Do Vision-and-Language Transformers Learn Grounded Predicate-Noun Dependencies?

Mitja Nikolaus[1,2]
mitja.nikolaus@univ-amu.fr

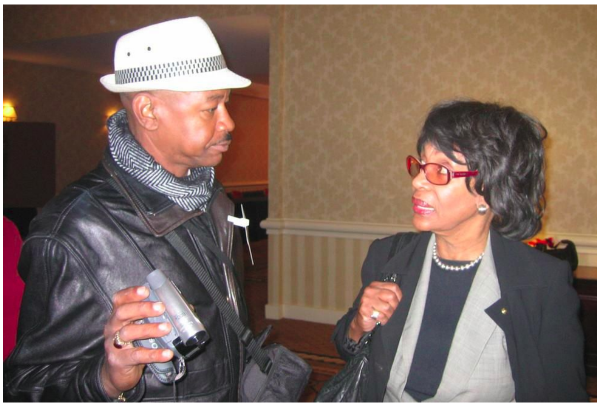
Emmanuelle Salin[1] and **Stephane Ayache**[1] and **Abdellah Fourtassi**[1] and **Benoit Favre**[1]

[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2]Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France

Abstract

Recent advances in vision-and-language modeling have seen the development of Transformer architectures that achieve remarkable performance on multimodal reasoning tasks. Yet, the exact capabilities of these black-box models are still poorly understood. While much of previous work has focused on studying their ability to learn meaning at the word-level, their ability to track syntactic dependencies between words has received less attention.

We take a first step in closing this gap by creating a new multimodal task targeted at evaluating understanding of predicate-noun dependencies in a controlled setup. We evaluate a range of state-of-the-art models and find that their performance on the task varies considerably, with some models performing relatively well and others at chance level. In an effort to explain this variability, our analyses indicate that the quality (and not only sheer quantity) of pretraining data is essential. Additionally, the best performing models leverage fine-grained multimodal pretraining objectives in addition to the standard image-text matching objectives. This study highlights that targeted and controlled evaluations are a crucial step for a precise and rigorous test of the multimodal knowledge of vision-and-language models.



Target sentence:
A man is wearing a hat.

Distractor Sentence:
A man is wearing glasses.

Figure 1: We evaluate V&L models on their ability to track predicate-noun dependencies that require a joint understanding of the linguistic and visual modalities. The task is to find the correct sentence (choosing between the target and distractor) that corresponds to the scene in the image. In this example, the models should connect the predicate "is wearing a hat" to "man". A model that does not track dependencies would judge the distractor sentence "A man is wearing glasses" as equally likely, as there is a man is the image, as well as a person that is wearing glasses.

1 Introduction

Vision-and-language (V&L) models have recently shown substantial improvement on a range of multimodal reasoning tasks. Taking inspiration from successes in text-only Natural Language Processing (Devlin et al., 2019; Brown et al., 2020), state-of-the-art V&L models are usually composed of a Transformer-based architecture pre-trained in a self-supervised manner on large-scale data, and then fine-tuned on downstream tasks.

While these models show remarkable performance on a range of tasks, more controlled and systematic analyses are necessary in order to obtain a better understanding of their exact multimodal knowledge.

A range of studies has investigated their ability to map words to their visual referents for nouns (Kazemzadeh et al., 2014; Mao et al., 2016; Shekhar et al., 2017) and verbs (Ronchi and Perona, 2015; Yatskar et al., 2016; Pratt et al., 2020; Hendricks and Nematzadeh, 2021), but there are only a few studies on whether recent V&L models can capture multimodal syntactic dependencies between words and concepts.

In this paper, we explore how well V&L models learn predicate-noun dependencies across modalities (see example in Figure 1). To this end, we create an evaluation set that contains carefully selected images and pairs of sentences with minimal

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1538–1555
December 7-11, 2022 ©2022 Association for Computational Linguistics

differences. Given an image and two predicate-noun sentences, the models need to find the correct sentence corresponding to the image. Crucially, they can only succeed by taking into account the dependencies between the visual concepts in the image corresponding to the noun and predicate in the sentence.

As it has been shown that visual reasoning performance in several tasks can be spuriously augmented by capitalizing on textual biases in the training data (Goyal et al., 2017; Agrawal et al., 2018; Hendricks et al., 2018; Cao et al., 2020), we counter-balance our evaluation dataset in a way that controls for such exploitation of linguistic biases.

