Visual Spatial Reasoning

Fangyu Liu Guy Emerson Nigel Collier
University of Cambridge, UK
{fl399, gete2, nhc30}@cam.ac.uk

Abstract

Spatial relations are a basic part of human cognition. However, they are expressed in natural language in a variety of ways, and previous work has suggested that current vision-and-language models (VLMs) struggle to capture relational information. In this paper, we present Visual Spatial Reasoning (VSR), a dataset containing more than 10k natural text-image pairs with 66 types of spatial relations in English (e.g., under, in front of, facing). While using a seemingly simple annotation format, we show how the dataset includes challenging linguistic phenomena, such as varying reference frames. We demonstrate a large gap between human and model performance: The human ceiling is above 95%, while state-of-the-art models only achieve around 70%. We observe that VLMs' by-relation performances have little correlation with the number of training examples and the tested models are in general incapable of recognising relations concerning the orientations of objects.[1]

1 Introduction

Multimodal NLP research has developed rapidly in recent years, with substantial performance gains on tasks such as visual question answering (VQA) (Antol et al., 2015; Johnson et al., 2017; Goyal et al., 2017; Hudson and Manning, 2019; Zellers et al., 2019), vision-language reasoning or entailment (Suhr et al., 2017, 2019; Xie et al., 2019; Liu et al., 2021), and referring expression comprehension (Yu et al., 2016; Liu et al., 2019). Existing benchmarks, such as NLVR2 (Suhr et al., 2019) and VQA (Goyal et al., 2017), define generic paradigms for testing vision-language models (VLMs). However, as we further discuss in § 2, these benchmarks are not ideal for probing VLMs as they typically conflate multiple sources of error and do not allow controlled analysis on specific linguistic or cognitive properties, making it difficult to categorize and fully understand the model failures. In particular, spatial reasoning has been found to be particularly challenging for current models, and much more challenging than capturing properties of individual entities (Kuhnle et al., 2018; Cirik et al., 2018; Akula et al., 2020), even for state-of-the-art models such as CLIP (Radford et al., 2021; Subramanian et al., 2022).

Another line of work generates synthetic datasets in a controlled manner to target specific relations and properties when testing VLMs, e.g., CLEVR (Liu et al., 2019) and ShapeWorld (Kuhnle and Copestake, 2018). However, synthetic datasets may accidentally overlook challenges (such as orientations of objects which we will discuss in § 5), and using natural images allows us to explore a wider range of language use.

To address the lack of probing evaluation benchmarks in this field, we present VSR (Visual Spatial Reasoning), a controlled dataset that explicitly tests VLMs for spatial reasoning. We choose spatial reasoning as the focus because it is one of the most fundamental capabilities for both humans and VLMs. Such relations are crucial to how humans organize their mental space and make sense of the physical world, and therefore fundamental for a grounded semantic model (Talmy, 1983).

The VSR dataset contains natural image-text pairs in English, with the data collection process explained in § 3. Each example in the dataset consists of an image and a natural language description that states a spatial relation of two objects presented in the image (two examples are shown in Figure 1 and Figure 2). A VLM needs to classify the image-caption pair as either true or false, indicating whether the caption is correctly describing the spatial relation. The dataset covers 66 spatial relations and has >10k data points, using 6,940 images from MS COCO (Lin et al., 2014).

Situating one object in relation to another requires a *frame of reference*: a system of coordinates against which the objects can be placed.

[1]Data and code: github.com/cambridgeltl/visual-spatial-reasoning.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 635–651, 2023. https://doi.org/10.1162/tacl_a_00566
Action Editor: Mohit Bansal. Submission batch: 8/2022; Revision batch: 12/2022; Published 6/2023.

Figure 1: Caption: *The potted plant is at the right side of the bench.* Label: `True`. Image source: Antoine K. ‘‘Texting’’, uploaded November 5, 2010. `https://www.flickr.com/photos/ktoine/5149301465/` (CC BY-SA 2.0).

Figure 2: Caption: *The cow is ahead of the person.* Label: `False`. Image source: ccarlstead. ‘‘Holy cow’’, uploaded March 24, 2023. `https://www.flickr.com/photos/cristic/6863977248/` (CC BY-NC-ND 2.0).

Drawing on detailed studies of more than forty typologically diverse languages, Levinson (2003) concludes that the diversity can be reduced to three major types: intrinsic, relative, and absolute. An intrinsic frame is centered on an object, e.g., *behind the chair*, meaning at the side with the backrest. A relative frame is centered on a viewer, e.g., *behind the chair*, meaning further away from someone’s perspective. An absolute frame uses fixed coordinates, e.g., *north of the chair*, using cardinal directions. In English, absolute frames are rarely used when describing relations on a small scale, and they do not appear in our dataset. However, intrinsic and relative frames are widely used, and present an important source of variation. We discuss the impact on data collection in § 3.2, and analyse the collected data in § 4.

We test four popular VLMs, i.e., VisualBERT (Li et al., 2019), LXMERT (Tan and Bansal, 2019), ViLT (Kim et al., 2021), and CLIP (Radford et al., 2021) on VSR, with results given in § 5. While the human ceiling is above 95%, all four models struggle to reach 70% accuracy. We conduct comprehensive analysis on the failures of the investigated VLMs and highlight that (1) positional encodings are extremely important for the VSR task; (2) models’ by-relation performance barely correlates with the number of training examples; (3) in fact, several spatial relations that concern orientation of objects are especially challenging for current VLMs; and (4) VLMs have extremely poor generalization on unseen concepts.

2 Related Work

2.1 Comparison with Synthetic Datasets

Synthetic language-vision reasoning datasets, e.g., SHAPES (Andreas et al., 2016), CLEVR (Liu et al., 2019), NLVR (Suhr et al., 2017), and ShapeWorld (Kuhnle and Copestake, 2018), enable full control of dataset generation and could potentially benefit probing of spatial reasoning capability of VLMs. They share a similar goal to us, to diagnose and pinpoint weaknesses in VLMs. However, synthetic datasets necessarily simplify the problem as they have inherently bounded expressivity. In CLEVR, objects can only be spatially related via four relationships: ‘‘left’’, ‘‘right’’, ‘‘behind’’, and ‘‘in front of’’, while VSR covers 66 relations.

Synthetic data does not always accurately reflect the challenges of reasoning in the real world. For example, objects like spheres, which often appear in synthetic datasets, do not have orientations. In real images, orientations matter and human language use depends on that. Furthermore, synthetic images do not take the scene as a context into account. The interpretation of object relations can depend on such scenes (e.g., the degree of *closeness* can vary in open space and indoor scenes).

