Universidad del País Vasco · Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Master Thesis

Master in Language Analysis and Processing

Grounding Language Models for Spatial Reasoning

Author
Julen Etxaniz

Advisors
Oier Lopez de Lacalle
Aitor Soroa

Departments
Computer Systems and Languages
Computational Science and Artificial Intelligence

2023

Abstract

Humans are able to learn to understand and process the distribution of space, and one of the initial tasks of Artificial Intelligence has been to show machines the relationships between space and the objects that appear in it. Humans naturally combine vision and textual information to acquire spatial relationships among objects, and when reading a text, we are able to mentally depict the spatial relationships that may appear in it. Thus, the visual differences between images depicting "a person sits and a dog stands" and "a person stands and a dog sits" are obvious for humans, but still not clear for automatic systems. In this project, we propose to build grounded Neural Language models that are able to perform this kind of spatial reasoning. Neural Language models (LM) have shown impressive capabilities on many NLP tasks but, despite their success, they have been criticized for their lack of meaning. Vision-and-Language models (VLM), trained jointly on text and image or video data, have been offered as a response to such criticisms, but recent work has shown that these models struggle to ground spatial concepts properly. In the project we propose to build spatially-aware language models that ground spatial concepts in images. We propose to use a variety of methods that involve the creation of synthetic datasets specially focused on spatial reasoning capabilities, as well as the use of multi-task learning. We expect the new models to improve the state of the art in spatial reasoning.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

This chapter is an introduction of the master thesis and includes background, objectives and research questions.

1.1 Background

Neural Language Models (LM) have shown **impressive capabilities** on many Natural Language Processing (NLP) tasks [1, 2, 3]. LMs are pretrained on large corpora in order for them to learn universal language representations, which are beneficial for downstream NLP tasks and can avoid training a new model from scratch. The pretrained models are then fine-tuned in specific downstream tasks, using annotated data that is orders of magnitude smaller than the text used in the pretraining phase. Following this transfer learning methodology, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English [4, 1].

Despite the impressive results of LMs for different language-related tasks, many authors criticize them for their **lack of meaning** [5, 6]. In their opinion, language models trained exclusively on linguistic form (i.e. words) are unable to learn meaning. Those authors suggest that **grounding is one of the key elements to bring human-like language understanding**. However, language grounding is a very broad area that covers a great diversity of techniques, modalities and concepts. In this project, we will focus on spatial reasoning, that is, **grounding LMs with spatial concepts**. We choose spatial reasoning because it is one of the most fundamental capabilities for both humans and LMs. Such relations are crucial to how humans organize the mental space and make sense of the physical world, and therefore fundamental for a grounded theory of semantics [7]. However, spatial reasoning has been found to be particularly challenging (much more challenging than capturing properties of individual entities) for current models [8].

Vision Language Models (VLM), which are trained jointly on text and image, have been proposed as a general solution to the lack of grounding in language models [9, 10, 11, 12]. VLMs have been used in tasks that require grounding spatial concepts, such as VQA [13] or NLVR2 [14], but recent work has shown that **VLMs struggle to ground spatial concepts properly** [15]. Large generative VLMs trained on massive amounts of data like DALLE-2 [11] or IMAGEN [12] are known to possess visual-reasoning skills [16], but they are not publicly available and only accessible to large companies.

There are several **works that try to ground language models to spatial relations**. For example, [17, 18] focus on the acquired commonsense knowledge of models about object scales, e.g. do they know that a person is bigger than an ant? However, they ask about generic object scale relations, without providing any context. Some other authors [19, 20] work on implicit and explicit spatial relations of objects, given some descriptive texts. The proposed benchmark datasets are designed for object bounding box generation.

Multimodal training datasets with images and corresponding textual descriptions that include explicitly spatial relations tend to be small. A very recent work proposes a method called Pseudo-Q to **automatically create synthetic datasets that can be used to train visually grounded models** [21]. Their method consists of leveraging an off-the-shelf object detector to identify visual objects from unlabeled images, and then creating language queries for these objects that are obtained in an unsupervised fashion with a pseudo-query generation module. We propose to follow a similar approach, and create synthetic datasets that are specially tailored to acquire spatial relations.

With the objective of **evaluating spatial relations**, a recent work provides new unified datasets

[15]. 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 concepts (they assume the grounding occurs in a previous training process). Interestingly, authors find that VLMs, and more concretely text-to-image systems, perform much better than text-only LMs.

CLEVR was one of the pioneering works on testing compositional language and elementary visual reasoning [22]. However, it presents 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. In a similar fashion, SpartQA provides a synthetic question-answering dataset that is specially focused on spatial reasoning capabilities. However, it contains only text and no images, and therefore it does not provide any means to ground spatial concepts.

The Winoground dataset [23] is focused on **evaluating visio-linguistic compositional reasoning** in VLMs. Each instance in the dataset is composed of two images and two captions, but crucially, both captions contain a completely identical set of words, only in a different order. The task is then to match them correctly, which requires the systems to properly deal with composition in natural language.

Another very recent dataset named Visual Spatial Reasoning (VSR) [24], whose objective is to test spatial grounding capabilities by covering 65 different spatial relations over natural images collected from COCO [25]. Given an image, VSR provides 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 with respect to the given image. Another advantage of this dataset is that it is annotated by humans. Given its features, **we believe VSR is a good candidate to evaluate spatial grounding in LMs**.

1.2 Objectives

Despite the impressive performance of pretrained vision and language models (VLMs) on a wide variety of multimodal tasks, they remain poorly understood. One important question is to what extent such models are able to conduct unimodal and multimodal compositional reasoning and spatial reasoning. For example, the visual differences between images depicting "a person sits and a dog stands" and "a person stands and a dog sits" are clamorously obvious for humans, but still not clear for current state-of-the-art VLMs. To perform well on tasks where compositional and spatial reasoning is required, the models do not only need a proper encoding of text and images, but also to be able to **ground meaning across the two modalities** (spatial grounding).

Thus the main objective of the project is to **learn language models for spatial reasoning via the grounding of LMs with spatial concepts and relations**. One of the main goals of the project is to investigate ways to acquire grounded representation for spatial reasoning. In that sense, we will define suitable ways to incorporate spatial information into pre-trained vision and language models. Towards this goal, this project will focus on using the latest advances in deep-learning techniques, pre-trained LMs for effective zero and few-shot transfer learning.

We have defined the following specific objectives in the scope of spatial reasoning:

1. **Investigate the use of synthetic datasets to overcome the lack of annotated datasets for spatial grounding**. As to avoid the scarcity of multimodal datasets that explicitly describe spatial relations, we propose to automatically construct synthetic datasets on spatial relations and use them to train existing language models in a self-supervised way, with the final aid of obtaining spatially grounded language models. In particular, we propose two alternatives to produce the synthetic datasets:

    a) **Explicit verbalization** of spatial relations in images. Given an image in an existing dataset, we propose to use an object detector to identify the entities in the images, as well as handdesigned verbalization templates to automatically generate textual descriptions of the spatial relations among them.

b) **Using large generative VLMs**, which are known to obey spatial relations as described in the text, to obtain realistic images with entities that are arranged following certain spatial relations.

2. **Investigate the use of multi-tasking and multi-sourcing to improve generalization properties**. In a multi-task training paradigm, the model is forced to learn more than one task simultaneously, therefore improving its generalization capabilities. We will investigate multi-task settings to combine the verbalized dataset, the images produced by the generative VLMs, as well as traditional training data to obtain spatial-aware language models.

3. **Improve zero-shot and few-shot generalization of VLM models** to obtain effective models in small data regimes of the spatial reasoning domain without the necessity of explicitly annotating big quantities of spatial relations.

