrendan dago definituta eta I irudian betetzen da, non $j \in \{1, \ldots, m\}$. Horrek r erlazioa f_r erregela heuristiko baten bidez definitu daitekeela esan nahi du. Kasu honetan, heuristiko honek bi objektuen kaxa inguratzaileak ($bb_s etabb_o$) hartuko ditu kontutan eta bi objektuen artean erlazioa betetzen den ala ez itzuliko du f_r funtzioak (ikus 5.1. Ekuazioa). f_r funtzioak Johnson *et al.* (2018) lana jarraituz definitu ditugu, bi objektuen kaxa inguratzaileen arteko erlazio espazial ez anbiguoak zehaztuz. Begiratu C.1.2. Eranskina funtzio hauen definizioak eta adibideak ikusteko.

$$t_j = \langle l_s, r, l_o \rangle \in T_I \longleftrightarrow f_r(bb_s, bb_o) \tag{5.1}$$

Erabili ditugun COCO datu multzoko I eta O_I pare kopurua artifizialki handitzeko hainbat estrategia jarraitu ditugu (ausazko mozketak eta iraulketa horizontalak eginik). Ondoren, etiketa ezberdinak dituzten bi objektu aukeratzen ditugu ausaz: o_s eta o_o. Horrela, bi objektuen arteko balizko erlazio espazialak zeintzuk diren begiratzen dugu f_r funtzioak erabiliz eta ausaz aukeratzen dugu horietako bat, j-garren balizko erlazioa eraikiz T_I zerrenda osoa kalkulatu gabe: $t_j = \langle l_s, r, l_o \rangle$. Azkenik, t_j hirukotea berbalizatzen dugu ebaluazioko goiburukoetan erabilitako txantiloi berak erabiliz. Txantiloi hauek C.1.1. Eranksinean aurki daitezke.

Bertsioa	Irudiak	Goiburuko ezberdinak			I/G Pareak
		Entrenamendua	Garapena	Ebaluazioa	
Main	103,4K	60,8K	2,5K	60,8K	9,9M
Unseen	83,6K	46,9K	2,5K	8,0K	4,8M

5.1 Taula – SR4G bertsioen estatistikak. *Irudiak* zutabean entrenamenduan zehar erabilitako irudi kopurua zehazten da. *Goiburuko ezberdinak* zutabean, berriz, azpimultzo bakoitzean ager daitezkeen goiburuko ezberdinak definitzen dira. Azkenik, *I/G pareak* sor ditzakegun irudi/goiburuko pareak zehazten ditu.

SR4G-ren Bertsioak

SR4G datu multzoko bi bertsio eraiki ditugu, *main* eta *unseen* bertsioak hain zuzen ere. *Main* bertsioan entrenamenduko goiburukoak exekuzio unean sortzen ditugu murrizketarik gabe. Honek hirukote bera entrenamendu (*train*), garapen (*val*) edota ebaluazioan (*test*) zehar ager daitekeela esan nahi du. *Unseen* bertsioan, berriz, COCO datu multzoko 80 objektuak hiru azpimultzotan banatu ditugu honako banaketarekin: $|O_{\text{train}}| = 45$, $|O_{\text{val}}| = 5$ and $|O_{\text{test}}| = 30$. Xehetasunetan sartuz, entrenamenduan zehar goiburukoak dinamikoki sortzen ditugunean O_{train} azpimultzoko objektuak bakarrik hartzen ditugu kontuan. Garapenerako, O_{val} objektuekin konbinazio gutxi eraiki ditzakegunez, gutxienez O_{val} objektuetako bat duten hirukote guztiak hartzen ditugu kontutan, O_{test} multzoko objekturik ez dituzten bitartean. Ebaluaziorako goiburukoak sortzeko O_{test} azpimultzoan azaltzen diren objektuak bakarrik erabiltzen ditugu. 5.1. Taulan aipatutako bertsioen irudi eta goiburuko kopuruak zehazten ditugu (xehetasun gehiago C.1.3. Eranskinean).

5.2.2 Ebaluazioa

Erlazio espazialetan testu bidezko irudi sortzaileen errendimendua ebaluatzeko hiru ebaluazio metrika erabili ditugu, Gokhale *et al.* (2023) lanean aurkeztuak.

Object Accuracy: l_a eta l_b objektu etiketak eta I' sortutako irudi bat emanik, *object accuracy* metrikak etiketa horiei dagozkien bi objektuak irudian agertzen diren ala ez neurtzen du. Objektu detektore bat erabiliz, I' irudian agertzen diren objektuen zerrenda eskuratzen dugu $L_{I'} = \{l_1, \ldots, l_n\}$, 5.2. Ekuazioa erabiliz asmatze-tasa hau definitzeko. Ebaluazio metrika hau objektuak sortzeko gaitasuna neurtzeko erabilgarria da, ez baitu hirukotearen r erlazioa kontuan hartzen.

$$\text{OA}(I, l_a, l_b) = \begin{cases} 1 & \text{if } l_a, l_b \in L_{I'} \\ 0 & \text{else} \end{cases} \tag{5.2}$$

VISOR: I' sortutako irudia eta $t = \langle l_a, r, l_b \rangle$ hirukote espaziala emanda, VISOR metrikak bi objektuak sortu direla eta bien arteko erlazio espaziala zuzena dela neurtzen du (ikus 5.3. Ekuazioa). f_r funtzioak bi objektuen kaxa inguratzaileak hartzen ditu kontutan (bb_a and bb_b) eta hirukotea balizkoa den aztertzen du. 5.2. Ekuazioan bezala, kaxa inguratzaile hauek objektu detektore baten bitartez lortzen dira. Beraz, VISOR metrikan balio altuagoak lortzen dira goiburukoetan azaltzen diren objektuen sorrera kontsistenteago batekin eta haien arteko kokapen erlatibo zuzenagoekin, goiburukoan aipatzen diren hirukote espazialak irudikatzeko ahalmena erakutsiz.

$$\text{VISOR}(I, t) = \begin{cases} 1 & \text{if } l_a, l_b \in L_{I'} \wedge \\ & f_r(bb_a, bb_b) \\ 0 & \text{else} \end{cases} \tag{5.3}$$

VISOR$_{\text{Cond}}$: Zuzen irudikatu diren hirukote espazialen proportzioa adierazten du metrika honek ere, baina bi objektuak agertzen diren irudiak bakarrik hartzen ditu kontuan. Gure ekarpenak arrazoinamendu espazialean zentratzen direnez, VISOR$_{\text{Cond}}$ ebaluazio metrika aztertuko dugu gehienbat, honek kuantifikatzen baitu erlazio espazialen irudikapen zuzena objektuen sorrera alde batera utziz. Sistema ezberdinek objektuak sortzeko ahalmen ezberdinak izan ditzaketenez, gure esperimentuetarako informazio baliagarriena VISOR$_{\text{Cond}}$ metrikak ematen digu, erlazio espazialen ulermena isolatzen baitu.

Kapitulu honetan zehar hiru metriken balioak ematen ditugu eta, gainera, erabilitako erlazio espazialen kopurua 4tik 14ra handitu dugu.

5.2.3 Ikasketa Algoritmoa

Kapitulu honetan erabilitako ereduei ez diegu aldaketarik egin arkitektura edota galera funtzio aldetik. Azken finean, gure helburua erlazio espazialak dituzten goiburukoekin difusio ereduak doitzea da eta eredu hauen arrazoinamendu espazialean dakartzaten onurak aztertu nahi ditugu. Gure esperimentuen oinarritzat SD eredua hartu dugu, erlazio espazialen sorkuntzan errendimendu altuena ematen baitu bitarteko pausorik eman gabe (Gokhale *et al.* 2023). Hala ere, SD ereduak

ez dira erlazio espazialak irudikatzen kontsistenteenak. Artearen egoera *Pipeline* ereduek zehazten dute, eta hauekin konparaketa 5.3.3. Atalean egin dugu.[2]

SD ereduen bi bertsio erabili ditugu lan honetan zehar: SD v1.4 eta SD v2.1, irudiak 512x512 eta 768x768ko pixel kopuruarekin sortzen dituztenak, hurrenez hurren. Esan bezala, eredu hauek SR4G datu multzoan doitzeko Rombach *et al.* (2022) lanean erabili zuten $\mathcal{L}_{CLDM}$ galera funtzioa erabili dugu, hots, testu bidezko irudi sorkuntza burutzeko erabili zutena.

$$\mathcal{L}_{CLDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\Big] \tag{5.4}$$

5.4. Ekuazioan zehazten da $\mathcal{L}_{CLDM}$ funtzioa (*loss for conditional latent diffusion models*). 2. Kapituluko 2.2. Ekuazioaren antzekoa da, baina hainbat aldaketa ditu. Alde batetik, SD ereduek espazio latentean egiten dute lan, x irudiaren errepresentazio bektorial baten gainean aplikatuz bai difusioa eta baita zarata-ezabatze prozesua ere, $\mathcal{E}(x)$-ren gainean alegia. Difusio prozesua t aldiz aplikatuz, z_t bektoreak duen ϵ zarata zein den aurresaten saiatzen da SD eredua. Beste aldetik, berreskuratu nahi den x irudia y goiburukoaz baldintzatzeko τ_θ eredua erabiltzen dugu, SD ereduen kasuan CLIP ereduaren testu kodetzailea izanik (Radford *et al.* 2021). Hortaz, SD ereduak aurresandako zarata $\epsilon_\theta(z_t, t, \tau_\theta(y))$ da eta, 5.4. Ekuazioari esker, ϵ zarata errealarekin duen diferentzia minimizatzen ikasten da.

Entrenamenduan zehar garapen azpimultzoan ebaluatzen ditugu ereduak zehaztutako entrenamendu pauso kopuru bat pasatzen den bakoitzeko. Entrenamendua bukatu ostean, garapen azpimultzoan $\text{VISOR}_{\text{Cond}}$ balio altuena lortu duen ereduaren bertsioarekin geratzen gara. Gokhale *et al.* (2023) jarraituz, lau irudi sortzen ditugu hirukote espazial bakoitzeko emaitza tenteagoak lortzeko.

5.3 Esperimentuak

Atal honetan, difusio ereduek erlazio espazialak irudikatzeko duten ahalmena handitzen dugu eredu hauek SR4G datu multzo sintetikoan doituz. Gainera, eredu doituak ikasi duena ikusi ez dituen objektuetara orokortzeko ahalmena duela erakusten dugu. Artearen egoera osatzen duten *pipeline* sistemekin konparaketa burutzen dugu baita ere, eta hainbat analisi burutzen ditugu emaitza nagusietan lortutako ondorioak sendotzeko.

[2]Laburbilduz, *pipeline* ereduek lehenengo objektu-kokapenak sortzen dituzte, ondoren irudi sorkuntza objektu-kokapen horietan baldintzatuz. Sistema hauek hizkuntza-eredu handien menpekoak dira, *pipeline* ereduen konplexutasuna handiagoa izanik.

Hyperparameter	Value
Training steps	100k
Batch size	64
Learning Rate	10^{-5}
Optimizer	AdamW
Adam β_1	0,9
Adam β_2	0,999
Adam ϵ	10^{-8}
Weight decay	0,01
Mixed-precision	bf16

5.2 Taula – Difusio ereduen doikuntzan erabilitako hiperparametroak.

5.3.1 Esperimentazio Ezarpenak

Entrenamendurako ezarpenak: 5.2. Taulan entrenamenduan zehar erabilitako hiperparametroak definitu ditugu. Erabilitako ikasketa-tasa eta optimizatzailea SD ereduen aurrentrenamenduan erabilitako berberak dira, eta gainerako hiperparametroak eskura daukagun azpiegituraren arabera egokitu ditugu. Gainera, batezbesteko mugikor esponentziala (ingelesez *exponential moving average*) (Kingma and Ba 2015) erabiltzen dugu parametroak eguneratzeko AdamW optimizatzailearekin (Loshchilov and Hutter 2019) eta ikasketa-tasa planifikatzailerik gabe. Ebaluazioak 5K entrenamendu pauso bakoitzeko burutzen ditugu entrenamendu pauso guztiak egin arte.

Datu gehikuntza estrategiak gehitu ditugu gure entrenamenduan zehar, hala nola, iraulketa horizontalak eta ausazko mozketak. C.3. Eranskinean xehetasun gehiago aurki daitezke.

Memoria behar ezberdinak direla eta, 2 eta 4 NVIDIA A100 GPU erabili ditugu SD v1.4 eta SD v2.1 ereduak doitzeko, hurrenez hurren. Bi kasuetan 64ko sorta tamaina efektiboa erabili dugu. Doikuntza bakoitza burutzeko 3-4 egun behar izan ditugu.

Ebaluazio ezarpenak: Erabili ditugun ebaluazio metrikak objektu detektore bat erabiltzen dute goiburukoetan azaltzen diren objektuak irudian kokatzeko. Gokhale *et al.* (2023) jarraituz, hiztegi irekiko OWL-ViT objektu detektorea erabili dugu (Minderer *et al.* 2022), CLIP eta ViT-B/32 ereduetan oinarritzen dena (Radford *et al.* 2021; Zhai *et al.* 2022a). 0,1-eko konfiantza atalasea zehaztu dugu gure es-

Eredua	$\text{VISOR}_{\text{Cond}} \uparrow$	VISOR ↑	OA ↑
	Main split		
SD v1.4	60,9	17,6	29,0
SD v2.1	64,0	27,4	42,8
SD_{SR4G} v1.4	69,0	26,8	38,9
SD_{SR4G} v2.1	69,5	31,7	45,6
	Unseen split		
SD v1.4	60,1	17,3	28,7
SD v2.1	64,0	28,4	44,4
SD_{SR4G} v1.4	68,9	23,7	34,4
SD_{SR4G} v2.1	69,4	29,4	42,4

5.3 Taula – *Main* eta *unseen* bertsioetan lortutako emaitzak. SD v1.4 eta v2.1 ereduak dagozkien SD_{SR4G} eredu doituekin konparatzen ditugu.

perimentuetan zehar, irudiaren eskualde konkretu batean objektu bat dagoen ala ez zehazteko erabiltzen delarik. Gainera, hiztegi irekiko objektu detektorearen sarreran detektatu nahi dugun objektua zehazteko ondoko txantiloia erabiltzen dugu Minderer *et al.* (2022) lanean emandako gomendioak jarraituz: *"a photo of a* $\langle$*OBJ*$\rangle$*."*.

SD ereduek sortutako irudien aldakortasuna dela eta, ebaluazio goiburuko bakoitzeko 4 irudi sortzen ditugu. Hortaz, 10K irudi sortzen ditugu garapeneko azpimultzoan ebaluatzerako garaian, guk egindako hainbat probetan emaitza kontsistenteak lortu ditugu eta irudi kopuru honekin. Goiburuko bakoitzeko 4 irudi sortzeko joera jarraituz, ebaluazio azpimultzoan 243,3K eta 32,1K irudi sortzen ditugu *main* eta *unseen* bertsioetan, hurrenez hurren.

5.3.2 Ikasketa Espazialaren Eragina

5.3. Taulan SD eta SD_{SR4G} ereduekin lortutako emaitzak erakusten ditugu, metrika bakoitzeko emaitza onenak beltzaranez markatuz. Doitutako ereduei SD_{SR4G} deritzogu.

***Main* bertsioa:** SD_{SR4G} ereduek ebaluazio metrika guztietan lortzen dituzte hobekuntzak SD ereduekiko, objektu eta erlazio espazialen sorkuntza ahalmenak indartuz. Emaitza hauek gure hasierako hipotesiarekin bat datoz, erlazio espazial esplizituak dituzten irudi-goiburuko pareekin doitzea lagungarria dela erakutsiz hauen sorkuntzarentzat. Gure esperimentuetan SD_{SR4G} v1.4 eta v2.1 ereduek pareko errendimendua erakusten dute, baina v2.1-ek objektu sorkuntza hobeto menderatzen du. Aipatzekoa da ere v1.4 ereduen arteko diferentzia handiagoa dela v2.1-ko eredueekin lortutakoa baino.

***Unseen* bertsioa:** SD_{SR4G}-ren hobekuntzak nondik datozen aztertzeko, *unseen* bertsioan doitzen eta ebaluatzen ditugu eredu hauek baita ere. Izan ere, *main* bertsioko entrenamenduan objektu pare bakoitzeko alborapen edo joera espazialak ikasi ditzakete, entrenamenduko objektu eta erlazioen arteko korrelazioak ikasiz orokortze gaitasunik gabe. *Unseen* bertsioak entrenamendu eta ebaluaziorako objektu ezberdinak erabiltzen dituenez, orokortze ahalmena isolatu dezakegu gure azterketan. 5.3. Taulan eredu doituek datu multzoko bi bertsioetan $\text{VISOR}_{\text{Cond}}$ eta VISOR metriketan kontsistenteki hobekuntzak lortzen dituztela antzeman daiteke. $\text{VISOR}_{\text{Cond}}$ metrikaren kasua bereziki interesgarria da, *main* bertsioan bezain altua baita, baina kasu honetan objektu sorkuntzak ez du baldintzatzen lortutako puntuazioa. Honek doikuntza ondoren gure ereduek ikusi ez dituen objektuetara orokortzeko ahalmena duela erakusten du. Azkenik, esan beharra dago probatu ditugun bi SD ereduek pareko joera erakusten dutela gure datu multzoko bi bertsioetan.

Irudien kalitatea: Entrenamenduan goiburuko sintetikoak erabiltzen ari garenez, doikuntzan zehar sortutako irudien kalitatea okertzen ez dela aztertzen dugu. Beraz, Fréchet Inception Distance (FID) (Heusel *et al.* 2017) metrika erabili dugu giza goiburukoetatik sortutako irudien kalitatea neurtzeko. Ebaluazio hau burutzeko COCO datu multzoko 2017 bertsioko garapenean dauden 5.000 irudi-goiburuko pare erabili ditugu. Gure esperimentu guztietan zehar FID balioak konstante mantentzen direla antzeman dugu. Sortutako irudien hainbat adibide 5.3.4. Atalean aztertu daitezke.

5.3.3 Artearen Egoerarekin Konparaketa

Artearen egoerako bi *pipeline* sistemekin konparatu ditugu gure ereduak: LayoutGPT (Feng *et al.* 2023) eta VPGen (Cho *et al.* 2023b). VPGen ereduak erabiltzen duen hizkuntza-eredu handia objektu-kokapenak sortzeko doituta dago,[3] beraz eredu hau dagoen bezala erabiltzen dugu. Kontuan izan behar dugu hizkuntza-eredu hau COCO datu multzoan doituta dagoela. Hortaz, VPGen-ek gure ebaluazioko irudi eta objektu-kokapen pareak ikusi ditu, kontaminazio arazoak izanik *unseen* bertsioko esperimentuetan.

LayoutGPT-ek 7B parametroko Llama-2 eredua erabiltzen du eta testuinguru bidezko ikasketa bidez objektu-kokapenak sortzeko egokitzen da, eredua doitu beharrean. Hortaz, instantzia multzo bat osatzen dugu hizkuntza-ereduaren testuingurua osatzeko. 400 goiburuko eta objektu-kokapen pare biltzen ditugu ausaz erlazio espazial bakoitzeko, 5,6K paretako datu multzoa eraikiz. Inferentziak egiteko $k = 8$ adibide aukeratzen ditugu, bildutako goiburukoen eta sarrerako goiburukoaren arteko CLIP bidezko antzekotasuna kalkulatuz (Radford *et al.* 2021) eta top-k adibide antzekoenak aukeratzen ditugu hizkuntza-ereduaren testuingurua osatzeko eta goiburuko objektuen objektu-kokapenak sortzeko.

5.4. Taulan bi SR4G bertsioetan lortutako emaitzak aurki ditzakegu. Bertan joera berdina ikus dezakegu, hots, SD_{SR4G} v2.1 ereduak *pipeline* sistemek osatzen duten artearen egoera gainditzen du $\text{VISOR}_{\text{Cond}}$ metrikan, erlazio espazialen sorrera ebaluatzen duena goiburukoan aipatutako bi objektuak irudian agertzen diren kasuetan. Hobekuntza hau bereziki garrantzitsua da, aztertu ditugun bi *pipeline* sistemak parametro kopuru aldetik 6 eta 11 aldiz handiagoak baitira. Gainera, hizkuntza-eredu handiez baliatutako arkitektura konplexuagoak dituzte gure difusio ereduekin konparatuta.

Bestalde, 5.2.2. Azpiatalean azaldu diren beste bi ebaluazio metriken balioak ikus ditzakegu 5.4. Taulan. Ebaluatutako ereduetatik VPGen-ek lortzen ditu emaitza hoberenak VISOR eta *object accuracy* metriketan. Espero genituen emaitzak dira hauek. Izan ere, VPGen-ek objektu sorrera sustatzen du bere entrenamenduan zehar eta VISOR balioak ereduaren objektu sorreraren ahalmenarekin korrelatua baitaude. Gainera, VISOR-en hobekuntzak objektu sorrera ahalmenagatik bakarrik hobetu direla ikus dezakegu, VPGen-en emaitzak okertu egiten baitira $\text{VISOR}_{\text{Cond}}$ metrikan. Gogoratu $\text{VISOR}_{\text{Cond}}$ metrikak objektuak sortzeko ahalmena ez duela kontuan hartzen.

[3] Hiru datu multzo ezberdin erabiltzen dituzte entrenamendurako goiburuko eta objektu-kokapen pareak lortzeko: Flickr30K entities (Plummer *et al.* 2015), COCO instances (Lin *et al.* 2014), eta PaintSkills (Cho *et al.* 2023a).

Eredua	Param.	VISOR$_{\text{Cond}}$ ↑	VISOR ↑	OA ↑
		Main split		
LayoutGPT	8,1B	64,7	24,7	38,1
VPGen	14,1B	67,7	34,5	51,0
SD v2.1	1,3B	64,0	27,4	42,8
SD$_{SR4G}$ v2.1	1,3B	69,5	31,7	45,6
		Unseen split		
LayoutGPT	8,1B	64,7	24,7	38,1
VPGen †	14,1B	68,4	37,0	54,1
SD v2.1	1,3B	64,0	28,4	44,4
SD$_{SR4G}$ v2.1	1,3B	69,4	29,4	42,4

5.4 Taula – Artearen egoerarekin konparaketa bi SR4G bertsioetan, ereduen parametro kopuruak zehaztuz. † VPGen kontaminatuta dago *unseen* bertsioan, ebaluazioan erabili diren hirukote espazialak ikusi baititu objektu-kokapenen sorkuntza ikasterakoan.

5.3.4 Analisia

Difusio ereduak SR4G-en doitzeak dakartzan ondorioen azterketa sakon bat egin dugu. Bertan, ereduen errendimendua aztertzen dugu erlazioka, baita hauen alborapena aurkako esanahia duten erlazioetan ere. Gainera, SR4G-ko hirukote espazialen maiztasuna eta ereduek hauetan dituzten errendimenduen arteko korrelazioa aztertu dugu. Azkenik, sortutako irudien analisi kualitatiboa burutu dugu.

Ereduen errendimendua erlazioka

5.5. Taulan VISOR$_{\text{Cond}}$ balioak erlazioka banatzen ditugu. Azterketa hau SD$_{SR4G}$ v2.1 ereduaren gainean egin dugu datu multzoko bi bertsioetan.

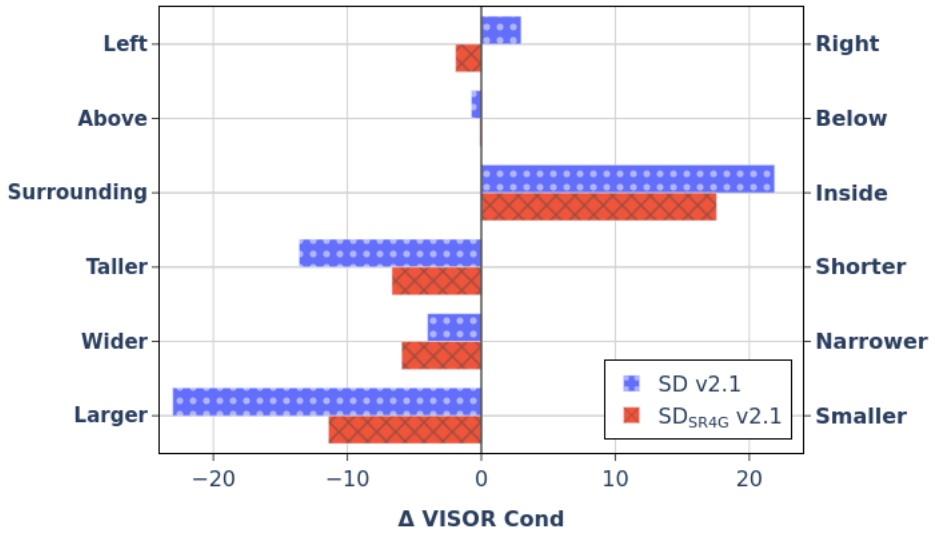
Lehenik eta behin, erlazio proiektibo guztietan hobekuntzak antzeman ditugu. Hobekuntza hau handiagoa da *left of* eta *right of* erlazioetan. Izan ere, SD ereduen entrenamenduan zehar burutzen diren irudien iraulketa horizontalak direla eta, ezin dituzte ardatz horizontalari dagozkien erlazio espazialak ikasi. Gure doikuntzan irudi eta goiburukoak beti ondo lerrokatuta daudenez, ereduek erlazio horiek ikasteko ahalmena erakusten dute, *above* eta *below* erlazioen pareko errendimendua lortuz doikuntza burutu ondoren.

Mota	Erlazioa	*Main* Bertsioa	*Unseen* Bertsioa
Proiektiboa	*Left of*	70,3 (+7,0)	69,8 (+8,8)
	Right of	72,4 (+8,0)	67,9 (+3,9)
	Above	72,0 (+4,5)	70,4 (+2,2)
	Below	71,4 (+4,5)	70,3 (+2,8)
Topologikoa	*Overlapping*	86,9 (-4,9)	84,0 (-5,2)
	Separated	79,5 (+17,0)	84,8 (+18,5)
	Surrounding	29,8 (+2,3)	21,7 (-2,1)
	Inside	43,4 (-7,4)	39,2 (-6,4)
Tamaina	*Taller*	71,2 (+1,6)	75,6 (+5,0)
	Shorter	67,5 (+8,5)	69,0 (+11,9)
	Wider	71,6 (+4,3)	73,0 (+6,9)
	Narrower	69,3 (+9,3)	67,1 (+5,0)
	Larger	71,5 (+0,5)	74,7 (+1,9)
	Smaller	65,2 (+12,7)	63,3 (+13,5)

5.5 Taula – SD_{SR4G} v2.1 ereduaren $\text{VISOR}_{\text{Cond}}$ balioak erlazioka. SD v2.1 eta SD_{SR4G} v2.1 ereduen arteko diferentzia parentesi artean ematen da.

Erlazio topologikoek aldakortasun handiagoa erakusten dute. *Separated* erlazioaren kasua berezia da, bi objekturen gainezartzea behartzen ez duen erlazio topologiko bakarra baita. Erabilitako erlazio topologikoetatik doikuntza ondoren okertzen ez den erlazio bakarra da, $18{,}5$ puntuko hobekuntza lortuz $\text{VISOR}_{\text{Cond}}$ metrikan. *Overlapping*-en kasua aurkakoa da, doikuntza ez baita onuragarria. SD v2.1 ereduak jada badaki erlazio hau zuzen sortzen, $91{,}8$ eta $89{,}2$ puntu lortuz bi bertsioetan. Beste bi erlazio topologikoak bereziki zailak dirudite. Erlazio hauen $\text{VISOR}_{\text{Cond}}$ balioak oso txikiak dira SD ereduetan, eta gure doikuntza espazialak emaitzak okertzen ditu, batez ere *inside*-n kasuan. Hau gure entrenamenduaren ahulezia bat da, eta entrenamendu estrategia ezberdinak planteatu beharko lirateke arazo honi ekiteko.

Azkenik, SD_{SR4G} ereduak hobekuntzak lortzen ditu tamaina erlazio guztietan. *Taller*, *wider* eta *larger* erlazioak sortzeko errendimendua altuagoa da hauen aurkako erlazioena baino, doikuntzan lortutako hobekuntzak txikiagoak badira ere. Honek SD ereduek erlazio espazialak sortzeko alborapen handiak dituztela iradokitzen du.