Last but not least, the vast majority of spatial relationships cannot be determined by rules. Even for the seemingly simple relationships like ‘‘left/right of’’, the determination of two objects’

spatial relationships can depend on the observer's viewpoint, whether the object has a *front*, if so, what are their orientations, etc.

2.2 Spatial Relations in Existing Vision-language Datasets

Several existing vision-language datasets with natural images also contain spatial relations (e.g., NLVR2, COCO, and VQA datasets). Suhr et al. (2019) summarize that there are 9 prevalent linguistic phenomena/challenges in NLVR2 (Suhr et al., 2019) such as coreference, existential quantifiers, hard cardinality, spatial relations, etc., and 4 in VQA datasets (Antol et al., 2015; Hudson and Manning, 2019). However, the different challenges are entangled in these datasets. Sentences contain complex lexical and syntactic information and can thus conflate different sources of error, making it hard to identify the exact challenge and preventing categorised analysis. Yatskar et al. (2016) extract 6 types of visual spatial relations directly from MS COCO images with annotated bounding boxes. But rule-based automatic extraction can be restrictive as most relations are complex and cannot be identified relying on bounding boxes. Recently, Rösch and Libovický (2022) extract captions that contain 28 positional keywords from MS COCO and swap the keywords with their antonyms to construct a challenging probing dataset. However, the COCO captions also have the error-conflation problem. Also, the number of examples and types of relations are restricted by COCO captions.

Visual Genome (Krishna et al., 2017) also contains annotations of objects' relations including spatial relations. However, it is only a collection of true statements and contains no negative ones, so cannot be framed as a binary classification task. It is non-trivial to automatically construct negative examples since multiple relations can be plausible for a pair of object in a given image. Relation classifiers are harder to learn than object classifiers on this dataset (Liu and Emerson, 2022).

Parcalabescu et al. (2022) propose a benchmark called VALSE for testing VLMs' capabilities on various linguistic phenomena. VALSE has a subset focusing on ''relations'' between objects. It uses texts modified from COCO's original captions. However, it is a zero-shot benchmark without training set, containing just 535 data points. So, it is not ideal for large-scale probing on a wide spectrum of spatial relations.

2.3 Spatial Reasoning Without Grounding

There has also been interest in probing models' spatial reasoning capability without visual input. For example, Collell et al. (2018), Mirzaee et al. (2021), and Liu et al. (2022) probe pretrained text-only models or VLMs' spatial reasoning capabilities with text-only questions. However, a text-only dataset cannot evaluate how a model relates language to grounded spatial information. In contrast, VSR focuses on the joint understanding of vision and language input.

2.4 Spatial Reasoning as a Sub-component

Last but not least, some vision-language tasks and models require spatial reasoning as a subcomponent. For example, Lei et al. (2020) propose TVQA+, a spatio-temporal video QA dataset containing bounding boxes for objects referred in the questions. Models then need to simultaneously conduct QA while detecting the correct object of interest. Christie et al. (2016) propose a method for simultaneous image segmentation and prepositional phrase attachment resolution. Models have to reason about objects' spatial relations in the visual scene to determine the assignment of prepositional phrases. However, if spatial reasoning is only a sub-component of a task, error analysis becomes more difficult. In contrast, VSR provides a focused evaluation of spatial relations, which are particularly challenging for current models.

3 Dataset Creation

In this section we detail how VSR is constructed. The data collection process can generally be split into two phases: (1) contrastive caption generation (§ 3.1) and (2) second-round validation (§ 3.2). We then discuss annotator hiring and payment (§ 3.3), dataset splits (§ 3.4), and the human ceiling and agreement of VSR (§ 3.5).

3.1 Contrastive Template-based Caption Generation (Figure 3)

In order to highlight spatial relations and avoid annotators frequently choosing trivial relations (such as ''near to''), we use a contrastive caption generation approach. Specifically, first, a pair of images, each containing two concepts of interest,

Category	Spatial Relations
Adjacency	Adjacent to, alongside, at the side of, at the right side of, at the left side of, attached to, at the back of, ahead of, against, at the edge of
Directional	Off, past, toward, down, deep down*, up*, away from, along, around, from*, into, to*, across, across from, through, down from
Orientation	Facing, facing away from, parallel to, perpendicular to
Projective	On top of, beneath, beside, behind, left of, right of, under, in front of, below, above, over, in the middle of
Proximity	By, close to, near, far from, far away from
Topological	Connected to, detached from, has as a part, part of, contains, within, at, on, in, with, surrounding, among, consists of, out of, between, inside, outside, touching
Unallocated	Beyond, next to, opposite to, after*, among, enclosed by

Table 1: The 71 available spatial relations; 66 of them appear in our final dataset ($*$ indicates not used).

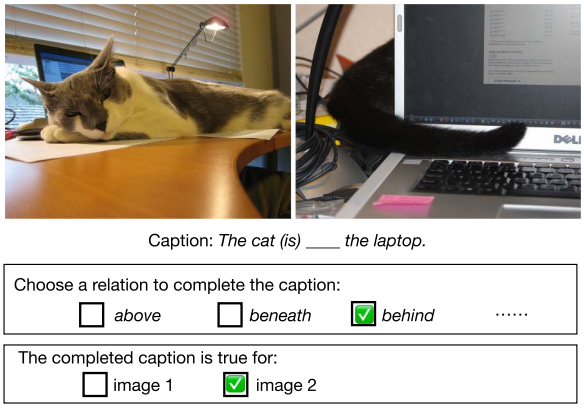


Figure 3: An annotation example of concepts ‘‘cat’’ & ‘‘laptop’’ in contrastive caption generation. The example generates two data points for our dataset: One ‘‘True’’ instance when the completed caption is paired with image 2 (right) and one ‘‘False’’ instance when paired with image 1 (left). Figure 3a source: Jeremy Zawodny. ‘‘Thunder-Man’’, uploaded October 16, 2007. `https://www.flickr.com/photos/jzawodn/1590039572/` (CC BY-NC 2.0). Figure 3b source: Chris Jobling. ‘‘Day 42: Cat and mouse?’’, uploaded September 30, 2008. `https://www.flickr.com/photos/51214457@N00/2901947727` (CC BY-SA 2.0).

would be randomly sampled from MS COCO (we use the train and validation sets of COCO 2017). Second, an annotator would be given a template containing the two concepts and is required to choose a spatial relation from a pre-defined list (Table 1) that makes the caption correct for one image but incorrect for the other image. We will detail these steps and explain the rationales in the following.

