Things not Written in Text: Exploring Spatial Commonsense from Visual Signals

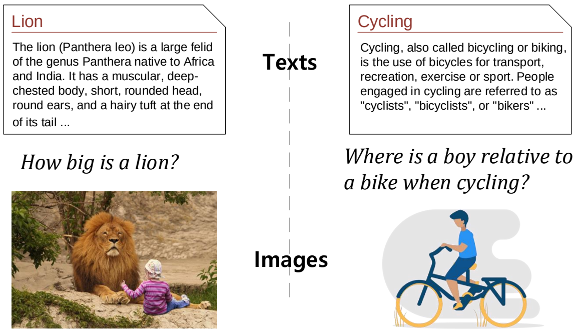
Anonymous ACL submission

Abstract

Spatial commonsense, the knowledge about spatial position and relationship between objects (like *the relative size of a lion and a girl*, and *the position of a boy relative to a bicycle when cycling*), is an important part of commonsense knowledge. Although pretrained language models (PLMs) succeed in many NLP tasks, they are shown to be ineffective in spatial commonsense reasoning. Starting from the observation that images are more likely to exhibit spatial commonsense than texts, we explore whether models with visual signals learn more spatial commonsense than text-based PLMs. We propose a spatial commonsense benchmark that focuses on the relative scales of objects, and the positional relationship between people and objects under different actions.[1] We probe PLMs and models with visual signals, including vision-language pretrained models and image synthesis models, on this benchmark, and find that image synthesis models are more capable of learning *accurate* and *consistent* spatial knowledge than other models. The spatial knowledge from image synthesis models also helps in natural language understanding tasks that require spatial commonsense.

1 Introduction

Spatial perception, the ability to detect the spatial position and to infer the relationship between visual stimuli (Donnon et al., 2005; Saj and Barisnikov, 2015), is basic but important for human beings (Pellegrino et al., 1984). It is of everyday use, from understanding the surrounding environment, like *when seeing a woman sitting in a car with her hands on the steering wheel, we know she is probably driving*, to processing spatial information and performing reasoning, like *navigating through a dense forest*. We regard the knowledge needed in spatial perception as spatial commonsense. Humans start to develop spatial perception and acquire spatial commonsense from infancy, and apply the commonsense through lifetime (Kuipers et al., 1990; Poole et al., 2006).



Lion

The lion (Panthera leo) is a large felid of the genus Panthera native to Africa and India. It has a muscular, deep-chested body, short, rounded head, round ears, and a hairy tuft at the end of its tail ...

How big is a lion?

Cycling

Cycling, also called bicycling or biking, is the use of bicycles for transport, recreation, exercise or sport. People engaged in cycling are referred to as "cyclists", "bicyclists", or "bikers" ...

Where is a boy relative to a bike when cycling?

Figure 1: Texts and images related to *lion* and *cycling*. Images are shown to contain more spatial knowledge.

Although text-based Pretrained Language Models (PLMs) achieve great performance on various commonsense reasoning tasks (Davison et al., 2019; Zhou et al., 2020), they are shown to be ineffective when dealing with spatial commonsense. Zhang et al. (2020) and Aroca-Ouellette et al. (2021) show that current PLMs lack the ability to reason about object scales. Bhagavatula et al. (2020) find that BERT (Devlin et al., 2019) underperforms on instances involving spatial locations. The struggle of PLMs with spatial commonsense is partly because spatial commonsense is rarely expressed explicitly in texts. We may write sentences like *lions are big animals*, but we seldom explicitly mention how big lions are; we also rarely write about the spatial relationship between a boy and a bicycle when he is cycling.

Spatial commonsense is exhibited in images more commonly. As shown in Figure 1, the two Wikipedia articles provide little spatial information, but a picture of *a lion and a girl* provides a reference to the size of a lion; and a painting of *a boy riding a bicycle* depicts that he sits *on* the bicycle. Hence, a natural idea is to elicit spatial knowledge from models with visual signals.

[1] Code and data are available in supplementary materials.

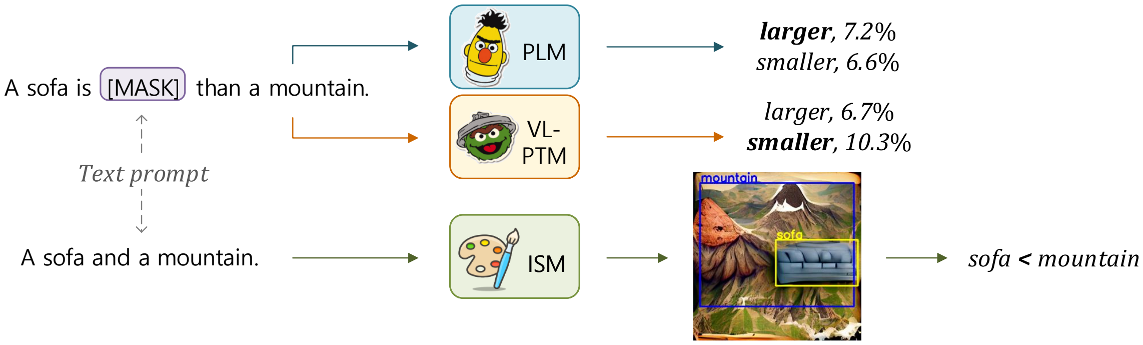


Figure 2: The probing process. We take the size comparison between $sofa$ and $mountain$ as an example.