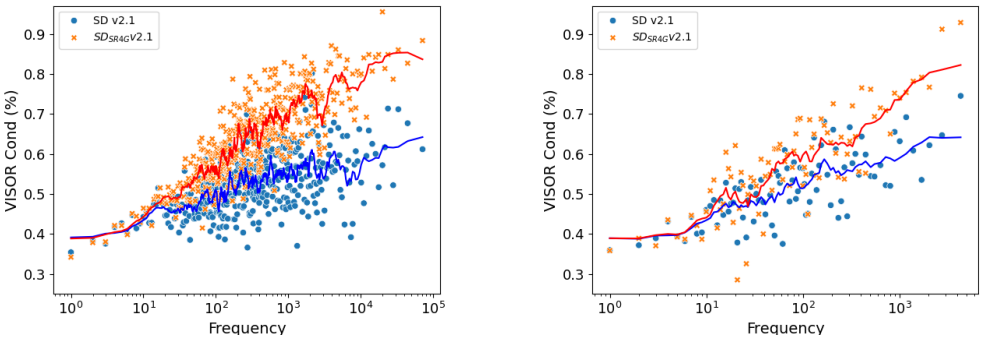


5.2 Irudia – Ardatz horizontalak bi erlazio espazialen arteko VISOR$_{\text{Cond}}$ balioen diferentziak zehazten ditu, ardatz bertikalean aurkako esanahia duten erlazio pareak zehaztuz. Emaitzak *unseen* bertsioan doitu diren SD v2.1 and SD$_{SR4G}$ v2.1 ereduenak dira.

Aurkako erlazioen alborapena

Gure erlazio gehienek aurkako esanahia duen erlazio bat dute. Adibidez, *right of* erlazioaren aurkakoa *left of* izango litzateke. Guztira, sei pare aurki daitezke gure erlazio multzoan, 5.2. Irudian zerrendatzen direnak. Bertan, haien arteko errendimendu diferentzia zehazten da doikuntza aurretik eta ondoren. *Unseen* bertsioan lortutako emaitzak erakusten ditugu.

Aurkako esanahiak dituzten erlazioen arteko alborapenak eta baita doikuntza burutu ondoren hauek murriztu diren ere aztertu nahi ditugu. 5.2. Irudian SD v2.1 ereduaren alborapen sendoak antzematen ditugu. C.2. Eranskinean alborapen hauek SD ereduen aurrentrenamenduan erabilitako datuetan erlazioek duten agerpen proportzioarekin korrelatua daudela erakusten dugu. Gainera, SD$_{SR4G}$ v2.1 ereduak erlazio pare guztien alborapenak txikitzen dituela ikus dezakegu, *wider* eta *narrower* parearen kasua kenduta. Honek gure doikuntzak ereduaren berezko alborapenak gutxitzen dituela erakusten du.



(a) *Main* bertsioko emaitzak. **(b)** *Unseen* bertsioko emaitzak.

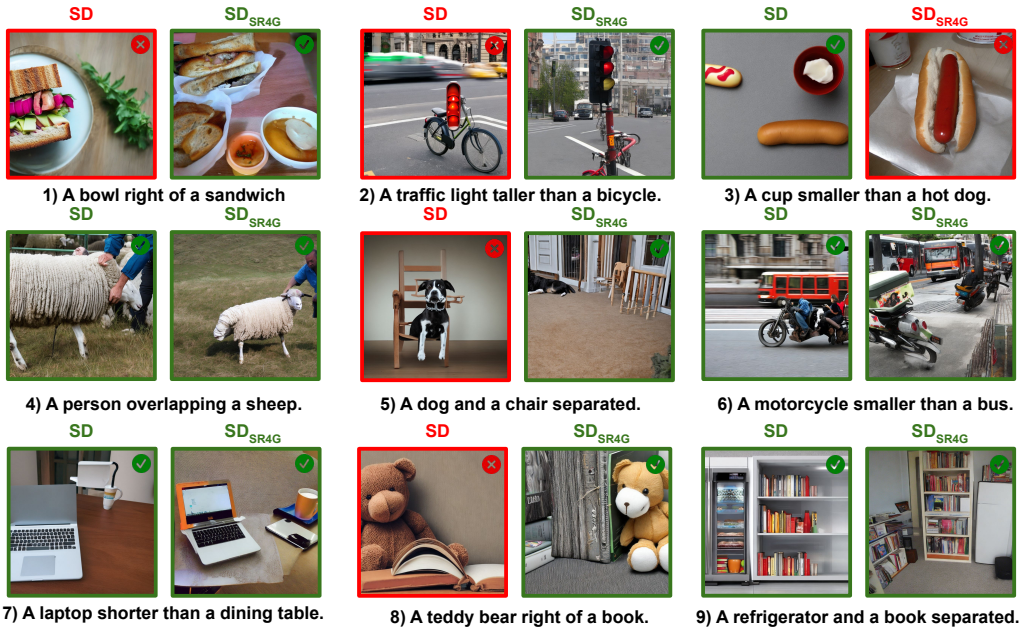
5.3 Irudia – Ardatz horizontal logaritmikoan SR4G-ko hirukoteen agerpenen maiztasuna zehazten dugu COCO-ko entrenamendu irudietan, eta ardatz bertikalean, berriz, hirukote horiekin lortutako VISOR$_{\text{Cond}}$ balioak. SD v2.1 eta SD$_{SR4G}$ v2.1 ereduen emaitzak ezberdintzen ditugu, eta hirukoteak maiztasunaren arabera multzokatzen ditugu ikusgarritasunagoa izateko.

Errendimendua hirukote espazialen maiztasunen arabera

SR4G irudi naturaletan oinarritutako goiburukoekin osatu dugu eta, hortaz, hirukote espazial batzuk besteak baino maizago azaltzen dira. Horregatik, entrenamenduan zehar ikusitako erlazioen maiztasunak gure emaitzetan zenbateraino eragiten duen aztertu nahi izan dugu. 5.3. Irudian COCO datu multzoko entrenamendu instantzietako hirukote espazialen maiztasunak hauen VISOR$_{\text{Cond}}$ balioekiko konparatzen ditugu, bai SD v2.1 eta baita SD$_{SR4G}$ v2.1 eredurako ere.

Alde batetik, 5.3a. Irudiak *main* bertsioan lortutako emaitzak erakusten ditu. Kasu honetan, irudi sortzaileak hirukote berdinak ikusi ditu entrenamenduan eta ebaluazioan. Espero den bezala, maiztasun handiko hirukoteen gaineko hobekuntza handiagoa da eredua doitu ostean. Gainera, SD ereduek ez dituzte COCO irudiak ikusi beraien aurrentrenamenduan zehar, baina SR4G datuekin doitu ez diren SD ereduek korrelazio hau erakusten dute ere bai.

Beste aldetik, 5.3b. Irudiak pareko korrelazioak erakusten ditu bi ereduetan *unseen* bertsioko esperimentuen emaitzekin. Hala ere, kasu honetan entrenamenduan zehar ikusi ez diren objektuekin egin dugu ebaluazioa. Emaitza hauek ikasitako erlazioak hirukote arruntagoetara transferitzea errazagoa dela adierazten dute, hirukoteetako objektuak doikuntzan zehar ikusi ez badira ere.



5.4 Irudia – SD v2.1 eta SD$_{SR4G}$ v2.1 ereduekin sortutako irudien hainbat irudi, *main* bertsioan doituak. Gure erregela heuristikoak jarraituz, goiburukoan azaltzen den erlazioa zuzen agertzen bada irudian marka berde batekin zehazten dugu. Bestela, ixa gorri bat ipintzen dugu irudiaren goi-eskuinaldeko eskualdean.

Analisi Kualitatiboa

Sortutako irudiak kualitatiboki ebaluatzeko, oinarrizko SD v2.1 eredua eta gure *main* bertsioan doitutako SD$_{SR4G}$ v2.1 eredua aukeratu ditugu. Ebaluazioa burutzeko hirukote espazial ohikoenak eta arraroenak baztertu ditugu. Azken finean, hirukote ohikoenak ondo sortzeko oso errazak dira, ⟨*truck, larger, dog*⟩ kasua bezala, bi objektuak ondo sortzearekin erlazioa betetzea oso erraza baita (irudi gehienetan kamioiak handiagoak baitira txakurrak baino). Hirukote arraroenek, berriz, ez daukate logika handirik eta ez dirudite naturalak, ⟨*bus, shorter, traffic light*⟩ hirukotearen kasua, adibidez. Horregatik, COCO-ko instantzietan 100 eta 1,000 agerpen arteko hirukoteak bakarrik hartu ditugu kontuan.[4]

Analisia egiteko, beraz, irudiak sortzen ditugu ausaz aukeratutako goiburukoak erabiliz, eredu bakoitzeko irudi bat sortuz. Goiburukoan aipatzen diren bi

[4]Maiztasun hauek 5.3. Irudian erakutsitako analisitik erauzi dugu.

objektuak ondo sortzen dituzten lehenengo bederatzi irudi pareekin gelditu gara. Bederatzi irudi pare horiek 5.4. Irudian azaltzen dira. Bertan, irudi bakoitzeko hirukote espaziala ondo sortzen den ala ez zehazten dugu.

5.4. Irudian azaltzen diren hainbat hirukote espazialen sorkuntza *erraza* da: 2, 3, 6, 7 eta 9 zenbakiarekin identifikatzen direnak, alegia. Izan ere, goiburukoan azaltzen diren objektuak sortzean arruntena dagokien erlazioa betetzea da. SD_{SR4G} ereduak denak ondo sortzen ditu 3-ren kasua izan ezik, kasu horretan katilua edo *mug* objektua ez baitago guztiz ikusgarri sortutako irudian (erabaki hau eztabaidagarria izan daiteke). SD ereduak, berriz, bigarren irudia sortzen du gaizki, semaforoa (*traffic light*) ez baitago ondo sortua.

1, 4, 5 eta 8 goiburukoak sortzea, aldiz, zailagoa da. SD_{SR4G} ereduak lau erlazioak zuzen sortzen ditu (*right of* bi aldiz, *overlapping* eta *separated*), baina SDek hirutan oker egiten du. Gaizki sortutako erlazioak interesgarriak dira. 1 eta 8 irudietarako, erlazio espazialak ez dira ohikoenak sarean aurki daitezkeen irudi naturaletan, SD-ek hauek sortzeko arazoak izanik. Hala ere, 5 irudiko txakurra eta aulkia separatuta aurkitzea ohikoa da irudietan, baina SD-ek ez du goiburukoan azaltzen den erlazioa jarraitzen. Honek SD ereduak *separated* erlazioa ezagutzen ez duela iradokitzen du, 5.5. Taulako emaitzak kontrastatuz.

5.4 Ondorioak

Lan honetan erlazio espazial esplizituak dituzten goiburuko eta irudi pareak sortzeko prozesu bat definitu dugu. Prozesu hau erabiliz, SR4G datu multzo sintetikoa sortu dugu, artearen egoerako irudi sortzaileak erlazio hauetan doitzea eta ebaluatzea ahalbidetzen duelarik. Difusio eredu aurrentrenatuen gainean doikuntza burutzeak hobekuntzak dakartzala erakutsi dugu, egungo artearen egoera gaindituz erlazio espazialen sorkuntzan. Doitutako ereduek entrenamenduan ikusi ez diren objektuetara orokortzeko gaitasunak erakusten dituzte. Gainera, analisi sakonagoetan tamaina erlazioen eta erlazio proiektiboen sorkuntzan hobekuntza nabarmenagoak lortzen direla antzeman da, difusio ereduen hasierako alborapenak txikituz eta hobeto orokortuz irudi errealetan ohikoagoak diren erlazioetara.

Etorkizunean sakonera kontuan hartzen duten erlazioak landu nahi ditugu, *in front of* eta *behind* besteak beste. Gainera, erlazio espazialak dituzten goiburuko naturalak jasotzeko metodo berriak aztertu nahi ditugu, artearen egoera giza goiburukoekin ebaluatzeko. Azkenik, interesgarria iruditzen zaigu difusio ereduetan objektuen kokapena kodetzen den eragiketen gaineko azterketa egitea eta galera funtzio berriak definitzea horiek zuzendu eta hobetzeko.

6. CHAPTER

Conclusions and Future Research

In this thesis, we have explored alternative approaches for tackling two limitations of current vision-and-language models (VLMs): world knowledge integration and spatial reasoning. On the one hand, by verbalizing images, we have learned to leverage better the implicit knowledge found in LMs for visual question-answering (VQA) tasks. On the other hand, we have shown the relevance of training image-caption pairs with explicit spatial relations for better spatial reasoning in current VLMs, both for text and image generation. More in-depth, the **main contributions** of this thesis are the following:

- We have developed a VQA system that first generates captions from images, and then works only with textual data. We show that such a system performs surprisingly well in knowledge-intensive tasks like OK-VQA, where questions cannot be answered with images alone, requiring access to external knowledge. Our analysis has concluded that the loss of information is compensated by the better inference ability of text-only pre-trained LMs when summarizing the image into a short description. We have also shown the importance of an LM's capacity when leveraging its implicit knowledge, outperforming contemporary state-of-the-art VLMs by a large margin and matching a 15-times larger ensemble model. The higher capacity of contemporary LMs appears to be an advantage for knowledge-intensive tasks compared to VLMs. Finally, we have noticed that both modalities are complementary, which we have analyzed by fusing visual and textual signals. This first contribution is aligned to research line **RL1**.

- We have presented a novel way to ground spatial relations in text-only language models through location tokens. To make LMs learn the grounding between spatial relations and location tokens, we have also proposed the Synthetic Spatial Training Dataset (SSTD), a textual dataset with unambiguous spatial relations between objects automatically derived from existing images. We have run experiments on a verbalized version of the Visual Spatial Reasoning dataset, where spatial grounding can be tested, showing that our approach to ground spatial relations in LMs is effective, generalizing even to spatial relations not present in SSTD. Our approach obtains better results than contemporary VLMs and, by scaling up our LMs, we get the new state-of-the-art in the VSR dataset. Nevertheless, we observe diminishing returns with the capacity increase, which may suggest that to ground better those spatial relations, the scale of LMs is not determinant. This contribution is associated with research line **RL2**, specifically with **RL2.1**.

- We have defined a dataset generation pipeline to build synthetic captions containing explicit spatial relations from COCO images and their object annotations. This way, we have created the Spatial Relations for Generation dataset (SR4G), containing millions of image-caption pairs for training and thousands of captions for evaluation. Fine-tuning Stable Diffusion models with these image-caption pairs (SD_{SR4G}) outperforms the original diffusion model, as well as surpassing state-of-the-art pipeline models that rely on layout generation. Further experiments have shown that SD_{SR4G} generalizes to unseen objects during fine-tuning. We observe that SD_{SR4G} learns to depict projective and scale relations better, reduces the bias that the original model has for opposite relations like *above* and *below*, and generalizes better to spatial triplets that are more frequent in real images. This contribution is related to research line **RL2**, specifically with **RL2.2**.

In terms of **publications**, parts of this dissertation have been published in peer-reviewed journals. Notably, two papers were published in *Journal Citation Reports (JCR) Q1* ranked journals, and the third one is currently under review.

Apart from them, during my PhD and outside this dissertation, I authored 5 other peer-reviewed papers presented at international conferences (1 ECAI and 2 IkerGazte) or workshops (1 SemEval and 1 SIGUL), as well as a book chapter. The 2 papers presented at IkerGazte were written in Basque, and one of them received the *most relevant research for the development of the Basque Country*

award. Finally, another paper is currently under review at an international top conference (NeurIPS 2024).

Future Research: In the course of this thesis, the main paradigm used in NLP tasks has shifted from using task-specific systems toward developing general-purpose large VLMs and LLMs. This has been enabled thanks to the recent advances in computational power and available data quantity. The latest models have higher capacity and, therefore, show better zero-shot and few-shot capabilities, which enables in-context learning approaches and better generalization capabilities.

Regarding the two limitations that have been tackled in this dissertation, world knowledge integration has been researched far more than spatial reasoning. On the one hand, retrieval augmented generation (RAG) has recently emerged as a key component of LLMs to avoid hallucinations and complement the LLM's strong implicit knowledge with relevant knowledge. Even though this approach already existed before the start of this thesis, it did not kick off until the advent of LLMs (Gao *et al.* 2023), and, nowadays, new works regarding RAG are published or made publicly available every week. On the other hand, spatial reasoning is a more niche domain for researchers and current state-of-the-art LLMs/VLMs still suffer from barely reasoning with spatial relations. Nevertheless, there are still new research lines that we would like to explore on both sides:

- **Use of richer and question-specific descriptions for knowledge-intensive VQA tasks.** As we have seen, verbalizing the input helps the model to better benefit from its implicit knowledge, but this process prunes most information from the image. Selecting the relevant information for each question is key to retrieving the implicit knowledge we need to answer the question. Wu *et al.* (2019) explores this idea, but this approach should be tested again with current state-of-the-art systems. Moreover, integrating new verbalization methods by generating scene graphs or longer captions would also help in this task.

- **Multimodal Retrieval Augmented Generation.** Even though this dissertation focuses on leveraging implicit knowledge, RAG approaches rely on retrieving explicit knowledge found mainly in textual documents. However, RAG has evolved beyond its original focus on text-based question-answering, now enabling it to encompass several modalities. This growth has led to the creation of groundbreaking multimodal models that apply

RAG principles across multiple fields (Yasunaga *et al.* 2023). As some knowledge is not well represented in text corpora (such as common-sense knowledge), exploring approaches that can retrieve this knowledge directly from domain-specific knowledge graphs is an open research line.

- **Usage of location tokens in text-only spatial reasoning tasks.** Many tasks that LMs solve require spatial reasoning, even though they are not grounded in the real world and the meaning of spatial relations is beyond what they can learn. Therefore, we aim to transition to text-only spatial reasoning tasks like SpartQA (Mirzaee *et al.*, 2021) and RESQ (Mirzaee and Kordjamshidi, 2022), where we plan to transform the natural language scene descriptions with explicit spatial relations of those tasks to our textual scene descriptions based on location tokens.

- **Creation of an evaluation dataset containing natural captions with spatial relations.** There is a lack of proper evaluation datasets for spatial reasoning in image generation. That is why the creation of a dataset where each instance is composed of a natural image-caption pair with annotated spatial relations mentioned in the caption would enable more robust automatic evaluations. Currently, existing evaluation datasets are either synthetic or too small to draw statistically relevant conclusions.

- **Development of spatially relevant training objectives for text-to-image generation.** Intermediate layout generation for text-to-image generation has defined the state-of-the-art regarding spatial reasoning. Nevertheless, layout generation requires expensive inferences with LLMs. Hertz *et al.* (2022) observe that diffusion models encode layout information in its cross-attention layers. Therefore, we seek to remove the intermediate layout generation step and keep their performance by defining training objectives that force the model to depict objects in proper positions. Following our work, we could use object annotations as ground truth labels in cross-attention layers to signal proper layouts without explicitly encoding object locations in the input itself.

Bibliography

Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F.L., Almeida D., Altenschmidt J., Altman S., Anadkat S., *et al.*. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alayrac J.B., Donahue J., Luc P., Miech A., Barr I., Hasson Y., Lenc K., Mensch A., Millican K., Reynolds M., *et al.*. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022.

Anderson P., He X., Buehler C., Teney D., Johnson M., Gould S., and Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6077–6086. IEEE Computer Society, 2018.

Antol S., Agrawal A., Lu J., Mitchell M., Batra D., Zitnick C.L., and Parikh D. Vqa: Visual question answering. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2425–2433, 2015.

Assran M., Duval Q., Misra I., Bojanowski P., Vincent P., Rabbat M., LeCun Y., and Ballas N. Self-supervised learning from images with a joint-embedding predictive architecture. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15619–15629, June 2023.

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., and Ives Z. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, 722–735. Springer, Berlin, Heidelberg, 2007.

Ba L.J., Kiros J.R., and Hinton G.E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bagherinezhad H., Hajishirzi H., Choi Y., and Farhadi A. Are elephants bigger than butterflies? reasoning about sizes of objects. *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Bender E.M., Gebru T., McMillan-Major A., and Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. URL https://doi.org/10.1145/3442188.3445922.

Bengio Y., Ducharme R., and Vincent P. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Betker J., Goh G., Jing L., Brooks T., Wang J., Li L., Ouyang L., Zhuang J., Lee J., Guo Y., *et al.*. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Bhakthavatsalam S., Richardson K., Tandon N., and Clark P. Do dogs have whiskers? a new knowledge base of haspart relations. *arXiv preprint arXiv:2006.07510*, 2020.

Bordes F., Pang R.Y., Ajay A., Li A.C., Bardes A., Petryk S., Mañas O., Lin Z., Mahmoud A., Jayaraman B., *et al.*. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., and Amodei D. Language models are few-shot learners. In Larochelle H., Ranzato M., Hadsell R., Balcan M.F., and Lin H., editors, *Advances in Neural Information Processing Systems*, 33 lib., 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., and Zagoruyko S. End-to-end object detection with transformers. *European conference on computer vision*, 213–229. Springer, 2020.

Changpinyo S., Sharma P., Ding N., and Soricut R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3558–3568, 2021.

Chefer H., Alaluf Y., Vinker Y., Wolf L., and Cohen-Or D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Cherti M., Beaumont R., Wightman R., Wortsman M., Ilharco G., Gordon C., Schuhmann C., Schmidt L., and Jitsev J. Reproducible scaling laws for contrastive language-image learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829, 2023.

Chiang W.L., Li Z., Lin Z., Sheng Y., Wu Z., Zhang H., Zheng L., Zhuang S., Zhuang Y., Gonzalez J.E., *et al.*. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Cho J., Zala A., and Bansal M. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3043–3054, 2023a.

Cho J., Zala A., and Bansal M. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023b.

Cho K., van Merriënboer B., Bahdanau D., and Bengio Y. On the properties of neural machine translation: Encoder–decoder approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103–111, 2014.

Chowdhery A., Narang S., Devlin J., Bosma M., Mishra G., Roberts A., Barham P., Chung H.W., Sutton C., Gehrmann S., *et al.*. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Collell G., Van Gool L., and Moens M.F. Acquiring common sense spatial knowledge through implicit spatial templates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 lib., 2018.

Devlin J., Chang M.W., Lee K., and Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Ding M., Zheng W., Hong W., and Tang J. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., *et al.*. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2020.

Driess D., Xia F., Sajjadi M.S., Lynch C., Chowdhery A., Ichter B., Wahid A., Tompson J., Vuong Q., Yu T., *et al.*. Palm-e: An embodied multimodal language model. *International Conference on Machine Learning*, 8469–8488. PMLR, 2023.

Elazar Y., Mahabal A., Ramachandran D., Bedrax-Weiss T., and Roth D. How large are lions? inducing distributions over quantitative attributes. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3973–3983, 2019.

Elu A., Azkune G., de Lacalle O.L., Arganda-Carreras I., Soroa A., and Agirre E. Inferring spatial relations from textual descriptions of images. *Pattern Recognition*, 113:107847, 2021.

Esser P., Rombach R., and Ommer B. Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883, 2021.

Feng W., Zhu W., Fu T.j., Jampani V., Akula A., He X., Basu S., Wang X.E., and Wang W.Y. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023.

Gao L., Biderman S., Black S., Golding L., Hoppe T., Foster C., Phang J., He H., Thite A., Nabeshima N., *et al.*. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y., Dai Y., Sun J., and Wang H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Gardères F., Ziaeefard M., Abeloos B., and Lecue F. ConceptBert: Concept-aware representation for visual question answering. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 489–498, Online, November 2020. Association for Computational Linguistics.

Gokhale T., Palangi H., Nushi B., Vineet V., Horvitz E., Kamar E., Baral C., and Yang Y. Benchmarking spatial relationships in text-to-image generation, 2023.

Goyal Y., Khot T., Summers-Stay D., Batra D., and Parikh D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 6325–6334. IEEE Computer Society, 2017.

Gui L., Wang B., Huang Q., Hauptmann A.G., Bisk Y., and Gao J. Kat: A knowledge augmented transformer for vision-and-language. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 956–968, 2022.

Gurari D., Li Q., Stangl A.J., Guo A., Lin C., Grauman K., Luo J., and Bigham J.P. Vizwiz grand challenge: Answering visual questions from blind people. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

He K., Zhang X., Ren S., and Sun J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society, 2016.

He P., Liu X., Gao J., and Chen W. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020.

Hertz A., Mokady R., Tenenbaum J., Aberman K., Pritch Y., and Cohen-Or D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Hessel J., Holtzman A., Forbes M., Le Bras R., and Choi Y. CLIPScore: A reference-free evaluation metric for image captioning. In Moens M.F., Huang X., Specia L., and Yih S.W.t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.595.

Heusel M., Ramsauer H., Unterthiner T., Nessler B., and Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hu Y., Hua H., Yang Z., Shi W., Smith N.A., and Luo J. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

Hudson D.A. and Manning C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709, 2019.

Ilievski F., Szekely P., and Zhang B. Cskg: The commonsense knowledge graph. *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, 680–696. Springer, 2021.

Jaegle A., Gimeno F., Brock A., Vinyals O., Zisserman A., and Carreira J. Perceiver: General perception with iterative attention. *International conference on machine learning*, 4651–4664. PMLR, 2021.

Jia C., Yang Y., Xia Y., Chen Y.T., Parekh Z., Pham H., Le Q., Sung Y.H., Li Z., and Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. *International conference on machine learning*, 4904–4916. PMLR, 2021.

Jiang H., Misra I., Rohrbach M., Learned-Miller E., and Chen X. In defense of grid features for visual question answering. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10267–10276, 2020.

Johnson J., Gupta A., and Fei-Fei L. Image generation from scene graphs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1219–1228, 2018.

Johnson J., Hariharan B., Van Der Maaten L., Fei-Fei L., Lawrence Zitnick C., and Girshick R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2901–2910, 2017.

Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., and Amodei D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kawar B., Zada S., Lang O., Tov O., Chang H., Dekel T., Mosseri I., and Irani M. Imagic: Text-based real image editing with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017, 2023.

Kim W., Son B., and Kim I. Vilt: Vision-and-language transformer without convolution or region supervision. *International Conference on Machine Learning*, 5583–5594. PMLR, 2021.

Kingma D.P. and Ba J. Adam: A method for stochastic optimization. In Bengio Y. and LeCun Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Koh J.Y., Salakhutdinov R., and Fried D. Grounding language models to images for multimodal inputs and outputs. *International conference on machine learning*, 2023.

Krishna R., Zhu Y., Groth O., Johnson J., Hata K., Kravitz J., Chen S., Kalantidis Y., Li L.J., Shamma D.A., *et al.*. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Krizhevsky A., Sutskever I., and Hinton G.E. Imagenet classification with deep convolutional neural networks. In Pereira F., Burges C., Bottou L., and Weinberger K., editors, *Advances in Neural Information Processing Systems*, 25 lib. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.t., Rocktäschel T., *et al.*. Retrieval-augmented generation

for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Li J., Li D., Savarese S., and Hoi S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International conference on machine learning*, 19730–19742. PMLR, 2023a.

Li L.H., Yatskar M., Yin D., Hsieh C.J., and Chang K.W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Li X., Yin X., Li C., Zhang P., Hu X., Zhang L., Wang L., Hu H., Dong L., Wei F., *et al.*. Oscar: Object-semantics aligned pre-training for vision-language tasks. *European Conference on Computer Vision*, 121–137. Springer, 2020.

Li Y., Liu H., Wu Q., Mu F., Yang J., Gao J., Li C., and Lee Y.J. Gligen: Open-set grounded text-to-image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521, 2023b.

Lin T.Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., and Zitnick C.L. Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer, 2014.

Lin Y., Xie Y., Chen D., Xu Y., Zhu C., and Yuan L. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571, 2022.

Liu F., Eisenschlos J.M., Piccinno F., Krichene S., Pang C., Lee K., Joshi M., Chen W., Collier N., and Altun Y. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*, 2022a.

Liu F., Emerson G., and Collier N. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Liu H., Li C., Wu Q., and Lee Y.J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Liu X., Yin D., Feng Y., and Zhao D. Things not written in text: Exploring spatial commonsense from visual signals. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2365–2376, 2022b.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., and Stoyanov V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

Loshchilov I. and Hutter F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Loshchilov I. and Hutter F. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lu J., Batra D., Parikh D., and Lee S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32:13–23, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Luo J., Khandelwal S., Sigal L., and Li B. Emergent open-vocabulary semantic segmentation from off-the-shelf vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4029–4040, June 2024.