4. **Improve the state of the art in spatial reasoning**. Improve the state of the art in spatial reasoning. The final goal is to apply the findings learnt from previous objectives to improve the state-of-the-art in multiple datasets. We plan to evaluate our models at least on two vision and language datasets. The first one is the Winoground dataset [23], which presents a novel task for evaluating the ability of vision and language models to conduct visio-linguistic compositional reasoning. The second one is the VSR benchmark [24] for investigating VLMs capabilities in recognising 65 types of spatial relationships in natural text-image pairs.

1.3 Research Questions

Research Tasks (RT) and Questions (RQ) are based on the objectives from the previous section.

RT0. Prepare the research scenario. The initial task is related to **gathering corpora, exploring different datasets**, **Language Models** (LM) and **building a baseline prototype**. We have already identified some important datasets on spatial reasoning but we will check if there is any new appropriate dataset to evaluate our models. At the same time, we will examine and reimplement (if needed) state-of-the-art systems in order to further understand the task to be solved. This leads us to the following research questions: **RQ0.A) Are the available datasets appropriate to evaluate the spatial abilities of current LMs? RQ0.B) Which is the best pre-trained LMs for spatial reasoning?** We will conduct a quantitative and qualitative analysis of the existing text-only LMs and vision-language LMs in order to 1) measure the appropriateness of probing evaluations of the datasets and 2) explain the limitations of different types of pre-trained LMs.

RT2: Perform synthetic data generation using generative models to learn spatial grounding. We will focus on using large generative VLMs to construct high quality synthetic images that depict a fixed set of spatial relations. In that sense, we want to answer the following research questions. **RQ2.A) Which is the right way to make explicit the implicit information encoded in generative VLMs? RQ2.B) Can we improve the state-of-the-art of vision and language models in tasks that require spatial reasoning?**

RT3: Perform multi-task and multi-source learning in few-shot settings. In this task we will focus on finding ways of applying multi-task learning using multiple sources of information in order to force LMs to ground spatial relations into text without the necessity of explicitly annotating big quantities of spatial relations. This leads us to the following research questions: **RQ3.A) What kind of tasks and information sources are relevant to learn spatial information effectively? RQ3.B) What is the best way to combine the task in a multi-task setting? RQ3.C) Can we effectively minimize annotated data to obtain state-of-the-art results in tasks that require spatial reasoning?**

2 Related Work

3 Datasets

This chapter introduces the datasets and metrics we used.

3.1 Winoground

3.1.1 Dataset

3.1.2 Metrics

3.1.2.1 Score

Performance on Winoground [23] is computed according to three different metrics that evaluate different aspects of the models' visio-linguistic reasoning abilities.

The first metric is the **text score**, which measures whether a model can select the correct caption, given an image. Given images I_0 and I_1 and captions C_0 and C_1, the text score for an example (C_0, I_0, C_1, I_1) is computed according to:

$$ts(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \quad \text{and } s(C_1, I_1) > s(C_0, I_1) \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where $s(\cdot)$ is the model's score for the image/caption pair.

The second metric is the **image score**, which measures whether a model can select the correct image, given a caption. The image score for an example is computed according to:

$$is(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \quad \text{and } s(C_1, I_1) > s(C_1, I_0) \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

Our final metric **group score** combines the previous two, which measures if every combination for a given example is correctly scored by the model. The group score for an example is computed according to:

$$gs(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } ts(C_0, I_0, C_1, I_1) \\ & \quad \text{and } is(C_0, I_0, C_1, I_1) \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

3.1.2.2 Accuracy

We also add three additional accuracy metrics. These are similar to the previous ones, but accuracy is 0.5 when one of the pairs is correct.

Given images I_0 and I_1 and captions C_0 and C_1, the **text accuracy** for an example (C_0, I_0, C_1, I_1)

is computed according to:

$$ta(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \quad \text{and } s(C_1, I_1) > s(C_0, I_1) \\ 0.5 & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \quad \text{xor } s(C_1, I_1) > s(C_0, I_1) \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

where $s(\cdot)$ is the model's score for the image/caption pair.

The **image accuracy** for an example is computed according to:

$$ia(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \quad \text{and } s(C_1, I_1) > s(C_1, I_0) \\ 0.5 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \quad \text{xor } s(C_1, I_1) > s(C_1, I_0) \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

The **group accuracy** in our framework is computed according to:

$$ga(C_0, I_0, C_1, I_1) = (ta(C_0, I_0, C_1, I_1) + ia(C_0, I_0, C_1, I_1))/2 \tag{3.6}$$

4 Methods

This chapter explains the methods we used for evaluation.

4.1 Winoground

4.1.1 Models

We evaluate various configurations of the following multimodal transformers: OFA [26], BLIP [27], CLIP [28], FLAVA [29], LXMERT [10], UniT [30], UNITER [31], VILLA [32], VinVL [33], ViLT [34], VisualBERT [35] and ViLBERT [9]. We also evaluate several configurations of two types of RNN-based models: VSE++ [36] and VSRN [37].

We provide a high-level overview of the differences between the models in Table 4.1 which includes pretraining datasets, architecture, and attention mechanisms between the modalities. We omit datasets that were only used to train backbones. We exclude the language embedding from this table as every model uses a pretrained BERT tokenizer, except CLIP, VSE++, and VSRN. The pretraining datasets include COCO [25], Visual Genome (VG) [38], Conceptual Captions (CC) [39], SBU Captions [40], Flickr30k [41], VQA 2.0 [42], VCR [43], NLVR2 [44], SNLI-VE [45], QNLI [46], MLNI-mm [47], QQP [48], Localized Narratives (LN) [49], Wikipedia Image Text (WIT) [50], Conceptual Captions 12M (CC 12M) [51], Red Caps (RC) [52], YFCC100M [53], SST-2 [54], and LAION [55]. CLIP uses their own dataset for pretraining.

Model	Datasets	# Images, Captions	Architecture	Attention
VinVL [33]	VQA, GQA, VG-QA, COCO, Flickr30k, CC, SBU	1.89, 4.87	single-stream	merged
UNITER [31]	COCO, VG, CC, SBU	4.20, 9.58	single-stream	merged
ViLLA [32]	COCO, VG, CC, SBU	4.20, 9.58	single-stream	merged
VisualBERT [35]	COCO, NVLR2	0.30, 0.52	single-stream	merged
ViLT [34]	COCO, VG, SBU, CC	4.10, 9.85	single-stream	merged
LXMERT [10]	COCO, VG	0.18, 9.18	dual-stream	modality-specific, co-attn, merged
ViLBERT [9]	CC	3.30, 3.30	dual-stream	modality-specific, co-attn, merged
UniT [30]	COCO detect., VG detect., VQAv2, SNLI-VE QNLI, MNLI-mm, QQP, SST-2	0.69, 1.91	dual-stream	modality-specific, merged
FLAVA $_{ITM}$ [29]	COCO, SBU, LN, CC, VG, WIT, CC 12M, RC, YFCC100M	70.00, 70.00	dual-stream	modality-specific, merged
FLAVA $_{Contrastive}$ [29]	COCO, SBU, LN, CC, VG, WIT, CC 12M, RC, YFCC100M	70.00, 70.00	dual-stream	modality-specific
CLIP [28]	–	400.00, 400.00	dual-stream	modality-specific
VSE++ and VSRN $_{COCO}$ [36, 37]	COCO	0.11, 0.57	dual-stream	–
VSE++ and VSRN $_{Flickr30k}$ [36, 37]	Flickr30k	0.03, 0.16	dual-stream	–
OFA [26]	CC 12M, CC 3M, SBU, COCO, VG-Cap	20.00, 20.00	single-stream	modality-specific, merged
BLIP$_{ITM}$ 14M [27]	COCO, VG, SBU, CC, CC 12M	14.00, 15.00	dual-stream	modality-specific, merged
BLIP$_{ITC}$ 14M [27]	COCO, VG, SBU, CC, CC 12M	14.00, 15.00	dual-stream	modality-specific
BLIP$_{ITM}$ 129M [27]	COCO, VG, SBU, CC, CC 12M, LAION	129.00, 130.00	dual-stream	modality-specific, merged
BLIP$_{ITC}$ 129M [27]	COCO, VG, SBU, CC, CC 12M, LAION	129.00, 130.00	dual-stream	modality-specific

Table 4.1: A high-level overview of the differences between the models we evaluate by the pretraining datasets, architecture, and attention mechanisms between the modalities.