We first study *whether models with visual signals learn more spatial knowledge than text-only models*. We select Vision-Language PreTrained Models (VL-PTMs) and Image Synthesis Models (ISMs) for investigation. VL-PTMs encode text and images together, fusing their features to deal with downstream tasks. ISMs take text as input, and generate images based on the text. To evaluate the spatial commonsense in PLMs and models with visual signals, we design a benchmark. It involves two subtasks: 1) comparing sizes and heights of different objects (like *a lion and a girl*), and 2) determining the positional relationship between a person and an object when a certain action happens (like *a boy's position when riding a bicycle*). The subtasks are designed to examine the model's capability to master two kinds of spatial commonsense: understanding spatial scales, and the relationship between surrounding objects and ourselves.

As shown in Figure 2, we probe models with text prompts on this benchmark. We feed text prompts with masks to PLMs and VL-PTMs, and take the possible word with the highest probability as its prediction. We probe ISMs in a similar way: we first feed the text prompts to ISMs and then evaluate the generated images. We evaluate the images with two methods: automatically comparing bounding boxes of objects and conducting human evaluation. Results show that models with visual signals learn more accurate spatial commonsense than PLMs.

Besides the performance comparison, we are also interested in *how is the quality of spatial commonsense learned by different models?* We investigate how consistent the spatial knowledge in different models is, like whether it can manifest *a lion is larger than a girl* and *a girl is smaller than a lion* simultaneously; and to what extent models can generalize the knowledge when uncommon scenarios like *an enchantress lights the sparkler* appear. We observe that ISMs are capable of generating consistent spatial knowledge and the performance is robust in uncommon scenarios.

The following problem is *how to benefit natural language understanding tasks with the spatial knowledge from ISMs?* We investigate this in the question answering scenario. Take a question like *A boy is riding a bicycle. Is he on the bicycle?* We generate an image about the question context *a boy who is riding a bicycle* with a text prompt using ISMs, and feed both the question and the generated image into vision-language models to predict an answer. This framework outperforms strong question answering models pretrained on texts only. While this is a simplified scenario of spatial commonsense reasoning, it manifests a possible way to employ the spatial knowledge learned by ISMs in natural language understanding.

Motivated by the observation that images contain more spatial commonsense than texts, we 1) design a framework, including the data and probing methods, to compare the spatial commonsense reasoning ability of models with different modalities; 2) propose methods to evaluate the quality of learned spatial commonsense, and find that models with visual signals, especially ISMs, learn more *precise* and *robust* spatial knowledge than PLMs; and 3) demonstrate the improvement in spatial commonsense question answering with the help of ISMs.

2 Related Works

2.1 Spatial Commonsense Reasoning

Object Scales. Bagherinezhad et al. (2016) build a dataset for objects' size comparison, and Elazar et al. (2019) provide distributional information about objects' lengths. Forbes and Choi (2017) also involve spatial comparison but are criticized for ill-defined comparison (Elazar et al., 2019). Aroca-

Ouellette et al. (2021) design a physical reasoning dataset that requires not only spatial commonsense but also a complex reasoning process, which is extremely challenging for existing models. We choose the formulation of object comparison in pairs as this kind of knowledge is easy to be probed from different models.

Spatial Relationship. Collell et al. (2018) introduce a dataset of spatial templates for objects under different relations, but the spatial relations are represented as relative positions of bounding boxes, which are hard to express in language. Mirzaee et al. (2021) design a textual spatial reasoning benchmark, and Johnson et al. (2017) and Hudson and Manning (2019) involve spatial reasoning in images, but they focus on logical reasoning rather than commonsense. Contrast to them, we build a dataset to describe the spatial relationship between people and objects in certain actions with preposition words.

2.2 Knowledge Probing

Early attempts in probing PLMs (Liu et al., 2019a; Hewitt and Manning, 2019) mainly train a classifier on the task of interest with the encoded representations. However, the probing performance is highly influenced by the probe design (Pimentel et al., 2020), thus is hard to reflect the ability of PLMs.

Recently, prompt-based methods (Petroni et al., 2019; Zhou et al., 2020) become more prevalent to study what knowledge PLMs already encode. PLMs take a prompt as input, and generate the continuation (for generative PLMs) or predict masked words (for discriminative PLMs). This does not need additional training, and only a small development set is used to choose optimal prompts and answers (Jiang et al., 2020). In this work, we probe PLMs and VL-PTMs with prompts. Prompt-based methods are also used in model training (Schick and Schütze, 2021; Zhou et al., 2021), while we focus on the knowledge already learned by models.

Basaj et al. (2021); Oleszkiewicz et al. (2021) try to apply the probing methods into the computer vision domain, but they focus on probing representations of visual models. In contrast, we probe ISMs by evaluating the generated images.

3 Benchmark Construction

3.1 Datasets

Size and Height. Inspired by the cognitive discovery (Hersh and Caramazza, 1976) that people tend to categorize objects scales into fuzzy sets, we select 25 common objects in daily life, and categorize them into 5 groups as shown in Table 1a to construct the dataset for size comparison. Typical objects in the former group are smaller than those in the latter group. We form 250 pairs with objects from different groups, like ⟨*ant, bird*⟩, where the first object is smaller than the second in commonsense. Models are asked to compare the size of objects in pairs. To avoid an imbalance of answer distribution, we also consider the reversed pairs like ⟨*bird, ant*⟩, so there are 500 instances in total.