Marino K., Chen X., Parikh D., Gupta A., and Rohrbach M. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14111–14121, 2021.

Marino K., Rastegari M., Farhadi A., and Mottaghi R. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3190–3199. IEEE, 2019.

Mikolov T., Karafiát M., Burget L., Černockỳ J., and Khudanpur S. Recurrent neural network based language model. *Interspeech 2010*, 2010.

Minderer M., Gritsenko A., Stone A., Neumann M., Weissenborn D., Dosovitskiy A., Mahendran A., Arnab A., Dehghani M., Shen Z., *et al.*. Simple open-vocabulary object detection. *European Conference on Computer Vision*, 728–755. Springer, 2022.

Mirzaee R., Faghihi H.R., Ning Q., and Kordjamshidi P. Spartqa: A textual question answering benchmark for spatial reasoning. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4582–4598, 2021.

Mirzaee R. and Kordjamshidi P. Transfer learning with synthetic corpora for spatial role labeling and reasoning. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6148–6165, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.413.

Nichol A.Q., Dhariwal P., Ramesh A., Shyam P., Mishkin P., Mcgrew B., Sutskever I., and Chen M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *International Conference on Machine Learning*, 16784–16804. PMLR, 2022.

Oord A.v.d., Li Y., and Vinyals O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ordonez V., Kulkarni G., and Berg T. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor J., Zemel R., Bartlett P., Pereira F., and Weinberger K., editors, *Advances in Neural Information Processing Systems*, 24 lib. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Parikh A., Wang X., Gehrmann S., Faruqui M., Dhingra B., Yang D., and Das D. ToTTo: A controlled table-to-text generation dataset. In Webber B., Cohn T., He Y., and Liu Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1173–1186, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.89.

Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Kopf A., Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., and Chintala S. Pytorch: An imperative style, high-performance deep learning library. In Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., and Garnett R., editors, *Advances in Neural Information Processing Systems*, 32 lib. Curran Associates, Inc.,

2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Patel R. and Pavlick E. Mapping language models to grounded conceptual spaces. *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gJcEM8sxHK.

Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., and Zettlemoyer L. Deep contextualized word representations. In Walker M., Ji H., and Stent A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1202.

Petroni F., Rocktäschel T., Riedel S., Lewis P., Bakhtin A., Wu Y., and Miller A. Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473, 2019.

Plummer B.A., Wang L., Cervantes C.M., Caicedo J.C., Hockenmaier J., and Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Proceedings of the IEEE international conference on computer vision*, 2641–2649, 2015.

Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., *et al.*. Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763. PMLR, 2021.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., and Liu P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Rajbhandari S., Rasley J., Ruwase O., and He Y. Zero: Memory optimizations toward training trillion parameter models. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16. IEEE, 2020.

Ramesh A., Dhariwal P., Nichol A., Chu C., and Chen M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., and Sutskever I. Zero-shot text-to-image generation. In Meila M. and Zhang T., editors, *Proceedings of the 38th International Conference on Machine Learning*, 139 lib. of *Proceedings of Machine Learning Research*, 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Ren S., He K., Girshick R.B., and Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In Cortes C., Lawrence N.D., Lee D.D., Sugiyama M., and Garnett R., editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 91–99, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

Rombach R., Blattmann A., Lorenz D., Esser P., and Ommer B. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695, 2022.

Ronneberger O., Fischer P., and Brox T. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer, 2015.

Salimans T., Goodfellow I., Zaremba W., Cheung V., Radford A., and Chen X. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Schuhmann C., Beaumont R., Vencu R., Gordon C., Wightman R., Cherti M., Coombes T., Katta A., Mullis C., Wortsman M., *et al.*. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Schwenk D., Khandelwal A., Clark C., Marino K., and Mottaghi R. A-okvqa: A benchmark for visual question answering using world knowledge. *European conference on computer vision*, 146–162. Springer, 2022.

Shah S., Mishra A., Yadati N., and Talukdar P.P. Kvqa: Knowledge-aware visual question answering. *Proceedings of the AAAI conference on artificial intelligence*, 33 lib., 8876–8884, 2019.

Shannon C.E. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

Shao Z., Yu Z., Wang M., and Yu J. Prompting large language models with answer heuristics for knowledge-based visual question answering. *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 14974–14983, 2023.

Sharma P., Ding N., Goodman S., and Soricut R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565, 2018.

Shevchenko V., Teney D., Dick A., and van den Hengel A. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, 1–18, Kyiv, Ukraine, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.lantern-1.1.

Simonyan K. and Zisserman A. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

Singh A., Hu R., Goswami V., Couairon G., Galuba W., Rohrbach M., and Kiela D. Flava: A foundational language and vision alignment model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650, 2022.

Speer R., Chin J., and Havasi C. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31 lib., 4444–4451, 2017.

Su W., Zhu X., Cao Y., Li B., Lu L., Wei F., and Dai J. Vl-bert: Pre-training of generic visual-linguistic representations. *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygXPaEYvH.

Sutskever I., Vinyals O., and Le Q.V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Szegedy C., Ioffe S., Vanhoucke V., and Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the AAAI conference on artificial intelligence*, 31 lib., 2017.

Tan H. and Bansal M. LXMERT: learning cross-modality encoder representations from transformers. In Inui K., Jiang J., Ng V., and Wan X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 5099–5110. Association for Computational Linguistics, 2019.

Tan H. and Bansal M. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2066–2080, 2020.

Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., Bashlykov N., Batra S., Bhargava P., Bhosale S., *et al.*. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tulshan A.S. and Dhage S.N. Survey on virtual assistant: Google assistant, siri, cortana, alexa. *International symposium on signal processing and intelligent recognition systems*, 190–201. Springer, 2018.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., and Polosukhin I. Attention is all you need. In Guyon I., von Luxburg U., Bengio S., Wallach H.M., Fergus R., Vishwanathan S.V.N., and Garnett R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008, 2017.

Wallace B., Dang M., Rafailov R., Zhou L., Lou A., Purushwalkam S., Ermon S., Xiong C., Joty S., and Naik N. Diffusion model alignment using direct preference optimization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238, 2024.

Wang H., Li J., Wu H., Hovy E., and Sun Y. Pre-trained language models and their applications. *Engineering*, 25:51–65, 2023. ISSN

2095-8099. URL https://www.sciencedirect.com/science/article/pii/S2095809922006324.

Wang P., Wu Q., Shen C., Dick A., and van den Hengel A. Explicit knowledge-based reasoning for visual question answering. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 1290–1296, 2017a.

Wang P., Wu Q., Shen C., Dick A., and Van Den Hengel A. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017b.

Wang P., Yang A., Men R., Lin J., Bai S., Li Z., Ma J., Zhou C., Zhou J., and Yang H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *International Conference on Machine Learning*, 23318–23340. PMLR, 2022a.

Wang Z., Li M., Xu R., Zhou L., Lei J., Lin X., Wang S., Yang Z., Zhu C., Hoiem D., *et al.*. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35: 8483–8497, 2022b.

Wang Z., Yu J., Yu A.W., Dai Z., Tsvetkov Y., and Cao Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Wei J., Bosma M., Zhao V., Guu K., Yu A.W., Lester B., Du N., Dai A.M., and Le Q.V. Finetuned language models are zero-shot learners. *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=gEZrGCozdqR.

Wei J., Tay Y., Bommasani R., Raffel C., Zoph B., Borgeaud S., Yogatama D., Bosma M., Zhou D., Metzler D., *et al.*. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022b.

Wei J., Wang X., Schuurmans D., Bosma M., Xia F., Chi E., Le Q.V., Zhou D., *et al.*. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022c.

Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., von Platen P., Ma C., Jernite Y., Plu J., Xu C., Scao T.L., Gugger S., Drame M., Lhoest Q., and Rush A.M. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45, Online, October 2020. Association for Computational Linguistics.

Wu J., Hu Z., and Mooney R. Generating question relevant captions to aid visual question answering. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3585–3594, 2019.

Wu J., Lu J., Sabharwal A., and Mottaghi R. Multi-modal answer validation for knowledge-based vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 lib., 2712–2721, 2022.

Xie N., Lai F., Doran D., and Kadav A. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Xie S., Girshick R., Dollár P., Tu Z., and He K. Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500, 2017.

Xu T., Zhang P., Huang Q., Zhang H., Gan Z., Huang X., and He X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324, 2018.

Yang Z., Gan Z., Wang J., Hu X., Lu Y., Liu Z., and Wang L. An empirical study of gpt-3 for few-shot knowledge-based vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 lib., 3081–3089, 2022.

Yang Z., Wang J., Gan Z., Li L., Lin K., Wu C., Duan N., Liu Z., Liu C., Zeng M., *et al.*. Reco: Region-controlled text-to-image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14246–14255, 2023.

Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R.R., and Le Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Yasunaga M., Aghajanyan A., Shi W., James R., Leskovec J., Liang P., Lewis M., Zettlemoyer L., and Yih W.t. Retrieval-augmented multimodal language modeling. *Proceedings of the 40th International Conference on Machine Learning*, 39755–39769, 2023.

Yu L., Shi B., Pasunuru R., Muller B., Golovneva O., Wang T., Babu A., Tang B., Karrer B., Sheynin S., *et al.*. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2206.10789*, arXiv–2309, 2023.

Zeng A., Attarian M., Choromanski K.M., Wong A., Welker S., Tombari F., Purohit A., Ryoo M.S., Sindhwani V., Lee J., *et al.*. Socratic models: Composing zero-shot multimodal reasoning with language. *The Eleventh International Conference on Learning Representations*, 2022a.

Zeng Y., Zhang X., and Li H. Multi-grained vision language pre-training: Aligning texts with visual concepts. In Chaudhuri K., Jegelka S., Song L., Szepesvari C., Niu G., and Sabato S., editors, *Proceedings of the 39th International Conference on Machine Learning*, 162 lib. of *Proceedings of Machine Learning Research*, 25994–26009. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/zeng22c.html.

Zeng Y., Zhang X., Li H., Wang J., Zhang J., and Zhou W. $X^2$2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3156–3168, 2024.

Zhai X., Kolesnikov A., Houlsby N., and Beyer L. Scaling vision transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12104–12113, 2022a.

Zhai X., Wang X., Mustafa B., Steiner A., Keysers D., Kolesnikov A., and Beyer L. Lit: Zero-shot transfer with locked-image text tuning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18123–18133, 2022b.

Zhang C., Van Durme B., Li Z., and Stengel-Eskin E. Visual commonsense in pretrained unimodal and multimodal models. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5321–5335, 2022.

Zhang P., Li X., Hu X., Yang J., Zhang L., Wang L., Choi Y., and Gao J. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529:5579–5588, 2021.

Glosategia

aro *epoch*

artearen egoera *state of the art*

asmatze-tasa *accuracy*

aurrentrenatu *pretrain*

ausazko mozketa *random crop*

azpimultzo *split*

batezbesteko mugikor esponentziala *exponential moving average*

bateratze *ensemble*

bektore geruza *embedding layer*

bizkarrezur eredua *backbone model*

datu gehikuntza *data augmentation*

datu-multzo *dataset*

doikuntza *fine-tuning*

doitu *to fine-tune*

entropia gurutzatu bitarra *binary cross-entropy*

ereduaren gaitasuna *model capacity*

erro bilaketa *stemming*

eskalatze legea *scaling law*

etiketa, klase *label*

ezagutza behar handiko *knowledge intensive*

ezagutza iturri *knowledge base*

fusio goiztiar *early fusion*

fusio berantiar *late fusion*

ikusizko galdera-erantzute *visual question-answering*

galdera ireki *open ended question*

geruza anitzeko perzeptroi *multilayer perceptron*

geruza ezkutu *hidden layer*

goiburuko sortzaile *caption generator*

hirukote espaziala *spatial triplet*

hizkuntza eredu *language model*

ikasketa gidatu *teacher forcing*

ikasketa tasa *learning rate*

ikusizko inferentzia *visual entailment*

ikusizko hizkuntza-ereduak *vision-and-language models*

ikusmen-testu *visio-linguistic*

iraulketa horizontala *horizontal flip*

irudi eskualdeen ezaugarriak *visual region features*

irudi goiburuko sorkuntza *image captioning*

kaxa inguratzailea *bounding box*

klase leuneko entropia gurutzatua *soft cross entropy*

kokapen token *location token*

modalitate anitz *multimodal*

munduko ezagutza *world knowledge*

objektu-kokapenen sorrera *layout generation*

oinarritu *to ground*

oinarritze *grounding*

posizio bektore *position embedding*

sailkapen buru *classification head*

sailkapen geruza *classification layer*

sorta tamaina *batch size*

sorta tamaina efektiboa *effective batch size*

testu bidezko irudi sorkuntza *text-to-image generation*

testuinguru bidezko ikasketa *in-context learning*

xede orokor *general purpose*

zarata-ezabatze prozesua *denoising process*

A. APPENDIX

Original papers

In this appendix, we present the original papers presented in the manuscript of this thesis in the recommended reading order.

Image Captioning for Effective Use of Language Models in Knowledge-Based Visual Question Answering

Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, Eneko Agirre
HiTZ Center, University of the Basque Country (UPV/EHU)
{ander.salaberria, gorka.azcune, oier.lopezdelacalle,
a.soroa, e.agirre}@ehu.eus

Abstract

Integrating outside knowledge for reasoning in visio-linguistic tasks such as visual question answering (VQA) is an open problem. Given that pretrained language models have been shown to include world knowledge, we propose to use a unimodal (text-only) train and inference procedure based on automatic off-the-shelf captioning of images and pretrained language models. Our results on a visual question answering task which requires external knowledge (OK-VQA) show that our text-only model outperforms pretrained multimodal (image-text) models of comparable number of parameters. In contrast, our model is less effective in a standard VQA task (VQA 2.0) confirming that our text-only method is specially effective for tasks requiring external knowledge. In addition, we show that our unimodal model is complementary to multimodal models in both OK-VQA and VQA 2.0, and yield the best result to date in OK-VQA among systems not using external knowledge graphs, and comparable to systems that do use them. Our qualitative analysis on OK-VQA reveals that automatic captions often fail to capture relevant information in the images, which seems to be balanced by the better inference ability of the text-only language models. Our work opens up possibilities to further improve inference in visio-linguistic tasks.[1]

1 Introduction

Most visio-linguistic tasks are framed in such a way that all the necessary information to solve them is in the images and texts provided in the dataset. That is the case of visual question-answering (VQA) (Antol et al., 2015) or visual entailment (Xie et al., 2019), to name a few. In addition, some tasks require access to external knowledge in order to solve them. An example is *Outside Knowledge VQA* (OK-VQA) (Marino et al., 2019), where the image content is not sufficient to answer the questions. Contrary to self-contained VQA tasks, which can be solved grounding images and text alone, these tasks require methods that leverage external knowledge resources and are able to do inference on that knowledge.

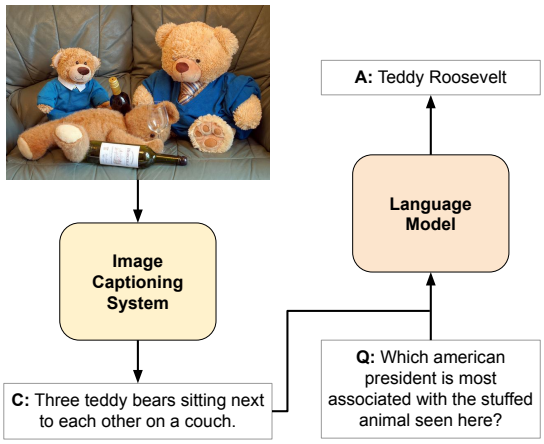


Figure 1: Given a question and image, we verbalize the contents of the image and apply a pretrained language model for inference. We show that current text-only models are better in generalization and inference than multimodal models for knowledge-based QA.

External knowledge useful for OK-VQA can be broadly classified into two categories, according to (Marino et al., 2020): (i) symbolic knowledge, which can be represented using graphs, for example ConceptNet (Speer et al., 2017), and (ii) implicit knowledge, which is encoded in the weights of neural networks trained in different datasets. Supporting the later case, transformer-based language models (LM) pretrained in large corpora like BERT (Devlin et al., 2019) have been successfully used as implicit knowledge bases (Petroni et al., 2019). In any case, the best results on the OK-VQA dataset have been reported by systems that use both pretrained models and symbolic knowledge, usually integrating external knowledge sources (Gardères

[1] Our code will be publicly available soon.

et al., 2020; Marino et al., 2020; Wu et al., 2021; Shevchenko et al., 2021).

In this paper we focus on the use of implicit knowledge in the form of pretrained LMs. While using LMs is relatively common in OK-VQA, they are usually integrated into multimodal transformers by diverse means, so as to integrate the visual and textual inputs of the task. Given that LMs were originally designed to process textual input and are extensively trained in textual corpora, we hypothesized that a system that relies exclusively on text will allow LMs to better leverage their implicit knowledge. Because OK-VQA is a visio-linguistic task, we propose to use automatic image captioning as a way to verbalize the information in the image, where the captions are descriptions of the images which are used as input to the LMs. Once the captions are generated, all the inference in our method is done using text-only models. We are aware that captions do not contain all the information in an image, and want to check whether the text-only models can compensate for that initial loss of information.

The approach proposed in this paper, named caption-based model, can be seen in Figure 1.

To validate our hypothesis, we present an extensive experimentation on the OK-VQA dataset, comparing our proposed caption-based model with the *de facto* standard of visio-linguistic tasks, i.e. multimodal transformers, which are widely used in VQA tasks to process the questions (text) and the images. We also analyze the compatibility between images and captions based on two different fusion strategies. As a result of our experiments, we find that:

- Captions are more effective than images for OK-VQA when pretrained language and multimodal models are used as is, and achieve similar results when both are fine-tuned on additional VQA datasets.

- The combination of the two approaches improves results further, showing that the text-only and multimodal models make complementary inferences.

- The larger contribution of captions on OK-VQA with respect to results on a regular VQA dataset (Goyal et al., 2017) show that our text-only system is specially effective when external knowlede is needed.

- Our combined system is best among systems using implicit knowledge only, and nearly matches the results of state-of-the-art systems that integrate symbolic knowledge graphs.

2 Related Work

There are many **visual question-answering datasets** in the literature (Antol et al., 2015; Goyal et al., 2017; Johnson et al., 2017), where given an image and a question about the contents of that image, a system has to provide a textual answer. Some VQA datasets also demand leveraging external knowledge to infer the answer and, thus, they are known as knowledge-based VQA tasks. Good examples are KB-VQA (Wang et al., 2017b), KVQA (Sanket Shah and Talukdar, 2019), FVQA (Wang et al., 2017a) and OK-VQA (Marino et al., 2019). KVQA requires knowledge about named entities (e.g. Barack Obama, White House, United Nations) and that knowledge is already provided as a graph. FVQA annotates questions by selecting a fact from a fixed knowledge base but its size is relatively small. KB-VQA is even smaller, presenting template-based questions whose answers can be obtained reasoning over commonsense resources or Wikipedia. In contrast, OK-VQA requires knowledge from unspecified external resources and, although smaller than KVQA in terms of the number of images and question-answer pairs, it is considerably bigger than the other knowledge-based VQA datasets.

Currently, **multimodal transformers** are the most successful systems for VQA and can be broadly classified into two types: single-stream and double-stream transformers. A good example of the former is VisualBERT (Li et al., 2019), where the BERT architecture (Devlin et al., 2019) is used, adding visual features obtained by an object detector as input and using visio-linguistic pretraining tasks, such as image-text matching. OSCAR (Li et al., 2020) also follows a very similar philosophy, adding object tags to the input and proposing different pretraining strategies. Among double-stream transformers, VilBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) use a dedicated transformer for each modality (text and image) to fuse them with a cross-modal transformer. Their differences lie mainly on some architectural choices and pretraining task selection.

Regarding **OK-VQA systems**, multimodal transformers have also been used to provide im-

plicit knowledge from pretraining tasks. For example, VilBERT uses a pretrained BERT to encode the questions, so it uses the implicit knowledge that BERT acquired during its pretraining. Additionally, VilBERT is further trained on Conceptual Captions (Sharma et al., 2018), a very large image-caption dataset from where additional knowledge can be acquired. Those multimodal transformers are the backbone of the best performing systems for OK-VQA, which also use symbolic knowledge to bring some extra performance.

ConceptBert (Gardères et al., 2020) was the first system to use multimodal transformers and symbolic knowledge for OK-VQA. It is based on a combination of a pretrained BERT to encode questions, a graph convolutional neural network to encode triples extracted from the ConceptNet knowledge graph (Speer et al., 2017) and a multimodal transformer (VilBERT) to jointly represent and reason over image features and encoded question tokens.

A similar approach was followed by KRISP (Marino et al., 2020), combining again a multimodal transformer with symbolic knowledge. In this case, the multimodal transformer, called MMBERT, is based on VisualBert (Li et al., 2019) and initialized with the weights of a pretrained BERT. Additionally, authors built a knowledge graph fusing DBPedia (Auer et al., 2007), ConceptNet (Speer et al., 2017), VisualGenome (Krishna et al., 2017) and hasPart KB (Bhakthavatsalam et al., 2020). They used different image feature encoders and the question tokens to obtain a subset of the full graph relevant to the target question and image. Finally, using a graph convolutional neural network, they combined the symbolic and implicit knowledge to predict the final answer.

Some recent approaches, named MAVEx (Wu et al., 2021) and RVL (Shevchenko et al., 2021) showed different ways to combine implicit and symbolic knowledge. MAVEx used a pretrained VilBERT to generate various candidate answers which were later validated using answer-specific knowledge retrieval. Authors used both textual and visual knowledge resources, including images searched using Google, sentences from Wikipedia articles, and concepts from ConceptNet. On the other hand, RVL trained the two-stream multimodal transformer LXMERT (Tan and Bansal, 2019) with an auxiliary objective that aligned its representations with knowledge graph embeddings retrieved from ConceptNet and Wikidata.

Regarding the use of **captions for VQA**, to the best of our knowledge, Mucko (Zhu et al., 2020) is the only system that explores this idea. Mucko uses dense captions (Johnson et al., 2016) to query a knowledge graph to extract relevant information to answer the question. The reported results on OK-VQA are well below the state-of-the-art. Dense captions describe different regions of an image using short sentences. Our method differs in the use of a single caption which is the input to the LM, and does not require any knowledge graph.

3 Implemented models

In this section we describe the implemented models. We use Pytorch (Paszke et al., 2019) and the Transformers library (Wolf et al., 2020) for all the implementation work.

3.1 Caption-based model (CBM)

Our caption-based model, denoted by CBM, is divided in two steps: (i) a caption generation system that generates a short description of a given image and (ii) a language model that takes this caption and a question in order to answer it.

We use **OSCAR** (Li et al., 2020) to generate captions from images, a transformer encoder that produces state-of-the-art results on several multimodal tasks including image captioning. As it is common in multimodal transformers, OSCAR uses a pretrained object detector called FasterRCNN (Ren et al., 2015) to obtain region features from images and their respective labels. Both features and labels alongside manually annotated captions are then fed to the transformer during pretraining, following the work of (Anderson et al., 2018). The performance on image-captioning of both base and large models is similar, so we use OSCAR-base as our image-captioning system for all of our experiments.

During OSCAR's fine-tuning step on image captioning, some of OK-VQA's test split images and gold captions are used. In order to ensure fairness and avoid any contamination in our experiments, we fine-tune a pretrained OSCAR model on image-captioning removing these instances from its training process.

On the other hand, the LM we use in all the experiments is a pretrained **BERT-base** model (Devlin et al., 2019). We feed sequences of tokenized captions and questions $T^{(0)} = \{\mathbf{t}_i^{(0)} | i = 1, \ldots, n_t\}$ to BERT, and take the output of the

$[CLS]$ or first token of the sequence $\mathbf{t}_1^{(n_l)}$, where n_t is the number of tokens in the sequence and n_l is the number of transformer layers.

Although VQA (Antol et al., 2015; Goyal et al., 2017) and OK-VQA (Marino et al., 2019) were defined with open-ended answers, recent state-of-the-art models (Zhang et al., 2021; Marino et al., 2020) cast these tasks as classification problems, building a fixed vocabulary of answers from the training dataset. In order to fine-tune the language model for VQA, we add a **classification head** to the $[CLS]$ embedding. Our classification head is a multilayer perceptron (MLP) with one hidden layer after $\mathbf{t}_1^{(n_l)}$. We define our MLP in Eq. 1.

$$\begin{aligned}\mathbf{h} &= \text{LayerNorm}(\text{GELU}(\mathbf{W}_h\mathbf{t}_1^{(n_l)} + \mathbf{b}_h)) \\ \hat{\mathbf{y}} &= \text{Softmax}(\mathbf{W}_{\hat{y}}\mathbf{h} + \mathbf{b}_{\hat{y}})\end{aligned} \tag{1}$$

We use a GELU activation function as well as layer normalization (Ba et al., 2016). The trainable parameters are $\mathbf{W}_h \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_h \in \mathbb{R}^{d_h}$, $\mathbf{W}_{\hat{y}} \in \mathbb{R}^{d_h \times n_{label}}$ and $\mathbf{b}_{\hat{y}} \in \mathbb{R}^{n_{label}}$, where n_{label} equals to the number of labels on a given classification task and d_h equals to 768.

3.2 Question-only baseline (BERT_Q)

In order to assess the contribution of captions, we also trained a model which only had the question in the input, without any information about the image or caption, denoted as BERT_Q. This model can be seen as an ablation of CBM.

3.3 Multimodal transformer (MMBERT)

We compare our CBM model with the multimodal transformer-based MMBERT (Marino et al., 2020), a variant of BERT that uses the question text and image region features as input. While BERT is designed to only process textual inputs, MMBERT adapts its embedding layer in order to be able to process features from images.

We use a FasterRCNN with a ResNeXt-152 (Xie et al., 2016) as its backbone to extract a total of n_v region features $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{n_v}\}$ per image. Each of these $\mathbf{v}_i \in \mathbb{R}^{d_v}$ features represents an object that appears in the image, where d_v equals to 2048. $\mathbf{V}$ lacks the positional information between objects, which can be solved concatenating the corresponding bounding box coordinates to each feature. Upon some initial experiments, we concluded that this extra information does not improve performance in any of VQA 2.0 and OK-VQA. We use MMF Multimodal Framework (Singh et al., 2020) to extract the image region features that are fed into MMBERT.

In order to allow for easier comparison between our CBM and MMBERT we use the output representation for $[CLS]$ to feed into the classification multilayer perceptron (see Section 3.1). Note that this is slightly different from the original MMBERT (Marino et al., 2020), which uses the average of all token representations in the last transformer layer.

3.4 Loss function

Contrary to previous works in VQA, we do not use binary cross-entropy loss, as initial experiments showed that cross-entropy loss with soft labels (SCE) converges faster with similar results. SCE loss is defined in Eq. 2, where $\mathbf{y}$ is the ground truth vector with probabilities proportional to the VQA evaluation metric (Eq. 3) assigned to each class.

$$\mathcal{L}_{SCE}(\mathbf{y}, \hat{\mathbf{y}}) = -\mathbf{y} \cdot \log \hat{\mathbf{y}} \tag{2}$$

3.5 Combining both modalities

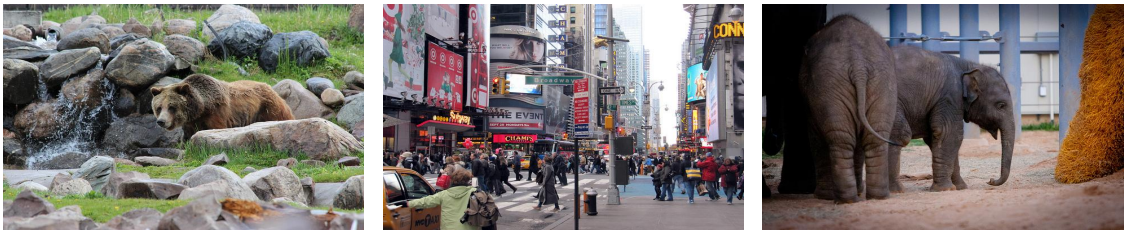
We are also interested in analyzing the complementarity of both models, i.e. the text-only modality using questions and captions, and the image-text modality with image region features and questions. Therefore, we define two different approaches to check how they complement each other.