5 Results

This chapter introduces baseline results and our results.

5.1 Winoground

5.1.1 Compared To Humans

5.1.1.1 Baseline

We show baseline results in Table 5.1, which includes the following multimodal transformers: CLIP [28], FLAVA [29], LXMERT [10], UniT [30], UNITER [31], VILLA [32], VinVL [33], ViLT [34], VisualBERT [35] and ViLBERT [9]. They also evaluate several configurations of two types of RNN-based models: VSE++ [36] and VSRN [37].

Model	Score Text	Score Image	Score Group	Accuracy Text	Accuracy Image	Accuracy Group
MTurk Human	**89.50**	**88.50**	**85.50**	**93.75**	**93.88**	**93.81**
Random Chance	25.00	25.00	16.67	50.00	50.00	50.00
VinVL	**37.75**	17.75	14.50	**62.75**	**57.75**	**60.25**
UNITER$_{large}$	**38.00**	14.00	10.50	**63.25**	**55.75**	**59.50**
UNITER$_{base}$	**32.25**	13.25	10.00	**60.62**	**55.50**	**58.06**
ViLLA$_{large}$	**37.00**	13.25	11.00	**62.62**	**55.25**	**58.94**
ViLLA$_{base}$	**30.00**	12.00	8.00	**59.62**	**55.00**	**57.31**
VisualBERT$_{base}$	15.50	2.50	1.50	**50.50**	49.88	**50.19**
ViLT (ViT-B/32)	**34.75**	14.00	9.25	**60.50**	**55.38**	**57.94**
LXMERT	19.25	7.00	4.00	**52.12**	**51.88**	**52.00**
ViLBERT$_{base}$	23.75	7.25	4.75	**57.25**	**52.50**	**54.87**
UniT$_{ITM Finetuned}$	19.50	6.25	4.00	**50.25**	**50.75**	**50.50**
FLAVA$_{ITM}$	**32.25**	20.50	14.25	**62.75**	**59.13**	**60.94**
FLAVA$_{Contrastive}$	**25.25**	13.50	9.00	**59.25**	**55.12**	**57.19**
CLIP (ViT-B/32)	**30.75**	10.50	8.00	**60.38**	**53.25**	**56.81**
VSE++$_{COCO}$ (ResNet)	22.75	8.00	4.00	**51.38**	**50.88**	**51.12**
VSE++$_{COCO}$ (VGG)	18.75	5.50	3.50	**50.38**	49.75	**50.06**
VSE++$_{Flickr30k}$ (ResNet)	20.00	5.00	2.75	**51.50**	**50.25**	**50.88**
VSE++$_{Flickr30k}$ (VGG)	19.75	6.25	4.50	**52.75**	**51.00**	**51.88**
VSRN$_{COCO}$	17.50	7.00	3.75	**50.38**	**51.12**	**50.75**
VSRN$_{Flickr30k}$	20.00	5.00	3.50	**53.25**	**51.75**	**52.50**

Table 5.1: Results on the Winoground dataset across the text, image and group score and accuracy metrics. Results above random chance in **bold**.

5.1.1.2 Ours

We show our results in Table 5.2, which includes various configurations of the following multimodal transformers: OFA [26], BLIP [27], CLIP [28], FLAVA [29] and ViLT [34].

We test 4 different versions of ViLT. The first one is the pre-trained only version, without finetuning. Two others are finetuned for retrieval on COCO and Flickr30k. The last one is finetuned for visual reasoning on NLVR2. The best one is the one trained on NLVR2, which shows that finetuning on that task helps perform better on Winoground. Finetuning for retrieval is also helpful and improves the results of the pre-trained model. The score of the pre-trained model is lower than the baseline one.

For FLAVA and CLIP we manage to replicate baseline results. We also test 3 other CLIP models with different configurations and find that they all perform similar to the baseline configuration.

We test the 5 model sizes of OFA. Taking into account that this model gets state-of-the-art performance on many tasks, the performance is not very good. Even the biggest model is not better than the best baseline model. OFA is trained to generate "yes" or "no" when given an image and the text "Does the image describe <caption>?". This might explain why it does not perform that well on retrieval and Winoground.

We test many configurations of BLIP, which include different training sizes, scoring, vision transformer sizes and finetuning datasets. ITM score is better than ITC score in all the cases. Even the 14M pretrained only model is better than all the previously tested models. Finetuning for retrieval on COCO and Flickr30k improves the results even more, reaching nearly above random performance in text, image and group scores.

However, even the best model is still far from human performance in text, image and group scores. If we look at accuracy metrics, the gap is reduced, but the difference is still very big. Image score remains much lower than text score for all the models.