	Size
1	ant, coin, nut, bullet, dice
2	bird, cup, shell, bottle, wallet
3	tyre, chair, microwave, dog, suitcase
4	human, sofa, bookshelf, tiger, bed
5	house, cinema, mountain, truck, plane

(a) Objects of different levels of sizes.

	Height
1	ant, insect, water drop, bullet, dice
2	bird, cup, shoe, bottle, mobile phone
3	table, chair, trash can, sofa, suitcase
4	human, horse, bookshelf, camel, door
5	apartment, theatre, giraffe, truck, street lamp

(b) Objects of different levels of heights.

Table 1: The dataset of object scales.

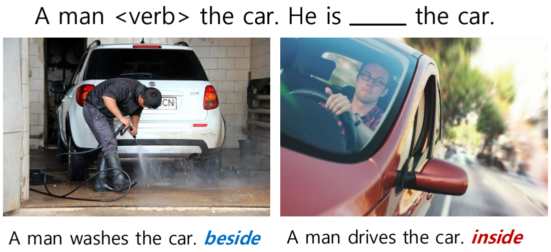

A man <verb> the car. He is _____ the car.
A man washes the car. *beside*
A man drives the car. *inside*

Figure 3: Example of two positional relations between *man* and *car*.

The dataset for comparing objects' heights is constructed similarly, as shown in Table 1b. We also form 500 instances with the objects. The comparison between objects is validated by 5 human annotators for both datasets.

Positional Relationship. The positional relationship dataset consists of human actions regarding objects and the most likely positional relation between the person and the object. We consider four types of positional relations: *above, below, inside, beside*, as they do not overlap with each other.

We select common objects, and write actions between people and the objects. The actions do *not* contain prepositions, like *sit on the chair*. Each object is accompanied by two actions with different positional relations. Take Figure 3 as an example. The man is *beside* the car when washing the car, whereas he is *inside* the car when driving it. Therefore, the relation cannot be easily inferred from collocations between the person and the object. The relations are validated by 5 annotators. The dataset contains 224 instances.

3.2 Probing Tasks

We probe PLMs and VL-PTMs through masked word prediction. Given a text prompt with masks and a set of possible words, a model calculates the probability of each possible word filling the masked position. The word with the highest possibility is regarded as the prediction.

We also probe ISMs through text prompts. The input is a piece of descriptive text, and the output is the image generated by an ISM. We assess the image with two methods as described in 3.3.

PLMs are found to perform poorly in scenarios involving complex reasoning over spatial knowledge (Aroca-Ouellette et al., 2021), and we want to investigate whether they even fail in early stages, like the acquisition of spatial knowledge. So we probe models with simple tasks. In the subtask of size and height, the prompt for PLMs and VL-PTMs is in the form of O_a *is [MASK] than* O_b, where $\langle O_a, O_b\rangle$ is an object pair. The possible answer set is $\{larger, smaller\}$ for size and $\{taller, shorter\}$ for height. The prompt for ISMs is in the form of O_a *and* O_b, and the objects in generated images are compared for size and height.

In the subtask of positional relationship, the prompt for PLMs and VL-PTMs contains an event scenario and a masked token for the positional relationship, like *A woman washes the car. She is [MASK] the car.* The possible answer set is $\{above, below, inside, beside\}$. The prompt for ISMs describes the scenario only, like *A woman washes the car.*

3.3 Evaluation

We assess the images generated by ISMs with two methods. We first use the spatial information of bounding boxes (referred to as ISM (Box)). For each object mentioned in the prompt, we select the classified bounding box with the highest confidence. To mitigate the effect of viewpoint (an object closer to the camera may appear larger in the image), we compute the average depth of the box as the object's depth. The object detector is from Zhang et al. (2021), and the depth estimator is from Godard et al. (2019). When probing the relative size, we compare $area \times depth^2$ of the two objects' boxes; and when probing the relative height, we compare $height \times depth$. When classifying positional relations, we use the mapping rules between spatial relations and image regions from Visual Dependency Grammar (VDG) (Elliott and Keller, 2013). The rules are in Appendix A.1.

Some generated images are vague, and object detection models are trained to process clear pictures, so a number of objects are not recognized. To precisely assess the generated images, we conduct human evaluation on all images (referred to as ISM (Human)). Annotators are asked to compare the size/height of the objects in the images (in the first subtask) and classify the positional relationship between the person and the object (in the second subtask). Each image is evaluated by two annotators, and the average performance is reported. Specifically, we report the accuracy and macro F1 between models' predictions and correct answers. Besides the performance of ISMs on the subset of recognized instances, we also report the performance on the full dataset, giving the unrecognized instances a random guess.

4 Probing Spatial Commonsense

4.1 Models

We take BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b) as examples of text-only PLMs. For VL-PTMs, we choose VinVL (Zhang et al., 2021), which performs well in various vision-language tasks. As it preserves the masked word prediction objective like PLMs, it can also be probed with prompts. We choose VQGAN+CLIP[2] as a representative of ISMs. It uses CLIP (Radford et al., 2021) to guide VQGAN (Esser et al., 2021) to generate images that best match the given text. To make a fair comparison regarding model size, we select BERT-large, RoBERTa-large, and VinVL-large. We use VQGAN with codebook size $Z = 16384$ and downsampling factor $f = 16$, and CLIP with ViT-B/32 (Dosovitskiy et al., 2020) architecture. All four models are of similar sizes.