Image Pair Sampling. MS COCO 2017 contains 123,287 images and has labeled the segmentation and classes of 886,284 instances (individual objects). Leveraging the segmentation, we first randomly select two concepts (e.g., ‘‘cat’’ and ‘‘laptop’’ in Figure 3), then retrieve all images containing the two concepts in COCO 2017 (train and validation sets). Then, images that contain multiple instances of any of the concept are filtered out to avoid referencing ambiguity. For the single-instance images, we also filter out any of the images with instance pixel area size $< 30,000$, to prevent extremely small instances. After these filtering steps, we randomly sample a pair in the remaining images. We repeat such a process to obtain a large number of individual image pairs for caption generation.

Fill in the Blank: Template-based Caption Generation. Given a pair of images, the annotator needs to come up with a valid caption that makes it a correct description for one image but incorrect for the other. In this way, the annotator should focus on the key difference between the two images (which should be a spatial relation between the two objects of interest) and choose a caption that differentiates the two. Similar paradigms are also used in the annotation of previous vision-language reasoning datasets such as NLVR(2) (Suhr et al., 2017, 2019) and MaRVL (Liu et al., 2021). To regularize annotators from writing modifiers and differentiating the image pair with things beyond accurate spatial relations, we opt for a template-based classification task instead of free-form caption writing.[2] Additionally, the template-generated dataset can

[2]Hendricks and Nematzadeh (2021) propose a zero-shot probing benchmark of similar spirit for *verb* understanding. All captions are simplified as subject-verb-object triplets.

be easily categorized based on relations and their categories. Specifically, the annotator would be given instance pairs as shown in Figure 3.

The caption template has the format of ‘‘*The* `ENT1` *(is)* _____ *the* `ENT2`.’’, and the annotators are instructed to select a relation from a fixed set to fill in the slot. The copula ‘‘is’’ can be omitted for grammaticality. For example, for ‘‘contains’’ and ‘‘has as a part’’, ‘‘is’’ should be discarded in the template when extracting the final caption.

The fixed set of spatial relations enable us to obtain the full control of the generation process. The full list of used relations are listed in Table 1. The list contains 71 spatial relations and is adapted from the summarized relation table of Marchi Fagundes et al. (2021). We made minor changes to filter out clearly unusable relations, made relation names grammatical under our template, and reduced repeated relations. In our final dataset, 66 out of the 71 available relations are actually included (the other 6 are either not selected by annotators or are selected but the captions did not pass the validation phase).

3.2 Second-round Human Validation

In the second-round validation, every annotated data point is reviewed by at least 3 additional human annotators (validators). Given a data point (consisting of an image and a caption), the validator gives either a `True` or `False` label as shown in Figure 4 (the original label is hidden). In our final dataset, we exclude instances with fewer than 2 validators agreeing with the original label.

Design Choice on Reference Frames. During validation, a validator needs to decide whether a statement is true or false for an image. However, as discussed in § 1, interpreting a spatial relation requires choosing a *frame of reference*. For some images, a statement can be both true and false, depending on the choice. As a concrete example, in Figure 1, while the potted plant is on the left side from the viewer’s perspective (relative frame), the potted plant is at the right side if the bench is used to define the coordinate system (intrinsic frame).

In order to ensure that annotations are consistent across the dataset, we communicated to the annotators that, for relations such as ‘‘left’’/‘‘right’’ and ‘‘in front of’’/‘‘behind’’, they should consider both possible reference frames, and assign the label `True` when a caption is true from either the intrinsic or the relative frame. Only when a caption is incorrect under both reference frames (e.g., if the caption is ‘‘*The potted plant is under the bench.*’’ for Figure 1) should a `False` label be assigned.

Caption: *The pizza is at the edge of the dinning table.*

The caption is: ☑ True ☐ False

Figure 4: A second-round validation example. Image source: Marisa McClellan. ‘‘Becky’s grilled pizza’’, uploaded May 31, 2011. `https://www.flickr.com/photos/marusula/5779127081/` (CC BY-NC-ND 2.0).

On a practical level, this adds difficulty to the task, since a model cannot naively rely on pixel locations of the objects in the images, but also needs to correctly identify orientations of objects. However, the task is well-defined: A model that can correctly simulate both reference frames would be able to perfectly solve this task.

From a theoretical perspective, by involving more diverse reference frames, we are also demonstrating the complexity of human cognitive processes when understanding a scene, since different people approach a scene with different frames. Attempting to enforce a specific reference frame would be methodologically difficult and result in an unnaturally restricted dataset.

3.3 Annotator Hiring and Organization

Annotators were hired from `prolific.co`. We required them to (1) have at least a bachelor’s degree, (2) be fluent in English, and (3) have a >99% historical approval rate on the platform. All annotators were paid 12 GBP per hour.

For caption generation, we released the task with batches of 200 instances and the annotator was required to finish a batch in 80 minutes. An annotator could not take more than one batch per

split	train	dev	test	total
random	7,680	1,097	2,195	10,972
zero-shot	4,713	231	616	5,560

Table 2: Statistics of the *random* and *zero-shot* splits.

day. In this way we had a diverse set of annotators and could also prevent annotators from becoming fatigued. For second-round validation, we grouped 500 data points in one batch and an annotator was asked to label each batch in 90 minutes.

In total, 24 annotators participated in caption generation and 45 participated in validation. Four people participated in both phases, which should have minimally impacted the validation quality. The annotators had diverse demographic backgrounds: They were born in 15 countries, were living in 13 countries, and had 12 nationalities. Fifty annotators were born and living in the same country while others had moved to different ones. The vast majority of our annotators were residing in the UK (32), South Africa (9), and Ireland (7). The ratio for holding a bachelor/master/PhD as the highest degree was: 12.5%/76.6%/10.9%. Only 7 annotators were non-native English speakers while the other 58 were native speakers. In our sample, 56.7% of the annotators self-identified as female and 43.3% as male.

3.4 Dataset Splits

We split the 10,972 validated data points into train/dev/test sets in two different ways. The stats of the two splits are shown in Table 2. In the following, we explain how they are created. ***Random split*:** We split the dataset randomly into train/dev/test with a ratio of 70/10/20. ***Concept zero-shot split*:** We create another concept zero-shot split where train/dev/test have no overlapping concepts. That is, if ‘‘dog’’ appears in the train set, then it does not appear in dev or test sets. This is done by randomly grouping concepts into three sets with a ratio of 50/20/30 of all concepts. This reduces the dataset size, since data points involving concepts from different parts of the train/dev/test split must be filtered out. The concept zero-shot split is a more challenging setup since the model has to learn concepts and the relations in a compositional way instead of remembering the co-occurrence statistics of the two.