5.1.2 Results By Linguistic Tag

5.1.2.1 Baseline

See Table 5.3

5.1.2.2 Ours

See Table 5.4

5.1.3 Results By Visual Tag

5.1.3.1 Baseline

See Table 5.5

5.1.3.2 Ours

See Table 5.6

Model	Score Text	Score Image	Score Group	Accuracy Text	Accuracy Image	Accuracy Group
MTurk Human	**89.50**	**88.50**	**85.50**	**93.75**	**93.88**	**93.81**
Random Chance	25.00	25.00	16.67	50.00	50.00	50.00
ViLT (ViT-B/32)	**27.50**	8.75	6.00	**56.88**	**53.12**	**55.00**
ViLT$_{COCO}$ (ViT-B/32)	**32.75**	13.50	11.25	**61.88**	**56.00**	**58.94**
ViLT$_{Flickr30k}$ (ViT-B/32)	**35.00**	11.50	9.75	**61.62**	**54.50**	**58.06**
ViLT$_{NLVR2}$ (ViT-B/32)	**38.00**	15.25	12.00	**58.75**	**55.62**	**57.19**
FLAVA$_{ITM}$	**32.25**	20.50	14.25	**62.75**	**59.13**	**60.94**
FLAVA$_{ITC}$	**25.25**	13.50	9.00	**59.25**	**55.12**	**57.19**
CLIP (ViT-B/32)	**30.75**	10.25	8.25	**60.38**	**53.12**	**56.75**
CLIP (ViT-B/16)	25.00	10.25	7.00	**57.88**	**53.75**	**55.81**
CLIP (ViT-L/14)	**28.50**	11.00	8.00	**60.38**	**54.62**	**57.50**
CLIP (ViT-L/14-336)	**27.50**	12.00	8.00	**59.38**	**55.12**	**57.25**
OFA$_{Tiny}$	20.50	8.00	3.75	**53.50**	**52.00**	**52.75**
OFA$_{Base}$	**26.50**	10.50	7.00	**58.88**	**54.00**	**56.44**
OFA$_{Medium}$	22.75	9.00	5.50	**54.25**	**52.75**	**53.50**
OFA$_{Large}$	**26.00**	8.75	5.75	**58.38**	**52.88**	**55.62**
OFA$_{Huge}$	**36.25**	15.50	13.50	**64.38**	**56.62**	**60.50**
BLIP$_{ITM14M}$ (ViT-B/16)	**39.25**	19.00	15.00	**65.88**	**58.25**	**62.06**
BLIP$_{ITC14M}$ (ViT-B/16)	**32.25**	13.75	10.50	**62.25**	**56.50**	**59.38**
BLIP$_{ITM}$ (ViT-B/16)	**40.50**	20.50	16.50	**66.25**	**59.00**	**62.62**
BLIP$_{ITC}$ (ViT-B/16)	**29.75**	14.50	9.50	**59.88**	**56.12**	**58.00**
BLIP$_{ITM}$ (ViT-B/16) (CapFilt-L)	**37.50**	18.50	14.00	**65.00**	**59.13**	**62.06**
BLIP$_{ITC}$ (ViT-B/16) (CapFilt-L)	**31.50**	10.50	8.50	**61.38**	**53.62**	**57.50**
BLIP$_{ITM}$ (ViT-L/16)	**42.50**	18.25	15.50	**66.88**	**57.25**	**62.06**
BLIP$_{ITC}$ (ViT-L/16)	**33.25**	12.00	9.00	**61.75**	**55.00**	**58.38**
BLIP$_{ITMCOCO}$ (ViT-B/16)	**48.00**	24.50	**20.00**	**69.88**	**61.25**	**65.56**
BLIP$_{ITCCOCO}$ (ViT-B/16)	**37.75**	15.75	12.75	**65.00**	**56.88**	**60.94**
BLIP$_{ITMFlickr30k}$ (ViT-B/16)	**46.25**	24.25	**21.25**	**69.25**	**60.62**	**64.94**
BLIP$_{ITCFlickr30k}$ (ViT-B/16)	**38.25**	15.00	12.25	**65.38**	**56.12**	**60.75**
BLIP$_{ITMCOCO}$ (ViT-L/16)	**46.75**	24.00	**20.50**	**68.88**	**61.00**	**64.94**
BLIP$_{ITCCOCO}$ (ViT-L/16)	**37.75**	13.75	10.50	**64.88**	**55.75**	**60.31**
BLIP$_{ITMFlickr30k}$ (ViT-L/16)	**45.00**	24.75	**20.50**	**68.62**	**60.50**	**64.56**
BLIP$_{ITCFlickr30k}$ (ViT-L/16)	**36.00**	16.25	13.50	**63.38**	**56.75**	**60.06**

Table 5.2: Results on the Winoground dataset across the text, image and group score and accuracy metrics. Results above random chance in **bold**.

Model	Object			Relation			Both			1 Main Pred			2 Main Preds		
	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group
MTurk Human	**92.20**	**90.78**	**88.65**	**89.27**	**90.56**	**86.70**	**76.92**	**57.69**	**57.69**	**87.33**	**85.62**	**82.53**	**95.37**	**96.30**	**93.52**
VinVL	**36.88**	17.73	14.18	**37.77**	17.60	14.16	**42.31**	19.23	**19.23**	**39.38**	21.23	**17.47**	**33.33**	8.33	6.48
UNITER$_{large}$	**39.01**	12.77	9.93	**36.05**	14.16	9.87	**50.00**	19.23	**19.23**	**40.07**	16.44	13.36	**32.41**	7.41	2.78
UNITER$_{base}$	**34.04**	11.35	9.22	**30.04**	14.16	10.30	**42.31**	15.38	11.54	**35.27**	14.73	11.99	24.07	9.26	4.63
ViLLA$_{large}$	**36.88**	14.89	11.35	**37.34**	12.88	11.16	**34.62**	7.69	7.69	**39.73**	17.12	14.38	**29.63**	2.78	1.85
ViLLA$_{base}$	**33.33**	15.60	9.93	**27.04**	9.01	6.01	**38.46**	19.23	15.38	**33.22**	14.04	10.27	21.30	6.48	1.85
VisualBERT$_{base}$	19.15	2.13	0.71	12.88	2.15	1.72	19.23	7.69	3.85	16.44	2.74	1.71	12.96	1.85	0.93
ViLT (ViT-B/32)	**31.91**	15.60	9.22	**36.91**	11.59	8.15	**30.77**	**26.92**	**19.23**	**35.27**	17.12	11.64	**33.33**	5.56	2.78
LXMERT	22.70	9.22	6.38	17.60	5.58	2.58	15.38	7.69	3.85	19.18	8.56	5.14	19.44	2.78	0.93
ViLBERT$_{base}$	**29.08**	10.64	7.09	19.31	3.00	1.72	**34.62**	**26.92**	**19.23**	23.97	8.90	5.82	23.15	2.78	1.85
UniT$_{ITM finetuned}$	17.73	5.67	2.13	18.03	4.72	3.43	**42.31**	23.08	**19.23**	21.58	6.85	4.11	13.89	4.63	3.70
FLAVA$_{ITM}$	**31.91**	23.40	14.89	**30.04**	16.31	12.02	**53.85**	**42.31**	**30.77**	**36.30**	24.66	**17.81**	21.30	9.26	4.63
FLAVA$_{Contrastive}$	23.40	19.15	11.35	23.61	8.58	5.58	**50.00**	**26.92**	**26.92**	**26.37**	16.44	10.62	22.22	5.56	4.63
CLIP (ViT-B/32)	**34.75**	7.80	6.38	22.75	8.58	5.58	**80.77**	**42.31**	**38.46**	**35.27**	13.01	10.27	18.52	3.70	1.85
VSE++$_{COCO}$ (ResNet)	21.99	6.38	1.42	23.61	9.01	5.58	19.23	7.69	3.85	25.00	9.59	4.79	16.67	3.70	1.85
VSE++$_{COCO}$ (VGG)	17.73	2.13	2.13	18.45	7.30	3.86	**26.92**	7.69	7.69	18.49	4.79	2.74	19.44	7.41	5.56
VSE++$_{Flickr30k}$ (ResNet)	20.57	6.38	3.55	18.88	4.29	2.15	**26.92**	3.85	3.85	21.58	6.51	3.42	15.74	0.93	0.93
VSE++$_{Flickr30k}$ (VGG)	17.73	4.96	2.84	19.74	6.87	5.15	**30.77**	7.69	7.69	20.55	6.16	4.79	17.59	6.48	3.70
VSRN$_{COCO}$	15.60	4.96	2.13	18.88	7.73	4.72	15.38	11.54	3.85	17.12	7.19	3.77	18.52	6.48	3.70
VSRN$_{Flickr30k}$	16.31	4.96	2.13	21.03	4.29	3.86	**30.77**	11.54	7.69	20.89	5.82	3.77	17.59	2.78	2.78

Table 5.3: The results by linguistic tag. Results above chance are in **bold**.