Early fusion. For each question we feed both caption and image features alongside the question to the language model. This system can be seen as a MMBERT which processes a multimodal input composed by a question (text), a caption (text) and image region features. We initialize the weights of this model with the weights of the base language model (BERT-base) and fine-tune it on the target train data.

Late fusion. We train the caption-based model (Section 3.1) and MMBERT (Section 3.3) separately, each of them with their corresponding inputs, and combine their outputs in inference time to obtain the final answer. The combination is done by multiplying output probabilities of both models for each class and taking the answer with the highest value.

4 Datasets

The main dataset for our experiments is OK-VQA (Marino et al., 2019), since it allows us evaluating



VQA: What color is the bear? brown
OK-VQA: What species of bear is this? grizzly

VQA: What is the weather like? cloudy
OK-VQA: Why would one suspect that this is not chicago? sign

VQA: Are the animals in captivity? yes
OK-VQA: Which valuable material grows on this animal's face? ivory

Figure 2: Some examples of VQA 2.0 and OK-VQA datasets for the same images. VQA questions are about image contents, while OK-VQA questions require outside knowledge.

the usage of the implicit knowledge of LMs in a multimodal task. But we also run experiments on the VQA 2.0 dataset (Goyal et al., 2017) with a double motivation: (i) to use it as additional pre-training before applying the model to OK-VQA; (ii) to analyze the performance differences among models on a knowledge-based VQA dataset and a standard VQA dataset. Examples of both datasets can be found in Figure 2.

4.1 VQA 2.0

This dataset contains open-ended questions about images where questions focus on identifying objects in the image and their attributes, detecting relations between them, as well as counting those objects. The dataset is composed of 204K images taken from the COCO dataset (Lin et al., 2014) and 1.1M questions, each question having 10 (possibly repeated) annotations as accepted answers. Following the classification setting of VQA tasks, which is currently the dominant paradigm, VQA 2.0 has 3129 different possible answers, extracted from the most frequent answers of the training split.

VQA 2.0 is divided in three splits named train, dev and test. Some of the images from the development split of VQA 2.0 are reused in OK-VQA's test split. So, in order to avoid any contamination, we do not use the VQA 2.0 dev set for any training or hyper-parameter tuning.

(Antol et al., 2015) proposed a standard evaluation metric for VQA tasks where a system answer is considered totally correct if it appears at least three times in the ten ground-truth annotations. Considering that a given answer appears x times in a question's annotations, this accuracy metric is defined in Eq. 3.

$$\text{acc} = \min\left(\frac{x}{3}, 1\right) \tag{3}$$

4.2 OK-VQA

The OK-VQA dataset is built upon 14,031 images from the COCO dataset and 14,055 crowd-sourced questions. Each question has ten annotated answers (possibly repeated), and the evaluation metric is the same as in VQA 2.0 (Eq. 3). As a knowledge-based VQA dataset, OK-VQA requires outside knowledge to answer the questions. However, this outside knowledge is neither provided nor identified, i.e. there is not a list of available knowledge sources for this task, making the task more challenging.

There are two versions of this dataset, depending on how the stemming of the answers provided by the crowd-sourcers is handled. The stemming used in OK-VQA v1.0 results in some "non-word" answers (such as "poni tail" instead of "pony tail"). OK-VQA v1.1 applied a different stemming algorithm, resulting in a more coherent answer vocabulary. We use OK-VQA v1.1 through our experiments, except for the state-of-the-art comparison, as most published systems report results on the v1.0 version.

5 Experiments and results

This section provides results of the models defined in Section 3 and compare them with the state-of-the-art.

5.1 Experimental settings

We use the same hyperparameters as (Marino et al., 2020) for fine-tuning CBM, MMBERT, BERT_Q and Early fusion models both in VQA 2.0 and OK-VQA tasks. We train our models for 88K steps using AdamW optimizer (Loshchilov and Hutter, 2019). Our batch size is of 56 with a maximum learning rate of $5 \cdot 10^{-5}$ following a cosine schedule with a linear warmup of 2K steps. All experiments have been run in a single GPU with 12GB of vRAM

Model	Acc.
Without VQA pretraining	
BERT$_Q$	21.2 ±0.2
MMBERT	29.6 ±0.6
CBM (ours)	**32.5** ±0.4
With VQA pretraining	
BERT$_Q$	23.0 ±0.2
MMBERT	35.7 ±0.3
CBM (ours)	**36.0** ±0.4

Table 1: Performance on OK-VQA for the three models (respectively, question only, image-based and caption-based) withouth and with additional pretraining on VQA 2.0. Mean accuracy and standard deviation across 3 different runs.

and their runtimes are at most of 12 hours.

5.2 Images vs. captions

Table 1 shows the results for the three models presented in Section 3, which share the same architecture and initial parameters. Topmost rows for the models fine-tuned only on OK-VQA (tagged as "Without VQA pretraining"), and the bottom rows for the same models which have been fine-tuned on VQA 2.0 before being fine-tuned on OK-VQA.

We observe that the sole use of questions BERT$_Q$ offers poor performance compared to the other two systems, achieving up to 13 points less accuracy. This shows that having any representation of the image (captions or image region features) is key to answer questions correctly. This is further justified comparing the improvement that VQA pretraining entails, as BERT$_Q$ improves less than 2 points, whereas the other two improve their accuracy between 4-6 points.

Contribution of captions. When we compare the performance of CBM and MMBERT, we see that, when there is no visio-linguistic pretraining involved, CBM performs better in OK-VQA. However, when we pretrain these models in a similar multimodal task like VQA 2.0, their accuracy increases by 4-6 points and both obtain similar performance. As OK-VQA's training is comparatively smaller (9K instances vs. VQA's 410K instances), we hypothesize that training MMBERT on OK-VQA is not enough to adapt the model to the new input modality. However, as CBM uses only text, the fine-tuning with such small training is more effective.

Model	Acc.
Without VQA pretraining	
Early fusion	32.5 ±0.4
Late fusion	**34.0** ±0.4
With VQA pretraining	
Early fusion	38.2 ±0.8
Late fusion	**38.6** ±0.2

Table 2: Performance on OK-VQA for early and late fusion models. Mean accuracy and standard deviation across 3 different runs.

5.3 Combining CBM and MMBERT

Given the different nature of the inputs, we wanted to check whether CBM and MMBERT are complementary. Our hypothesis is that the former can take advantage of the implicit knowledge acquired by the language model, whereas the latter has access to more fine-grained information found in image regions. Following the approaches of early and late fusion defined in Section 3.5, we show their performance in Table 2.

These fusion models improve the performance of both CBM and MMBERT by 2-3 points in almost all cases. The only case where there is no improvement comparing to CBM is in the early fusion without VQA pretraining. This may be caused again by the small training split of OK-VQA, causing difficulties to learn how to ground textual and visual modalities. However, this is solved when VQA pretraining is added to the model, increasing vastly the amount of data seen by the models and showing similar performance on both early and late fusion models. The results validate our hypothesis, showing that image region features and captions are complementary.

5.4 Comparison with the state of the art

To compare our models with state-of-the-art models in OK-VQA, we had to repeat the experiments in OK-VQA v1.0. The results vary slightly, as can be seen in Table 3. In that table, we show the results of various models using only implicit knowledge and combining it with symbolic knowledge. As our models do not use symbolic knowledge, the corresponding column is empty.

The performance of KRISP (Marino et al., 2020), MAVEx (Wu et al., 2021) and RVL (Shevchenko et al., 2021) is very similar. But RVL has a contamination issue as images from OK-VQA's test split were used to train their multimodal transformer. In

Model	Imp.	+Sym.
ConceptBERT	31.4	33.7
KRISP	*36.3	38.4
RVL	†37.3	†39.0
CBM (ours)	36.3	-
Late fusion (ours)	**39.2**	-

(a) Results on OK-VQA v1.0.

Model	Imp.	+Sym.
MAVEx	35.2	38.7
KRISP	37.1	38.9
CBM (ours)	36.0	-
Late fusion (ours)	**38.6**	-

(b) Results on OK-VQA v1.1.

Table 3: Comparison to the state-of-the-art on OK-VQA. Results are divided in two tables, one per OK-VQA version. Topmost rows of each table are taken from respective papers, except *, computed by us. Imp. for implicit knowledge, +Sym. for systems additionally using symbolic knowledge. † for contaminated results (see main text).

Table 3 we observe that using symbolic knowledge improves the results around 2 points in average. The highest improvement is achieved by MAVEx with 3.5 points[2]. Notice that all four systems use different ways to integrate symbolic knowledge from different resources.

If we look at our caption-based model CBM, we see that its performance is on par with the multimodal transformers used by the other systems. We believe this is remarkable, since we do not use directly any visual features in our models. Furthermore, when we use late fusion, the results we obtain are comparable to the systems which also use symbolic knowledge. Notice that we only use implicit knowledge for our systems and match the performance of systems which combine implicit and symbolic knowledge.

6 Analysis

In this section we first contrast the results on OK-VQA with those obtained in VQA 2.0, discussing the reasons for the different performance. We then compare the performance of CBM with manually annotated captions or the ones generated by OSCAR (Li et al., 2020); and, finally, we present some qualitative analysis.

[2]An ensemble composed by 5 MAVEx models with the same multimodal transformer achieves an accuracy of 39.4%.

Model	Acc.
MMBERT	65.8
CBM (ours)	59.6
Early fusion (ours)	**67.8**
Late fusion (ours)	67.7

Table 4: Performance on dev split of VQA 2.0.

6.1 Results on VQA 2.0

Even though both unimodal and multimodal methods perform similarly in OK-VQA, we observed a different trend in VQA 2.0. Table 4 shows that CBM obtains 59.6, while MMBERT achieves 6 points more. We think this is due to the information loss when converting an image into a caption, as relevant information that is needed to answer the question can be lost. This is specially important for VQA 2.0, where the questions refer directly to image contents, spatial relations and object attributes (see Figure 2). Captions do not usually provide that additional information, and tend to focus on the description of the most relevant information. However, looking at the performance in OK-VQA, we see that captions contain enough information to effectively use the implicit knowledge of the BERT language model.

Regarding early and late fusion models, both of them improve the performance of MMBERT by 2-3 points, showing that our model is complementary to multimodal methods also in the VQA dataset.

6.2 Ground truth captions

In order to measure the effects of the image captioning system to our proposed CBM model, we check the gap of performance between the use of generated captions and gold captions. As OK-VQA is built upon images from the COCO dataset (Lin et al., 2014), each image has five different annotated captions. We use these captions and fine-tune CBM on OK-VQA without VQA pretraining following the same experimental settings. We repeat this experiment three times, as it is done through the entire work. On each run we select a different set of captions, that is, for each image we just choose one gold caption randomly and use it during the entire training process. As we also have several captions in all of OK-VQA's test split, we test each fine-tuned model three times following the same caption selection process.

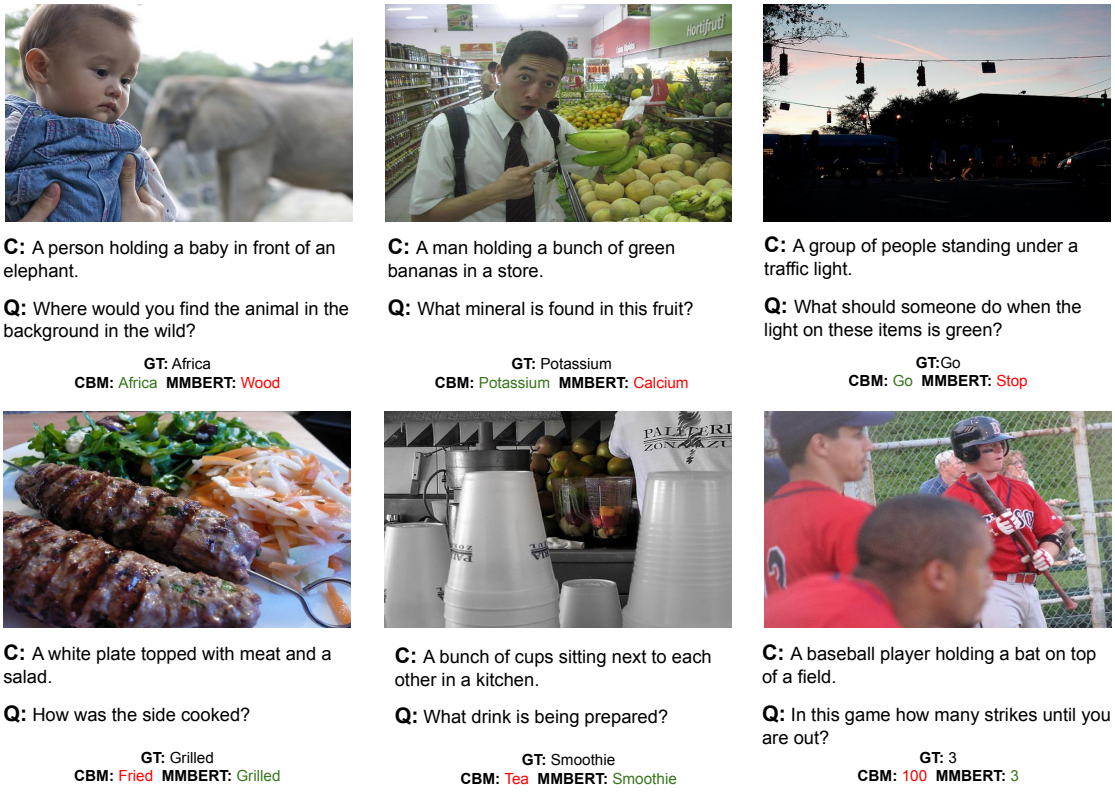


Figure 3: Examples of OK-VQA questions where only one of the two models (CBM or MMBERT) answers correctly according to the ground truth (GT). C refers to captions generated by OSCAR. Correct answer in green, incorrect in red.

Table 1 already shows that we achieve an accuracy and standard deviation of 32.5 ± 0.4 using generated captions on OK-VQA's test split. However, when we use gold captions we get an average accuracy of 32.3 ± 0.3 in all of our runs. In both cases we obtain similar results, showing that captions generated by OSCAR contain enough information for CBM to perform comparably on this specific task.

6.3 Qualitative Analysis on OK-VQA

Both unimodal and multimodal algorithms perform similarly (see Table 1), but in 38.7% of the test examples their output differs and only one of them is correct. Figure 3 shows some OK-VQA test examples together where the outputs of CBM and MMBERT with VQA pretraining differ.

Starting with the top-left example, CBM can infer that elephants are native to Africa whereas MMBERT does not. In fact, the generated caption includes the information that the animal found in the image is an elephant, performing the first step needed to answer the question. This way, the LM can focus on using its implicit knowledge in order to answer correctly.

The other two examples in the top row behave similarly. The caption facilitates the grounding between the question and the image. Whenever a question refers to the image ("this fruit" and "these items"), if the caption already mentions these objects ("bananas" and "traffic light", respectively), the LM seems to better leverage its implicit knowledge and reasoning capabilities to answer the question. The top-right example is interesting in this regard. While the image shows red traffic lights, the question asks about the effects of green lights. This may trick MMBERT into answering the effect that red lights produce, not the green ones.

The bottom row of Figure 3 shows three examples where the caption does not give enough information to infer the answer. In the first case CBM can not decide whether the meat is steamed, fried or grilled by only examining the caption, while MMBERT does have access to visual cues of the image, where we can see that the meat is grilled. This also happens in the second example, as the caption does not specify any ingredient of the bev-

erage while we can see fruits in the image. The rightmost example illustrates an example where the caption could support the inference, but where CBM is wrong: with the given caption, "this game" refers to baseball, however, CBM is unable to infer that three strikes are enough for a strikeout whereas MMBERT manages to give the correct answer.

All in all, these examples support our hypothesis that visual features and captions are complementary. They also show that our system has some advantages regarding the interpretability of the system, specially in the cases our method is wrong. In some cases like the two leftmost examples in the bottom row, the object or feature needed to answer the question is missing from the caption. In other cases, the required information is in the caption, but the inference is erroneous.

7 Conclusions

In this paper we present a VQA system which describes images with a caption to then ignore the image completely. We show that such a system performs surprisingly well in OK-VQA, where the questions cannot be answered with the image alone, requiring access to external knowledge. Our analysis indicates that the loss of information when summarizing the image into a caption is compensated by the better inference ability of text-only pretrained language models. In the future we would like to explore whether richer descriptions of images might improve results further, and whether text-only language models are more effective when incorporating symbolic knowledge graphs than current multimodal models.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086. IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. 2020. Do dogs have whiskers? a new knowledge base of haspart relations. *arXiv preprint arXiv:2006.07510*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language

tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2020. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. *arXiv e-prints*, pages arXiv–2012.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Naganand Yadati Sanket Shah, Anand Mishra and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *AAAI*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. 2021. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. In *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 1–18, Kyiv, Ukraine. Association for Computational Linguistics.

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017a. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R Dick, and Anton van den Hengel. 2017b. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2021. Multi-modal answer validation for knowledge-based vqa. *arXiv preprint arXiv:2103.12248*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *CoRR*, abs/2006.09073.

GROUNDING SPATIAL RELATIONS IN TEXT-ONLY LANGUAGE MODELS

Gorka Azkune, Ander Salaberria, Eneko Agirre
HiTZ Basque Center for Language Technologies - IXA NLP Group
University of Basque Country (UPV/EHU)
Donostia, Basque Country, Spain
{gorka.azcune, ander.salaberria, e.agirre}@ehu.eus

ABSTRACT

This paper shows that text-only Language Models (LM) can learn to ground spatial relations like *left of* or *below* if they are provided with explicit location information of objects and they are properly trained to leverage those locations. We perform experiments on a verbalized version of the Visual Spatial Reasoning (VSR) dataset, where images are coupled with textual statements which contain real or fake spatial relations between two objects of the image. We verbalize the images using an off-the-shelf object detector, adding location tokens to every object label to represent their bounding boxes in textual form. Given the small size of VSR, we do not observe any improvement when using locations, but pretraining the LM over a synthetic dataset automatically derived by us improves results significantly when using location tokens. We thus show that locations allow LMs to ground spatial relations, with our text-only LMs outperforming Vision-and-Language Models and setting the new state-of-the-art for the VSR dataset. Our analysis show that our text-only LMs can generalize beyond the relations seen in the synthetic dataset to some extent, learning also more useful information than that encoded in the spatial rules we used to create the synthetic dataset itself.

***K*eywords** Spatial relations · Grounding · Language Models

1 Introduction

Spatial relations like *left of* or *on top of* can be naturally grounded to images. Thus, Vision-and-Language Models (VLM) seem the most suitable option to ground the textual form to real world concept usage. However, general-purpose VLMs such as CLIP [Radford et al., 2021], VisualBERT [Li et al., 2019], LXMERT [Tan and Bansal, 2019] or ViLT [Kim et al., 2021] have been shown to struggle to ground spatial relations properly [Liu et al., 2022a,b]. The situation is even worse for text-only LMs, which lag behind VLMs for spatial grounding [Liu et al., 2022a].

Spatial grounding and reasoning are very interesting for text-only tasks, as shown by various works [Liu et al., 2022a, Mirzaee et al., 2021, Mirzaee and Kordjamshidi, 2022]. One alternative to solve those text-only tasks would be using VLMs and feed them only with textual inputs. However, some researchers already identified that the language used to train those VLMs is not as rich and varied as the language used for text-only tasks [Tan and Bansal, 2020], which hinders the potential of VLMs for text-only tasks.

In this paper, we explore another avenue and we focus on spatial grounding for text-only LMs. Following the current trend of translating visual information into textual information [Yang et al., 2022, Zeng et al., 2022, Wang et al., 2022, Liu et al., 2022c], we propose to use textual tokens in a novel way to represent real-world scenes and leverage pretrained LMs. More concretely, we propose to use location tokens to represent the positions and spatial extent of objects in a scene. Our hypothesis is that those location tokens offer a way to ground spatial relations in the LM.

To validate that hypothesis, we run experiments on a verbalized version of the multimodal Visual Spatial Reasoning (VSR) dataset [Liu et al., 2022b]. The dataset contains image-caption pairs, where the caption mentions a spatial relation between two objects of the image, plus a true/false label, depending if the caption is true for the image. To approach this task with a text-only LM, we use an off-the-shelf object detector, which returns object labels and their

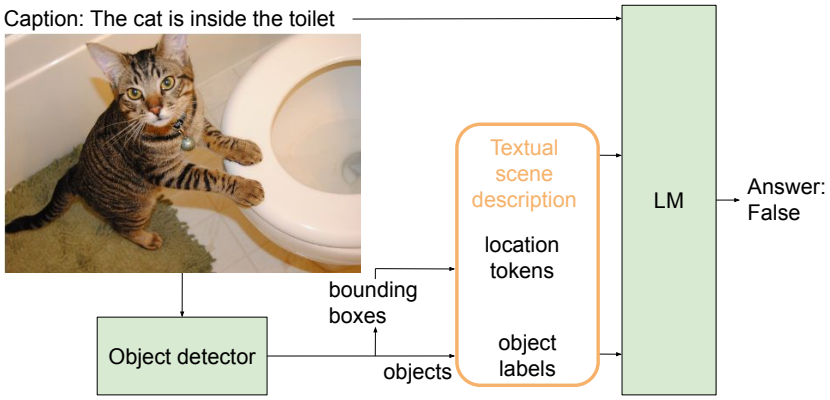


Figure 1: Given an image and a caption with a spatial relation, the task in VSR is to output whether the caption is true for the image. We propose a text-only alternative of the dataset, where an off-the-shelf object detector returns the labels and locations (derived from the bounding boxes), which are used as the textual description of the scene depicted in the image. The description and caption are input to a LM, to test its spatial grounding capabilities.

bounding boxes (BB). We convert the BB coordinates to four location tokens. We prepend the location tokens to the corresponding object label (e.g. *cat*), and build a **textual scene description** that represents the contents and locations in a given scene (Figure 1). Then, we only concatenate the provided caption with the aforementioned textual scene description and train a LM for binary classification (Figure 1). This way, we can test the spatial grounding capabilities of a text-only LM.

As a result of our experiments we show that:

1. Location tokens are effective to ground spatial relations, as shown by the positive results of our model.
2. The training set of VSR is too small for learning how to ground spatial relations to locations, but an automatically produced synthetic dataset of spatial relations allows to do so, while a LM without locations fails.
3. The LMs trained on the synthetic dataset can generalize to some extent to spatial relations that have not been observed in the synthetic data. Specially interesting is to see the performance boost for relations that require depth information.
4. Our text-only LMs outperform baseline VLMs for VSR, obtaining the best results for the VSR task to date.
5. Our text-only LMs clearly outperform a rule-based baseline, showing that the LMs learn more information than that encoded in the manually defined spatial rules.

Our code, models and datasets are freely available[1].

2 Related work

Some authors suggest that grounding is one of the key elements to bring human-like language understanding [Bender and Koller, 2020]. However, grounding covers a great diversity of techniques, modalities and concepts [van der Velde, 2015, Laflaquière et al., 2018]. Thus, this paper is focused on spatial relations and their grounding. In that sense, there are two major domains related to this paper: how spatial grounding can be evaluated (Section 2.1), and how spatial information is represented in current deep learning models, covering VLMs - which are the current paradigms of how to ground text on visual data - and text-only LMs (Section 2.2).

2.1 Datasets for spatial grounding

The spatial commonsense knowledge of current LMs and VLMs is evaluated from different angles. For example, [Bagherinezhad et al., 2016, Elazar et al., 2019] focus on the acquired commonsense knowledge of models about object scales, e.g. do they know that a person is bigger than an ant? In that sense, they do not provide a specific scene context, but rather ask about generic object scale relations, so the dataset they provide is not useful for our work.

[1] https://github.com/gazkune/SpatialLM

Some other authors, [Collell et al., 2018, Elu et al., 2021] propose datasets and methods to generate bounding boxes from textual descriptions. Although the evaluation approach is suitable to test spatial grounding, they focus on implicit spatial relations, whereas our focus is on explicit relations. Thus, the proposed datasets are not suitable for our analysis.

With the objective of evaluating both object scales and spatial relations, a recent work provides new unified datasets [Liu et al., 2022a]. As the objective of such work is to evaluate whether VLMs learn more spatial commonsense than LMs, the datasets are purely textual, so they do not provide any means to ground spatial relations (they assume the grounding occurs in a previous training process) and hence, they are not useful for our work. Interestingly, authors find that VLMs, and more concretely text-to-image systems, perform much better than text-only LMs.

There are other ways to test the spatial inference and reasoning capabilities of models, though. CLEVR was one of the pioneering works on testing compositional language and elementary visual reasoning [Johnson et al., 2017]. Using 3D rendered images of simple objects such as spheres, cones and cubes, different questions are generated automatically. A model has to process the image and the question to provide an answer. Although CLEVR can be used to test spatial grounding, it has two major drawbacks for the work presented in this paper: i) questions not only cover spatial grounding but some other concepts such as compositional language and attribute identification, and ii) spatial relations are limited to four, i.e. *left*, *right*, *behind* and *in front*. The natural extension of CLEVR is GQA [Hudson and Manning, 2019], which shares similar ideas but it is built on natural images. Although spatial grounding is very important for this task, compositional language is also evaluated. As both dimensions appear together, we believe this dataset is not the best option for our purposes.

In the text-only scenario, SpartQA provides another synthetic question-answering dataset (there is also a subset annotated by humans). Given a textual story (a spatial description of a scene using explicit relations), a model has to answer some spatial questions about that scene. The task is specially focused on spatial reasoning capabilities, such as transitivity, and it does not provide any means to ground spatial relations, as its target is the reasoning process. Recently, similar datasets haven been proposed as an extension and improvement of SpartQA [Mirzaee and Kordjamshidi, 2022].

In this paper, we use the recent Visual Spatial Reasoning (VSR) dataset [Liu et al., 2022b] to evaluate the spatial grounding capabilities of text-only LMs. VSR has been designed to test spatial grounding capabilities, covering 65 different spatial relations over natural images collected from COCO [Lin et al., 2014]. Given an image, they provide a caption which describes a spatial relation between two of the objects that appear in the image. That relation can be real or fake, and that is precisely what the model has to infer, i.e. whether the caption is correct respect to the given image. The dataset is fully annotated by humans. Given its features, we believe VSR is a good candidate to evaluate spatial grounding in LMs and thus, we use it in our experiments. However, as text-only LMs cannot process images, we propose a way to verbalize those images and run meaningful experiments.

2.2 Encoding spatial information

The most successful VLMs today are based on multimodal transformers [Tan and Bansal, 2019, Kim et al., 2021]. Although architectures may vary, the basic idea is to input the models with textual tokens and visual features. As transformers are feed-forward networks, they do not consider the order of the inputs, and thus, positional encodings are used to represent, for example, word order [Vaswani et al., 2017]. A similar idea is used also for visual features. LXMERT [Tan and Bansal, 2019], for instance, uses the x_0, y_0, x_1, x_2, W, H coordinates of a bounding box for a given visual feature, projects them linearly and sums it to the visual feature itself before inputting it to the transformer. Alternatively, ViLT [Kim et al., 2021] does not use any object detector, but works directly on image patches. They use positional embeddings to represent the order of those patches in the image, very similar to the positional embeddings of textual tokens.

Regarding text-only LMs, to the best of our knowledge, [Patel and Pavlick, 2022] is the only work where scenes are represented with textual tokens on which spatial grounding and reasoning can be performed. More concretely, they propose to create grid-like structures with textual tokens inside the vocabulary of the LM. Their proposal is interesting, but it is limited to toy experiments, since they can only represent *small* scenes and six spatial relations: *left*, *right*, *up*, *down*, *top* and *bottom*. In contrast, with our approach we cover complex scenes depicted in natural images and 23 spatial relations (Table 1).

3 The VSR dataset

The VSR dataset contains natural image-text pairs to test the spatial grounding capabilities of machine learning models. As can be seen in Figure 2, a textual description of an image is provided, where the spatial relation of two objects is explicitly described. The spatial relation can be true or false. To solve the task properly, models have to be able to

Caption: The person is ahead of the cow.
Label: True.

Caption: The cat is inside the toilet.
Label: False.

Figure 2: Two examples extracted from the VSR dataset.

ground around 65 different spatial relations, which are grouped in 7 categories: adjacency, directional, orientation, projective, proximity, topological and unallocated.

The dataset has two splits: the *random* split and the *zero-shot* split. The later is designed such that train/dev/test sets have no overlapping concepts and force models to learn concepts and the relations in a compositional way instead of memorizing co-occurrence statistics of the two. However, it is smaller than the random split, which has a total of 10,119 examples. The zero-shot split has 6,430 image-text pairs in total.