As language models are sensitive to the expressions in probing (Liu et al., 2021) (like changing

[2] Originated by Ryan Murdoch, @advadnoun on Twitter. Implementation details are in Appendix A.2.

Model	**Acc** (avg. / σ)	**F1** (avg. / σ)
BERT	49.8 / 2.66	47.7 / 2.48
RoBERTa	54.1 / 3.93	52.2 / 6.92
VinVL	**61.8** / 2.47	**54.4** / 3.06
Model	**Acc**	**F1**
Best PLM	54.1 (52.2)	52.2 (46.7)
VinVL	**61.8** (61.6)	54.4 (53.8)
ISM (Box)	54.8 (**81.6**)	**54.8** (**81.6**)
Best PLM	54.1 (52.9)	52.2 (51.0)
VinVL	61.8 (61.6)	54.4 (54.3)
ISM (Human)	**72.7** (**76.5**)	**72.6** (**76.4**)

(a) Comparing sizes of objects. Both objects are recognized by the object detection model in 15% images and are recognized by humans in 86% images.

Model	**Acc** (avg. / σ)	**F1** (avg. / σ)
BERT	50.8 / 2.29	50.3 / 0.25
RoBERTa	50.8 / 6.43	49.2 / 7.45
VinVL	**64.5** / 7.61	**61.5** / 10.5
Model	**Acc**	**F1**
Best PLM	50.8 (48.6)	50.3 (47.9)
VinVL	**64.5** (**69.3**)	**61.5** (65.2)
ISM (Box)	52.5 (68.1)	52.5 (**68.1**)
Best PLM	50.8 (48.5)	50.3 (47.5)
VinVL	64.5 (63.9)	61.5 (60.6)
ISM (Human)	**78.9** (**85.4**)	**78.8** (**85.3**)

(b) Comparing heights of objects. Both objects are recognized by the object detection model in 14% images and are recognized by humans in 82% images.

Table 2: Probing performance on object scales. The numbers are in percentages (%). In the last six lines, the first number is the performance on the whole dataset, and the number in parentheses indicates performance on the subset of instances where the generated images can be recognized by object detection models for lines 4-6, and on the subset recognized by humans for lines 7-9. Standard deviation of models on different folds is represented with σ.

an answer choice from *larger* to *bigger*, the predictions of BERT may differ a lot), we generate new prompts and answers based on those originally designed in the benchmark, and search for the optimal ones for PLMs and VL-PTMs. Similar to Jiang et al. (2020), we use back-translation to generate 10 candidates for prompts and answers, and filter out the repeated ones. To select prompts and answers, we split the dataset into 5 folds, where different folds do not share the same objects. For each run, one fold is used as the development set to choose the best candidate, and the model is probed on other folds with the chosen prompt. We report average performance of 5 runs.

4.2 Probing Results

Size and Height. Table 2 reports the probing performance of comparing the scales of objects. We also demonstrate probing results on RelativeSize (Bagherinezhad et al., 2016) in Appendix B. We observe that PLMs perform similarly. Even the best PLMs are slightly better than random guesses, indicating they are ineffective in predicting object scales. Although RoBERTa is trained on more texts and assumed to encode more knowledge, its performance is similar to BERT's. It shows that PLMs do not learn much spatial commonsense from texts even if the pretrained corpus greatly increases.

With the help of visual features in pretraining, VinVL greatly outperforms PLMs. ISM (Box), which simply compares bounding boxes in images generated by the ISM, also outperforms PLMs. Since only a small portion of instances are recognized with bounding boxes, if we only consider the predictions on these instances, the gap between ISM (Box) and PLMs is more than 15%. These indicate that models with visual signals learn accurate spatial commonsense knowledge from images.

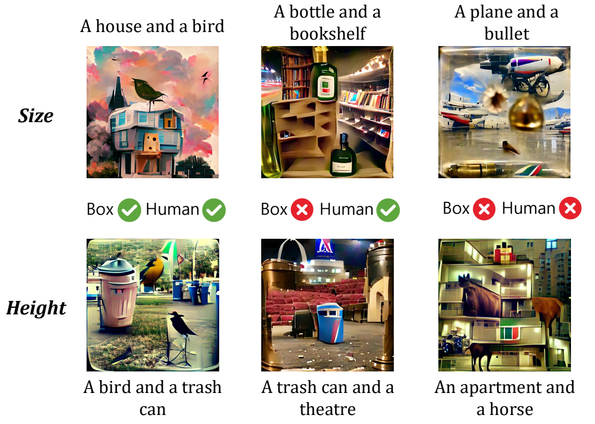


Figure 4: Images generated by ISM in scale comparison. Objects are successfully recognized by both the object detection model and humans in the left column, by humans but not the object detection model in the middle column, and by neither of them in the right column.

ISM (Box) performs better than VinVL on those recognizable instances, but underperforms on the whole dataset. We conduct human evaluation on the generated images for more precise assessment. More than 80% of images are recognized by humans and these images accurately reflect the spatial commonsense compared to PLMs and VinVL. [3] The gap between VinVL and ISM (Human) may

[3] The agreement between annotators is more than 90%.