Model	Object			Relation			Both			1 Main Pred			2 Main Preds		
	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group
MTurk Human	**92.20**	**90.78**	**88.65**	**89.27**	**90.56**	**86.70**	**76.92**	**57.69**	**57.69**	**87.33**	**85.62**	**82.53**	**95.37**	**96.30**	**93.52**
ViLT (ViT-B/32)	**29.08**	10.64	4.96	**26.18**	7.73	6.44	**30.77**	7.69	7.69	**30.14**	10.62	7.53	20.37	3.70	1.85
ViLT$_{COCO}$ (ViT-B/32)	**33.33**	15.60	12.77	**30.90**	10.73	9.01	**46.15**	**26.92**	**23.08**	**36.64**	15.75	14.04	22.22	7.41	3.70
ViLT$_{Flickr30k}$ (ViT-B/32)	**32.62**	14.89	11.35	**35.62**	8.15	7.73	**42.31**	23.08	**19.23**	**36.99**	14.38	11.99	**29.63**	3.70	3.70
ViLT$_{NLVR2}$ (ViT-B/32)	**39.01**	16.31	14.18	**36.48**	14.59	10.30	**46.15**	15.38	15.38	**39.73**	18.15	15.07	**33.33**	7.41	3.70
FLAVA$_{ITM}$	**31.91**	23.40	14.89	**30.04**	16.31	12.02	**53.85**	**42.31**	**30.77**	**36.30**	24.66	**17.81**	21.30	9.26	4.63
FLAVA$_{ITC}$	23.40	19.15	11.35	23.61	8.58	5.58	**50.00**	**26.92**	**26.92**	**26.37**	16.44	10.62	22.22	5.56	4.63
CLIP (ViT-B/32)	**35.46**	7.80	6.38	22.32	7.73	5.58	**80.77**	**46.15**	**42.31**	**35.62**	13.01	10.62	17.59	2.78	1.85
CLIP (ViT-B/16)	**27.66**	10.64	5.67	19.31	6.44	4.29	**61.54**	**42.31**	**38.46**	**30.14**	11.99	8.90	11.11	5.56	1.85
CLIP (ViT-L/14)	**27.66**	8.51	5.67	**25.75**	9.87	6.44	**57.69**	**34.62**	**34.62**	**30.14**	13.01	9.93	24.07	5.56	2.78
CLIP (ViT-L/14-336)	**32.62**	12.77	9.22	21.03	8.15	4.29	**57.69**	**42.31**	**34.62**	**30.48**	14.04	10.62	19.44	6.48	0.93
OFA$_{Tiny}$	22.70	6.38	2.13	17.17	6.87	3.43	**38.46**	**26.92**	15.38	23.97	8.22	4.45	11.11	7.41	1.85
OFA$_{Base}$	**25.53**	14.18	7.09	24.46	6.87	5.15	**50.00**	23.08	**23.08**	**28.77**	12.67	8.56	20.37	4.63	2.78
OFA$_{Medium}$	19.86	7.80	4.26	22.32	7.73	4.72	**42.31**	**26.92**	**19.23**	24.32	10.96	6.85	18.52	3.70	1.85
OFA$_{Large}$	**26.24**	10.64	5.67	24.03	5.15	3.86	**42.31**	**30.77**	**23.08**	**29.45**	10.96	7.53	16.67	2.78	0.93
OFA$_{Huge}$	**40.43**	18.44	15.60	**30.90**	11.59	9.87	**61.54**	**34.62**	**34.62**	**39.73**	19.18	**16.78**	**26.85**	5.56	4.63
BLIP$_{ITM14M}$ (ViT-B/16)	**41.84**	23.40	**17.73**	**36.05**	14.59	11.59	**53.85**	**34.62**	**30.77**	**43.84**	23.63	**18.49**	**26.85**	6.48	5.56
BLIP$_{ITC14M}$ (ViT-B/16)	**34.04**	13.48	9.93	**28.33**	12.02	9.44	**57.69**	**30.77**	**23.08**	**37.67**	16.44	13.01	17.59	6.48	3.70
BLIP$_{ITM}$ (ViT-B/16)	**46.10**	22.70	**17.73**	**35.62**	17.60	14.16	**53.85**	**34.62**	**30.77**	**45.89**	**25.34**	**20.55**	**25.93**	7.41	5.56
BLIP$_{ITC}$ (ViT-B/16)	**34.75**	14.18	9.22	**25.32**	13.73	8.58	**42.31**	23.08	**19.23**	**33.56**	16.10	10.62	19.44	10.19	6.48
BLIP$_{ITM}$ (ViT-B/16) (CapFilt-L)	**39.01**	19.86	12.77	**34.76**	15.88	12.45	**53.85**	**34.62**	**34.62**	**41.10**	22.60	**17.12**	**27.78**	7.41	5.56
BLIP$_{ITC}$ (ViT-B/16) (CapFilt-L)	**36.88**	12.77	9.22	**26.18**	8.58	7.30	**50.00**	15.38	15.38	**35.96**	13.36	10.96	19.44	2.78	1.85
BLIP$_{ITM}$ (ViT-L/16)	**41.84**	19.86	**17.02**	**40.77**	16.31	13.73	**61.54**	**26.92**	**23.08**	**45.55**	23.29	**20.21**	**34.26**	4.63	2.78
BLIP$_{ITC}$ (ViT-L/16)	**34.04**	14.18	11.35	**30.90**	9.01	6.01	**50.00**	**26.92**	**23.08**	**36.99**	14.04	10.96	23.15	6.48	3.70
BLIP$_{ITMCOCO}$ (ViT-B/16)	**42.55**	**26.95**	**19.15**	**49.79**	21.89	**19.31**	**61.54**	**34.62**	**30.77**	**48.97**	**29.79**	**24.66**	**45.37**	10.19	7.41
BLIP$_{ITCCOCO}$ (ViT-B/16)	**36.88**	19.15	14.18	**36.05**	11.59	10.30	**57.69**	**34.62**	**26.92**	**41.78**	18.84	15.07	**26.85**	7.41	6.48
BLIP$_{ITMFlickr30k}$ (ViT-B/16)	**49.65**	**28.37**	**22.70**	**42.49**	19.74	**18.45**	**61.54**	**42.31**	**38.46**	**51.03**	**28.42**	**26.03**	**33.33**	12.96	8.33
BLIP$_{ITCFlickr30k}$ (ViT-B/16)	**36.88**	17.02	10.64	**36.48**	12.02	11.16	**61.54**	**30.77**	**30.77**	**40.75**	17.12	13.70	**31.48**	9.26	8.33
BLIP$_{ITMCOCO}$ (ViT-L/16)	**48.94**	**25.53**	**20.57**	**44.64**	22.32	**20.60**	**53.85**	**30.77**	**19.23**	**51.03**	**28.42**	**23.97**	**35.19**	12.04	11.11
BLIP$_{ITCCOCO}$ (ViT-L/16)	**36.88**	14.18	11.35	**36.05**	11.16	7.30	**57.69**	**34.62**	**34.62**	**41.10**	16.44	13.36	**28.70**	6.48	2.78
BLIP$_{ITMFlickr30k}$ (ViT-L/16)	**46.10**	22.70	16.31	**42.06**	24.89	**21.46**	**65.38**	**34.62**	**34.62**	**50.34**	**29.11**	**24.66**	**30.56**	12.96	9.26
BLIP$_{ITCFlickr30k}$ (ViT-L/16)	**39.01**	19.86	15.60	**30.47**	11.59	9.44	**69.23**	**38.46**	**38.46**	**39.38**	20.55	**17.12**	**26.85**	4.63	3.70

Table 5.4: The results by linguistic tag. Results above chance are in **bold**.