3.5 Human Ceiling and Agreement

We randomly sample 500 data points from the final random split test set of the dataset for computing human ceiling and inter-annotator agreement. We hide the labels of the 500 examples and two additional annotators are asked to label `True`/`False` for them. On average, the two annotators achieve an accuracy of 95.4% on the VSR task. We further compute the Fleiss' kappa among the original annotation and the predictions of the two humans. The Fleiss' kappa score is 0.895, indicating near-perfect agreement according to Landis and Koch (1977).

4 Dataset Analysis

In this section we compute some basic statistics of our collected data (§ 4.1), analyze where human annotators have agreed/disagreed (§ 4.2), and present a case study on reference frames (§ 4.3).

4.1 Basic Statistics of VSR

After the first phase of contrastive template-based caption generation (§ 3.1), we collected 12,809 raw data points. In the phase of the second round validation (§ 3.2), we collected 39,507 validation labels. Every data point received at least 3 validation labels. In 69.1% of the data points, all validators agree with the original label. We find that 85.6% of the data points have at least $\frac{2}{3}$ annotators agreeing with the original label. We use $\frac{2}{3}$ as the threshold and exclude all instances with lower validation agreement. After excluding other instances, 10,972 data points remained and are used as our final dataset.

Here we provide basic statistics of the two components in the VSR captions: The concepts and the relations. Figure 5 demonstrates the relation distribution. ‘‘touching’’ is most frequently used by annotators. The relations that reflect the most basic relative coordinates of objects are also very frequent, e.g., ‘‘behind’’, ‘‘in front of’’, ‘‘on’’, ‘‘under’’, ‘‘at the left/right side of’’. Figure 6 shows the distribution of concepts in the dataset. Note that the set of concepts is bounded by MS COCO and the distribution also largely follows MS COCO. Animals such as ‘‘cat’’, ‘‘dog’’, and ‘‘person’’ are the most frequent. Indoor objects such as ‘‘dining table’’ and ‘‘bed’’ are also very dominant. In Figure 6, we separate

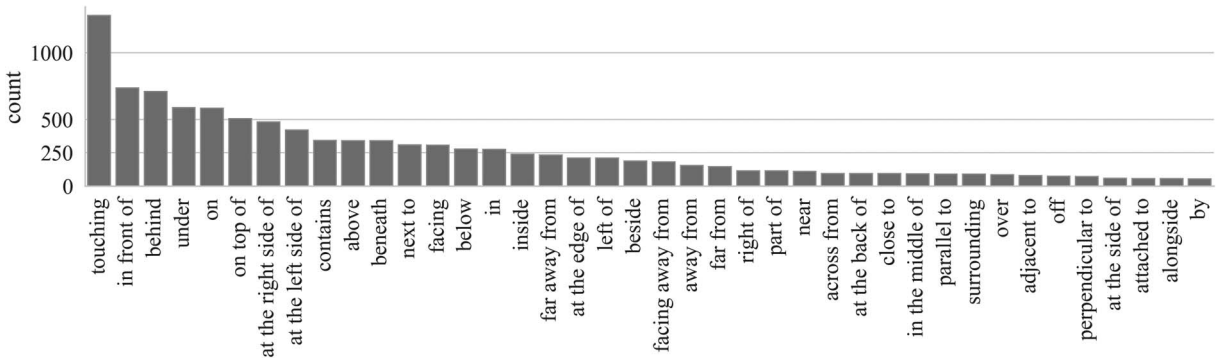


Figure 5: Relation distribution of the final dataset (sorted by frequency). Top 40 most frequent relations are included. It is clear that the relations follow a long-tailed distribution.

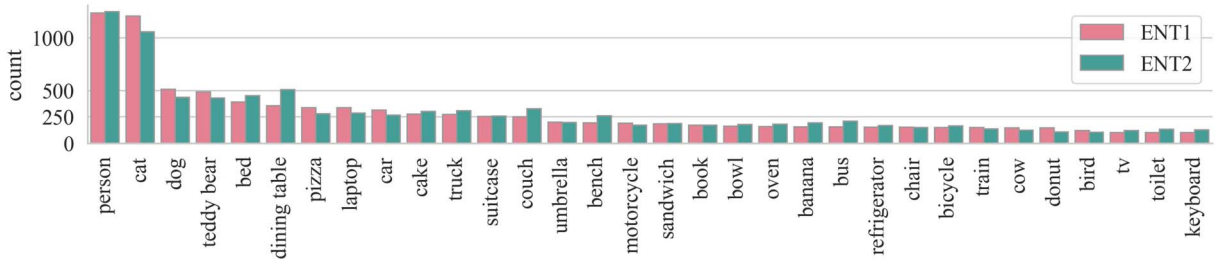


Figure 6: Concept distribution. Only concepts with > 100 frequencies are included.

the concepts that appear at `ENT1` and `ENT2` positions of the sentence and their distributions are generally similar.

4.2 Where Do Annotators Disagree?

While we propose using data points with high validation agreement for model evaluation and development, the unfiltered dataset is a valuable resource for understanding cognitive and linguistic phenomena. We sampled 100 examples where annotators disagree, and found that around 30 of them are caused by annotation errors but the rest are genuinely ambiguous and can be interpreted in different ways. This shows a level of intrinsic ambiguity of the task and variation among people.

Along with the validated VSR dataset, we also release the full unfiltered dataset, with annotators' and validators' metadata, as a second version to facilitate linguistic studies. For example, researchers could investigate questions such as where disagreement is more likely to happen and how people from different regions or cultural backgrounds might perceive spatial relations differently.

To illustrate this, the probability of two randomly chosen annotators disagreeing with each other is given for each relation in Figure 7. Some of the relations with high disagreement can be interpreted in the intrinsic reference frame, which requires identifying the orientations of objects, for example, ''at the side of'' and ''in front of''. Other relations have a high level of vagueness, e.g., for the notion of *closeness*: ''near'' and ''close to''. By contrast, part-whole relations, such as ''has as a part'', ''part of'', and in/out relations such as ''within'', ''into'', ''outside'', and ''inside'' have the least disagreement.

4.3 Case Study: Reference Frames

It is known that the relative reference frame is often preferred in English, at least in standard varieties. For example, Edmonds-Wathen (2012) compares Standard Australian English and Aboriginal English, as spoken by school children at a school on Croker Island, investigating the use of the relations ''in front of'' and ''behind'' when describing simple line drawings of a person and a tree. Speakers of Standard Australian English were found to prefer the relative frame, while speakers of Aboriginal English were found to prefer the intrinsic frame.