Model	**Acc** (avg. / σ)	**F1** (avg. / σ)
BERT	26.1 / 4.15	19.0 / 5.20
RoBERTa	31.0 / 15.4	20.1 / 9.29
VinVL	**56.1** / 7.09	**41.8** / 6.69
Model	**Acc**	**F1**
Best PLM	31.0 (32.5)	20.1 (17.6)
VinVL	**56.1** (**56.0**)	**41.8** (**36.0**)
ISM (Box)	33.0 (42.5)	26.5 (26.1)
Best PLM	31.0 (30.5)	20.1 (20.1)
VinVL	56.1 (56.4)	41.8 (42.9)
ISM (Human)	**73.4** (**75.4**)	**65.1** (**68.0**)

Table 3: Probing performance on positional relationship (%). The symbols are identical to those in Table 2. Both the person and the object are recognized with bounding boxes in 39% images and by humans in 93% images.

come from different ways of using visual signals in pretraining. A main training objective of VinVL, and other VL-PTMs, is aligning text with image regions. The discriminative features of objects are amplified, while other features may not receive as much attention. For instance, the shape and color are the discriminative features of an *apple*, and its size is not that important in recognition. In image synthesis, models need to learn comprehensive knowledge of objects in order to reconstruct them, and spatial knowledge may be learned implicitly in this process.

Figure 4 demonstrates images generated by the ISM given the prompts of object pairs. ISM grasps the main characteristics of the objects, including their scales. Some objects (like *theatre* at the bottom of the middle column) can be identified by humans but are difficult for the object detection model because they are obstructed by objects in the foreground. And some objects are generated in multiple fragments (like *plane* and *horse* in the right column), therefore cannot be recognized by either the object detection model or humans.

Positional Relationship. The probing performance on positional relationship is shown in Table 3. VinVL outperforms PLMs more than 20%, and ISM (Human) outperforms PLMs more than 35%, suggesting that models with visual signals learn more knowledge of the scenarios, especially the positions of objects relative to people.

The gap between PLMs and ISM (Box) is smaller compared to the gap in the subtask of size and height. One reason is that the rules defined in VDG cannot perfectly reflect the true positional relationship in images. For example, the man is *beside* the car in the left image of Figure 3, but he will be regarded as *inside* the car by the rules, as the region of car covers the region of man.

Model	Size		Height	
	Sym.	Trans.	Sym.	Trans.
Best PLM	37.5	71.9	25.9	73.1
VinVL	**43.5**	**95.0**	**43.0**	**93.2**
Best PLM†	36.6	72.2	26.1	72.3
VinVL†	44.4	**95.3**	32.2	**97.8**
ISM (Human)†	**82.5**	81.1	**83.2**	85.2

Table 4: The percentage (%) of predictions that meet consistency. Sym and Trans indicate symmetry and transitivity. † indicates performance on the subset of images recognized by humans.

Text-based PLMs tend to lean towards certain positional relations between a person and an object, without referring to the action. In 64% cases, RoBERTa chooses the same option for a ⟨*person, object*⟩ pair with different actions, while the proportion is 21% for VinVL, and 28% for ISM (Human).

5 Quality of Spatial Knowledge

5.1 Consistency

Models that master better spatial knowledge should be able to infer the relative scale of two objects from intermediate references. For example, if a model knows *a dog is larger than an ant* and *a sofa is larger than a dog*, it may learn *a sofa is larger than an ant*, even if it has not seen *sofa* and *ant* together. We inspect models on how consistent their probing results are.

The consistency is measured in two aspects: *symmetry* and *transitivity*. Symmetry implies that if a model predicts $A > B$, then it should also predict $B < A$, and vice versa: $A < B \implies B > A$. Here $>$ and $<$ are in terms of size or height. We enumerate the object pairs and count the percentage of predictions that meet the symmetry criterion. Transitivity implies that if a model predicts $A > B$ and $B > C$, then it should predict $A > C$. It also works for $<$, $A < B \land B < C \implies A < C$. We enumerate the triples $\langle A, B, C\rangle$ where the predicted relation between $\langle A, B\rangle$ is identical to the prediction between $\langle B, C\rangle$, and count the percentage that the prediction between $\langle A, C\rangle$ meets the transitivity criterion. Note that we only evaluate whether the predictions are consistent with each other, regardless of the gold answers.

We evaluate the consistency of predictions from PLMs that perform the best in the probing tasks