According to the experiments performed in the VSR dataset by authors [Liu et al., 2022b], the best VLMs are far from human performance. While humans obtain an accuracy of 95.4 for both splits, the best model for the random split, i.e LXMERT [Tan and Bansal, 2019], is around 70.1 and it performs worse in the zero-shot split (63.0). This performance gap between humans and VLMs shows that there is still much work to do to better ground spatial relations.

4 Learning to ground spatial relations in text-only LMs

In this paper, we propose to ground spatial relations in LMs introducing the concept of *location tokens*. These location tokens use numbers that are already in the vocabulary of the LM. Thus, we can represent any scene, using four location tokens to represent the position and the spatial extension of an object and combining them with the object name (and any other object attribute). This textual scene representation allows LMs to relate spatial relations like *left of* with specific arrangement of location tokens, providing a way to ground those relations.

To test our hypothesis, we verbalize the VSR dataset and use it for training and evaluation. As Figure 1 shows, we approach the problem stated in VSR in the following way: (i) we obtain textual scene descriptions using an object detector, (ii) we include in that description the location tokens derived from the object bounding boxes, (iii) we concatenate the caption and the textual scene description and input it to the LM, (iv) we fine-tune the LM on that input for binary classification. We also offer the possibility to previously train the LM in our Synthetic Spatial Training Dataset.

4.1 Textual scene descriptions

Given that VSR is a visio-linguistic dataset, the scene is defined by an image. We convert that scene to a textual description using a state-of-the-art object detector, VinVL [Zhang et al., 2021], which given an image, produces a list of objects with their name, attributes and bounding boxes. More concretely, an object detected by VinVL is represented as $O = \{name, attr_1, \ldots, attr_n, BB\}$, where $BB \in \mathbb{R}^4$ are the x_0, y_0, W, H coordinates of the bounding box.

To convert those BBs to location tokens, we follow this procedure (Figure 3): (i) normalize the image's width and height in the $[0, 1]$ range, (ii) divide the image in a regular grid of size $(G \times G)$, and (iii) find the grid cells for the BB coordinates (x_0, y_0, x_1, y_1) which we call $(\hat{x}_0, \hat{y}_0, \hat{x}_1, \hat{y}_1)$, i.e. discrete coordinates. Those discrete coordinates (after tokenization of the corresponding strings) are the location tokens. As a result, for every object detected, we get a sequence of four location tokens or discrete coordinates. Thus, our textual scene description $Descr(S)$ is a sequence of textual objects $\{O_0, O_1, \ldots O_N\}$, where each object is a string of the form: $O_i = \{\hat{x}^i_0, \hat{y}^i_0, \hat{x}^i_1, \hat{y}^i_1, name_i\}$. Notice that VinVL also returns a list of attributes for every object. Unless stated otherwise, we discard those attributes in the textual scene description.

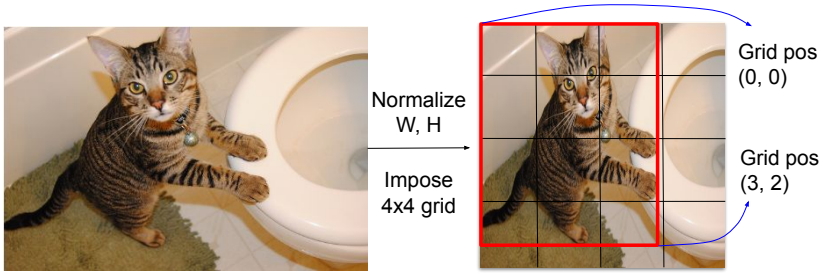


Figure 3: An illustrative example of how BB coordinates are converted to location tokens. In this case, with a grid size of 4×4, the location tokens for cat (red box) are *0 0 3 2*.

Category	Spatial Relations
Object position in the image	top left, bottom left, left, top right, bottom right, right, top, bottom, center
Object size comparison	wider, narrower, taller, shorter, larger, smaller
Two object positional relations	surrounding, inside, left of, above, right of, below, overlapping, separated

Table 1: The 23 relations in our Synthetic Spatial Training Dataset organized in three categories.

For the VSR task, we produce textual descriptions for all the images, concatenate them with the captions provided in the dataset and input it to the LM. Using positional embeddings, the LM can learn to interpret properly the order of location tokens and their correspondence with the object names. For example, for the image in Figure 3, the textual description of the object *cat* is: *0 0 3 2 cat*. Assuming that our grid size $G = 4$, this is interpreted as having a cat covering the left part of the image. We would do similarly for all the objects of the image to build our textual scene description.

Notice that for VSR, the textual scene descriptions are derived from images. But in the general case, we could derive them from other modalities like graphs or text. For instance, given a natural textual description of a scene (e.g. "a cat is on top of a table"), textual scene descriptions with location tokens could be obtained. However, as we could not find any suitable dataset for those cases, we leave them out of the scope of this paper (see Section 7).

4.2 The Synthetic Spatial Training Dataset

Multimodal training datasets with images and corresponding textual descriptions that include explicit spatial relations tend to be small. As a second ingredient of our approach we automatically construct a synthetic dataset with spatial relations named Synthetic Spatial Training Dataset (SSTD), which is used to teach LMs on how to relate location tokens and explicit spatial descriptions. Given an image in an existing dataset, an object detector is used to produce textual descriptions with object labels and location tokens. Given two objects and their bounding boxes, simple rules and templates are used to generate a positive or negative question about the spatial relation between the two objects (or alternatively, about a single object). Figure 4 shows such a generated example. The most important advantages of SSTD are: i) it can generate thousands and thousands of different examples, ii) it involves light human labour[2], iii) it can be easily extended to support new spatial relations, and iv) it can be used as a visio-linguistic or text-only dataset.

To build SSTD, we use the 2014 version of the COCO dataset [Lin et al., 2014]. We obtain SSTD training examples from the train set and validation examples from the validation set. Instead of using human annotated object detections, we use automatic VinVL detections, because the vocabulary size of VinVL is much larger than COCO (1848 classes against 80). In COCO, for example, we have the class "person", while VinVL detects more specific classes like "woman", "man", "boy" or "girl", among others, which add more diversity to SSTD. Although VinVL introduces errors in the object detection label or bounding box, this is not important for the text-only case, as we do not need matching visual and textual representations of the image. We are just interested in generating correct spatial relations for the detected object bounding boxes and labels.

In order to generate SSTD, we manually define a list of interesting and unambiguous spatial relations based on previous work [Johnson et al., 2018]. For example, given two bounding boxes, deciding whether an object is *left of* another

[2]We spent ~ 5 hours of work for our specific implementation including rules and templates.

object, is unambiguous. However, using only those bounding boxes, it is not possible to decide whether the objects are *close to* each other. Even though both BBs may be close, one of the objects can actually be very far in the depth dimension, so we need the context of the image to decide about the spatial relation. In that sense, notice that we did not have to adapt SSTD relations to VSR, just focus on what kind of relations we could unambiguously derive from bounding boxes. In consequence, SSTD should be useful for other tasks involving spatial grounding, not only VSR. In Table 1 we provide all the implemented relations and the category they belong to. All of them can be implemented following some simple rules based on object bounding boxes (more details in Appendix A). This is the process we follow to generate an example for SSTD:

1. We take an image and check the number of detected objects. As we implement one- or two-object relations, depending on the number of detections, we randomly select among the three categories of Table 1 (i.e. if we have only one detection, we select "object position in the image"). If we have two or more objects, we prioritize two-object relations (i.e, we assign 70% of probability to two-object relations and 30% to one-object relations). Given the category, we randomly sample the required objects (one or two depending on the relation) from all the detections.
2. We randomly decide between generating an affirmative or negative question. This way, we make sure that *yes* and *no* answers will be balanced. Using hand-designed verbalization templates, we generate the question corresponding to the spatial relation selected in the previous step (templates are provided in Appendix A).
3. We verbalize the scene in the image. We provide two kinds of verbalizations: i) generate the textual scene description as the concatenation of all objects detected by VinVL in the image, where each object is accompanied by its location; ii) use only the concatenation of object names, excluding location tokens. Notice that other image verbalization approaches could easily be added, such as captions[3].
4. A SSTD example is comprised by a question, a textual scene description and an answer. The image is discarded in the text-only version.

Following this procedure, we can generate many examples from each image. In that sense, SSTD does not have a fixed size: users can decide how many examples they want to extract from each image. In our case, during the spatial training phase of our models, we decide to produce random examples from the same images (COCO train set) in each epoch. That means that the models see an estimate of $num\_epochs \times 80K$ examples during the training process, where $80K$ corresponds to the number of images for COCO train set. Finally, as VSR is also based on COCO, to avoid any contamination, we do not include in the train set of SSTD the images that are already in VSR dev or test splits.

5 Experiments and results

We use the *random* split of the VSR dataset for the experiments, given its bigger size. For all the fine-tuning processes described, we train the models in the train set and select the best performing model in the validation set. That model is then evaluated in the test set. Following the recommendations of VSR authors, we provide the average results of three different runs, with the observed standard deviation. The hyperparameters of different models and GPU usage are specified in Appendix B.

5.1 The influence of the location tokens and spatial training

We want to assess the importance of two fundamental factors of our approach: i) the use of location tokens for LMs, and ii) the benefits of a spatial training phase using SSTD to better leverage those location tokens. For that purpose, we use BERT-base [Devlin et al., 2018] as our LM and train it in different ways, testing different combinations of using (or not) location tokens and previously training (or not) spatially with SSTD. We add a classification head on top of the *[CLS]* embedding ($\mathbf{t}_1^{(n_l)}$, where n_l is the index of the top layer) for binary classification. We define the head as a multilayer perceptron (MLP) of one hidden layer. We define our MLP in Eq. 1.

$$\begin{aligned} \mathbf{h} &= \text{LayerNorm}(\text{GELU}(\mathbf{W}_h \mathbf{t}_1^{(n_l)} + \mathbf{b}_h)) \\ \hat{\mathbf{y}} &= \text{Sigmoid}(\mathbf{W}_{\hat{y}} \mathbf{h} + \mathbf{b}_{\hat{y}}) \end{aligned} \tag{1}$$

In order to develop the spatial training phase using SSTD, we randomly built a validation set for SSTD (comprising 40,504 examples) and chose the model which performs best as the one to be used in the VSR experiments.

[3]We consider that for our experiments, those alternative verbalization approaches are not interesting, since we want to test how explicit spatial relations are grounded to location tokens.

Q: Is man right of horse?
Descr: 0 3 16 29 horse 14 7 26 31 man 22 6 31 31 baby 21 5 28 10 tree 0 5 23 31 building…
A: Yes.

Figure 4: An example of the SSTD validation set generated from the image, which includes question (Q), description (Descr) and answer (A), but not the image itself. Description partially shown, as it comprises 44 objects. Location tokens are discrete grid coordinates of the BB, e.g. $(0, 3)$ and $(16, 29)$ for horse.

Model	Locations	SSTD
BERT-base	No	76.96
BERT-base	Yes	94.49

Table 2: Results (accuracy) on the validation set of our synthetic SSTD dataset.

	Model	Locations	VSR acc
Language Models	BERT-base	No	62.11±0.88
	BERT-base	Yes	61.60±0.92
Spatially trained Language Models	BERT-base	No	61.83±0.28
	BERT-base	Yes	**73.69±0.88**

Table 3: Test results on VSR as mean accuracy with standard deviation. First block for language models with and without location tokens. Second block for spatially trained language models (using SSTD) which are then fine-tuned on the VSR training set.

Table 3 shows the obtained VSR test results for the mentioned combinations. The first block shows the performance of BERT-base fine-tuned on the VSR training set, with no significant differences between using or not location tokens. However, we do observe important differences in the second block, where both BERT-base models are previously trained on our Synthetic Spatial Training Dataset (SSTD) and only the model which uses location tokens improves over the previous models. The improvement with the use of spatial training and locations with respect to the other three options is notable, with ~ 12 absolute point improvement. The results show that location tokens are a good way to encode spatial information for language grounding, and that the spatial training step using SSTD is crucial to make the model learn how that grounding should be done.

On the other hand, Table 2 shows the results obtained in the validation split of SSTD. Although SSTD is used for spatial training and the obtained results are not the focus of this paper, it is interesting to see how using location tokens, the LM can achieve 94.49 of accuracy, whereas without location tokens, it cannot reach an accuracy of 77. The gap is of around 17 absolute points, which, once again, shows the importance of location tokens.

	Model	Parameters	VSR acc
Multimodal Systems	CLIP (w/ prompting)	632M	55.2±1.4
	VisualBert†	110M	57.4±0.9
	ViLT	87.4M	69.3±0.9
	LXMERT	240M	70.1±0.9
Our Spatially trained Language Models	BERT-base	110M	73.69±0.88
	BERT-large	336M	**74.44±0.73**
	T5-base	220M	73.09±0.59
	T5-large	770M	**74.49±0.36**
	T5-3B	3B	**74.52±0.25**

Table 4: Test results on VSR as mean accuracy with standard deviation. First block for multimodal systems, see text for references. † for models with no spatial information. Second block for our spatially trained language models.

5.2 Comparison with the state of the art

In this section we compare our results to the current state-of-the-art models for VSR, and, in addition, we explore whether scaling up LMs brings some extra performance. For that purpose, we use BERT-large as our LM (also adding a binary classification head as in Eq. 1), but we also explore the T5 family of encoder-decoder models [Raffel et al., 2020]. We include T5 models because the larger size of some models and in order to explore encoder-decoder models, as opposed to encoder-only models such as BERT. To use T5, we add text prefixes before each sentence, such as *'caption:'* for the VSR caption and *'context:'* for the textual scene description. This is done to mimic the input prompts used during the pretraining process of the T5 model, and help the LM to better leverage what it has learnt before. As T5 is a generative LM, it produces answers in an open-ended text generation manner. We select the answer (yes or no) with maximum probability. Thus we do not use any classifier head in this case.

Table 4 shows the obtained results for those experiments[4]. The best VLM, i.e. LXMERT, obtains an accuracy of 70.1. All our spatially trained LMs surpass that accuracy significantly, which is notable as our models only access bounding box labels and locations, losing potentially important information in the image. The best models are our three largest LMs, with over 74 accuracy, 4 absolute points ahead of LXMERT.

From those results, we can conclude that location tokens and the spatial training phase are good strategies to ground spatial relations in LMs. More importantly, LMs can handle spatial information, which opens the door for applications such as document layout tasks or textual spatial reasoning, for example. However, if we look at the benefits of scaling up the LMs, our experiments show diminishing returns for this specific task. It is true that our best model is a T5 of 3B parameters, however the difference with T5-large or BERT-large is quite small. Notice, though, that we did not perform any extensive hyperparameter tuning, so it could be the case that those larger LMs could actually perform better. Regarding sizes, we would like to note that we use the decoder part of T5 to generate one of Yes or No, and as such it would seem that the decoder is oversized.

6 Analysis of the results

In this section we analyse the results of individual spatial relations, we compare our systems with a rule-based baseline and a VLM, and we finally analyse the use of object attributes.

6.1 Analysis per spatial relation

As the 65 relations in VSR are of different nature, we compare the performance of our spatially trained LMs relation by relation. The objective is to see how spatial training and scale affect the performance. Figure 5 shows the accuracy of three LMs per relation. The selected models are BERT-base with location tokens but without any spatial training, the same BERT-base with spatial training and BERT-large, also with location tokens and spatial training. We only visualize the relations that appear 15 times or more in the test set.

In general, spatial training helps in almost all relations, with some exceptions. For instance, an orientation relation (*facing away*) and an adjacency relation (*at the edge of*). This could be expected, as SSTD does not cover those relations, because orientation cannot be inferred from BB information, and the object detector in use (VinVL) does not

[4]The results of VLMs are directly extracted from [Liu et al., 2022b].

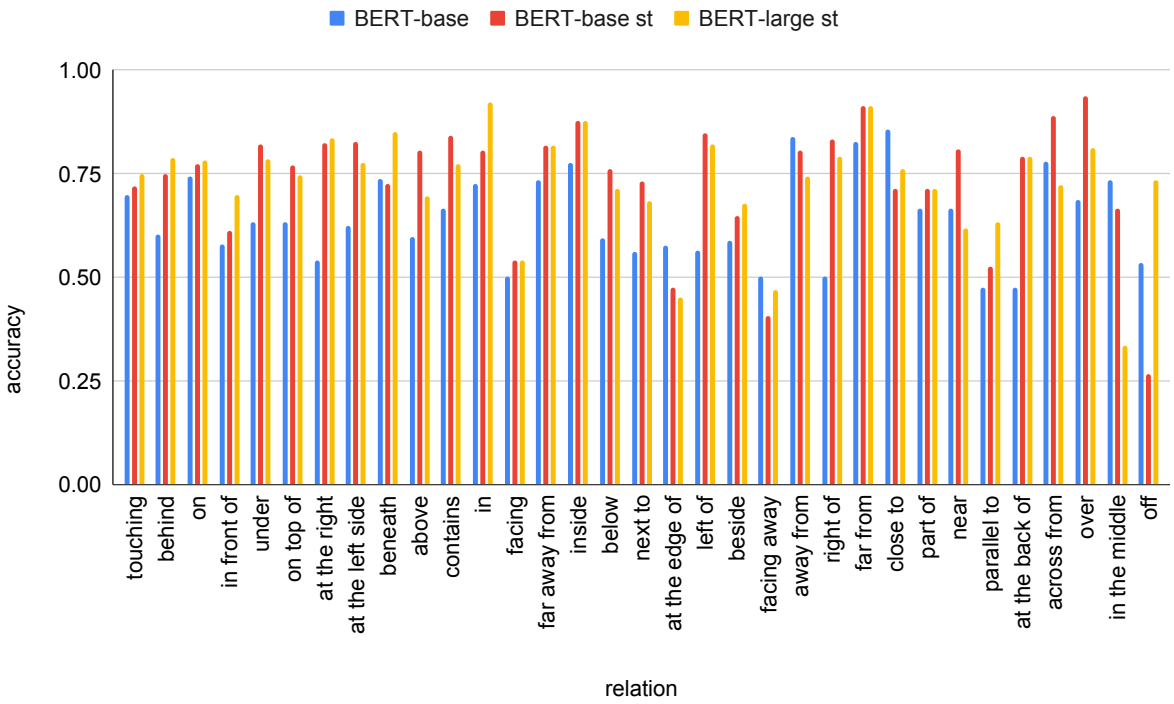


Figure 5: Comparison of three BERT models in terms of accuracy per spatial relation. Relations are ordered by frequency in descending order. For readability, we only show the relations that appear more than 15 times in the test set. All three models use location tokens. The "st" acronym in the model name indicates that the model has been spatially trained before the fine-tuning on VSR. Best viewed in color.

codify it in the attributes either. Orientation seems to be also difficult for VLMs [Liu et al., 2022b], so more work is needed in this regard.

There are also positive effects which show the generalization capabilities of the LMs to some extent. BBs do not provide any 3D information, so we did not include relations like *behind*, *in front of* and *at the back of* in SSTD, but spatial training performs very well for those relations. One of our hypothesis is that SSTD does include size relations (*wider*, *smaller* and so on), and thus the spatially trained models learn to combine BB information with typical size relations to infer depth (e.g. as persons are larger than cats, if a particular person is smaller than a cat, it has to be farther in the scene). We plan to further investigate those cases, since they provide hints of how spatially trained LMs can leverage location tokens to generalize to spatial relations that cannot be described unambiguously with arithmetic rules. We provide a preliminary qualitative analysis in Appendix C.

We also observe in Figure 5 that, in general, the performance for VSR relations covered in SSTD (*at the right side of*, *at the left side of*, *on top of*, *above* and so on) improves significantly. Knowledge transfer for those relations was expected, as they are semantically very similar to some SSTD relations. However, in one case, *beneath*, which is tightly related to the SSTD relation *below*, spatially trained BERT-base does not outperform BERT-base, but BERT-large does ($+12$ absolute points).

To add more context to this analysis, Table 5 provides the number of VSR relations per category, alongside the coverage in SSTD and the performance difference between a BERT-base model with and without spatial training (both with location tokens). Overall, SSTD covers only 17 out of the 65 relations in VSR, but there are some relations in SSTD which can be helpful for some other relations in VSR. For example, the VSR relation *detached to* is related to the SSTD relation *overlapping*. Depending on the image, overlapping BBs can be detached objects, but in general, BBs that do not overlap will be detached. Looking at the performance difference (3rd column of Table 5), we can see that spatial training is beneficial for all the categories, except for *topological*, where the difference is very small in any case. The *unallocated* category has an impressive performance gain ($+56.8$), but it is not very significant since there are only 51 examples in the test set. In general, we can say that those categories that are better represented in SSTD, consistently

VSR category	VSR Relations	In SSTD	Perf. gap
Adjacency	10	2	+4.7
Directional	11	2	+2.9
Orientation	4	0	+9.1
Projective	12	8	+14.4
Proximity	5	0	+1.1
Topological	18	5	-1.2
Unallocated	5	0	+56.8

Table 5: For every category in VSR, we show how many relations there are. In the second column, we show how many relations are already covered in SSTD. In the last column, the average performance difference between a spatially trained BERT-base against a BERT-base without spatial training is shown.

improve in VSR. That is the case of *projective* ($+14.4$), *adjacency* ($+4.7$) and *directional* ($+2.9$). In that sense, the performance gain of 9.1 absolute points for *orientation* relations is quite surprising.

Finally, in terms of LM size, the differences between BERT-base and BERT-large are irregular. In general, BERT-large performs better, but there are some cases where BERT-base outperforms it. We do not observe any remarkable behavior.

6.2 Comparison with a rule-based baseline

An interesting question that arises from our results is whether our spatially trained LMs learn more than the heuristic spatial rules represented in SSTD. To answer that question, we implemented a rule-based baseline, using the same spatial rules of SSTD to solve the VSR dataset (implementation details can be found at Appendix D). We found that around 38% of test instances could be solved using our spatial rules. However, due to caption-context object matching failures, only 25% of the instances are actually solved using rules. The obtained accuracy for those instances is 60.7, clearly below the performance of our spatially trained LMs. Indeed, if we solve randomly all the instances that cannot be solved by rules (around 75% of the test set), we obtain an overall accuracy of 52.4, whereas our best spatially trained LM has an accuracy of 74.5.

Figure 6 provides a detailed comparison between our rule-based baseline and the spatially trained BERT-large model for VSR test. As can be seen, for all those relations that can be solved using bounding boxes and heuristic rules, the spatially trained LM clearly outperforms the rule-based baseline for all the relations except three: *within* and *around*, where both approaches have the same performance, and *into*, where the rule-based baseline obtains better results (notice, though, that there are only 6 instances for that relation in VSR test, so the results are not very representative). From those results we can conclude that our text-only LMs learn more than the information encoded in the spatial rules of SSTD.

6.3 Comparison with a VLM

Even though it is not the main focus of the paper, it is also interesting to see how our spatially trained LMs compare to VLMs. For that analysis, we compare the results of our spatially trained BERT-large and LXMERT for every relation in VSR test.

Figure 7 shows the accuracy obtained by both models, grouped by categories. As can be observed, there are no important differences, except for the *unallocated* category, where BERT-large significantly outperforms LXMERT (92 vs 68). However, if we look at the performance relation by relation, there are interesting differences. In Figure 8, we show the accuracy obtained with both models for those relations where the difference is bigger than 4 absolute points (we consider that difference being significant, since it is approximately the overall difference of both models for VSR test). As can be seen, BERT-large outperforms LXMERT for the relations *in front of, at the left side of, in, far away from, inside, left of, far from, close to, at the back of* and *over*. Some relations only require two-dimensional information (*at the left side of, left of, over*) and thus, the better performance of BERT-large could be expected. However, it is curious to see that BERT-large is better than LXMERT for relations like *in front of, in, far away from, inside, far from, close to* and *at the back of*. Those relations should benefit from visual information, but it seems LXMERT cannot leverage that information properly. On the other hand, LXMERT only outperforms BERT-large significantly for the relations *on top of* and *in the middle of*. In the case of *on top of*, the difference is of 4 absolute points and we do not see any clear reason for that difference. For the relation *in the middle of*, BERT-large is specially bad, even worse than BERT-base, which is on par with LXMERT. We believe this behaviour is more related to the low number of instances for that relation in VSR test (only 15).

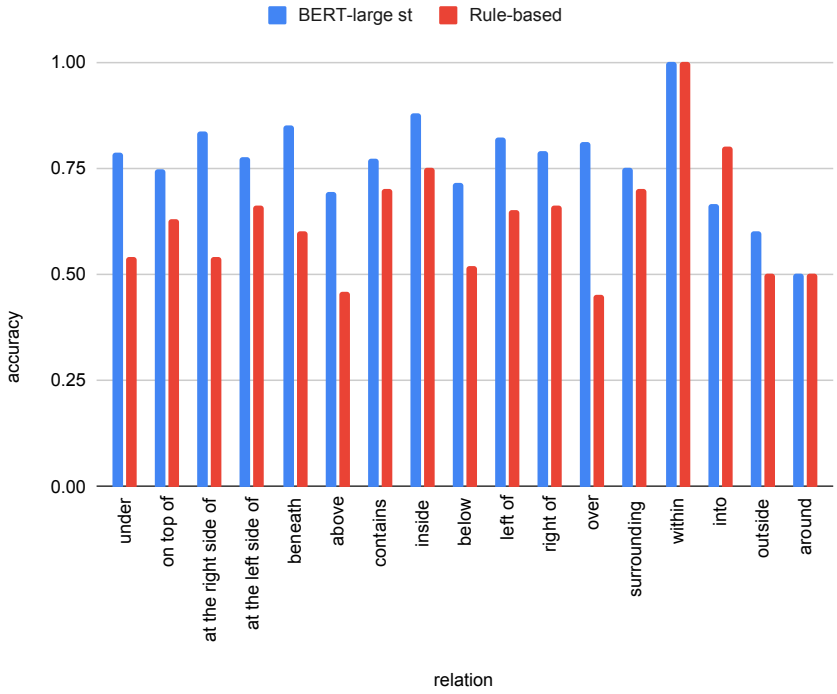


Figure 6: Comparison of our spatially trained BERT-large model and the rule-based baseline for the VSR test relations that can be solved using bounding boxes and heuristic rules. Best viewed in color.

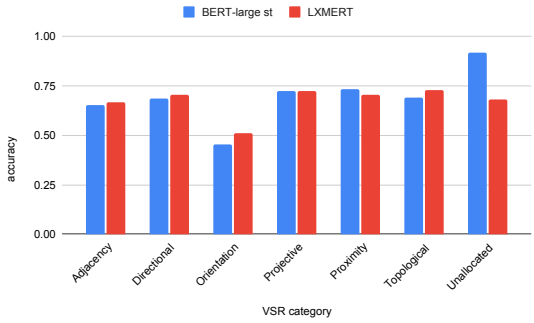


Figure 7: Comparison of our spatially trained BERT-large model and LXMERT for the VSR test categories. Best viewed in color.

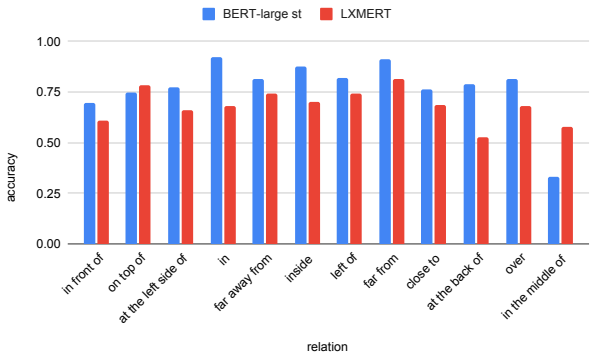


Figure 8: Comparison of our spatially trained BERT-large model and LXMERT for the VSR test relations, where the difference between both models is bigger than 4 absolute points. Best viewed in color.

6.4 Analysis of the use of object attributes

VinVL returns not only objects but also their attributes like colors, poses (*open hand*, *standing boy*), sizes, textures (*striped jacket*), materials (*brick wall*) and so on. We modified the spatial training phase to include the attributes in the textual scene description and trained a BERT-base model with the same hyperparameters as in Section 5.1. Afterwards, we fine-tune the best SSTD validation model on the VSR training set. Again, we add the attributes in the textual scene descriptions. The VSR test accuracy is of 74.14, which is inside the standard deviation of the BERT-base models shown in Table 3. We conclude that using object attributes as extracted by VinVL is not beneficial for this specific task,

although our analysis in the previous section showed that additional attributes non covered by VinVL like orientation or depth information, if extracted, could be of use.