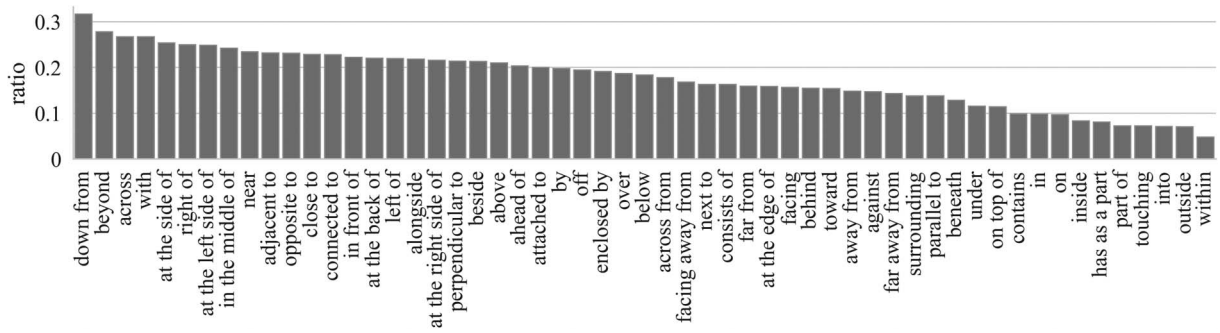


Figure 7: Per-relation probability of having two randomly chosen annotator disagreeing with each other (sorted from high to low). Only relations with > 20 data points are included in the figure.

Our methodology allows us to investigate reference frame usage across a wide variety of spatial relations, using a wide selection of natural images. To understand the frequency of annotators using relative vs. intrinsic frames, we label instances' reference frames and study their distributions. The majority of examples that can be interpreted differently under different reference frames are left/right-related relations (i.e., ''left/right of'' and ''at the left/right side of''). We find all left/right-related *true*[3] statements and classify them into three categories: (1) intrinsic, (2) relative, and (3) both (the caption is correct under either intrinsic and relative frames of reference). Among the 616 instances, 68 (11%) and 518 (84%) use intrinsic and relative frames respectively, 30 (5%) can be interpreted with both frames. Since the vast majority of our annotators were native English speakers (91%), and all were university-educated, our finding is consistent with previous work suggesting that the relative frame is the most common frame in standard varieties of English.

Besides the overall trend, the use of reference frames can vary with the circumstances. Related patterns have been studied in cognitive science. For example, Vukovic and Williams (2015) find a three-way interaction between linguistic cues, spatial configurations in an image, and a person's own preferences on reference frames.

We investigated whether reference to a person in the image might influence how annotators comprehend the scene. 198 out of the 616 instances involve ''person'' in the caption. And out of the 198 human-involved instances, 32 (16%) use an intrinsic frame and 154 (78%) use a relative frame (12, i.e., 6%, can be interpreted with both frames), while the proportions were 9% and 87% for instances not involving ''person''. This is a statistically significant difference (using two-tailed Fisher's exact test, $p = 0.0054$ if ignoring both-frame cases, and $p = 0.0045$ if grouping both-frame and intrinsic cases). In other words, this suggests that the involvement of a human can more likely prompt the use of the intrinsic frame.

5 Experiments

In this section, we test VLMs on VSR. We first introduce baselines and experimental configurations in § 5.1, then experimental results and analysis in § 5.2. Then we discuss the role of frame of reference using experiments in § 5.3 and finally conduct sample efficiency analysis in § 5.4.

5.1 Baselines and Experiment Configurations

Baselines. For finetuning-based experiments, we test three popular VLMs: VisualBERT (Li et al., 2019),[4] LXMERT (Tan and Bansal, 2019),[5] and ViLT (Kim et al., 2021).[6] All three models are stacked Transformers (Vaswani et al., 2017) that take image-text pairs as input. The difference mainly lies in how or whether they encode the position information of objects. We report only finetuned results but not direct inferences from off-the-shelf checkpoints since some of their pretraining objectives are inconsistent with the

[3] According to our guideline, false statements are interpreted as false under both frames.

[4] huggingface.co/uclanlp/visualbert-nlvr2-coco-pre.

[5] huggingface.co/unc-nlp/lxmert-base-uncased.

[6] huggingface.co/dandelin/vilt-b32-mlm.

model	lr	batch size	epoch	token length
VisualBERT	2e-6	32	100	32
LXMERT	1e-5	32	100	32
ViLT	1e-5	12	30	max

Table 3: A listing of hyperparameters used for all VLMs (''lr'': learning rate).

binary classification task of VSR, thus requiring additional engineering.

We additionally test the alt text pretrained dual-encoder CLIP (Radford et al., 2021) as an off-the-shelf baseline model (no finetuning).[7] We follow Booth (2023) to construct negation or antonym of each individual relation. For example, ''facing'' $\rightarrow$ ''facing away from'' and ''ahead of'' $\rightarrow$ ''not ahead of''. For each sample, we compare the embedding similarity of the image-caption pair and that of the negated caption. If the original pair has a higher probability then the model prediction is `True`, otherwise `False`. We call this method CLIP (w/ prompting). We only report direct prompting results without finetuning since CLIP finetuning is expensive.

Experimental Configurations. We save checkpoints every 100 iterations and use the best-performing checkpoint on dev set for testing. All models are run three times using three random seeds. All models are trained with the AdamW optimizer (Loshchilov and Hutter, 2019). The hyperparameters we used for training the three VLMs are listed in Table 3.

5.2 Experimental Results

In this section, we provide both quantitative and qualitative results of the four baselines. Through analyzing the failure cases of the models, we also highlight the key abilities needed to solve this dataset.

As shown in Table 4, the best-performing models on the random split are LXMERT and ViLT, reaching around 70% accuracy, while VisualBERT is just slightly better than the chance level. On the zero-shot split, all models' performance decline substantially and the best model, ViLT, only obtains 63.0% accuracy. The off-of-the-shelf CLIP model obtains around 55% on both sets, indicating its weaknesses in spatial reasoning echoing Subramanian et al.'s (2022) findings. Overall, these results lag behind the human ceiling by more than 25% and highlight that there is substantial room for improving current VLMs.