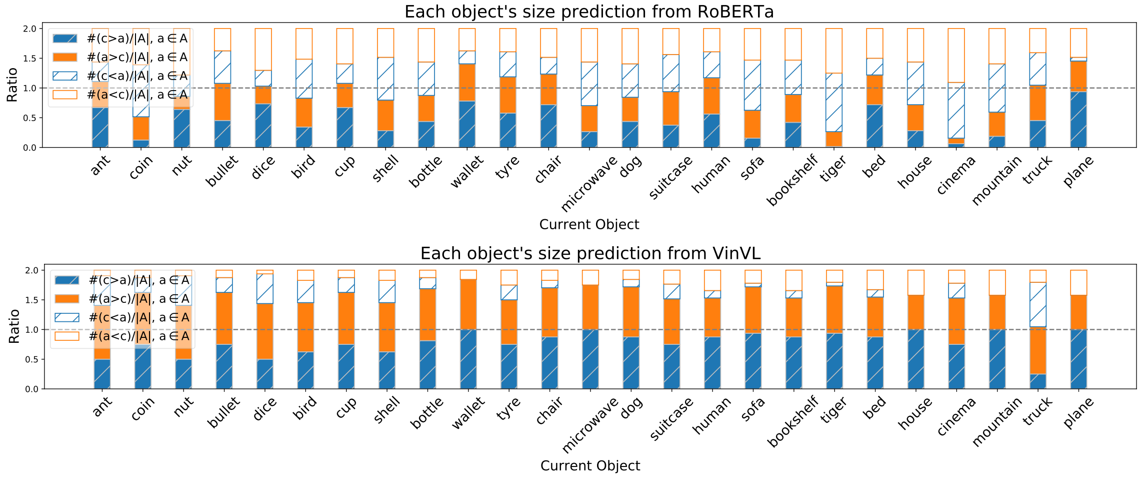


Figure 5: Predictions from RoBERTa and VinVL in the subtask of objects' sizes. c is the current object and A is the set of all other comparable objects. $\#(c > a)/|A|$ indicates the ratio of predicting the current object larger than others. As $c > a$ and $a > c$ should not appear simultaneously, the sum of the two solid bars is expected to be 1.

(RoBERTa for size and BERT for height), VinVL, and ISM (Human). The results are in Table 4.

VinVL outperforms the best PLM in both metrics, and the characteristics of them are similar: the transitive consistency is high, while the symmetric consistency is low. To further analyze this phenomenon, we exhibit each object's size predictions from RoBERTa and VinVL in Figure 5. The models exhibit different behaviors in recognizing object scales. As the objects (X-axis of Figure 5) are roughly listed from smaller to larger groups, the bottom blue bars are expected to follow a non-descending order from left to right, and the solid orange bars should be non-ascending. The predictions of VinVL are generally in line with this trend, while RoBERTa's predictions are disordered. For example, *ant* is predicted to be *larger than* other objects with high probability, and *cinema is larger than others* is unlikely to happen. On the other hand, if the model predictions are consistent, the two solid bars should sum to 1. However, the sum is far above 1 for most objects in VinVL's predictions. This bias towards words indicating the choice of *large* may come from the pretraining corpus. For example, *sofa* occurs twice as many times with words indicating large as with words indicating small in COCO (Lin et al., 2014), one of VinVL's pretraining datasets.

ISM's predictions comply with the symmetry criterion, outperforming other models by 40%, while also having good transitive consistency. The knowledge probed from ISM is more consistent. More images generated by ISM are in Appendix C.

Model	**Acc** (avg. / σ)	**F1** (avg. / σ)
BERT	27.4 / 3.17	19.7 / 7.25
RoBERTa	29.5 / 16.0	20.1 / 9.90
VinVL	**58.1** / 1.97	**44.4** / 1.63
Model	**Acc**	**F1**
Best PLM	29.5 (28.4)	20.1 (19.1)
VinVL	58.1 (52.3)	44.4 (41.0)
ISM (Human)	**66.5** (**74.8**)	**59.4** (**69.2**)

Table 5: Probing models on the generalized dataset of positional relationship. The symbols are identical to those in Table 2. The human recognition ratio is 81%.

5.2 Generalizability

ISM may learn positional relations from training images directly. For example, *a boy riding a bicycle* is a *common* scenario and may frequently exist in ISM's training set, so models can generate images more easily when being fed with the text prompts like *a boy rides a bicycle*. To further challenge ISM's capability, we make a generalized version of our original positional relationship dataset. It is designed to examine whether models are able to robustly reflect the spatial commonsense knowledge when facing *uncommon* scenarios.

A generalized scenario is built upon the original one by replacing the person and object in the text prompts. We select the new person and new object from the subterms of the original ones (those with *IsA* relation in ConceptNet (Speer et al., 2017), like *enchantress* is a *woman*). To ensure these newly constructed scenarios are not likely to appear in the training data of models, we search them in

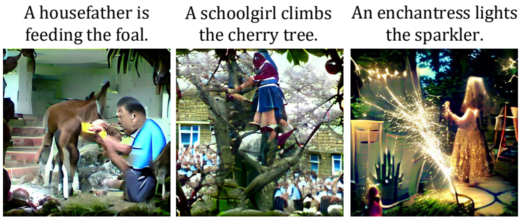

A housefather is feeding the foal.
A schoolgirl climbs the cherry tree.
An enchantress lights the sparkler.

Figure 6: Images generated by ISM with the generalized prompts.

Model	Size		Height		PosRel	
	Acc	F1	Acc	F1	Acc	F1
UnifiedQA	51.3	38.5	58.4	52.8	56.7	48.1
ISM w/ VinVL	**52.4**	**43.8**	**59.4**	**54.3**	**59.8**	**58.7**

Table 6: Performance of answering commonsense questions. Accuracy (%) and macro F1 (%) are reported. PosRel refers to positional relationship.

BookCorpus (Zhu et al., 2015) and remove the scenarios that have appeared. The newly generated scenarios are also validated by humans to ensure that they are reasonable.

Results of probing PLMs, VinVL, and ISM[4] on the generalized dataset are in Table 5. PLMs and VinVL achieve similar performance on both the generalized dataset and the original one, indicating that they behave robustly when facing uncommon scenarios. The performance gap between other models and ISM (Human) slightly narrows down, but ISM (Human) still outperforms VinVL more than 8%. Figure 6 exhibits images generated by ISM with the generalized prompts. Although it is difficult for ISM to generate unfamiliar objects, it is still capable of capturing the positional relations.

6 Solving Natural Language Questions

We investigate how to acquire spatial knowledge from ISMs and whether the knowledge is effective in natural language understanding scenarios. To our best knowledge, there is no appropriate task that focuses on spatial commonsense, so we create a toy task by transforming our probing benchmark into the form of question answering (QA).