7 Conclusions and future work

In this paper, we have presented a novel way to ground spatial relations in text-only language models through location tokens. To make LMs learn the grounding between spatial relations and location tokens, we also propose the Synthetic Spatial Training Dataset, a textual dataset with unambiguous spatial relations between objects automatically derived from existing images. We run experiments on a verbalized version of the Visual Spatial Reasoning dataset, where spatial grounding can be tested, showing that our approach to ground spatial relations in LMs is effective. Indeed, when compared with VLMs, we obtain even better results, which is another important indication that our spatial grounding approach is working.

Furthermore, scaling up our LMs we obtain the new state-of-the-art in VSR. However, we observe diminishing returns, which may suggest that to ground better those spatial relations, scale is not determinant. That opens the door for other techniques and approaches.

In the future, we want to deepen on spatial training, including categories like orientation and depth, for example. We also want to transition to text-only spatial reasoning tasks like SpartQA [Mirzaee et al., 2021] and RESQ [Mirzaee and Kordjamshidi, 2022], where we plan to transform the natural language scene descriptions with explicit spatial relations provided in those tasks, to our textual scene descriptions based on location tokens. We want to see whether those grounded representations do actually improve the spatial reasoning capabilities of LMs.

Acknowledgments

Ander is funded by a PhD grant from the Basque Government (PRE_2021_2_0143). This work is partially supported by the Ministry of Science and Innovation of the Spanish Government (AWARE project TED2021-131617B-I00, DeepKnowledge project PID2021-127777OB-C21), and the Basque Government (IXA excellence research group IT1570-22).

References

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. Things not written in text: Exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376, 2022a.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022b.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. Spartqa: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, 2021.

Roshanak Mirzaee and Parisa Kordjamshidi. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.413.

Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, 2020.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.

Andy Zeng, Maria Attarian, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, et al. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2022.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497, 2022.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*, 2022c.

Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020.

Frank van der Velde. Communication, concepts and grounding. *Neural networks*, 62:112–117, 2015.

Alban Laflaquière, J Kevin O'Regan, Bruno Gas, and Alexander Terekhov. Discovering space—grounding spatial topology and metric regularity in a naive agent's sensorimotor experience. *Neural Networks*, 105:371–392, 2018.

Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Are elephants bigger than butterflies? reasoning about sizes of objects. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, 2019.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Aitzol Elu, Gorka Azkune, Oier Lopez de Lacalle, Ignacio Arganda-Carreras, Aitor Soroa, and Eneko Agirre. Inferring spatial relations from textual descriptions of images. *Pattern Recognition*, 113:107847, 2021.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=gJcEM8sxHK`.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

A SSTD Implementation Details

We will present here the rules and heuristics followed to derive spatial relations from bounding boxes, grouped by category (see Table 1). We also present the templates we use to generate automatic questions for every case. We assume all BB coordinates are normalized between $[0, 1]$.

Object position in the image We first define a region for each of *top left*, *top right*, *bottom left* and *bottom right*. For example, $[0, 0, 0.5, 0.5]$ corresponds to *top left*. If the object BB is inscribed in one of those regions, we return that spatial relation. Otherwise, we check whether the object is in the following regions: *top*, *bottom*, *left* or *right*. An object is in the *left* region, for instance, if the object bounding box is inscribed in the $[0, 0, 0.5, 1]$ region. In all the other cases, the object is in the *center*. Given an object obj and a region reg, the template we use for question generation is: "is $\langle obj \rangle$ in $\langle reg \rangle$ region?"

Object size comparison Assuming two objects obj_1 and obj_2 and their bounding boxes, we calculate the functions $width(obj)$, $tall(obj)$ and $area(obj)$ for each object, using BB coordinates. If $width(obj_1) > width(obj_2)$, obj_1 is *wider* than obj_2 (and obj_2 is *narrower* than obj_1). We apply analogous rules for *taller/shorter* using the $length(obj)$ function and *larger/smaller* using the $area(obj)$ function. Given two objects obj_1, obj_2 and a size comparison relation rel, the template we use for question generation is: "is $\langle obj_1 \rangle$ $\langle rel \rangle$ than $\langle obj_2 \rangle$?"

Two object positional relations Assuming two objects obj_1 and obj_2 and their bounding boxes, if the BB of obj_1 is inscribed in the BB of obj_2, obj_1 is *inside* obj_2, and obj_2 is *surrounding* obj_1. For the relations *left of*, *right of*, *above* and *below*, we use the angle between the centers of both objects. If the center of obj_2 is between the angles $[\frac{-3}{4}\pi, \frac{3}{4}\pi]$, we say obj_2 is *left of* obj_1. Similarly, $[\frac{-3}{4}\pi, \frac{-1}{4}\pi]$ corresponds to *above*, $[\frac{-1}{4}\pi, \frac{1}{4}\pi]$ corresponds to *right of* and $[\frac{1}{4}\pi, \frac{3}{4}\pi]$ corresponds to *below*. Finally, using the Intersection over Union (IoU) of both BBs, we say that obj_1 and obj_2 are *separated* if their IoU is 0, and *overlapping* if IoU > 0. Given two objects obj_1, obj_2 and a positional relation rel, the template we use for question generation is: "is $\langle obj_1 \rangle$ $\langle rel \rangle$ $\langle obj_2 \rangle$?". In the case of the relation *separated* we use the following template: "are $\langle obj_1 \rangle$ and $\langle obj_2 \rangle$ separated?".

B Hyperparameters and GPU Usage

We always use a grid size $G = 32$ all over the experiments. For experiments with BERT-base, both for the spatial training and VSR fine-tuning, we train the models for 20K steps, with AdamW optimizer, a batch size of 56, a maximum learning rate of 5×10^{-5}, a warmup phase of 2K steps and cosine scheduler for learning rate decay. We use a single NVIDIA A30 GPU to perform all the experiments. Each of the experiments need around 5 hours.

We train BERT-large models for 20K steps, with a batch size of 32, maximum learning rate of 10^{-5}, AdamW optimizer, warmup phase of 2K steps and cosine scheduler. Using a NVIDIA A100 GPU, we need around 4 hours for the spatial training and additional 5 hours for fine-tuning on VSR. In the case of T5 we train the models spatially for 88K steps (T5-3B is trained for 20K steps due to its size) and fine-tune on VSR for 20K. We use a batch size of 32, AdamW optimizer, maximum learning rate of 5×10^{-5}, a warmup phase of 2K steps and cosine scheduler for learning rate decay. Regarding the T5 family: T5-base is trained on 1 NVIDIA A30 GPU: for spatial training it needs ~ 20 hours and for VSR fine-tuning ~ 3.5 hours. T5-large is trained on 1 NVIDIA A100 GPU: it needs 1 day and ~ 4 hours for spatial training, whereas VSR fine-tuning takes ~ 3.5 hours. Finally, T5-3B is also trained on a single NVIDIA A100 GPU: spatial training ~ 20 hours (20K steps) and VSR fine-tuning ~ 15 hours.

No hyperparameter search was performed.

C Qualitative analysis of generalization capabilities

We compare some examples of two text-only LMs: the BERT-base model with location tokens trained only on VSR (BERT for short) and the BERT-base model with location tokens trained on SSTD and fine-tuned on VSR (st-BERT for short). We want to see the effects of the spatial training on SSTD to better generalize in VSR. For that purpose, we focus on two relations that cannot be represented in SSTD, since they cannot be unambiguously defined with BB

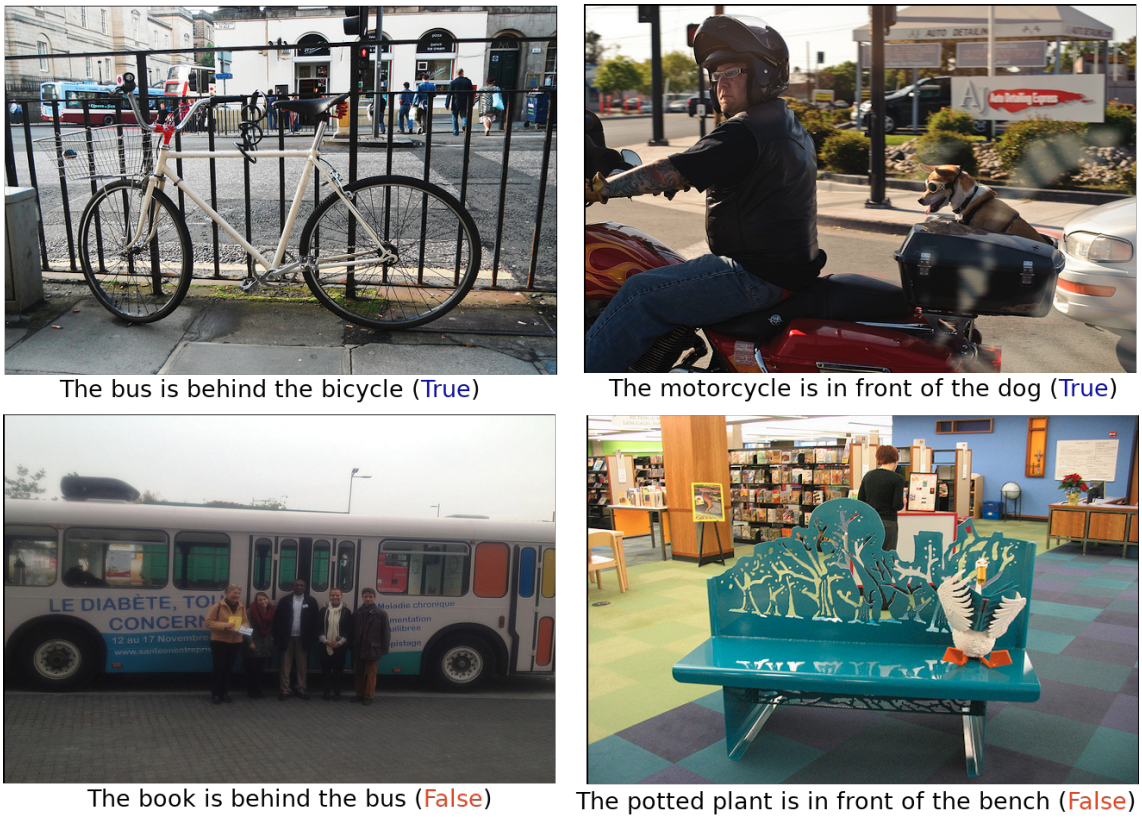


Figure 9: Comparison of the predictions of two BERT models for VSR test examples. The spatially trained BERT model predicts correctly the labels, whereas the BERT which has been trained only on VSR does not.

information and involve 3D arrangement of objects: *behind* and *in front of*. For *behind*, the accuracy of BERT is 0.6 and the accuracy of st-BERT is 0.75, calculated over 136 examples. For *in front of*, BERT scores 0.58 and st-BERT 0.61 (116 examples). Those results show that SSTD training helps even when the spatial relations are not represented in the dataset. Figure 9 offers some intuition of why this might be happening. For the first example, we see that the bus is much smaller than the bike. As SSTD includes relative size relations, we think the model has learned that buses are typically bigger than bikes. Thus while training on VSR, the model might be able to leverage that information and relate size differences with 3D arrangements of objects. A similar reasoning can be applied to the second (motorcycle and dog) and the last examples (bench and potted plant), but for the *in front of* relation. For the third example (bus and book), it seems st-BERT could leverage the fact that the book can only be visible if it is in front of the bus, given the arrangement of the bus BB. However, BERT could not predict the spatial relation correctly.

We also analyse two other relations that are not in SSTD, but are also related to relative object sizes: *next to* and *far from*. For *next to*, BERT obtains 0.56 and st-BERT 0.73 (over 41 examples). For *far from* BERT scores 0.83 and st-BERT 0.91 (over 23 examples). Notice that the relation *far away from* is very similar to *far from* and st-BERT clearly outperforms BERT also (0.88 vs 0.73 over 49 examples). For the first example (pizza and chair), given the small size of the chair, it can be inferred that it is far in the depth dimension. It seems st-BERT can leverage this information, whereas BERT cannot. For the second example (refrigerator and cat), both BBs overlap and it seems st-BERT infers that situation cannot lead to two objects far away given the typical sizes of those objects. The third example (backpack and cat) shows a case where both BBs are slightly overlapping. Again, the typical sizes of both objects could lead st-BERT to infer that they are actually next to each other. Finally, for the fourth example we see that the hot dog BB is inside the bowl BB. st-BERT infers that this is not the typical arrangement for *next to*, but BERT cannot do that, even though it has the same textual scene representation.

VSR relation	SSTD Relations
at the right side of	right of
at the left side of	left of
around	surrounding
into	inside
on top of	above
beneath	below
left of	left of
right of	right of
under	below
below	below
above	above
over	above
contains	surrounding
within	inside
surrounding	surrounding
inside	inside
outside	separated

Table 6: The mapping between VSR relations and SSTD relations.

D Implementation details of the rule-based baseline

To implement the rule-based baseline, we first defined manually a mapping between VSR relations and SSTD relations. As shown in Section 6.1, only 17 VSR relations out of 65 can be mapped to SSTD relations. That mapping is shown in Table 6. Given a VSR test instance, we check the spatial relation (provided in the annotations of the dataset) and if it can be mapped to a SSTD relation, we perform the following steps: a) from the VSR caption, we retrieve the subject and object using string manipulation; b) we find the same subject and object in the textual scene description, using string matching; c) if both subject and object are found, we retrieve their bounding boxes and apply SSTD rules to solve the instance; d) if any of subject or object are not found, or the relation cannot be mapped to a SSTD relation, we choose the answer randomly (50% of probability).

Figure 10: Comparison of the predictions of two BERT models for VSR test examples. The spatially trained BERT model predicts correctly the labels, whereas the BERT which has been trained only on VSR does not.

Improving Explicit Spatial Relationships in Text-to-Image Generation through an Automatically Derived Dataset

Ander Salaberria[1], **Gorka Azkune**[1], **Oier Lopez de Lacalle**[1], **Aitor Soroa**[1], **Eneko Agirre**[1], and **Frank Keller**[2]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{ander.salaberria, gorka.azcune, oier.lopezdelacalle, a.soroa, e.agirre}@ehu.eus
[2]University of Edinburgh
keller@inf.ed.ac.uk

Abstract

Existing work has observed that current text-to-image systems do not accurately reflect explicit spatial relations between objects such as *left of* or *below*. We hypothesize that this is because explicit spatial relations rarely appear in the image captions used to train these models. We propose an automatic method that, given existing images, generates synthetic captions that contain 14 explicit spatial relations. We introduce the Spatial Relation for Generation (SR4G) dataset, which contains 9.9 millions image-caption pairs for training, and more than 60 thousand captions for evaluation. In order to test generalization we also provide an *unseen* split, where the set of objects in the train and test captions are disjoint. SR4G is the first dataset that can be used to spatially fine-tune text-to-image systems. We show that fine-tuning two different Stable Diffusion models (denoted as SD_{SR4G}) yields up to 9 points improvements in the VISOR metric. The improvement holds in the *unseen* split, showing that SD_{SR4G} is able to generalize to unseen objects. SD_{SR4G} improves the state-of-the-art with fewer parameters, and avoids complex architectures. Our analysis shows that improvement is consistent for all relations. The dataset and the code are publicly available.[1]

1 Introduction

Text-to-image generators such as Midjourney, Stable Diffusion (Rombach et al., 2022) and Dalle-3 (Betker et al., 2023) have recently made rapid advances and generated a lot of interest. However, those systems are still far from being perfect and show some important weaknesses. For instance, as observed by (Gokhale et al., 2023) and (Cho et al., 2023b) among others, current text-to-image generators do not represent well explicit spatial relations like *left of* or *below*, which limits their capabilities for important applications like text-based image editing (Kawar et al., 2023).

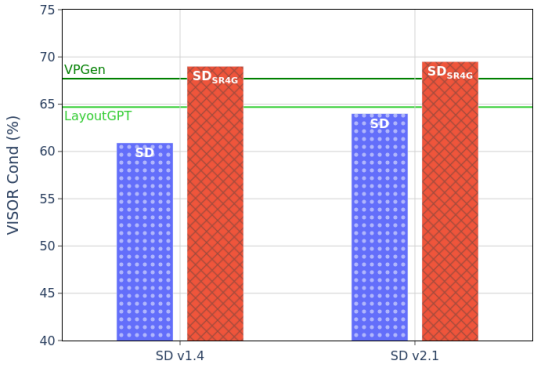


Figure 1: Fine-tuning Stable Diffusion on our SR4G dataset improves results significantly (two versions of SD shown), surpassing the state of the art in spatial-aware systems (see Section 4).

We hypothesize that the poor performance for explicit spatial relations is due to the lack of such relations in the datasets used to train those models. To support our hypothesis we analysed the LAION-2B dataset (Schuhmann et al., 2022), which has been used to train the state-of-the-art open source model Stable Diffusion. LAION-2B takes the captions from alt-text fields of images on the web. We automatically searched for explicit spatial relations (*left*, *right*, *below* and so on) and found that only 0.72% of cations contain the target words. Furthermore, 64.1% of these relations are *left* and *right*, which cannot be captured by image generators, as random horizontal flips are applied to images during training.

Motivated by the lack of captions with spatial relations, we focus on the training data to improve current end-to-end diffusion models; this is complementary to proposed architectural modifications on the system itself (Cho et al., 2023b; Feng et al., 2023). More concretely, we propose an approach

[1]Url: https://github.com/salanueva/sr4g

to automatically generate synthetic captions which contain explicit spatial relations with paired real images. Leveraging the object annotations in COCO (Lin et al., 2014) and heuristic rules to infer the spatial relation between two bounding boxes, we build a dataset of real images paired with synthetic captions, called Spatial Relations for Generation (SR4G).

We use SR4G to fine-tune two Stable Diffusion models, assuming that exposure to image-caption pairs with explicit spatial relations will enhance the capabilities of the models to represent those relations. To evaluate our fine-tuned models and compare to the unmodified base models, we use the recently proposed VISOR metric (Gokhale et al., 2023), which we extend to support more spatial relations.

The contributions of this paper are the following: (1) We release SR4G, the first benchmark that allows to fine-tune, develop and evaluate the spatial understanding capabilities of text-to-image models for 14 explicit relations; (2) Our experiments show that fine-tuning Stable Diffusion on SR4G improves the understanding of spatial relations and provides more accurate images; (3) The improvement holds even when tested on unseen objects, showing that the models are able to learn the relations, generalizing to unseen objects; (4) The results exceed the state-of-the-art in spatial understanding for image generation (Cho et al., 2023b; Feng et al., 2023) with fewer parameters and avoiding complex architectures or Large Language Models.

2 Related Work

Many text-to-image systems have been proposed in the last few years. In general, we can distinguish between those based on auto-regressive transformer architectures, such as the original Dall-E (Ramesh et al., 2021), the multi-task system OFA (Wang et al., 2022) or CogView2 (Ding et al., 2022); and those based on diffusion models, pioneered by GLIDE (Nichol et al., 2022), which evolved into current latent diffusion models such as Stable Diffusion (Rombach et al., 2022) and Attend-and-Excite (Chefer et al., 2023).

Although the results of text-to-image systems keep improving, recent work has shown that their performance for explicit spatial relations is low (Gokhale et al., 2023; Cho et al., 2023b); the models struggle to correctly draw textual descriptions like *a cat on top of a table*. To overcome these limitations, VPGen (Cho et al., 2023b) and LayoutGPT (Feng et al., 2023) propose pipeline systems, combining Large Language Models to generate layouts from textual prompts and layout-to-image generators such as GLIGEN (Li et al., 2023). The difference between both systems is that VPGen fine-tunes Vicuna-13B (Chiang et al., 2023) to generate layouts from textual descriptions, whereas LayoutGPT relies on Llama-2-7B (Touvron et al., 2023) and in-context learning for the same purpose.[2]

To avoid the use of complex and large pipeline systems, (Yang et al., 2023) propose ReCo, an end-to-end system which uses layout descriptions in the input. In this paper, we also focus on end-to-end systems, but we avoid inserting layout information into the input, as this imposes a substantial burden on users compared to simple text inputs.

To evaluate the performance of text-to-image generators for explicit spatial relations, dedicated datasets have been created, since commonly used datasets like COCO (Lin et al., 2014), CC12M (Changpinyo et al., 2021) or LAION (Schuhmann et al., 2022), contain very few examples of explicit spatial relations. For example, (Gokhale et al., 2023) propose the SR_{2D} dataset, composed of synthetic captions created combining two objects in the COCO object vocabulary and four explicit spatial relations. SR_{2D} only contains captions and it is thus not amenable for training. Similarly (Feng et al., 2023) published the Numerical and Spatial Reasoning dataset (NSR-1K) which does include caption-image pairs. The spatial part contains only 1021 image-caption pairs (738 for train and 283 for test, no development) for 4 relations, insufficient for accurate evaluation and too small for training.

Our paper proposes a new dataset with synthetic captions **and paired images** which can be used to train and evaluate spatial understanding of text-to-image generation systems, containing 14 different spatial relations and including 9.9 million image/caption pairs (Section 3). Finally, for evaluating the generated images, we follow (Gokhale et al., 2023; Feng et al., 2023; Cho et al., 2023b) and use an off-the-shelf object detector to extract bounding boxes and compute the spatial relation between detected objects.

[2]Originally they use LLMs from the OpenAI GPT family, but they have released a publicly available Llama-2 based variant of LayoutGPT, which we use in this work.

3 SR4G: A new synthetic dataset for explicit spatial relation generation

Given the shortcomings of previous datasets, we propose to generate meaningful synthetic captions for real images, and use them to build the SR4G dataset (Spatial Relations for Generation). We increase the number of spatial relations used in previous work (Gokhale et al., 2023; Cho et al., 2023b; Feng et al., 2023) including not only projective or scale relations, but also topological ones. The full list of unambiguous spatial relations we used is as follows:

Projective: *left of*, *right of*, *above* and *below*.
Topological: *overlapping*, *separated*, *surrounding* and *inside*.
Scale: *taller*, *shorter*, *wider*, *narrower*, *larger* and *smaller*.

Our objective is to build a dataset for training, development and evaluation. For training, we need image-caption pairs, but for evaluation, captions with spatial relations are enough, since, following previous work (Gokhale et al., 2023; Cho et al., 2023b), the outputs of the image generator are not evaluated against real images. The evaluation method is described in Section 3.4.

3.1 Captions for evaluation

We first generate a set of spatial triplets of the form ⟨*subject, relation, object*⟩. We build our initial set of triplets using all pairwise combinations of the 80 objects in the vocabulary of COCO (Lin et al., 2014), yielding $3,160$ object pairs, and combining each pair with all of our 14 spatial relations, resulting in 88,480 spatial triplets.

However, some spatial triplets in the initial set are not *natural*. For example, it is very difficult to find natural images for triplets like ⟨*skis, above, toothbrush*⟩ or ⟨*truck, inside, cat*⟩. We want to remove those *unnatural* triplets from our dataset to focus on triplets that appear in natural images. Therefore, we identify all triplets that appear at least once in the training split of the COCO dataset and use that subset to generate our evaluation captions, which consists of 68.8% of the entire set of triplets (60,836 triplets).

Using hand-designed templates to be as simple as possible (Appendix A.1), we generate the final evaluation captions from the set of spatial triplets (Figure 4 shows some examples). Those captions reflect only the spatial relations between two objects, avoiding to include any other textual details.

3.2 Image-caption pairs for training

For training, we need captions with explicit spatial relations and real images in which those relations are depicted. We use the COCO 2017 training split to collect real images with object annotations and define a methodology to generate first spatial triplets from those images, and then textual captions derived from those triplets.

Given an image I and a list of n objects $O_I = \{o_1, o_2, \ldots, o_n\}$ belonging to I, the goal is to generate a triplet with a valid spatial relation r between two objects in O_I: o_s and o_o, where $s, o \in \{1, \ldots, n\}$. For each object o_i, we know its respective label l_i and bounding box (*bbox*) $bb_i = \{x_i^0, y_i^0, x_i^1, y_i^1\}$, that is, four coordinates that define the position and size of o_i in the image.

Therefore, $t_j = \langle l_s, r, l_o \rangle$ is a triplet defined in SR4G that is represented in I. We call this set of valid triplets $T_I = \{t_1, \ldots, t_m\}$, where m is the number of valid spatial relations in the given image I. This implies that each relation r has to be linked to a heuristic rule f_r where, given the *bboxes* of two objects, it determines whether a given triplet is instantiated or not (see Eq. 1). We follow (Johnson et al., 2018) and define f_r functions, which represent unambiguous spatial relations between two object bounding boxes (see Appendix A.2).

$$t_j = \langle l_s, r, l_o \rangle \in T_I \longleftrightarrow f_r(bb_s, bb_o) \quad (1)$$

We apply data augmentation strategies (random crops and horizontal flips) to the original COCO images in order to obtain an image I and its object list O_I. Then, we randomly select two objects as o_s and o_o, compute the list of valid relations using our predefined f_r functions, and randomly select one of these relations, building the j-th valid relation of I without computing the entire T_I set: $t_j = (l_s, r, l_o)$. Finally, we verbalize the obtained triplet t_j using the same hand-designed templates as for the evaluation captions (Section 3.1).

3.3 Dataset splits

We build two different splits of SR4G, namely the *main* and the *unseen* splits. The *main* split consists of all the spatial triplets/captions of the SR4G test set (see Section 3.1). The training instances are generated on-the-fly without any restrictions on the triplets, which means that the same triplet can be in train, validation and test splits. For the *unseen* split, we randomly divide the COCO dataset's

Splits	Images	Unique Captions			I/C Pairs
		Train	Val	Test	
Main	103.4k	60.8k	2.5k	60.8k	9.9M
Unseen	83.6k	46.9k	2.5k	8.0k	4.8M

Table 1: SR4G dataset's statistics. *Images* column refers to the number of images used during training, *Unique triplets* column represents the amount of unique triplets, and *I/C pairs* refers to the number of unique image/caption pairs that can be generated.

80 objects into training, validation and test sets of $|O_{\text{train}}| = 45$, $|O_{\text{val}}| = 5$ and $|O_{\text{test}}| = 30$ objects, respectively. More specifically, during training we just take objects from O_{train} into account when randomly selecting *bboxes* to dynamically build spatial captions. For validation, as there are few combinations that can be built with O_{val}, we select triplets that contain one of these 5 objects at least once and do not contain any object that is set aside for the test split. For testing purposes we use triplets built by only using objects from O_{test}. Table 1 shows the relevant numbers of our splits (more details in Appendix A.3).

3.4 Evaluation metrics

To evaluate the performance of text-to-image systems for spatial relations, we use three evaluation metrics proposed by (Gokhale et al., 2023):

Object Accuracy: Given a generated image I' and two object labels l_a and l_b, object accuracy measures whether both objects appear in I'. We obtain a list of objects for I', i.e., $L_{I'} = \{l_1, \ldots, l_n\}$, by using an off-the-shelf open-vocabulary object detector, OWL-ViT (Minderer et al., 2022). This metric is useful for analyzing the object generation capabilities of an image generator, as it does not take the relation r into account.

$$\text{OA}(I, l_a, l_b) = \begin{cases} 1 & \text{if } l_a, l_b \in O_{I'} \\ 0 & \text{else} \end{cases} \quad (2)$$

VISOR: Given a generated image I' and a spatial triplet $t = (l_a, r, l_b)$, VISOR measures whether both objects appear and if the spatial relation r is valid between them. Function f_r takes the bounding boxes of both objects (bb_a and bb_b) and compares them to check if the triplet is valid. Bounding boxes are provided by the object detector. VISOR increases both when the model generates the requested objects and when the ratio of correctly generated relations increases, showing the ability of the model in visualising spatial triplets.

$$\text{VISOR}(I, t) = \begin{cases} 1 & \text{if } l_a, l_b \in L_{I'} \wedge \\ & f_r(bb_a, bb_b) \\ 0 & \text{else} \end{cases} \quad (3)$$

VISOR$_{\text{Cond}}$: This is the proportion of correctly generated spatial triplets, taking into account only images in which both objects are generated.