We evaluate pre-trained state-of-the-art V&L models in a zero-shot setting and find that the ability to track predicate-noun dependencies varies considerably from model to model. Of all models tested, UNITER (Chen et al., 2019) and LXMERT (Tan and Bansal, 2019) show the highest scores, but their performance is still far from optimal. Other models such as ViLBERT (Lu et al., 2019) and CLIP (Radford et al., 2021) perform at chance level. We discuss how differences in the models could explain their performance variability, highlighting the role of pretraining data quality and fine-grained multimodal pretraining objectives.

Code to reproduce the analyses and run the evaluation on new models is publicly available at `https://github.com/mitjanikolaus/multimodal-predicate-noun-dependencies`.

2 Related Work

Targeted evaluation of V&L models Recently, a growing number of tasks have been created for targeted evaluation of V&L models' abilities to perform various multimodal reasoning.

Shekhar et al. (2017) create sets of distractor captions to analyze whether V&L models are sensitive to single word replacements (with a focus on nouns). Similar targeted evaluation datasets have also been proposed for referring expressions (Chen et al., 2020), image-sentence matching (Hu et al., 2019), and Visual Question Answering (VQA; Bogin et al., 2021), with a focus on compositional reasoning.

Tasks such as visual semantic role labeling or situation recognition, typically involve classifying the primary activity depicted in an image, as well as the semantic roles of involved entities (Ronchi and Perona, 2015; Lu et al., 2016; Chao et al., 2015; Gupta and Malik, 2015; Yatskar et al., 2016; Pratt et al., 2020). While these studies demonstrate that V&L models can learn semantic roles to some degree in a supervised learning setup, such tasks do not allow for a controlled evaluation of models in a zero-shot setting.

In Hendricks and Nematzadeh (2021), the authors evaluate state-of-the-art V&L models in a controlled zero-shot setup and find that they still have more trouble understanding verbs compared to subjects or objects. They also observe that models trained on larger datasets with less descriptive captions perform worse than models trained on smaller, manually-annotated datasets.

Several works have also tried to shed more light on the precise multimodal semantic capabilities of V&L models using probing techniques. Salin et al. (2022) show that although state-of-the-art V&L models can grasp some multimodal concepts such as color, they still do not fully understand more difficult concepts such as object size and position in the image. Parcalabescu et al. (2021) use probing to demonstrate that such models still lack the capability to correctly count entities in an image.

Evaluation of grounded syntax Akula et al. (2020) tests for sensitivity to word order in referring expressions. Similarly, Thrush et al. (2022) studies the ability of V&L models to take word order into account by designing adversarial examples that require differentiating between similar image and text pairs, while the text pairs only differ in their word order. Their results suggest that state-of-the art models still lack precise compositional reasoning abilities.

Li et al. (2020a) studies so-called *syntactic grounding* of VisualBERT. They show that certain attention heads of the transformer architecture attend to entities that are connected via syntactic dependency relationships. However, such probing experiments do not necessarily indicate to what degree a model is actually *using* the encoded information when making predictions.

In our work, we test a range of state-of-the-art models specifically on their ability to track predicate-noun dependencies. Crucially, we test the models in a much more controlled setting compared to previous work: Our setup involves visual distractors as well as control task, disentangling the challenge of understanding syntactic dependencies from more simple object and predicate recognition. Additionally, we strictly control for any possible linguistic bias by counter-balancing all evaluation examples.

3 Methods

3.1 Evaluation Dataset

We construct an evaluation dataset that is suited for evaluating the sensitivity to visually grounded predicate-noun dependencies in a zero-shot setup.

The data consists of pairs of triplets, and each triplet consists of an Image I, a target sentence S_1, and a distractor sentence S_2. Target and distractor sentences are minimal pairs, i.e. one sentence differs from the other only with regard to either the noun (e.g., “A girl is sitting.” vs. “A man is sitting.”, Figure 2) or the predicate (e.g., “A man is wearing a hat.” vs. “A man is wearing glasses.”, Figure 1).

Crucially, the images always contain visual distractors, meaning that both the noun and the predicate of the distractor sentence are present in the image, but they do not have a noun-predicate relationship (e.g., for the distractor sentence “A man is wearing glasses”, there is a man in the image, who is not wearing glasses, and a person wearing glasses, who is not a man). Thus, it is necessary to take into account the *dependency* between noun and predicate to distinguish between the target and distractor sentence (Figure 1).