Model	Symbolic			Pragmatics			Same Image Series		
	Text	Image	Group	Text	Image	Group	Text	Image	Group
MTurk Human	**96.43**	**92.86**	**92.86**	**58.82**	**41.18**	**41.18**	**95.65**	**91.30**	**91.30**
VinVL	25.00	17.86	14.29	**29.41**	5.88	5.88	**34.78**	17.39	13.04
UNITER$_{large}$	**39.29**	**28.57**	**17.86**	**35.29**	0.00	0.00	4.35	8.70	0.00
UNITER$_{base}$	**46.43**	14.29	14.29	**29.41**	17.65	11.76	8.70	8.70	0.00
ViLLA$_{large}$	**39.29**	14.29	10.71	17.65	0.00	0.00	17.39	4.35	0.00
ViLLA$_{base}$	**42.86**	17.86	14.29	**29.41**	5.88	5.88	13.04	8.70	4.35
VisualBERT$_{base}$	**28.57**	0.00	0.00	5.88	0.00	0.00	13.04	0.00	0.00
ViLT (ViT-B/32)	**28.57**	17.86	10.71	**35.29**	0.00	0.00	**26.09**	0.00	0.00
LXMERT	**28.57**	3.57	3.57	17.65	5.88	0.00	8.70	4.35	0.00
ViLBERT$_{base}$	**28.57**	10.71	7.14	**29.41**	5.88	5.88	13.04	0.00	0.00
UniT$_{ITM finetuned}$	14.29	10.71	7.14	17.65	5.88	5.88	21.74	4.35	4.35
FLAVA$_{ITM}$	25.00	**28.57**	**17.86**	17.65	**29.41**	11.76	17.39	8.70	0.00
FLAVA$_{Contrastive}$	17.86	10.71	10.71	11.76	23.53	5.88	17.39	4.35	4.35
CLIP (ViT-B/32)	**39.29**	3.57	3.57	**35.29**	5.88	5.88	8.70	0.00	0.00
VSE++$_{COCO}$ (ResNet)	**32.14**	10.71	10.71	23.53	11.76	0.00	13.04	4.35	4.35
VSE++$_{COCO}$ (VGG)	17.86	14.29	7.14	17.65	0.00	0.00	13.04	4.35	4.35
VSE++$_{Flickr30k}$ (ResNet)	21.43	3.57	0.00	23.53	0.00	0.00	17.39	4.35	0.00
VSE++$_{Flickr30k}$ (VGG)	**28.57**	10.71	10.71	11.76	0.00	0.00	13.04	4.35	0.00
VSRN$_{COCO}$	7.14	3.57	0.00	11.76	0.00	0.00	13.04	0.00	0.00
VSRN$_{Flickr30k}$	21.43	3.57	3.57	**35.29**	11.76	5.88	8.70	4.35	4.35

Table 5.5: The results by visual tag. Results above chance are in **bold**.

Model	Symbolic			Pragmatics			Same Image Series		
	Text	Image	Group	Text	Image	Group	Text	Image	Group
MTurk Human	**96.43**	**92.86**	**92.86**	**58.82**	**41.18**	**41.18**	**95.65**	**91.30**	**91.30**
ViLT (ViT-B/32)	21.43	7.14	3.57	17.65	5.88	5.88	17.39	8.70	4.35
ViLT$_{COCO}$ (ViT-B/32)	21.43	10.71	10.71	**29.41**	17.65	5.88	21.74	8.70	4.35
ViLT$_{Flickr30k}$ (ViT-B/32)	**28.57**	7.14	7.14	23.53	0.00	0.00	**26.09**	4.35	4.35
ViLT$_{NLVR2}$ (ViT-B/32)	**42.86**	10.71	10.71	**41.18**	0.00	0.00	17.39	13.04	4.35
FLAVA$_{ITM}$	25.00	**28.57**	**17.86**	17.65	**29.41**	11.76	17.39	8.70	0.00
FLAVA$_{ITC}$	17.86	10.71	10.71	11.76	23.53	5.88	17.39	4.35	4.35
CLIP (ViT-B/32)	**35.71**	3.57	3.57	**35.29**	5.88	5.88	13.04	0.00	0.00
CLIP (ViT-B/16)	21.43	3.57	3.57	**29.41**	11.76	11.76	4.35	4.35	0.00
CLIP (ViT-L/14)	**28.57**	10.71	3.57	23.53	17.65	11.76	13.04	8.70	4.35
CLIP (ViT-L/14-336)	**28.57**	14.29	7.14	17.65	17.65	5.88	13.04	4.35	0.00
OFA$_{Tiny}$	21.43	7.14	7.14	11.76	17.65	0.00	21.74	8.70	0.00
OFA$_{Base}$	**28.57**	10.71	10.71	23.53	5.88	5.88	21.74	13.04	4.35
OFA$_{Medium}$	**28.57**	10.71	7.14	17.65	5.88	5.88	13.04	8.70	4.35
OFA$_{Large}$	**28.57**	14.29	10.71	**29.41**	0.00	0.00	13.04	0.00	0.00
OFA$_{Huge}$	**39.29**	14.29	14.29	11.76	11.76	5.88	17.39	4.35	4.35
BLIP$_{ITM14M}$ (ViT-B/16)	**46.43**	17.86	**17.86**	**35.29**	11.76	11.76	17.39	4.35	0.00
BLIP$_{ITC14M}$ (ViT-B/16)	**32.14**	14.29	10.71	**29.41**	0.00	0.00	13.04	0.00	0.00
BLIP$_{ITM}$ (ViT-B/16)	**50.00**	17.86	**17.86**	**29.41**	5.88	5.88	13.04	4.35	0.00
BLIP$_{ITC}$ (ViT-B/16)	**39.29**	10.71	7.14	5.88	11.76	0.00	4.35	8.70	0.00
BLIP$_{ITM}$ (ViT-B/16) (CapFilt-L)	**42.86**	17.86	14.29	23.53	17.65	**17.65**	17.39	4.35	0.00
BLIP$_{ITC}$ (ViT-B/16) (CapFilt-L)	**42.86**	0.00	0.00	17.65	0.00	0.00	4.35	0.00	0.00
BLIP$_{ITM}$ (ViT-L/16)	**53.57**	25.00	**25.00**	**29.41**	5.88	0.00	**26.09**	4.35	0.00
BLIP$_{ITC}$ (ViT-L/16)	**39.29**	17.86	14.29	**41.18**	11.76	11.76	8.70	4.35	4.35
BLIP$_{ITMCOCO}$ (ViT-B/16)	**53.57**	17.86	**17.86**	**58.82**	17.65	**17.65**	**39.13**	8.70	0.00
BLIP$_{ITCCOCO}$ (ViT-B/16)	25.00	10.71	7.14	**35.29**	5.88	5.88	17.39	8.70	4.35
BLIP$_{ITMFlickr30k}$ (ViT-B/16)	**53.57**	21.43	**21.43**	**35.29**	11.76	11.76	**26.09**	4.35	4.35
BLIP$_{ITCFlickr30k}$ (ViT-B/16)	**35.71**	10.71	10.71	23.53	17.65	11.76	17.39	4.35	0.00
BLIP$_{ITMCOCO}$ (ViT-L/16)	**39.29**	**35.71**	**25.00**	**58.82**	23.53	**17.65**	**26.09**	4.35	0.00
BLIP$_{ITCCOCO}$ (ViT-L/16)	**46.43**	14.29	14.29	17.65	5.88	5.88	13.04	0.00	0.00
BLIP$_{ITMFlickr30k}$ (ViT-L/16)	**39.29**	**28.57**	**25.00**	**47.06**	11.76	5.88	**30.43**	8.70	4.35
BLIP$_{ITCFlickr30k}$ (ViT-L/16)	**39.29**	14.29	14.29	**47.06**	5.88	5.88	21.74	13.04	13.04

Table 5.6: The results by visual tag. Results above chance are in **bold**.