Given that our contribution focuses on spatial understanding, we focus on VISOR$_{\text{Cond}}$, as it quantifies the ability of the model to represent spatial relations correctly without considering its object generation capability. It is the most informative measure, specially when comparing between systems which might have different object generation abilities, as it isolates the understanding of spatial relations. We thus use it as our main performance metric in the experiments, although we also report the other two metrics, while extending the number of spatial relations from 4 to 14,

4 Experiments and Results

In this section we show that end-to-end models improve their capability of depicting spatial relations when they are fine-tuned with synthetic training examples. Furthermore, we find that our fine-tuned models SD$_{SR4G}$ generalize to unseen objects during fine-tuning.

4.1 Experimental set-up

Models. We use Stable Diffusion (SD) as the base model, as it shows the best performance on spatial relation generation among publicly available end-to-end models (Gokhale et al., 2023). We use two different versions of Stable Diffusion: SD v1.4 and SD v2.1, which generate images of 512x512 and 768x768 pixels, respectively.

Training. To fine-tune SD models on SR4G, we use the original loss function proposed by (Rombach et al., 2022), i.e., the mean square error over latent noise representations. We fine-tune SD models for 100k training steps with an effective batch-size of 64 instances, evaluating on the validation split every 5k steps. After training is complete, we select the checkpoint with the highest VISOR$_{\text{Cond}}$ value on the validation split. Following (Gokhale et al., 2023), we generate four images per spatial relation in all of our evaluations for consistency. More details can be found in Appendixes B and C.

Model	VISOR$_{\text{Cond}}$ ↑	VISOR ↑	OA ↑
Main split			
SD v1.4	60.9	17.6	29.0
SD v2.1	64.0	27.4	42.8
SD$_{SR4G}$ v1.4	69.0	26.8	38.9
SD$_{SR4G}$ v2.1	**69.5**	31.7	45.6
Unseen split			
SD v1.4	60.1	17.3	28.7
SD v2.1	64.0	28.4	44.4
SD$_{SR4G}$ v1.4	68.9	23.7	34.4
SD$_{SR4G}$ v2.1	**69.4**	29.4	42.4

Table 2: Results obtained for the *main* and *unseen* splits of SR4G. Base models SD v1.4 and v2.1 are shown alongside with fine-tuned SD$_{SR4G}$ models.

Model	Par.	VISOR$_{\text{Cond}}$ ↑	VISOR ↑	OA ↑
Main split				
LayoutGPT	8.1B	64.7	24.7	38.1
VPGen	14.1B	67.7	34.5	51.0
SD v2.1	1.3B	64.0	27.4	42.8
SD$_{SR4G}$ v2.1	1.3B	**69.5**	31.7	45.6
Unseen split				
LayoutGPT	8.1B	64.7	24.7	38.1
VPGen †	14.1B	68.4	37.0	54.1
SD v2.1	1.3B	64.0	28.4	44.4
SD$_{SR4G}$ v2.1	1.3B	**69.4**	29.4	42.4

Table 3: Comparison to the state of the art, including model size for both splits. † VPGen is contaminated, as it was trained on layouts containing spatial triplets that appear in our test split.

4.2 Main results

Table 2 shows the results for our base and fine-tuned models for both SR4G splits, with the best results according to the main comparison metric in bold.

Main split: We observe that the SD$_{SR4G}$ models improve all metrics respect to the base SD models, increasing both object and spatial relation generation capabilities considerably. These results are in line with our initial hypothesis, proving that the exposure to image-caption pairs with explicit spatial relations improves spatial relation generation. Our results show that SD$_{SR4G}$ v1.4 and v2.1 have almost the same spatial capabilities, but v2.1 excels for object rendering. Notice that the differences of the base SD models are much bigger.

Unseen split: To analyse whether the improvements of SD$_{SR4G}$ on the *main* split come from learning specific correlations between pairs of objects, or between objects and spatial relations, we check the results on the *unseen* split. The *unseen* split uses different objects in train and test, and it is thus designed to decouple objects from spatial relations, allowing us to focus on the performance for spatial relations in isolation. In Table 2, we see that both versions of SD$_{SR4G}$ consistently improve the VISOR$_{\text{Cond}}$ and VISOR metrics over the base SD systems, also for the *unseen* split. It is specially interesting that VISOR$_{\text{Cond}}$, which is not influenced by object accuracy, is almost the same as for the *main* split. That means that our models are generalizing to unseen objects during the fine-tuning step. The behaviour of both versions is very similar to the *main* split.

Image quality: As we are using synthetic captions to train, we make sure that the image generation capabilities of these models do not worsen over training. Therefore, we monitor the Fréchet Inception Distance (FID) (Heusel et al., 2017) between the model's generated images from human annotated captions (retrieved from the COCO 2017 validation split) and their respective real images. During all of our experiments FID values have been constant and have not worsen after training. A random set of examples can be seen in Figure 4.

4.3 Comparison with the state of the art

We also compare against two recent state-of-the-art pipeline models: LayoutGPT and VPGen. The backbone Large Language Model (LLM) of VPGen has already been fine-tuned for layout generation,[3] so we use VPGen with no further adaptation. Note that the layout generation module of VPGen has been trained on COCO, and thus contains the objects underlaying our test sets. In the case of LayoutGPT, adaptation is performed with in-context learning. We thus define a set of instances that will be used as in-context examples to condition the 7B parameter Llama-2 LLM. For this purpose, we randomly extract 400 caption-layout pairs per different relation from our SR4G dataset, and build a set of 5.6k instances of caption-layout pairs. For inference, $k = 8$ examples are chosen by computing the CLIP-based similarity (Radford et al., 2021) between the input caption and the set of in-

[3]They use three different datasets to obtain caption-layout pairs to fine-tune the LLM: Flickr30K entities (Plummer et al., 2015), COCO instances 2014 (Lin et al., 2014), and PaintSkills (Cho et al., 2023a).

Type	Relation	Main Split	Unseen Split
Projective	*Left of*	70.3 (+7.0)	69.8 (+8.8)
	Right of	72.4 (+8.0)	67.9 (+3.9)
	Above	72.0 (+4.5)	70.4 (+2.2)
	Below	71.4 (+4.5)	70.3 (+2.8)
Topological	*Overlapping*	86.9 (-4.9)	84.0 (-5.2)
	Separated	79.5 (+17.0)	84.8 (+18.5)
	Surrounding	29.8 (+2.3)	21.7 (-2.1)
	Inside	43.4 (-7.4)	39.2 (-6.4)
Scale	*Taller*	71.2 (+1.6)	75.6 (+5.0)
	Shorter	67.5 (+8.5)	69.0 (+11.9)
	Wider	71.6 (+4.3)	73.0 (+6.9)
	Narrower	69.3 (+9.3)	67.1 (+5.0)
	Larger	71.5 (+0.5)	74.7 (+1.9)
	Smaller	65.2 (+12.7)	63.3 (+13.5)

Table 4: $\text{VISOR}_{\text{Cond}}$ values per relation obtained by SD_{SR4G} v2.1. The difference in $\text{VISOR}_{\text{Cond}}$ between SD v2.1 and fine-tuned SD_{SR4G} is given in brackets.

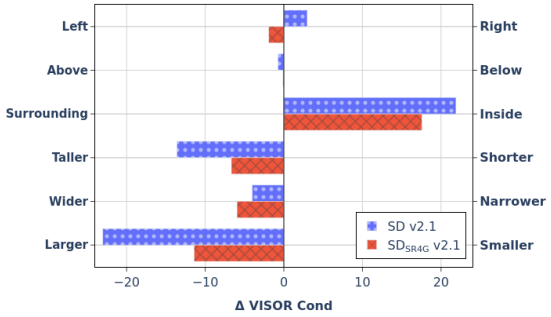


Figure 2: The horizontal axis depicts the difference of $\text{VISOR}_{\text{Cond}}$ values between relation pairs with opposing meanings defined on each side of the vertical axis. Results for SD and SD_{SR4G} v2.1 on the *unseen* split.

context examples, retrieving the top-k most similar examples and using them to condition the model to generate the proper layout.

Table 3 shows the obtained results for both SR4G splits. The same trend is observed, i.e. SD_{SR4G} v2.1 clearly outperforms both state-of-the-art pipeline systems in terms of $\text{VISOR}_{\text{Cond}}$, which measures the correctness of the spatial relation when both objects are generated. The improvement is especially important considering that both pipeline systems are significantly larger in terms of parameters, with a more complex architecture involving LLMs, and that both are specifically designed to generate scene layouts.

The table also shows the two auxiliary metrics, with VPGen obtaining the best results for object accuracy and VISOR. That is expected, since VPGen has been trained specifically for object generation, and VISOR is calculated over all the recognised objects. In fact, the better VISOR results are only due to better object accuracy, as our method produces better spatial configurations after factoring out object accuracy from VISOR ($\text{VISOR}_{\text{Cond}}$). Also note the contamination issue for the *unseen* split, as the text-to-layout step of VPGen has been fine-tuned on COCO. This implies that VPGen has seen text-layout pairs using the entire set of objects, having been trained on all the objects in our test set.

5 Analysis

We show an extensive analysis of the consequences of fine-tuning on SR4G, covering performance per relation, biases for opposite relations, performance by frequency of triplets and qualitative examples.

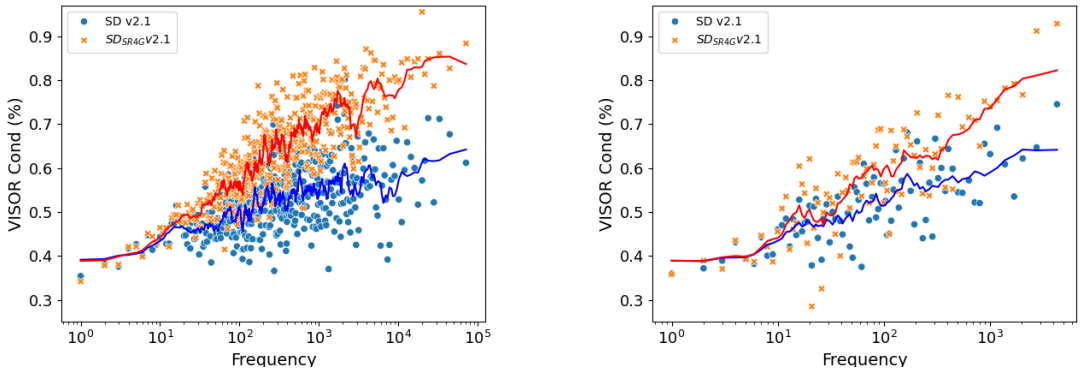
5.1 Analysing performance per relation

In Table 4 we show $\text{VISOR}_{\text{Cond}}$ values per spatial relation for SD_{SR4G} v2.1 (our best model), both in the *main* and *unseen* splits.

First, we observe that all projective relations significantly improve for both splits. The improvement is bigger for *left of* and *right of*. That might be due to random horizontal flips applied only to the images during the training of SD models, which are expected to damage the model's ability to correctly learn those relations.

Topological relations show a more variable behaviour. In the case of *separated*, our unique topological relation that does not involve generating overlapping objects, SD_{SR4G} is capable of improving its performance by up to 18.5 points $\text{VISOR}_{\text{Cond}}$. However, for *overlapping*, fine-tuning is not helpful. SD v2.1 already knows how to generate images with the *overlapping* relation, achieving $\text{VISOR}_{\text{Cond}}$ values of 91.8 and 89.2 in both test splits. On the other hand, *surrounding* and *inside* seem to be especially hard. The $\text{VISOR}_{\text{Cond}}$ values are low for the SD model and fine-tuning even makes them worse (especially for *inside*). This is a limitation of our current approach, and different training strategies must be explored to tackle this issue.

Finally, SD_{SR4G} improves for all scale relations. It is curious to observe that *taller*, *wider* and *larger* perform better than their opposites, even though the improvements over the base SD model are more modest. That suggests that the base SD model might have a bias towards those spatial relations.



(a) Results using *main* splits.

(b) Results using *unseen* splits.

Figure 3: Correlation between the frequency of SR4G triplets in COCO training instances (shown in the logarithmic horizontal axis) and their respective VISOR$_{\text{Cond}}$ results for SD v2.1 and SD$_{SR4G}$ v2.1. Triplets are grouped by frequency for visibility.

5.2 Analysing biases for opposite relations

Most of our relations have an opposite relation, i.e., *right of* is the opposite of *left of*. There are a total of six pairs of opposites in our relation set, which are listed in Figure 2 along with the difference in performance for these pairs before and after fine-tuning using the *unseen* split.

We want to see whether performance biases between opposites are reduced by fine-tuning. Figure 2 shows strong preferences of our base model SD v2.1 (in Appendix D, we show that those differences are correlated with the rate of appearance of each relation in the pretraining dataset of the SD models). We can also observe that SD$_{SR4G}$ v2.1 significantly reduces the difference in VISOR$_{\text{Cond}}$ between all relation pairs (except for *wider* and *narrower*), showing that fine-tuning reduces the inherent biases of the base model.

5.3 Performance by frequency of triplets

As SR4G is derived from natural images, some triplets are more frequent than others. To measure how the frequency of training triplets affects the results of our fine-tuned models, in Figure 3, we depict the VISOR$_{\text{Cond}}$ values of SD v2.1 and SD$_{SR4G}$ v2.1 depending on the frequency of each triplet in the COCO training set.

Figure 3a shows the results for the *main* split. In this case, the image generator has seen test triplets during training and, as expected, the more frequent these triplets, the greater the improvement after the fine-tuning. We can also observe that, even though SD models have not seen COCO images before, its performance is correlated with our computed frequencies.

On the other hand, Figure 3b shows a similar plot when training and evaluating on the *unseen* split. We observe similar correlations as in Figure 3a with both models. However, now we are evaluating on images generated from unseen triplets composed by objects that have not been seen during fine-tuning. Therefore, these results show that it is easier to transfer what is learnt to the most common triplets, even though we have not trained on them.

5.4 Qualitative Analysis

In order to visualize and qualitatively evaluate the generated images, we take SD v2.1 and SD$_{SR4G}$ v2.1 fine-tuned on the *main* split. We discard the most common and uncommon spatial triplets. The rationale is that the most common triplets often contain easy-to-generate relations (e.g., ⟨*truck, larger, dog*⟩) as generating both objects is enough to instantiate the relation itself, whereas the least frequent ones do not seem natural and would not be used in a prompt (e.g., ⟨*bus, shorter, traffic light*⟩). Therefore, we randomly pick triplets that occur between 100 and 1,000 times in COCO annotations (we obtain that range from the frequency analysis in Figure 3). We start generating images using random captions. We keep the first nine image pairs where both objects are generated correctly. Those nine pairs can be found in Figure 4, where we also indicate whether the spatial relation in the caption is depicted correctly or not.

Some of the captions of Figure 4 describe *easy* spatial relations, such as number 2, 3, 6, 7 and 9,

Figure 4: Image generation examples by SD v2.1 and SD$_{SR4G}$ v2.1 fine-tuned on the *main* split. Following our relation-specific heuristics, if the relation in the caption is correctly depicted, we indicate this with a green tick. Otherwise, there is a red cross in the top-right corner of the image.

where usually, if the correct objects are generated, the relation is also correct. SD$_{SR4G}$ generates those relations correctly, except for 3, which we denoted as a failure because the cup is not fully visible (the decision is arguable). SD fails for 2, rendering the traffic light very oddly. Captions 1, 4, 5 and 8 are more demanding: SD$_{SR4G}$ correctly depicts all the relations (*right of* twice, *overlapping* and *separated*), but SD fails for 1, 5 and 8. The failures are interesting: for 1 and 8, the spatial relations of the captions might not be the most typical ones in natural images, and SD struggles. However, for 5 it should be very common to see dogs and chairs separated, but SD does not follow the caption, which suggests that the relation *separated* is not known to SD.

6 Conclusions

In this work we define a dataset generation pipeline to build synthetic captions containing explicit spatial relations from COCO images and annotations. Fine-tuning diffusion models with these image-caption pairs outperforms the original diffusion models and also surpasses state-of-the-art pipeline models for spatial relation generation. We find that SD$_{SR4G}$ generalizes to unseen objects during fine-tuning. Further analysis shows that SD$_{SR4G}$ learns to better depict projective and scale relations, reduces the bias that the original model has for opposite relations, and generalizes better to spatial triplets that are more frequent in real images.

As future work, we plan to expand our relation set to include depth information with relations such as *in front of* and *behind*. We would also like to explore new ways to collect and annotate natural captions with spatial relations and evaluate state-of-the-art models with them.

7 Limitations

SR4G only contains captions in English, which limits its usage for non-English languages. To make it multi-lingual, caption generation scripts should be modified. On the other hand, SR4G is focused on unambiguous spatial relations defined over bounding box information, since they can be generated and evaluated automatically using off-the-shelf object detectors and heuristic rules. In that sense, orientation relations are discarded, even though their analysis is very interesting. Finally, we focus on 2D spatial relations. To introduce 3D relations should also be possible, using off-the-shelf depth estimation systems for images.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023a. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023b. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2023. Benchmarking spatial relationships in text-to-image generation.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and

Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. 2023. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113.

A Details on SR4G Dataset

In this appendix, we give more details about our *main* and *unseen* splits, as well as defining our hand designed templates and heuristics used to determine whether an image contains a given spatial relation between two objects.

A.1 Hand designed templates

The templates we use to generate captions from spatial triplets are shown in Table 5. As can be seen, those templates are designed to be as simple as possible, omitting attributes and verbs and focusing only on the objects and their spatial relation. This is very important to analyse spatial understanding in isolation.

Type	Relation	Template
Projective	*Left of*	⟨A⟩ to the left of ⟨B⟩.
	Right of	⟨A⟩ to the right of ⟨B⟩.
	Above	⟨A⟩ above ⟨B⟩.
	Below	⟨A⟩ below ⟨B⟩.
Topological	*Overlapping*	⟨A⟩ overlapping ⟨B⟩.
	Separated	⟨A⟩ and ⟨B⟩ separated.
	Surrounding	⟨A⟩ surrounding ⟨B⟩.
	Inside	⟨A⟩ inside of ⟨B⟩.
Scale	*Taller*	⟨A⟩ taller than ⟨B⟩.
	Shorter	⟨A⟩ shorter than ⟨B⟩.
	Wider	⟨A⟩ wider than ⟨B⟩.
	Narrower	⟨A⟩ narrower than ⟨B⟩.
	Larger	⟨A⟩ larger than ⟨B⟩.
	Smaller	⟨A⟩ smaller than ⟨B⟩.

Table 5: Templates used to generate synthetic captions.

A.2 Heuristic rules

We use heuristic rules to both build the dataset and evaluate the generated images. Assuming the spatial triplet $\langle l_s, r, l_o \rangle$ and the bounding boxes of its objects bb_s and bb_o that appear in an image, we define the heuristic rule f_r of relation r to determine whether the triplet is fulfilled in the image or not. We set $bb_i = \{x_i^0, y_i^0, x_i^1, y_i^1\}$ by defining the top-left $\{x_i^0, y_i^0\}$ and bottom-right coordinates $\{x_i^1, y_i^1\}$ of the bounding-box (*bbox*).

For *left of*, *right of*, *above* and *below*, we follow the heuristic rules defined in (Gokhale et al., 2023), by computing the centroid of each *bbox* $c_i = \{x_i^c, y_i^c\}$ and comparing their corresponding coordinates.

As we expand to 10 more relations, we follow the rules described in (Johnson et al., 2018). In our scale relations we compare either the height (*taller* and *shorter*), width (*wider* and *narrower*) or area (*larger*, *smaller*) difference between both *bboxes*. In the cases of *surrounding* and *inside*, we check whether bb_o is contained in bb_s or vice versa. Finally, using the Intersection over Union (IoU) of both *bboxes*, we say that both objects are *separated* if their IoU is 0, and *overlapping* if their IoU is positive.

O_{Train}
person, *car*, *motorcycle*, *airplane*, *train*, *boat*, *fire hydrant*, *bench*, *bird*, *elephant*, *bear*, *giraffe*, *handbag*, *tie*, *snowboard*, *baseball bat*, *baseball glove*, *surfboard*, *cup*, *knife*, *spoon*, *apple*, *sandwich*, *orange*, *broccoli*, *carrot*, *pizza*, *donut*, *chair*, *couch*, *potted plant*, *bed*, *dining table*, *toilet*, *laptop*, *mouse*, *remote*, *keyboard*, *oven*, *sink*, *book*, *clock*, *teddy bear*, *hair drier*, *toothbrush*
O_{Val}
umbrella, *cake*, *tv*, *refrigerator*, *vase*
O_{Test}
bicycle, *bus*, *truck*, *traffic light*, *stop sign*, *parking meter*, *cat*, *dog*, *horse*, *sheep*, *cow*, *zebra*, *backpack*, *suitcase*, *frisbee*, *skis*, *sports ball*, *kite*, *skateboard*, *tennis racket*, *bottle*, *wine glass*, *fork*, *bowl*, *banana*, *hot dog*, *cell phone*, *microwave*, *toaster*, *scissors*

Table 6: Objects used in train, val and test sets of our *Unseen split*.

A.3 Main and Unseen Splits

Table 6 shows the sets of objects used for training, validation and test in the *unseen* split, which we refer to as O_{train}, O_{val} and O_{test}, respectively.

There are few combinations that can be built with O_{val} for validation in the *unseen* split, so we select triplets that contain one object from O_{val} at least once and do not contain any object that is set aside for the test split. In other words, there are up to $(2 \cdot |O_{\text{train}}| \cdot |O_{\text{val}}| + \binom{|O_{\text{val}}|}{2}) \cdot 14 = 6,580$ triplets that fulfil this rule (around 5,326 that naturally occur in the COCO dataset).

Validation is computationally costly in both splits, as several images have to be generated to compute the evaluation metrics defined in Section 3.4. Preliminary experiments showed that generating just 10k images is enough to get consistent results. Thus, we randomly selected 2.5k spatial captions for the validation splits for both *main* and *unseen* splits (as we generate 4 images per caption).

B Training settings

Hyperparameters: In Table 7 we define the hyperparameters used for training. Learning rate and optimizer parameters are the ones used during the pretraining of SD models, the other listed hyperparameters have been adapted to our available infrastructure. We also take advantage of Exponential Moving Average (Kingma and Ba, 2015) to update the parameters of the models with an AdamW optimizer (Loshchilov and Hutter, 2019) and we do not use any learning-rate scheduler. We do validation runs every 5k steps and do not set any early-stopping mechanism.

Hyperparameter	Value
Training steps	100k
Batch size	64
Learning Rate	10^{-5}
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	10^{-8}
Weight decay	0.01
Mixed-precision	bf16

Table 7: Fine-tuning hyperparameters of the diffusion models.

GPU usage: Due to different memory needs, we use 2 and 4 NVIDIA A100 GPUs to fine-tune SD v1.4 and SD v2.1 models, respectively. In both cases we use an effective batch size of 64 by changing the amount of instances assigned to each GPU. Each of our fine-tunings need 3 days to be completed.

Data augmentation: During training we apply random horizontal flips and random crops to our images as a data augmentation strategy (resulting in I^* and O^j). Note that, random horizontal flips are common during the training of text-to-image models. This implies that spatial relations, such as *left of* and *right of*, can not be learnt correctly (as captions are not transformed according to those flips). Nevertheless, in our case we apply the same transformations to *bboxes*, which are used to generate captions synthetically, keeping this data augmentation strategy while maintaining the generated caption's spatial correctness.

Random crops might reduce the number of objects in O_{I^*}. If there are less than two objects after a given crop, we redo it up to max_iter times until there are at least two objects in the image.

We also define the hyperparameter k as the number of captions that can be concatenated to build the image-caption pairs built during training. Table 8 shows the results obtained by concatenating $k \in \{1, \ldots, 5\}$ captions. We observe that $k = 2$ obtains the best results, and we use this value of k

Nº Captions	$\text{VISOR}_{\text{Cond}}$ ↑	VISOR ↑	OA ↑
1	68.1	26.5	38.9
2	**69.4**	27.4	39.5
3	67.7	27.1	40.0
4	63.7	21.9	34.3
5	63.0	22.9	36.3

Table 8: We fine-tune SD v1.4 in the *main* split concatenating different amounts of captions in the input. These results correspond to the validation set of our *main* split.

during our entire work.

C Evaluation settings

The evaluation metrics used in this paper use an object detector to determine whether objects are generated correctly and where are located in the image. Following (Gokhale et al., 2023), we use OWL-ViT, an open-vocabulary object detector that uses a CLIP (Radford et al., 2021) backbone with a ViT-B/32 transformer architecture (Zhai et al., 2022). We also set 0.1 as the confidence threshold of OWL-ViT, which determines how sure the model must be for a given region of the image to contain a specific object.

As an open-vocabulary object detector, OWL-ViT takes as input the objects we want to detect and, in order to do so, we use their recommended template ("a photo of a ⟨OBJ⟩.") instead of the object label alone.

Due to the variability of images generated by Stable Diffusion, we generate 4 images per evaluation caption. Therefore, we generate 10k images per validation and a total of 243.3k and 32.1k images to test each model in the *main* and *unseen* splits, respectively.

D LAION Dataset and Spatial Relations

Figure 2 shows that Stable Diffusion models have a strong bias towards some spatial relations, preferring *taller* to *shorter*, for instance. To complete those results, we also show the same graphic but in the *main* split, which exhibits a very similar behaviour (Figure 5). To understand the origin of those biases, we checked the frequency of each spatial relation in the LAION-2B dataset (English subset), used to train SD models. Table 9 shows the appearances of 12 relations, divided in 6 relation pairs with opposite meanings. Every relation has its number of appearances in LAION into brackets. For each opposite relation pair, the first column contains the relation that best works with SD. The third column shows the ratio of appearance between the preferred relation and its opposite (>1 indicates that the preferred relation appears more times in LAION than its opposite relation). The results indicate that there is a clear correlation between the ratio of appearance of a relation and the bias of SD models. The only exception is the *right* and *left* pair, but both appear similar times and the bias towards *right* is very small.

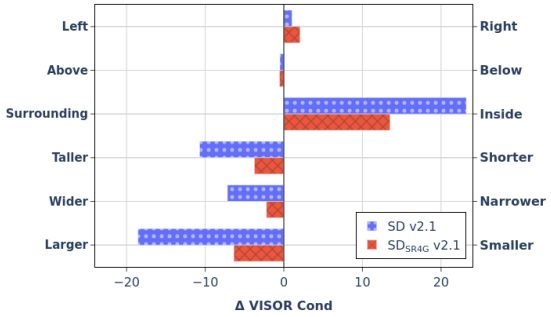


Figure 5: The horizontal axis depicts the difference of $\text{VISOR}_{\text{Cond}}$ values between relation pairs with opposing meanings defined on each side of the vertical axis. These results correspond to SD and SpaD v2.1 trained and evaluated using *main* splits.

Preferred Rel.	Opposite Rel.	Ratio of Appearance
Right (5M)	Left (5.6M)	0.91
Above (1.6M)	Below (0.7M)	2.47
Inside (2M)	Surrounding (0.3M)	7.61
Taller (49.3K)	Shorter (29.4K)	1.86
Wider (54.6K)	Narrower (5.7K)	9.62
Larger (0.8M)	Smaller (0.2M)	3.17

Table 9: Ratio in which the first relation appears more than the other. The relation in the first column is the preferred one by SD.

B. ERANSKINA

Arrazonamendu Espaziala Ikasten Hizkuntza-ereduetan

B.1 SSTD-ren Inplementazioa

Eranskin honetan kaxa inguratzaileetatatik erlazio espazialak inferitzeko erregelak eta heuristikoak azaltzen ditugu, 4.2. Taulako kategoriatan banatuta. Gainera, SSTDeko galderak sortzeko erabili ditugun txantiloiak zehaztu ditugu. Lan egiten ditugun kaxa inguratzaileak normalizatuta daude $[0, 1]$ tartean.

Objektu baten posizioa. Lehenik eta behin, *top left*, *top right*, *bottom left* eta *bottom right* eskualdeak definitu ditugu. Adibidez, $(0, 0, 0{,}5, 0{,}5)$ *top left* eskualdeari dagokio. Kaxa inguratzaile jakin bat eskualde horren badago, erlazio espazial hau betetzen duela deritzogu. Bestela, ondorengo eskualdeetan dagoen be-

giratzen dugu: *top*, *bottom*, *left* edo *right*. Adibide gisa, objektu bat *left* eskualdean dagoela diogu bere kaxa inguratzailea $(0, 0, 0{,}5, 1)$ tartean badago. Aurreko kondizioak betetzen ez dituenean, objektua irudiaren zentroan dagoela deritzogu (*center*). obj objektuak reg eskualdean dagoen galdera egiteko ondorengo txantiloia erabiltzen dugu: "is $\langle obj \rangle$ in $\langle reg \rangle$ region?"