[7] `huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K`.

model↓	random split	zero-shot split
human ceiling	95.4	
CLIP (w/ prompting)	56.0	54.5
VisualBERT	$55.2_{\pm 1.4}$	$51.0_{\pm 1.9}$
ViLT	$\mathbf{69.3}_{\pm 0.9}$	$\mathbf{63.0}_{\pm 0.9}$
LXMERT	$\mathbf{70.1}_{\pm 0.9}$	$61.2_{\pm 0.4}$

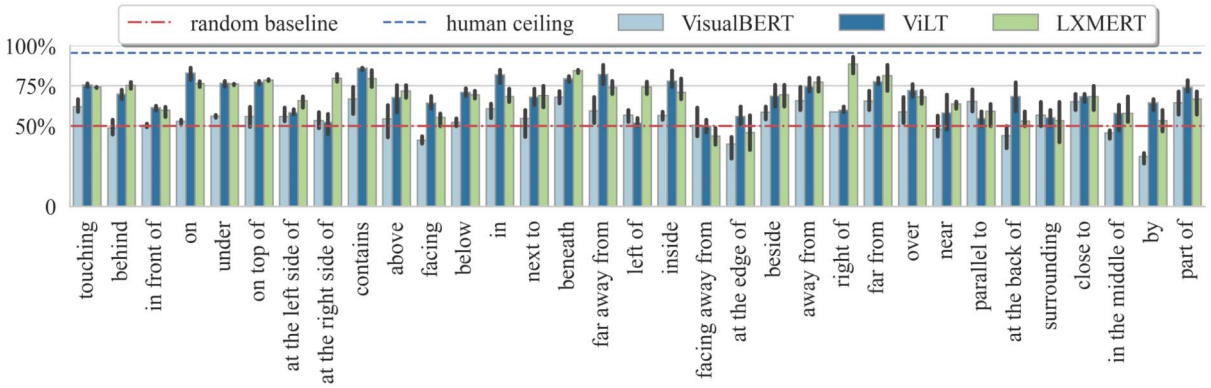
Table 4: Model performance on VSR test set. CLIP is applied without finetuning but with carefully engineered prompts while the other three smaller models are finetuned on the training set.

Explicit Positional Information Matters. Both LXMERT and ViLT outperform VisualBERT by large margins (>10%) on both splits. This is expected since LXMERT and ViLT encode explicit positional information while VisualBERT does not. LXMERT has position features as part of the input which encode the relative coordinates of objects within the image. ViLT slices an image into patches (instead of object regions) and uses positional encodings to signal the patches' relative positions. VisualBERT, however, has no explicit position encoding. Bugliarello et al. (2021) and Rösch and Libovický (2022) also highlight the importance of positional encodings of VLMs, which agrees with our observations.

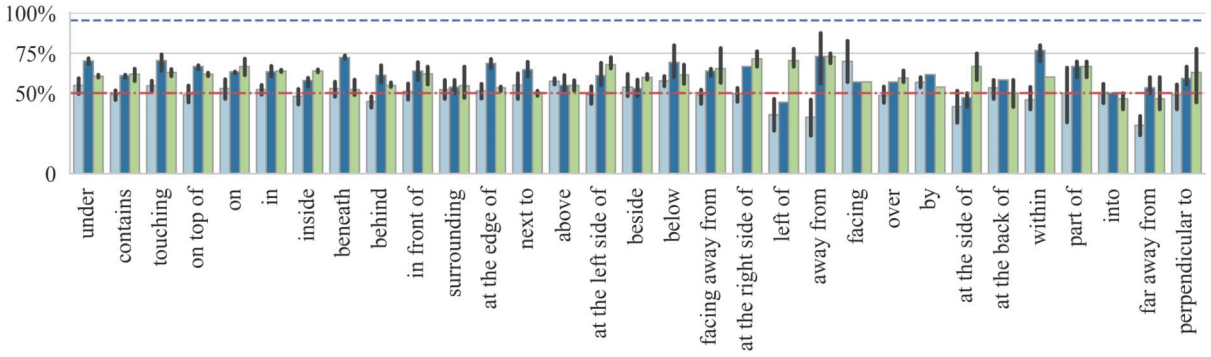
Random Split vs. Zero-shot Split. It is worth noting that the performance gap between the random and zero-shot splits is large. As we will show in § 5.4, the underlying cause is not likely to be the number of training examples, but rather that concept zero-shot learning is fundamentally a challenging task. The gap suggests that disentangling representations of concepts and relations is challenging for current models.

Sensitiveness to Random Seeds. Model performance varies by about one to two percentage points. These fluctuations illustrate the importance of always reporting the average performance of multiple runs to make sure the conclusion is reliable.

Performance by Relation. We give performance by relation for all three finetuned models



(a) random split



(b) zero-shot split

Figure 8: Performance (accuracy) by relation on the random (a) and zero-shot (b) split test sets. Relation order sorted by frequency (high to low from left to right). Only relations with more than 15 and 5 occurrences on the random and zero-shot tests, respectively, are shown.

on both random and zero-shot splits in Figure 8. The order from left to right is sorted by the frequency of relations in the dataset (within each split). Interestingly, there does not seem to be any correlation between performance and frequency of the relation, hinting that specific relations are hard not due to an insufficient number of training examples but because they are fundamentally challenging for current VLMs. Any relation that requires recognising orientations of objects seems to be hard, for example, ‘‘facing’’, ‘‘facing away from’’, ‘‘parallel to’’, and ‘‘at the back of’’. As an example, LXMERT failed on the two examples in Figure 9 which require understanding the front of a hair drier and a person respectively. In this regard, left-right relations such as ‘‘at the left/right side of’’ and ‘‘left/right of’’ are difficult because the intrinsic reference frame requires understanding the orientation of objects. As an example, in Figure 1, all three models predicted `False`, but in the intrinsic frame (i.e., from the bench’s point of view), the potted plant is indeed at the right.

To get a more high-level understanding of the relations’ performance, we group model performance by the categories of Marchi Fagundes et al. (2021): ‘‘Adjacency’’, ‘‘Directional’’, ‘‘Orientation’’, ‘‘Projective’’, ‘‘Proximity’’, ‘‘Topological’’, and ‘‘Unallocated’’ (also shown in Table 1). The results are shown in Figure 10. ‘‘Orientation’’ is the worst performing group on the random split, and on average all models’ performance is close to the chance level. When comparing random and zero-shot splits, performance has declined to some extent for almost all categories and models. The decrease in ‘‘Proximity’’ is particularly drastic across all models—it declined from close to 75% accuracy in random split to chance level in zero-shot split. ‘‘Proximity’’ contains relations such as ‘‘close to’’, ‘‘near’’, and ‘‘far from’’. We believe it is due to the fact that the notion of proximity is relative and very much dependent on the nature of the concept and its frequent physical context. For example, for a ‘‘person’’ to be ‘‘near’’ an indoor

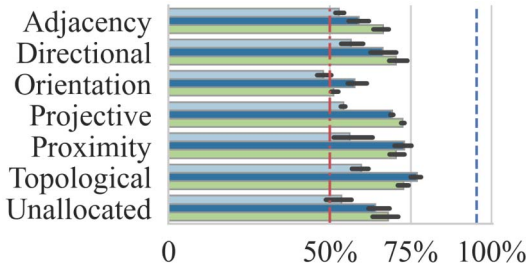
(a) Caption: *The hair drier is facing away from the person*. Label: `False`.