6 Discussion

6.1 Winoground

6.1.1 Capabilities of Encoders

6.1.1.1 Baseline

See Table 6.1

	Perplexity		Caption Length					
	Text-Image		Text		Image		Group	
Model	Corr.	p-value	Corr.	p-value	Corr.	p-value	Corr.	p-value
MTurk Human	0.05	0.07	**0.11**	**0.03**	**0.20**	**0.00**	**0.20**	**0.00**
VinVL	**-0.05**	**0.04**	**-0.11**	**0.03**	**-0.18**	**0.00**	**-0.20**	**0.00**
UNITER$_{large}$	-0.01	0.57	-0.08	0.13	-0.06	0.20	**-0.16**	**0.00**
UNITER$_{base}$	-0.03	0.22	**-0.15**	**0.00**	**-0.11**	**0.03**	**-0.14**	**0.00**
ViLLA$_{large}$	-0.02	0.39	-0.05	0.32	**-0.13**	**0.01**	**-0.12**	**0.01**
ViLLA$_{base}$	-0.04	0.13	**-0.14**	**0.01**	**-0.12**	**0.01**	**-0.11**	**0.03**
VisualBERT$_{base}$	-0.04	0.15	-0.09	0.07	-0.07	0.14	-0.06	0.22
ViLT (ViT-B/32)	-0.04	0.16	-0.09	0.06	**-0.20**	**0.00**	**-0.16**	**0.00**
LXMERT	-0.04	0.12	-0.00	0.97	-0.05	0.32	**-0.11**	**0.02**
ViLBERT$_{base}$	-0.04	0.11	-0.09	0.09	**-0.15**	**0.00**	**-0.14**	**0.00**
UniT$_{ITM Finetuned}$	-0.01	0.73	-0.03	0.53	-0.05	0.32	-0.02	0.73
FLAVA$_{ITM}$	-0.03	0.22	**-0.21**	**0.00**	**-0.22**	**0.00**	**-0.23**	**0.00**
FLAVA$_{Contrastive}$	**-0.06**	**0.01**	**-0.15**	**0.00**	**-0.25**	**0.00**	**-0.19**	**0.00**
CLIP (ViT-B/32)	-0.04	0.09	**-0.27**	**0.00**	**-0.19**	**0.00**	**-0.22**	**0.00**
VSE++$_{COCO}$ (ResNet)	**-0.05**	**0.04**	-0.03	0.60	-0.02	0.74	0.01	0.90
VSE++$_{COCO}$ (VGG)	-0.04	0.08	-0.02	0.65	0.03	0.50	0.03	0.56
VSE++$_{Flickr30k}$ (ResNet)	-0.02	0.43	-0.01	0.80	0.01	0.91	0.02	0.67
VSE++$_{Flickr30k}$ (VGG)	0.01	0.74	-0.09	0.07	-0.07	0.18	**-0.10**	**0.04**
VSRN$_{COCO}$	**-0.07**	**0.01**	-0.03	0.60	-0.05	0.30	-0.05	0.36
VSRN$_{Flickr30k}$	-0.02	0.32	-0.03	0.60	-0.10	0.06	-0.05	0.29

Table 6.1: (left) The correlation between model image-caption scores and the caption perplexity from GPT2. (right) The correlation between the model text, image and group scores and the caption length.

6.1.1.2 Ours

See Table 6.2

6.1.2 By Multimodal Pretraining Dataset Size

6.1.2.1 Baseline

See Table 6.3 and Fig. 6.1

6.1.2.2 Ours

See Table 6.4 and Fig. 6.2

Model	Perplexity Image-Caption Corr.	Perplexity Image-Caption p-value	Caption Length Text Corr.	Text p-value	Image Corr.	Image p-value	Group Corr.	Group p-value
MTurk Human	0.05	0.07	**0.11**	**0.03**	**0.20**	**0.00**	**0.20**	**0.00**
ViLT (ViT-B/32)	-0.04	0.08	**-0.12**	**0.02**	-0.07	0.17	-0.05	0.35
ViLT$_{COCO}$ (ViT-B/32)	-0.05	0.06	**-0.21**	**0.00**	**-0.16**	**0.00**	**-0.17**	**0.00**
ViLT$_{Flickr30k}$ (ViT-B/32)	**-0.05**	**0.03**	**-0.11**	**0.03**	**-0.17**	**0.00**	**-0.14**	**0.01**
ViLT$_{NLVR2}$ (ViT-B/32)	0.00	0.95	**-0.13**	**0.01**	**-0.11**	**0.03**	**-0.12**	**0.02**
FLAVA$_{ITM}$	-0.03	0.22	**-0.21**	**0.00**	**-0.22**	**0.00**	**-0.23**	**0.00**
FLAVA$_{ITC}$	**-0.06**	**0.01**	**-0.15**	**0.00**	**-0.25**	**0.00**	**-0.19**	**0.00**
CLIP (ViT-B/32)	-0.04	0.10	**-0.28**	**0.00**	**-0.21**	**0.00**	**-0.23**	**0.00**
CLIP (ViT-B/16)	-0.04	0.11	**-0.26**	**0.00**	**-0.22**	**0.00**	**-0.23**	**0.00**
CLIP (ViT-L/14)	-0.03	0.22	**-0.22**	**0.00**	**-0.17**	**0.00**	**-0.18**	**0.00**
CLIP (ViT-L/14-336)	-0.04	0.11	**-0.23**	**0.00**	**-0.22**	**0.00**	**-0.23**	**0.00**
OFA$_{Tiny}$	-0.01	0.66	**-0.17**	**0.00**	-0.06	0.24	**-0.12**	**0.02**
OFA$_{Base}$	-0.02	0.43	**-0.15**	**0.00**	**-0.12**	**0.02**	**-0.10**	**0.05**
OFA$_{Medium}$	-0.01	0.77	**-0.11**	**0.03**	**-0.14**	**0.00**	**-0.12**	**0.01**
OFA$_{Large}$	**-0.16**	**0.00**	**-0.18**	**0.00**	**-0.20**	**0.00**	**-0.17**	**0.00**
OFA$_{Huge}$	0.01	0.75	**-0.15**	**0.00**	**-0.17**	**0.00**	**-0.16**	**0.00**
BLIP$_{ITM14M}$ (ViT-B/16)	-0.00	0.85	**-0.22**	**0.00**	**-0.23**	**0.00**	**-0.21**	**0.00**
BLIP$_{ITC14M}$ (ViT-B/16)	-0.00	0.97	**-0.24**	**0.00**	**-0.17**	**0.00**	**-0.17**	**0.00**
BLIP$_{ITM}$ (ViT-B/16)	**-0.05**	**0.04**	**-0.24**	**0.00**	**-0.23**	**0.00**	**-0.22**	**0.00**
BLIP$_{ITC}$ (ViT-B/16)	**-0.06**	**0.02**	**-0.19**	**0.00**	**-0.17**	**0.00**	**-0.13**	**0.01**
BLIP$_{ITM}$ (ViT-B/16) (CapFilt-L)	**-0.10**	**0.00**	**-0.20**	**0.00**	**-0.28**	**0.00**	**-0.23**	**0.00**
BLIP$_{ITC}$ (ViT-B/16) (CapFilt-L)	**-0.10**	**0.00**	**-0.25**	**0.00**	**-0.17**	**0.00**	**-0.15**	**0.00**
BLIP$_{ITM}$ (ViT-L/16)	**-0.07**	**0.01**	**-0.17**	**0.00**	**-0.21**	**0.00**	**-0.19**	**0.00**
BLIP$_{ITC}$ (ViT-L/16)	**-0.08**	**0.00**	**-0.22**	**0.00**	**-0.17**	**0.00**	**-0.17**	**0.00**
BLIP$_{ITMCOCO}$ (ViT-B/16)	-0.04	0.11	**-0.17**	**0.00**	**-0.26**	**0.00**	**-0.22**	**0.00**
BLIP$_{ITCCOCO}$ (ViT-B/16)	**-0.06**	**0.02**	**-0.18**	**0.00**	**-0.26**	**0.00**	**-0.22**	**0.00**
BLIP$_{ITMFlickr30k}$ (ViT-B/16)	-0.04	0.11	**-0.25**	**0.00**	**-0.28**	**0.00**	**-0.28**	**0.00**
BLIP$_{ITCFlickr30k}$ (ViT-B/16)	**-0.07**	**0.00**	**-0.20**	**0.00**	**-0.19**	**0.00**	**-0.18**	**0.00**
BLIP$_{ITMCOCO}$ (ViT-L/16)	**-0.06**	**0.02**	**-0.24**	**0.00**	**-0.23**	**0.00**	**-0.23**	**0.00**
BLIP$_{ITCCOCO}$ (ViT-L/16)	**-0.10**	**0.00**	**-0.21**	**0.00**	**-0.21**	**0.00**	**-0.21**	**0.00**
BLIP$_{ITMFlickr30k}$ (ViT-L/16)	**-0.05**	**0.04**	**-0.27**	**0.00**	**-0.25**	**0.00**	**-0.23**	**0.00**
BLIP$_{ITCFlickr30k}$ (ViT-L/16)	**-0.09**	**0.00**	**-0.24**	**0.00**	**-0.19**	**0.00**	**-0.16**	**0.00**

Table 6.2: (left) The correlation between model image-caption scores and the caption perplexity from GPT2. (right) The correlation between the model text, image and group scores and the caption length.