Dataset. We construct a QA dataset of yes/no questions. Questions of objects' sizes are in the form of *Is O_a larger/smaller than O_b?* And questions of objects' heights are like *Is O_a taller/shorter than O_b?*, where O_a and O_b are two objects. Questions about positional relationship are accompanied with the action: for instance, *A man washes the car. Is the man inside the car?* To avoid bias in answer distribution, the numbers of *yes* and *no* are equal in gold answers. There are 500 questions for size, 500 for height, and 448 for positional relationship.

Models. We use VinVL-base together with our image synthesis model VQGAN+CLIP to answer spatial commonsense questions. The VinVL here is finetuned on the VQA (Goyal et al., 2017) task. It takes images generated from ISM with textual prompts from questions, and predicts the answer based on the question and image together. Note that the VQA training corpus does not contain commonsense reasoning questions.

We choose UnifiedQA (Khashabi et al., 2020) as a text-based QA model for comparison. Based on the pretrained T5 model (Raffel et al., 2019), UnifiedQA is continually trained on various QA tasks, including three yes/no datasets. We use UnifiedQA-large, which is comparable with our synthesis and reasoning model (ISM w/ VinVL) in size.

Results. As shown in Table 6, ISM w/ VinVL outperforms UnifiedQA on all subtasks, showing that spatial knowledge from ISMs can be directly used by vision-language models without additional training. Although some images cannot be precisely recognized by object detection models, vision-language models may find regions that are related to the objects mentioned in questions, and make decisions based on the features of these regions. The results on the simple natural language task show that it is beneficial to tackle natural language tasks with vision-language methods, and ISMs can be *a bridge between the two modalities*. With the development of ISMs and object detection techniques, we believe the generated images will help more.

7 Conclusion

We propose a new spatial commonsense probing framework to investigate object scales and positional relationship knowledge in text-based pretrained models and models with visual signals. Experimental results show that models with visual signals, especially ISMs, learn more accurate and consistent spatial commonsense than text-only models. Integrating ISMs with visual reasoning models outperforms PLMs in answering spatial questions. This manifests the potential of using spatial knowledge from ISMs in natural language reasoning.

[4] We do not consider ISM (Box) because many new objects we used are unfamiliar to object detection models. Only 17% of the objects are in the object detection classes.

References

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning of objects through space and time. *arXiv preprint arXiv:2106.03634*.

Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. Are elephants bigger than butterflies? reasoning about sizes of objects. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Dominika Basaj, Witold Oleszkiewicz, Igor Sieradzki, Michał Górszczak, B Rychalska, T Trzcinski, and B Zielinski. 2021. Explaining self-supervised image representations with visual probing. In *International Joint Conference on Artificial Intelligence*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *ICLR*.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-second AAAI conference on artificial intelligence*.

Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1178.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Tyrone Donnon, Jean-Gaston DesCôteaux, and Claudio Violato. 2005. Impact of cognitive imaging and sex differences on the development of laparoscopic suturing skills. *Canadian journal of surgery*, 48(5):387.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Harry M Hersh and Alfonso Caramazza. 1976. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General*, 105(3):254.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

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

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1896–1907.

Benjamin Kuipers et al. 1990. Commonsense knowledge of space: Learning from experience. *Advances in Spatial Reasoning*, 2:199.

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.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319.

Witold Oleszkiewicz, Dominika Basaj, Igor Sieradzki, Michał Górszczak, Barbara Rychalska, Koryna Lewandowska, Tomasz Trzciński, and Bartosz Zieliński. 2021. Visual probing: Cognitive framework for explaining self-supervised image representations. *arXiv preprint arXiv:2106.11054*.

James W Pellegrino, David L Alderton, and Valerie J Shute. 1984. Understanding spatial ability. *Educational psychologist*, 19(4):239–253.

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

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Carla Poole, Susan A Miller, and Ellen Booth Church. 2006. Development: Ages & stages–spatial awareness. *Early Childhood Today*, 20(6):25–30.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Arnaud Saj and Koviljka Barisnikov. 2015. Influence of spatial perception abilities on reading in school-age children. *Cogent Psychology*, 2(1):1049736.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 292–299.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Implementation Details

Relation	Definition
X *inside* Y	The entirety of region X overlaps with Y.
X *beside* Y	The angle between the centroid of X and the centroid of Y lies between 315° and 45° or 135° and 225°.
X *above* Y	The angle between X and Y lies between 225° and 315°.
X *below* Y	The angle between X and Y lies between 45° and 135°.

Table 7: Spatial relations between image regions in Visual Dependency Grammar (VDG).

A.1 Spatial Relations in Visual Dependency Grammar

We use the rules defined in Visual Dependency Grammar (Elliott and Keller, 2013) to determine the positional relationship between bounding boxes. The rules used are listed in Table 7. If two bounding boxes meet the *inside* standard, they will be predicted as *inside*. Otherwise, the angle between the centers of the boxes is calculated to determine whether the prediction is *above*, *below*, or *beside*.

A.2 Image Synthesis

We generate images of 512×512 pixels with text prompts. We use 1) VQGAN (Esser et al., 2021), which takes in a vector, and outputs a high-resolution image; and 2) CLIP (Radford et al., 2021), which can encode both text and images, and map them into a multi-modal embedding space. Image synthesis is the process of finding the optimal vector $\boldsymbol{v}$ inputted to VQGAN. In each iteration, the vector is fed into VQGAN to generate an image $img = \text{VQGAN}(\boldsymbol{v})$. CLIP encodes the image into $\boldsymbol{c} = \text{CLIP}(img)$, and encodes the text prompt into $\boldsymbol{t} = \text{CLIP}(text)$, respectively.