(b) Caption: *The bench is in front of the person*. Label: `True`.

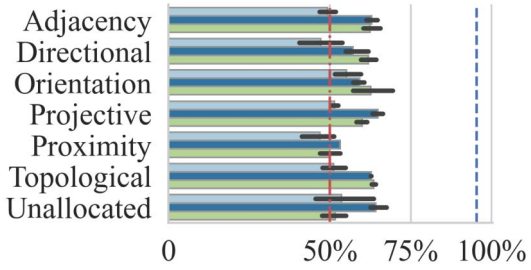
Figure 9: LXMERT failed on both examples. Figure 9a source: Austin & Zak. ‘‘three minutes with dryer on high’’, uploaded February 23, 2008. `https://www.flickr.com/photos/zakh/2285744646/` (CC BY-NC-SA 2.0). Figure 9b source: Carrick. ‘‘Elsa and Maia working hard sanding down the bench’’, uploaded April 17, 2012. `https://www.flickr.com/photos/carrickg/6941414780/` (CC BY-NC 2.0).

concept such as ‘‘oven’’ is very different from a person being ‘‘near’’ a frequent outdoor object such as ‘‘train’’ or ‘‘truck’’. Since the zero-shot split prevents models from seeing test concepts during training, the models have a poor grasp of what counts as ‘‘close to’’ or ‘‘far from’’ for these concepts, thus generalizing poorly.

Other Errors. While certain relations are intrinsically hard, we have observed other types of errors that are not bounded to specific relations. Here we give a few examples. Some instances require complex reasoning. In Figure 11, the model needs to recognize that both the cow and the back of the car are in the car’s side mirror and also infer the relative position of the back of the car and the cow. It is perhaps no surprise that two of the three



(a) random split



(b) zero-shot split

Figure 10: Performance by categories of relations, on the random (a) and zero-shot (b) split test sets. For legend information, see Figure 8.

models predicted wrongly. Some other examples require common sense. For example, in Figure 2, we can infer the person and the cow’s moving direction and can then judge if the cow is ahead of the person. LXMERT failed on this example. In Figure 3 (right), the model needs to infer that the main body of the cat is hidden behind the laptop. Interestingly, all three models predicted this example correctly.

5.3 Case Study on Reference Frames

As discussed in § 4.3, different frames of reference can be used in natural language and it would be helpful to understand whether our models recognise them. We argue that the task of identifying frame of reference itself is very hard for current models. However, learning to recognize frames of reference helps the task of visual spatial reasoning.

Firstly, we conduct a case study on left/right-related relations. We additionally label the reference frames of all true statements containing any of the left/right-related relations. We exclude all data points that can be interpreted in both intrinsic and relative frames to slightly reduce the complexity of the task. Then we finetune a ViLT checkpoint to predict the reference frame based on the true statement and the image. The model’s

Figure 11: Caption: *The cow is at the back of the car*. Label: `True`. LXMERT and VisualBERT predicted `False`. Image source: shorty76. ''Side Mirror View'', uploaded December 26, 2008. `https://www.flickr.com/photos/shorty_76/3136942358/` (CC BY-NC-ND 2.0).

Reference frame prediction task			
model↓	Precision	Recall	F1
ViLT	$59.2_{\pm 3.7}$	$59.7_{\pm 5.8}$	$56.9_{\pm 4.4}$
VSR task (left/right subset)			
model↓		Accuracy	
ViLT		$54.2_{\pm 0.6}$	
ViLT + rf_trained		**59.2**$_{\pm 1.8}$	

Table 5: ViLT model performance on the reference frame prediction task (upper half; we report macro-averaged Precision/Recall/F1 since the binary classification task is imbalanced); and VSR task using original pretrained checkpoint or the reference frame prediction task trained checkpoint (accuracy reported).

performance on test set is shown in the upper half of Table 5. We can see that reference frame prediction is an extremely hard task for the model. This is presumably because it requires taking into account a 3D viewpoint and simulating transformations between different viewpoints.

Secondly, we use this model trained with reference frame labels to initialize the VSR task model and further finetune it on the VSR task (only the left/right relations). The test results are shown in the lower part of Table 5.[8] We see a clear positive transfer from reference frame prediction task to the VSR task. This suggests that learning to recognise reference frames can indeed help downstream visual spatial reasoning. This makes sense since simulating the transformation of intrinsic/relative frames could be an intermediate reasoning step in detecting whether a statement is true/false.

5.4 Sample Efficiency

In order to understand the correlation between model performance and the number of training examples, we conduct sample efficiency analysis on VSR. The results are plotted in Figure 12. For the minimum resource scenario, we randomly sample 100 shots from the training sets of each split. Then we gradually increase the number of training examples to be 25%, 50%, and 75% of the whole training sets. Both LXMERT and ViLT have a reasonably good few-shot capability and can be quite performant with 25% of training data. LXMERT, in particular, reaches above 55% accuracy with 100 shots on both splits. The zero-shot split is substantially harder and most models appear to have already plateaued at around 75% of the training set. For the random split, all models are increasing performance with more data points, though improvement slows down substantially for LXMERT and ViLT after 75% of training data. The fact that LXMERT has the best overall few-shot capability may be suggesting that LXMERT's pretrained object detector has a strong inductive bias for the VSR dataset as it does not need to learn to recognise concept boundaries and classes from scratch. However, this advantage from LXMERT seems to fade away as the number of training examples increases.

[8]Note that the reference frame train/dev/test sets are derived from the VSR task split—so no data leakage is possible from train to dev and test sets even after the intermediate pretraining.