Bi objekturen arteko tamaina. Izan bedi obj_1 eta obj_2 objektuak dagozkien kaxa inguratzaileekin. Objektu bakoitzeko ondorengo funtzioak kalkulatzen ditugu, hauen kaxa inguratzaileak erabiliz: $width(obj)$, $tall(obj)$ eta $area(obj)$. $width(obj_1) > width(obj_2)$ baldintza betetzen bada, obj_1 obj_2 baino zabalagoa dela (*wider*) zehazten dugu. Adibide berarekin obj_2 obj_1 baino estuagoa (*narrower*) dela erraz ondorioztatu dezakegu ere bai. Pareko erregelak aplikatzen ditugu *taller/shorter* erlazioekin $length(obj)$ funtzioa erabiliz eta *larger/smaller*rekin $area(obj)$ funtzioa erabiliz. Galderaren txantiloiari begira, obj_1, obj_2 objektuak eta tamaina konparatzen duen rel erlazioa izanik, ondorengoa erabiltzen dugu: "is $\langle obj_1 \rangle$ $\langle rel \rangle$ than $\langle obj_2 \rangle$?"

Bi objekturen arteko posizioa. Izan bedi obj_1 eta obj_2 objektuak dagozkien kaxa inguratzaileekin. obj_1 objektuaren kaxa inguratzailea obj_2-ren kaxaren barruan badago, obj_1 obj_2-ren barne dagoela diogu (*inside*), baita obj_2-k obj_1 inguratzen duela ere (*surrounding*). *left of*, *right of*, *above* eta *below* erlazioetarako, bi kaxen arteko zentroek osatzen duten angelua erabiltzen dugu. Beste hitzetan, obj_1-en kaxaren zentroa jatorri gisa hartuz, lau koadrantetan zatitzen dugu irudia. obj_2-ren zentroa $[\frac{-3}{4}\pi, \frac{3}{4}\pi]$ angeluen artean badago, obj_2 obj_1-en ezkerretara dagoela

deritzogu (*left of*). Antzekoa burutzen dugu beste 3 erlazioekin: $[\frac{-3}{4}\pi, \frac{-1}{4}\pi]$ *above* erlazioari dagokio, $[\frac{-1}{4}\pi, \frac{1}{4}\pi]$ *right of* erlazioari and $[\frac{1}{4}\pi, \frac{3}{4}\pi]$, berriz, *below* erlazioari. Azkenik, bi kaxa inguratzaileen *IoU* balioa kalkulatuz (*Intersection-over-union*, obj_1 eta obj_2 separatuta daudela diogu hauen *IoU* baioa zerokoa bada (*separated*), edota teilakatuta daudela *IoU*> 0 bada (*overlapping*). Galderaren txantiloiari begira, obj_1, obj_2 objektuak eta posizioa konparatzen duen rel erlazioa izanik, ondorengoa erabiltzen dugu: "is $\langle obj_1 \rangle$ $\langle rel \rangle$ $\langle obj_2 \rangle$?". *separated* erlazioaren kasuan aldaera bat erabiltzen dugu: "are $\langle obj_1 \rangle$ and $\langle obj_2 \rangle$ separated?".

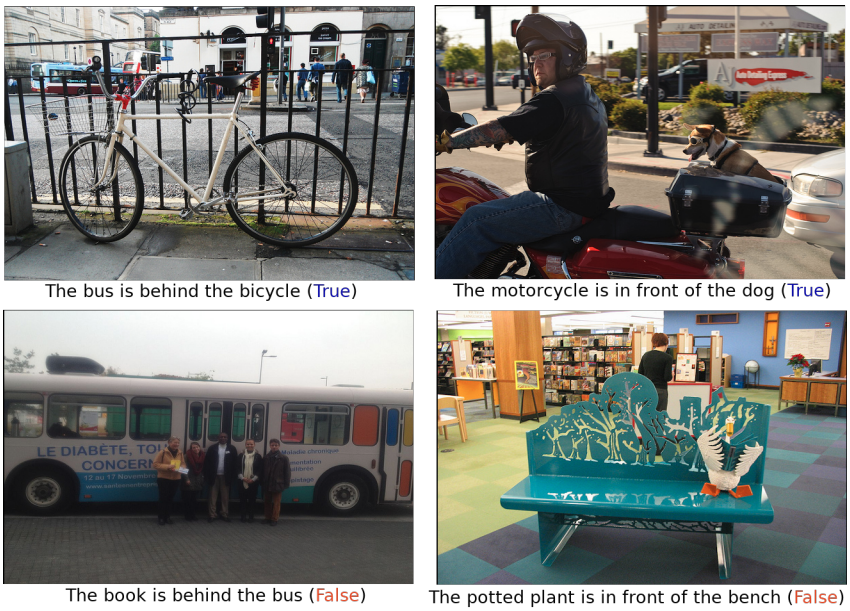
B.2 Orokortze Ahalmenaren Analisi Kualitatiboa

Bi hizkuntza-ereduren hainbat adibide konparatzen ditugu: i) VSRen bakarrik doitu den BERT-base eredua, kokapen tokenak erabiltzen dituena (BERT bezala izendatuko dugu hemendik aurrera), eta ii) aurretik ikasketa espaziala jasan duen BERT-base eredua, kokapen tokenak erabiltzen duelarik ere bai (st-BERT deritzoguna).

Aipatu dugun bezala, atal honetan SSTD datu-multzoarekin doikuntza burutzeak ematen duen orokortze ahalmena aztertu nahi dugu. Horretarako, SSTD-en agertzen ez diren bi erlaziotan zentratu gara: *behind* eta *in front of*. Erlazio hauek ezin dira guk erabilitako kaxa inguratzaileekin zehaztu anbiguetaterik sartu gabe, sakonera sartzen dute ekuazioan eta. *Behind*-en kasuan 136 instantzia ditugu VSR-ko ebaluazio azpimultzoan. BERT-ek 60 puntuko asmatze-tasa lortzen du eta st-BERT ereduak, berriz, 75. *In front of* erlazioaren kasuan 116 instantzia

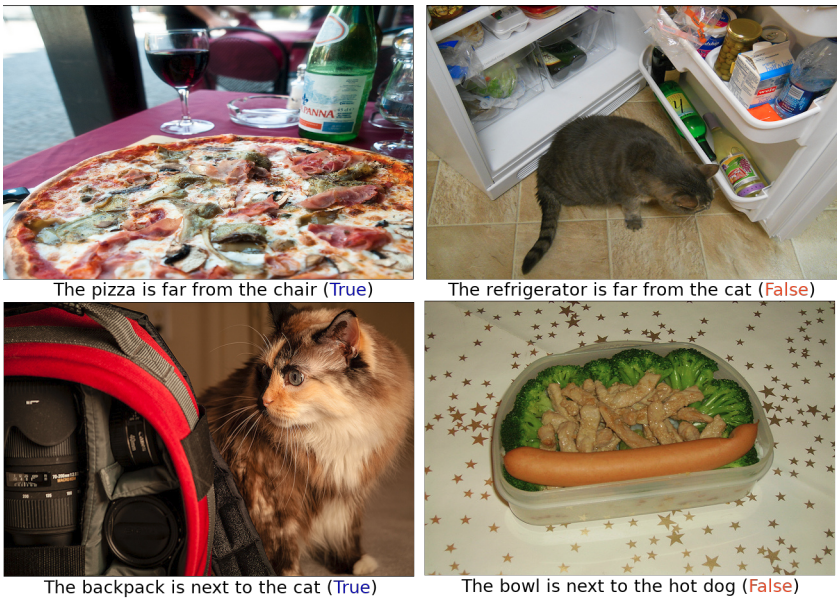
ditugu. BERTek 58 puntuko asmatze-tasa lortzen du eta st-BERT ereduak, berriz, 61. Emaitza hauek SSTD-ko ikasketak laguntzen duela erakusten du, erlazio hauek ikasketa horretan agertzen ez badira ere. B.1. Irudiak gertatzen ari denaren intuizioa ekar dezake. Irudiko lehen adibidean, autobusa bizikleta baino askoz txikiagoa dela ikus daiteke. SSTD-k bi objekturen arteko tamaina erlazionatzen dituzten erlazioak ditu eta, gure ustez, autobusak normalean bizikletak baino handiagoak direla ikasi du ereduak. Horrela, VSR-ko doikuntzan zehar ereduak informazio hau erabili dezake objektuen sakonera zein den inferitzeko. Antzeko arrazoinamendua egin daiteke bigarren (txakurra eta motoa) eta laugarren (landarea eta bankua) adibideen kasuan, instantzia hauek *in front of* erlazioa dutelarik. Hala ere, hirugarren adibiderako (autobusa eta liburua), st-BERT ereduak, bi objektuen kaxen informazioa erabiliz, liburua ikusgarri dagoela inferitzeko kapaza dela ikus dezakegu. Azken finean, liburuaren kaxa inguratzailearen informazioa lortu ahal izateko liburua autobusaren aurrean egon behar du. BERT-ek, ordea, ezin izan du ondo aurreikusi kasu hau.

SSTD-en ez dauden beste bi erlaziorekin berdina egin dugu. Kasu honetan objektuen arteko tamaina konparatzen dituzte: *next to* eta *far from*. *next to* erlazioaren kasuan, BERT-ek 56 puntu lortzen ditu eta st-BERT-ek, berriz, 73 (41 adibide konparatzen ditugu). *far from*-en kasuan, BERT-ek 83 eta st-BERT ereduak 91 lortzen dituzte (23 adibideren gainean kalkulatuta). Kontuan hartzekoa da *far away from* eta *far from* erlazioak oso antzekoak direla, eta st-BERT ereduak askoz puntuazio hobea lortzen duela (88 vs. 73, 49 instantzia erabiliz). B.2. Irudiko lehenengo adibidean (pizza eta aulkia), aulkiaren tamaina txikia dela eta,



B.1 Irudia – Bi BERT ereduren iragarpenen arteko konparaketa VSR-ko ebaluazio instantzietan. Ikasketa espaziala jasan duen BERT ereduak zuzen ebazten ditu kasu hauek, baina VSR-n bakarrik doitu den BERT ereduak ez ditu ondo asmatzen.

irudian aulkia oso sakon dagoela inferitu dezakegu. st-BERT ereduak informazio hau erabili dezakeela dirudi, BERT-ek ez bezala. Bigarren adibiderako (hozkailua eta katua), bi kaxa inguratzaileak teilakatzen dira eta st-BERT-ek bi objektuak oso urrun ez daudela ondorioztatzen du, bi objektuen arteko tamaina ezberdintasunak normaltzat hartu dituelako dirudienez. Hirugarren adibidean antzeko arrazoinamendua erabiltzen du ere bai (motxila eta katua). Azkenekoan, berriz, saltxitxa tuperraren barruan dagoela ikusi du. st-BERT ereduak *next to* erlazioan kasu hau ezohikoa dela ondorioztatzen du, baina BERT-ek ezin du berdina burutu, biek irudiaren deskribapen bera erabiltzen badute ere.



B.2 Irudia – Bi BERT ereduren iragarpenen arteko konparaketa VSR-ko ebaluazio instantzietan. Ikasketa espaziala jasan duen BERT ereduak zuzen ebazten ditu kasu hauek, baina VSR-n bakarrik doitu den BERT ereduak ez ditu ondo asmatzen.

B.3 Erregela Bidezko Sistemaren Inplementazioa

Erregeletan oinarritutako sistema inplementatzeko, VSR eta SSTD erlazioen arteko mapaketa bat burutu behar dugu. 4.3.4. Atalean azaldu den bezala, VSR-ko 65 erlazioetatik 17 bakarrik mapatu daitezke. Mapaketa hau B.1. Taulan azaltzen da. VSR-ko ebaluazio instantzia bat hartuta, erlazio espaziala zein den begiratzen dugu eta SSTD-ko beste erlazio batera mapatu dezakegun begiratzen dugu. Mapaketa burutu badaiteke, ondorengoa burutzen dugu: i) VSR-ko goiburukotik izena eta objektua erauzten ditugu karaktere kateko manipulazioa burutuz; ii) izen eta objektu berdinak bilatzen ditugu deskribapen testualean; iii) biak aurkitzen ba-

dira, hauen kaxa inguratzaileak erauzi eta SSTD-ko erregelak aplikatzen ditugu inferentzia bukatzeko; iv) mapaketak huts egiten badu, erantzun bitarra ausaz aukeratzen dugu ($\%50$-ko probabilitatearekin).

VSR relation	**SSTD Relations**
at the left side of	*left of*
at the right side of	*right of*
around	*surrounding*
into	*inside*
on top of	*above*
beneath	*below*
left of	*left of*
right of	*right of*
under	*below*
below	*below*
above	*above*
over	*above*
contains	*surrounding*
within	*inside*
surrounding	*surrounding*
inside	*inside*
outside	*separated*

B.1 Taula – VSR eta SSTD erlazioen arteko mapaketa.

C. ERANSKINA

Erlazio Espazialek Baldintzatutako Irudien Sorrera

C.1 SR4G Datu-multzoa

Eranskin honetan, *main* eta *unseen* azpimultzoen inguruko detaile gehiago ematen ditugu, baita hirukote espazialak berbalizatzeko erabili ditugun txantiloiak eta hirukote hauek irudian betetzen diren ebaluatzeko heuristikoak ere.

C.1.1 Eskuz Zehaztutako Txantiloiak

Hirukote espazialetatik goiburukoak sortzeko erabili diren txantiloiak C.1. Taulan azaltzen dira. Ikus daitekeenez, goiburuko hauek ahal den sinpleenak izatea nahi dugu, beharrezkoak ez diren atributu edota aditzak alde batera utziz eta hirukote espazialean bakarrik zentratuz. Azken hau oso garrantzitsua da ulermen espaziala isolatuta aztertzeko.

Mota	Erlazioa	Txantiloia
Proiektiboa	*Left of*	⟨A⟩ to the left of ⟨B⟩.
	Right of	⟨A⟩ to the right of ⟨B⟩.
	Above	⟨A⟩ above ⟨B⟩.
	Below	⟨A⟩ below ⟨B⟩.
Topologikoa	*Overlapping*	⟨A⟩ overlapping ⟨B⟩.
	Separated	⟨A⟩ and ⟨B⟩ separated.
	Surrounding	⟨A⟩ surrounding ⟨B⟩.
	Inside	⟨A⟩ inside of ⟨B⟩.
Tamaina	*Taller*	⟨A⟩ taller than ⟨B⟩.
	Shorter	⟨A⟩ shorter than ⟨B⟩.
	Wider	⟨A⟩ wider than ⟨B⟩.
	Narrower	⟨A⟩ narrower than ⟨B⟩.
	Larger	⟨A⟩ larger than ⟨B⟩.
	Smaller	⟨A⟩ smaller than ⟨B⟩.

C.1 Taula – Goiburuko sintetikoak sortzeko txantiloiak.

C.1.2 Erregela Heuristikoak

Erregela heuristikoak datu-multzoa sortzeko eta sortutako irudiak ebaluatzeko erabiltzen ditugu. Hirukote espazial bat $\langle l_s, r, l_o \rangle$ eta hirukotea osatzen duten bi objektuen kaxa inguratzaileak izanik (bb_s eta bb_o), f_r funtzioa erabiltzen dugu s eta o objektuen arteko r erlazioa irudian betetzen den ala ez zehazteko. $bb_i = \{x_i^0, y_i^0, x_i^1, y_i^1\}$ kaxa inguratzailea kaxaren goi-ezker $\{x_i^0, y_i^0\}$ eta behe-eskuinalde $\{x_i^1, y_i^1\}$ erpinen koordenatuekin definitzen dugu.

left of, *right of*, *above* eta *below* erlazioetarako Gokhale *et al.* (2023) lanean definitutako f_r funtzioak erabili ditugu, kaxa bakoitzaren $c_i = \{x_i^c, y_i^c\}$ zentroideak kalkulatuz eta erlazio bakoitzari dagozkion koordenatuak konparatuz.

C.1 Irudia – Bi objektu detektatuta dituen irudi bat. Objektuen kaxa inguratzaileak marraztuta azaltzen dira.

Erlazioa	$f_r(bb_m, bb_e)$	$f_r(bb_e, bb_m)$
Left of	✓	
Right of		✓
Above	✓	
Below		✓
Overlapping		
Separated	✓	✓
Surrounding		
Inside		
Taller		✓
Shorter	✓	
Wider		✓
Narrower	✓	
Larger		✓
Smaller	✓	

C.2 Taula – C.1. Irudiko objektuen kaxaak kontuan hartuta f_r erabili dugu zein erlazio betetzen diren jakiteko. Lehenengo zutabean gizona da subjektua eta bigarrenean, berriz, elefantea.

Definitu ditugun beste 10 erlazioen erregelak Johnson *et al.* (2018) lanetik erauzi ditugu. Tamaina erlazioetan objektuen kaxen altuerak (*taller* eta *shorter*), zabalerak (*wider* and *narrower*) edota azalerak (*larger* eta *smaller*) konparatu ditugu. *Surrounding* erlazioaren kasuan bb_o kaxa bb_s-ren barruan dagoen begiratzen dugu, eta *inside*-n kasuan, berriz, alderantziz. Azkenik, bi kaxa inguratzaileen Intersection-over-Union metrika (edo IoU) erabili dugu gelditzen zaizkigun bi erlazioetarako. Bi kaxen IoU balioa 0-koa bada bi objektuak banatuta daudela diogu (*separated*), eta IoU balioa positiboa bada, berriz, gainezarrita daudela (*overlapping*).

Adibide gisa, C.1. Irudiko bi objektuen arteko zein erlazio espazial betetzen diren aztertu dugu. Horretarako, bi objektuen kaxa inguratzaileak lortu ditugu: $bb_m = \{120, 220, 152, 377\}$ gizonarena, eta $bb_e = \{160, 84, 351, 524\}$ elefantearena. Behin kaxa inguratzaileak izanik, atal honetan zehaztutako f_r funtzioak erabili ditugu $f_r(bb_m, bb_e)$ eta $f_r(bb_e, bb_m)$ konbinazio guztiak kalkulatzeko.

Konbinazio hauen emaitzak C.2. Taulan azaltzen dira. Erlazio gehienetan irudi hau bakarrik begiratuz zein erlazio betetzen diren ikustea erraza da, kalkulurik egin gabe, *above* eta *below* kasuetan izan ezik. Bi kaxen zentroideen alturak oso parean daude eta ez da oso naturala egiten bata bestea baino gorago dagoela esatea. Hala ere, datu hauetan entrenatutako ereduek irudi gehiago ikusten dituzten heinean, *above* eta *below*-ren esanahia ikasten joango dira.

C.1.3 *Main* eta *Unseen* Bertsioak

C.3. Taulan datu-multzoaren *unseen* bertsioan erabilitako objektu multzoak azaltzen dira. Objektu multzo hauek entrenamendurako, garapenerako eta ebaluaziorako daude pentsatuta, O_{train}, O_{val} eta O_{test} deritzogunak, hurrenez hurren.

O_{val} azpimultzoko objektu kopurua txikia denez, hirukote ezberdin gutxi eraiki daitezke. Horregatik, gutxienez O_{val}-eko objektu bat duten hirukoteak aukeratzen ditugu, beste objektua $O_{\text{train}} \cup O_{\text{val}}$ multzokoa izanik. Guztira, $(2 \cdot |O_{\text{train}}| \cdot |O_{\text{val}}| + \binom{|O_{\text{val}}|}{2}) \cdot 14 = 6.580$ hirukote ezberdin lor ditzakegu murriztapen honekin, non 5.326 COCO datu-multzoko irudietan agertzen diren.

Balidazioa konputazionalki garestia da bi bertsioetan, milaka irudi sortu behar baitira 5.2.2. Ataleko metrikak kalkulatzeko. Hainbat esperimentu burutuz, 10K

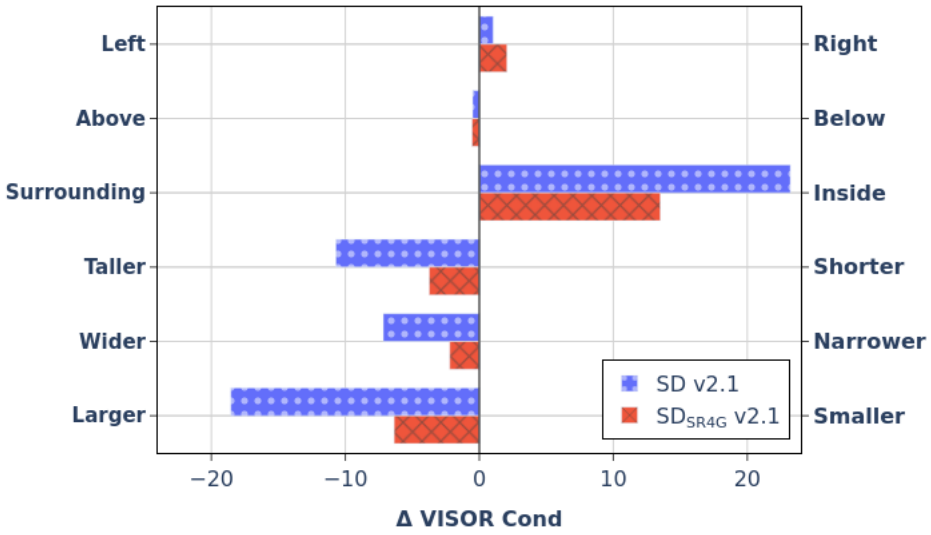
O_{train}
person, *car*, *motorcycle*, *airplane*, *train*, *boat*, *fire hydrant*, *bench*, *bird*, *elephant*, *bear*, *giraffe*, *handbag*, *tie*, *snowboard*, *baseball bat*, *baseball glove*, *surfboard*, *cup*, *knife*, *spoon*, *apple*, *sandwich*, *orange*, *broccoli*, *carrot*, *pizza*, *donut*, *chair*, *couch*, *potted plant*, *bed*, *dining table*, *toilet*, *laptop*, *mouse*, *remote*, *keyboard*, *oven*, *sink*, *book*, *clock*, *teddy bear*, *hair drier*, *toothbrush*
O_{val}
umbrella, *cake*, *tv*, *refrigerator*, *vase*
O_{test}
bicycle, *bus*, *truck*, *traffic light*, *stop sign*, *parking meter*, *cat*, *dog*, *horse*, *sheep*, *cow*, *zebra*, *backpack*, *suitcase*, *frisbee*, *skis*, *sports ball*, *kite*, *skateboard*, *tennis racket*, *bottle*, *wine glass*, *fork*, *bowl*, *banana*, *hot dog*, *cell phone*, *microwave*, *toaster*, *scissors*

C.3 Taula – *Unseen* bertsioan zehaztu diren entrenamendu, garapen eta ebaluazio azpimultzoetan erabilitako objektuak.

irudi sortzearekin emaitza kontsistenteak lortzen ditugula antzeman dugu. Horrela, 2,5K goiburuko espazial aukeratu genituen garapen azpimultzoetarako, bai *main* eta baita *unseen* bertsioan ere. 4 irudi sortzen ditugu goiburuko bakoitzeko, 10K irudi sortuz guztira.

C.2 LAION Datu-multzoa eta Erlazio Espazialak

5.2. Irudian Stable Diffusion (SD) ereduek erlazio espazialen arteko alborapen handia dutela antzeman dugu, *taller* hobetsiz *shorter* beharrean, adibidez. Emaitza hauek osatzeko esperimentu bera errepikatu dugu SR4G datu-multzoaren *main* bertsioarekin, joera antzekoak antzemanik (ikus C.2. Irudia). Alborapen hauen ja-



C.2 Irudia – Ardatz horizontalak VISOR$_{Cond}$ balioen diferentziak zehazten ditu aurkako esanahiak dituzten erlazio pareen artean, erlazio pare bakoitza ardatz bertikalean zehaztuta dagoelarik. Emaitza hauek SD v2.1 eta SD$_{SR4G}$ v2.1 ereduei dagokie, *main* bertsioan entrenatuta eta ebaluatuta daudenak.

Erlazio hobetsia	Aurkako erlazioa	Agerpen-proportzioa
Right (5M)	Left (5,6M)	0,91
Above (1,6M)	Below (0,7M)	2,47
Inside (2M)	Surrounding (0,3M)	7,61
Taller (49,3K)	Shorter (29,4K)	1,86
Wider (54,6K)	Narrower (5,7K)	9,62
Larger (0,8M)	Smaller (0,2M)	3,17

C.4 Taula – Aurkako erlazio pare bakoitzaren agerpen-proportzioa LAION-2B-en datu-multzoan. Lehenengo zutabean agertzen den erlazioa SD ereduak hobesten duena da.

torria ulertzeko, erlazio espazialen agerpen kopuruak aztertu ditugu LAION-2B-en datu-multzoan, SD ereduak entrenatzeko erabili den datu-multzoa hain zuzen ere. C.4. Taulak 12 erlazioren agerpen kopuruak erakusten ditu, aurkako esanahia duten 6 erlazio pareetan banatuta. Erlazio bakoitzak LAION-2B-en datu-multzoko agerpen kopurua parentesi artean dauka. Erlazio pare bakoitzeko, lehenengo zutabean agertzen dena SD v2.1 ereduak hobesten duena da. Hirugarren zutabeak, berriz, erlazio hobetsia bere erlazio parea baino zenbat aldiz gehiagotan agertzen den zehazten du, agerpen-proportzioa deritzoguna. Bat baino handiagoa den agerpen-proportzioak erlazio hobetsia bere parea baino gehiagotan agertzen dela erakusten du. Gure emaitzek korrelazio argi bat erakusten dute agerpen-proportzio eta SD v2.1 ereduaren preferentzien artean. Salbuespen bakarra *right* eta *left* parearen kasua da, non biak kopuru antzekoetan agertzen diren eta *right* erlazioarekin duen preferentzia oso txikia den.

C.3 Entrenamenduan Egindako Datu Gehikuntza

Entrenamenduan zehar hainbat datu gehikuntza estrategia aplikatzen ditugu, besteak beste ausazko mozketak eta iraulketa horizontalak. Ausazko iraulketa horizontalak ohikoak dira testu bidezko irudi sortzaileen entrenamenduan, baina iraulketa hauek irudiak bakarrik eraldatzen dituzte, goiburukoak berdin utziz. Horregatik, *left of* eta *right of* bezalako erlazio espazialen ikasketa ez da ahalbidetzen. Dena den, gure kasuan transformazio horiek ere irudiko objektuei aplikatzen diegu ere bai, goiburukoak automatikoki sortzeko erabiltzen ditugunak hain zuzen

Nº Captions	$\text{VISOR}_{\text{Cond}} \uparrow$	VISOR ↑	OA ↑
1	68,1	26,5	38,9
2	69,4	27,4	39,5
3	67,7	27,1	40,0
4	63,7	21,9	34,3
5	63,0	22,9	36,3

C.5 Taula – SD v1.4 doitu dugu SR4G-ko *main* bertsioan goiburuko kopuru ezberdinak konkatenatuz ereduaren sarreran. Emaitza hauek *main* bertsioko garapenean lortu dira.

ere. Horrela, datu gehikuntza hau erabili dezakegu erlazio espazialen oinarritze zuzena mantenduz.

Ausazko mozketak irudietako objektu kopurua murriztu dezakete, O_{I^*} zerrenda txikituz. Bi objektu baina gutxiago gelditzen diren mozketetan mozketa hau errepikatzen dugu $|O_{I^*}| \geq 2$ izan arte, gehienez max_iter saiakera eginik.

Ereduari elikatzen diogun sarrera hainbat goiburukoen arteko konkatenazioa izan daitekeenez, k hiperparametro bat ere definitu dugu konkatenatzen diren goiburuko kopuruak zehazteko. C.5. Taulak $k \in \{1, \ldots, 5\}$ goiburuko konkatenatuz lortzen ditugun emaitzak erakusten ditu. Ikus dezakegunez, $k = 2$ kasuan lortzen ditugu emaitza hoberenak $\text{VISOR}_{\text{Cond}}$ metrikan. Beraz, $k = 2$ erabili dugu gure lanean zehar.