Controlling for linguistic biases V&L models have shown to rely sometimes on textual bias instead of using visual information (Goyal et al., 2017; Agrawal et al., 2018; Hendricks et al., 2018; Cao et al., 2020). For example, if a training dataset contains more often the phrase “a girl is sitting” than “a man is sitting”, a model might prefer the caption “a girl is sitting” during evaluation only based on linguistic co-occurrence heuristics, irrespective of the visual content. In our evaluation dataset, we control for potential linguistic biases in the training datasets by pairing every triplet with a corresponding counter-balanced example where target and distractor sentence are flipped. More specifically, for every triplet (I_1, S_1, S_2), there exists a corresponding triplet (I_2, S_2, S_1), as depicted in Figure 2. In that way, a model that does not take into account the visual modality cannot succeed in the task (see also Nikolaus and Fourtassi, 2021).

Automatic pre-filtering Our evaluation dataset is based on Open Images (Kuznetsova et al., 2020). We pre-filter the images based on existing human-annotated object and relationship labels and bounding boxes. The objects refer to persons, animals, as well as inanimate objects. The relationships can either describe an action that an object is engaged in (e.g., WOMAN - SIT), or an action linking two objects (e.g., MAN - WEAR - GLASSES). All nouns in the selected relationships for our dataset refer to persons, due to lack of sufficient annotations for other kinds of agents.

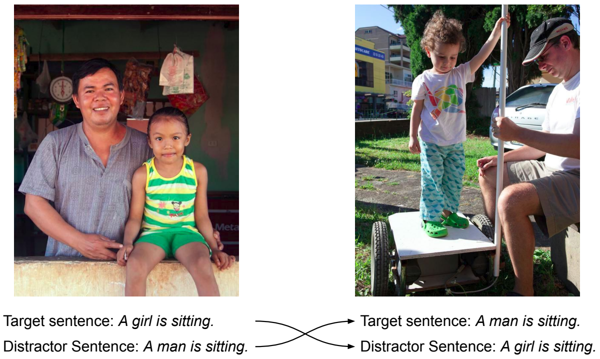

Target sentence: *A girl is sitting.*
Distractor Sentence: *A man is sitting.*
Target sentence: *A man is sitting.*
Distractor Sentence: *A girl is sitting.*

Figure 2: Counter-balanced evaluation: Each triplet has a corresponding counter-example, where target and distractor sentence are flipped.

We look for images that contain a target object-relationship pair as well as a distractor object-relationship pair for which either the target and distractor object are the same, but the relationships differ, or vice versa (as in the example in Figure 1). Additional details on the pre-filtering can be found in Appendix A.1.

Manual selection We manually select suitable images after the automated pre-filtering, in order to ensure high quality of each example and in particular to verify that the distractor sentences are indeed incorrect given the images. This step is crucial, because many of the annotations in Open Images are incomplete, and an image may contain, for example, a woman that is sitting but not annotated as such (in this case, we disregard the image for our evaluation set).

We select pairs of examples and counter-examples and ensure that there are no duplicate images within the set of images for each object-relationship pair.

Sentence generation We generate target and distractor sentences based on the verified object and relationship annotations from Open Images.

We construct English sentences using a template-based approach. Given an object and a relationship, we add the indefinite article (a/an) in front of each noun and use all verbs in present progressive tense as this is most frequent in image-text datasets.[1] For example, from WOMAN - IS - SIT we generate "a woman is sitting."; and from MAN - HOLD - CAMERA "a man is holding a camera.".

This template-based approach is necessary for our controlled evaluation. As the choice of the exact template for the construction of the sentences may influence the results[2], we evaluate the models, additionally, using a slightly different template, and we show that the overall result patterns remain largely similar (see Appendix A.4.2).

Final evaluation set The final evaluation set contains 2584 triplets. For 1486 of these triplets, the distractor sentence contains an incorrect predicate and for the other 1098 triplets, the distractor contains an incorrect noun. More detailed statistics regarding the number of triplets concerning specific concepts are provided in Appendix A.2.

A note on perceived gender annotations Our evaluation dataset uses annotations from the Open Images dataset, which rely on the physical appearance of persons to annotate their perceived gender. We use the provided annotations, and the resulting biases are unfortunately reproduced in our evaluation set. We discuss this issue in further detail in the Ethics Statement (Section 8).

In Salminen et al. (2018) gender classification from face pictures by human annotators shows an inter annotator agreement greater than 95%. True gender cannot be classified, and high inter-annotator agreement does not imply a correct gender choice, but we expect the gender annotations of Open Images to be reliable enough to be used as a basis for our analyses.