Pretraining	Score	Corr.	p-value
Image	Text	**0.84**	**0.00**
	Image	**0.76**	**0.00**
	Group	**0.75**	**0.00**
Caption	Text	**0.77**	**0.00**
	Image	**0.75**	**0.00**
	Group	**0.71**	**0.00**

Table 6.3: Correlations between the number of pretraining images and captions and the model text, image, and group scores. CLIP and FLAVA are excluded as outliers.

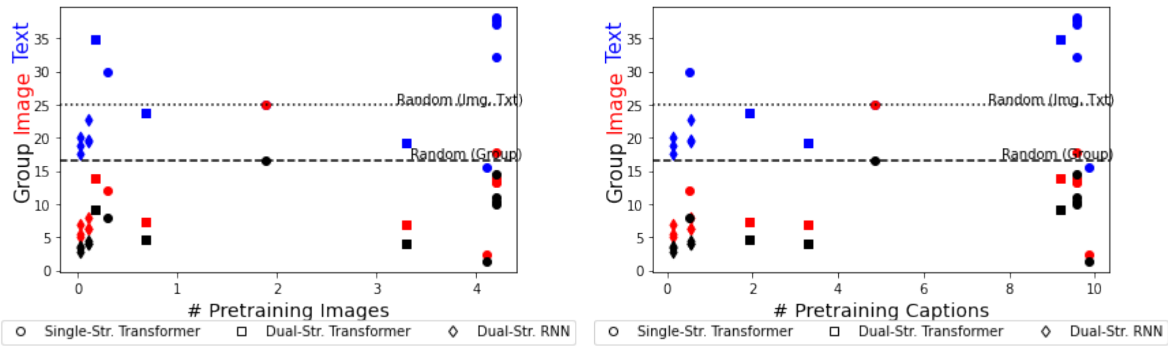

Figure 6.1: Graphs of the model performance on Winoground for each model by the number of pretraining images (left) and pretraining captions (right).

Pretraining	Score	Corr.	p-value
Image	Text	-0.09	0.65
	Image	-0.16	0.44
	Group	-0.13	0.53
Caption	Text	-0.09	0.66
	Image	-0.15	0.44
	Group	-0.12	0.54

Table 6.4: Correlations between the number of pretraining images and captions and the model text, image, and group scores. Only BLIP, CLIP and FLAVA are included.

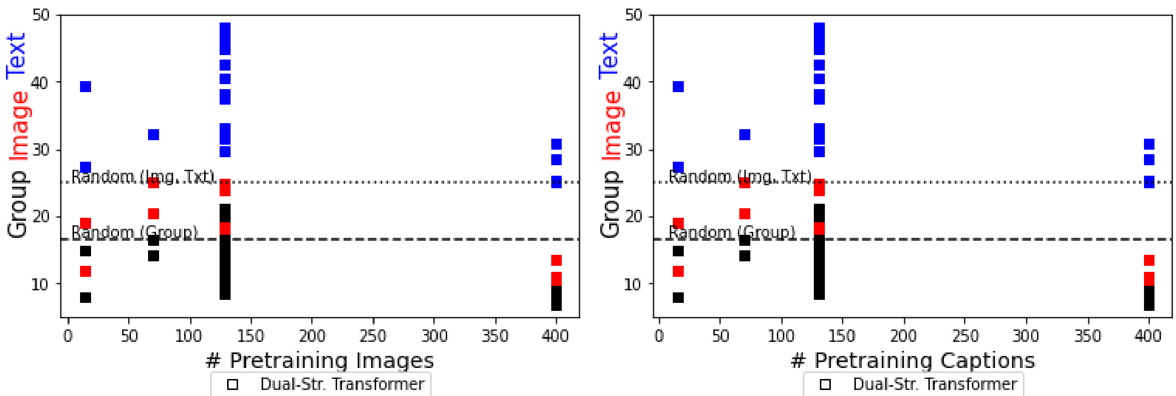

Figure 6.2: Graphs of the model performance on Winoground for each model by the number of pretraining images (left) and pretraining captions (right).

7 Conclusions

Appendix

Bibliography

[1] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019. See page 1.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. See page 1.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. See page 1.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. See page 1.

[5] 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. See page 1.

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. See page 1.

[7] Stephen C Levinson and Stephen C Levinson. *Space in language and cognition: Explorations in cognitive diversity*. Number 5. Cambridge University Press, 2003. See page 1.

[8] Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *ACL*, 2020. See page 1.

[9] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 2019. See pages 1, 9, and 11.

[10] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2020. See pages 1, 9, and 11.

[11] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. See page 1.

[12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. See page 1.

[13] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. See page 1.

[14] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. See page 1.

[15] Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*, 2022. See pages 1, 2.

[16] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022. See page 1.

[17] 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. See page 1.

[18] Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. How large are lions? inducing distributions over quantitative attributes. *arXiv preprint arXiv:1906.01327*, 2019. See page 1.

[19] 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. See page 1.

[20] 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. See page 1.

[21] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15513–15523, 2022. See page 1.

[22] 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 *CVPR*, 2017. See page 2.

[23] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. See pages 2, 3, and 7.

[24] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022. See pages 2, 3.

[25] 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 *ECCV*, 2014. See pages 2, 9.

[26] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. See pages 9, 11.

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. See pages 9, 11.

[28] 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 *ICML*, 2021. See pages 9, 11.

[29] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. See pages 9, 11.

[30] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *arXiv preprint arXiv:2102.10772*, 2021. See pages 9, 11.

[31] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. See pages 9, 11.

[32] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. See pages 9, 11.

[33] 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 *CVPR*, 2021. See pages 9, 11.

[34] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. See pages 9, 11.

[35] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *arXiv preprint arXiv:1908.03557*, 2019. See pages 9, 11.

[36] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. See pages 9, 11.

[37] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019. See pages 9, 11.

[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv preprint arXiv:1602.07332*, 2016. See page 9.

[39] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. See page 9.

[40] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. See page 9.

[41] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, 2014. See page 9.

[42] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. See page 9.

[43] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. See page 9.

[44] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017. See page 9.

[45] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. In *arXiv preprint arXiv:1811.10582*, 2018. See page 9.

[46] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *arXiv preprint arXiv:1606.05250*, 2016. See page 9.

[47] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *arXiv preprint arXiv:1704.05426*, 2017. See page 9.

[48] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. First quora dataset release: Question pairs, 2017. See page 9.

[49] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. See page 9.

[50] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *arXiv preprint arXiv:2103.01913*, 2021. See page 9.

[51] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. See page 9.

[52] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people. In *NeurIPS Datasets and Benchmarks*, 2021. See page 9.

[53] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. In *Communications of the ACM*, 2016. See page 9.

[54] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013. See page 9.

[55] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. See page 9.