6 Conclusion and Future Directions

We have presented Visual Spatial Reasoning (VSR), a controlled probing dataset for testing the capability of vision-language models (VLMs) of recognising and reasoning about spatial relations in natural image-text pairs. We made a series of linguistic observations on the variability of spatial language when collecting VSR. We highlighted the diverse use of reference frames among annotators, and also the ambiguous nature of certain spatial relations. We tested four popular VLMs on VSR, and found they perform more than 25% below the human ceiling. On a more challenging concept zero-shot split, the tested VLMs

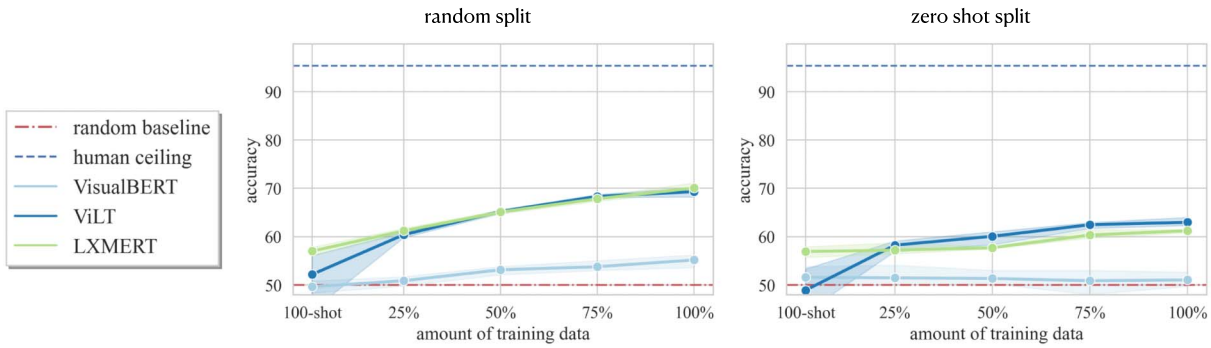


Figure 12: Sample efficiency analysis: model performance under different amounts of training data (100-shot, 25%, 50%, 75%, and 100% of training set). Results on both the random and zero-shot split test sets are shown. As training data increases, the performance plateaus on both sets but the flattening trend is more obvious on the zero-shot split.

struggled to reach 60% accuracy and their performance plateaued even with increased training examples. Among the finetuning-based VLMs, ViLT, and LXMERT outperformed VisualBERT, and we noted out that the explicit positional information in the former two models is crucial in the task. CLIP with prompt engineering achieved slightly better than random performance, suggesting poor capability in spatial reasoning. We also performed a by-relation analysis and found that the models' performance on certain relations have little correlation with the number of training examples, and certain relations are inherently more challenging. We identified orientation as the most difficult category of relations for VLMs. Proximity is another challenging category, especially in the zero-shot setup as this relation is highly concept-dependent. We hope the task serves as a useful tool for testing and probing future VLMs.

In future work, we plan to more extensively investigate whether large-scale pretrained dual-encoders such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and LiT (Zhai et al., 2022) can properly recognize spatial relations, especially in the finetuning setup. A comparison of dual- and cross-encoders' performance on each spatial relation might guide future model design. Recently, Alayrac et al. (2022), Chen et al. (2023), and Huang et al. (2023) proposed ultra-large-scale VLMs. It would be interesting to see if VLMs have better spatial reasoning capability when scaled up. Another direction is extending VSR to cover more languages and cultures (Liu et al., 2021; Bugliarello et al., 2022) and test multilingual VLMs. Along the same line, since we have also collected the metadata of annotators, the VSR corpus can be used as a resource for investigating research questions such as: How is ''space'' described among different dialects of English? How is ''space'' perceived among different populations? We hope that the annotation process of VSR can also serve as a basis for future cross-lingual and cross-cultural sociolinguistic research.

Acknowledgments

We thank the TACL reviewers and the action editor for their thoughtful comments. We thank Qian Wang and Rongtian Ye for helping trial the annotation scheme; Zihao Fu for helping set up the annotation server. The project is funded by Cambridge Language Sciences Incubator Fund. FL is supported by Grace & Thomas C.H. Chan Cambridge Scholarship.

References

Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.586

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han,

Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, November 28 – December 9, 2022, New Orleans, LA, USA*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016*, pages 39–48. 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 *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015*, pages 2425–2433. IEEE Computer Society. https://doi.org/10.1109/ICCV.2015.279

Joe Booth. 2023. CLIP visual spatial reasoning. https://github.com/Sohojoe/CLIP_visual-spatial-reasoning. GitHub repository.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994. https://doi.org/10.1162/tacl_a_00408

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. 2016. Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1503, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1156

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2123

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018*, pages 6765–6772. AAAI Press.

Cris Edmonds-Wathen. 2012. False friends in the multilingual mathematics classroom. In *12th International Congress on Mathematical Education Topic Study Group 28, 8–15 July, 2012, Seoul, Korea*, pages 5857–5866.

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. https://doi.org/10.1109/CVPR.2017.670

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.318

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE. https://doi.org/10.1109/CVPR.2019.00686

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*, pages 1988–1997. IEEE Computer Society. https://doi.org/10.1109/CVPR.2017.215

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73. https://doi.org/10.1007/s11263-016-0981-7

Alexander Kuhnle and Ann Copestake. 2018. Deep learning evaluation using deep linguistic processing. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 17–23, New Orleans, Louisiana. Association for Computational Linguistics.

Alexander Kuhnle, Huiyuan Xie, and Ann Copestake. 2018. How clever is the FiLM model, and how clever can it be? In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 162–172. https://doi.org/10.1007/978-3-030-11018-5_15

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. https://doi.org/10.2307/2529310, PubMed: 843571

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.730

Stephen C. Levinson. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press. https://doi.org/10.1017/CBO9780511613609

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. *ArXiv preprint*, abs/1908.03557.

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 *European Conference on Computer Vision*, pages 740–755. Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. 2019. CLEVR-Ref+: Diagnosing visual reasoning with referring expressions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*, pages 4185–4194. Computer Vision Foundation / IEEE. https://doi.org/10.1109/CVPR.2019.00431

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. 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, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.168

Yinhong Liu and Guy Emerson. 2022. Learning functional distributional semantics with visual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3976–3988, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.

Cristiane Kutianski Marchi Fagundes, Kristin Stock, and Luciene Delazari. 2021. A cross-linguistic study of spatial location descriptions in New Zealand English and Brazilian Portuguese natural language. *Transactions in GIS*, 25(6):3159–3187. https://doi.org/10.1111/tgis.12815

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. 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, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.364

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.567

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Philipp J. Rösch and Jindřich Libovický. 2022. Probing the role of positional information in vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1031–1041, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.77

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. 2022. ReCLIP: A strong zero-shot

baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5198–5215, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.357

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2034

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Leonard Talmy. 1983. How language structures space. In Herbert L. Pick and Linda P. Acredolo, editors, *Spatial Orientation: Theory, Research, and Application*, pages 225–282. Plenum Press. https://doi.org/10.1007/978-1-4615-9325-6_11

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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1514

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Nikola Vukovic and John N. Williams. 2015. Individual differences in spatial cognition influence mental simulation of language. *Cognition*, 142:110–122.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *ArXiv preprint*, abs/1901.06706. An earlier version of this paper was published at the NeurIPS 2018 ViGIL workshop.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1023

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE. https://doi.org/10.1109/CVPR.2019.00688

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133. https://doi.org/10.1109/CVPR52688.2022.01759