3.2 Metric

We evaluate pre-trained models on their image-text matching performance in a zero-shot setting, i.e. without any further training. For each triplet, we test whether the models give a higher similarity score for the correct sentence than for the distractor sentence. We calculate accuracy for each pair, i.e. the model needs to succeed for both the example and the counter-balanced example triplet.

For each pair of triplets $(t_1, t_2) = ([I_1, S_1, S_2], [I_2, S_2, S_1])$, we calculate the following score:

$$f(t_1, t_2) = \begin{cases} 1, & \text{if } s(I_1, S_1) > s(I_1, S_2) \\ & \text{and } s(I_2, S_2) > s(I_2, S_1) \\ 0, & \text{otherwise} \end{cases}$$

where $s(I, S)$ denotes the similarity between an image I and a sentence S. To obtain the similarity score, we use the softmaxed output of the image-text matching pretraining heads of the models.[3]

The final accuracy is the average score over all pairs in the evaluation set. Chance performance is at 25%.[4]

As the dataset was manually filtered and requires only rather simple understanding of the images, we assume human performance to be close to 100%. To verify this claim, we had a one person annotate a randomly sampled subset of 500 triplets. For each triplet, the annotator was asked to judge which of the two sentences describes the image better. The resulting performance was at 100%.

A topline: the cropped task In order to explore the effect of the visual distractors on this noun-predicate dependency task, we additionally evaluate all models in a cropped task: We reduce the image to the bounding box of the target object. Thus, the cropped image usually[5] only contains the target object, and no more visual distractors (i.e., the referent of the noun or the predicate in the distractor sentence is no longer present in the cropped image). To succeed at this (simpler) task, the model no longer needs to capture the predicate-noun dependency, it just needs to ground the single words correctly. We use this task to estimate how much the performance of the models is affected by the ability to ground nouns and predicates in our evaluation dataset, in comparison to the (more

[1] In cases where multiple connecting predicates between a verb and a noun are plausible (e.g. "a man wearing glasses" vs. "a man with glasses"), we choose the construction that occurs most frequently in the Conceptual Captions training data (Sharma et al., 2018). This dataset is most commonly used for training V&L transformers.

[2] For example, Ravichander et al. (2020) found that results of some probing experiments can vary substantially with slight changes in wording.

[3] For the model CLIP, we feed the image and both sentences at the same time, and obtain a similarity score for both sentences, where $s(I_1, S_1) = 1 - s(I_1, S_2)$.

[4] The model succeeds if the similarity scores fall into one of four possible configurations: $s(I_1, S_1) > s(I_1, S_2) \wedge s(I_2, S_1) > s(I_2, S_2)$; $s(I_1, S_1) < s(I_1, S_2) \wedge s(I_2, S_1) < s(I_2, S_2)$; $s(I_1, S_1) > s(I_1, S_2) \wedge s(I_2, S_1) < s(I_2, S_2)$; $s(I_1, S_1) < s(I_1, S_2) \wedge s(I_2, S_1) > s(I_2, S_2)$.

[5] If the bounding boxes of the target and visual distractor object overlap to a high degree, the cropped image might still contain (parts of) the distractor object.

sophisticated) ability of understanding predicate-noun *dependencies*.

3.3 Models

We consider a range of state-of-the-art V&L models that are pre-trained using text, image, and multimodal pretraining objectives on corpora of parallel image and text data. All models use the transformer architecture (Vaswani et al., 2017), but vary in terms of pretraining data and objectives, image encoders, and multimodal fusion approaches.

In addition to their image and text pretraining objectives, the models commonly make use of an image-text matching objective, where the models are asked to predict whether a given sentence describes an image or not. We leverage the output of the corresponding pretraining head for calculating image-text similarities for our task.[6]

We evaluate LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2019), ViLBERT (Lu et al., 2019), Oscar (Li et al., 2020b), VinVL (Zhang et al., 2021), ViLT (Kim et al., 2021), and CLIP (Radford et al., 2021). We could not evaluate the original VL-BERT (Su et al., 2020), because it was not pre-trained using an image-text matching loss. We also did not evaluate the original VisualBERT (Li et al., 2019), as their implementation of the image-text matching loss requires one correct caption (in addition to a possibly faulty caption), which is not available for the Open Images dataset. Both models were however evaluated in the controlled conditions using VOLTA (see Section 4.2).