The optimization goal is to bring $\boldsymbol{c}$ and $\boldsymbol{t}$, the representation of the image and text encoded by CLIP closer. The vector $\boldsymbol{v}$ is randomly initialized and optimized for 600 iterations. We use Adam optimizer with a learning rate of 0.5. This process looks like a normal model "training", but here both VQGAN and CLIP are pretrained and their parameters are frozen; only the vector $\boldsymbol{v}$ is optimized from randomness for every prompt.

A.3 Prompt Candidates Generation

When probing PLMs, we follow Jiang et al. (2020) to generate prompt and answer candidates with back-translation. Manually designed prompts and answers are translated from English to German and then backward. It is used to construct candidates with similar meanings. We leverage the translation model designed in Ng et al. (2019).

Model	**Acc** (avg. / σ)	**F1** (avg. / σ)
BERT	49.0 / 4.11	43.7 / 8.25
RoBERTa	48.9 / 1.71	43.4 / 5.42
VinVL	**60.6** / 1.47	**51.2** / 2.22
Model	**Acc**	**F1**
Best PLM	49.0 (47.5)	43.7 (40.5)
VinVL	**60.6** (60.8)	51.2 (49.8)
ISM (Box)	58.5 (**71.5**)	**58.5** (**71.4**)
Best PLM	49.0 (48.5)	43.7 (43.5)
VinVL	60.6 (65.5)	51.2 (55.7)
ISM (Human)	**72.5** (**76.5**)	**71.8** (**75.7**)

Table 8: Probing performance on RelatizeSize. Accuracy and macro F1 are reported. The numbers are in percentages (%). In the last six lines, the first number is the performance on the whole dataset, and the number in parentheses indicates performance on the subset of instances where the generated images can be recognized by object detection models and humans, respectively. The standard deviation on different folds is represented with σ. Both objects are recognized with bounding boxes in 40% images and are recognized by humans in 85% images.

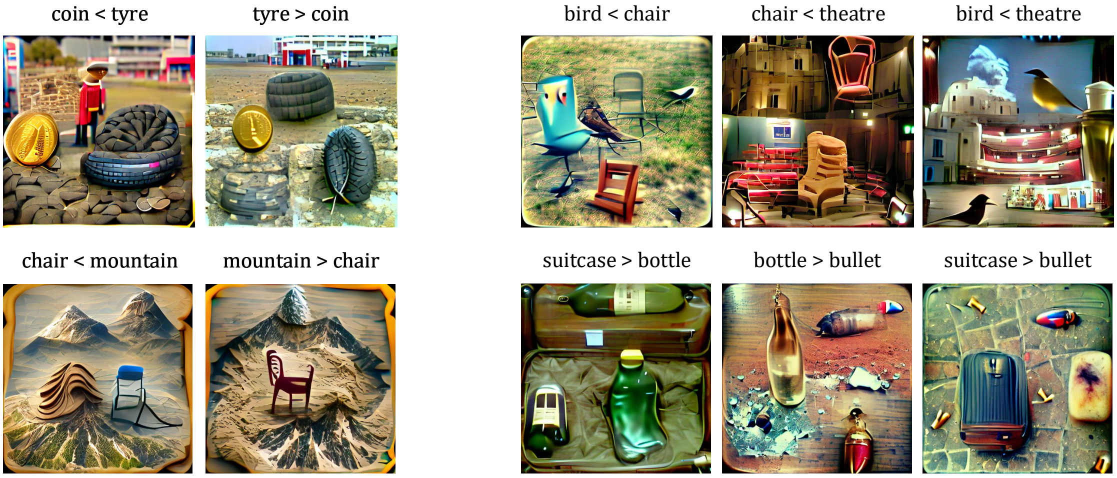
A.4 Computing Infrastructure

Experiments are conducted on NVIDIA GeForce RTX 3090 GPU. It takes 8 hours to generate 500 images on one GPU, and all other experiments can be executed in a few minutes.

B Probing Results on RelativeSize

RelativeSize (Bagherinezhad et al., 2016) is another dataset for comparing objects' sizes. Table 8 demonstrates the probing results on it. The results are consistent with those on our datasets: ISM probing, both evaluated with bounding boxes and evaluated by humans, outperforms PLM probing.

The methods used in Bagherinezhad et al. (2016) are all retrieval-based. They execute search engine queries and download images from Flickr to make the comparisons. So we do not compare with their results directly. However, it is worth noticing that our ISM probing is comparable to the image retrieval-based baseline (its accuracy is 72.4%). It exhibits that ISM learns sufficient knowledge from images.



(a) Two groups of generated images. The sizes of objects are consistent with each other.

(b) Two groups of generated images. The heights of objects meet the transitivity criterion.

Figure 7: Examples of the symmetric and transitive consistency of images generated by ISM.

C Consistency of Images Generated by ISM

Figure 7 exhibits the symmetric and transitive consistency of images generated by ISM. In Figure 7a, the relationship between the sizes of objects is consistent; in Figure 7b, objects' heights comply with the transitivity criterion. The consistency of scale knowledge makes the predictions more convincing, and gives models a chance to learn new comparisons between objects.