Pretraining datasets ViLBERT and VL-BERT are pretrained on Conceptual Captions (Sharma et al., 2018). UNITER, LXMERT, ViLT, and VinVL[7] make use of additional publicly available datasets such as COCO (Lin et al., 2014), SBU captions (Ordonez et al., 2011), Flickr30K (Young et al., 2014), VisualGenome (Krishna et al., 2017), and VQA datasets (Goyal et al., 2017; Zhu et al., 2016; Hudson and Manning, 2019). The original VisualBERT is only pre-trained on COCO. Appendix A.3 details the pretraining datasets for all models as well as their sizes. The pretraining data for CLIP has not been publicly released, the authors state that it consists of 400M image-text pairs from the internet (an order of magnitude more data than for most of the other models, which do not surpass 10M image-text pairs in size).

4 Results

4.1 Original Implementations

We test all models using the evaluation methods and data described above. We make use of pre-trained models made publicly available by the authors.

Resulting accuracies are shown in Table 1. We find that only some models perform substantially above chance, notably ViLT, UNITER and LXMERT. In the `cropped` task, performance is much higher for all models, with VinVL and ViLT reaching the highest performance. This gap in performance between the `full` and `cropped` tasks indicates that while those models can match nouns and predicates in the image with the corresponding words rather well, they struggle to take into account the dependencies between them.

	Accuracy	
Model	Full	Cropped
LXMERT	0.57	0.69
UNITER	0.54	0.64
ViLBERT	0.28	0.66
ViLT	0.40	0.75
Oscar	0.32	0.67
VinVL	0.30	0.76
CLIP	0.20	0.59
Chance	0.25	0.25

Table 1: Accuracy of models trained in original conditions when provided the full images and when only exposed to the target object in the `cropped` task.

4.2 Controlled Training Conditions

We additionally evaluate models that are trained in controlled (and therefore more directly comparable) conditions as proposed in the VOLTA framework (Bugliarello et al., 2021). In this setup, all

[6] The multimodal pretraining objective of Oscar does not involve matching and mismatching images and descriptions, but only matching and mismatching sequences of object tags. Therefore, we evaluate the checkpoint that has been fine-tuned for image-text retrieval.

[7] VinVL is pre-trained on parts of the Open Images dataset, and we found that most images used in our evaluation set are indeed part of the VinVL pretraining dataset. Because of this confound, results of VinVL are not directly comparable to the other models (even though we test the model here with novel sentences with respect to the images). Our results show however that VinVL compares very closely to Oscar, which has the same architecture, suggesting that the access to the images during training does not substantially affect the model's performance. The same confound is possibly present for CLIP, for which the training data is not public

models are trained on Conceptual Captions using the same pretraining objectives (masked language modeling, masked object classification, and image-text matching) and use the same image features, extracted from a Faster R-CNN.

We evaluate all models for which pretrained weights are available. Resulting accuracy scores are presented in Table 2.

	Accuracy	
Model	Full	Cropped
CTRL_UNITER	0.24	0.63
CTRL_LXMERT	0.20	0.56
CTRL_ViLBERT	0.27	0.66
CTRL_VL-BERT	0.24	0.66
CTRL_VisualBERT	0.20	0.64
Chance	0.25	0.25

Table 2: Accuracy of models trained in controlled conditions when provided the full images and when only exposed to the target object in the cropped task.

We find that under these controlled conditions, all models perform comparably and generally around chance level. It is therefore not straightforward to draw any conclusions regarding the effect of model architecture from these results.

In the cropped task, performance is much higher, with ViLBERT and VL-BERT reaching the highest performance. The performance gap between the two tasks (i.e., full vs. cropped) is substantially larger than for the original implementations, suggesting that the models are even less sensitive to predicate-noun dependencies under these controlled training conditions.

5 Analyses and Discussion

5.1 Comparing Model Performances

The role of pretraining data Within the set of the evaluated models, we do not find evidence for a correlation between the size of the pretraining dataset and the model's ability to capture predicate-noun dependencies (see also Appendix A.3). Despite being trained on comparable or even larger amounts of data, ViLT, Oscar and VinVL perform substantially worse than LXMERT and UNITER. CLIP performs below chance level, despite having by far the largest pretraining dataset.

The pretraining data of CLIP is not publicly available, but as it was automatically scraped from the internet we believe the quality (i.e descriptiveness) of its captions to be comparable to that of Conceptual Captions. In additional experiments (see Appendix A.4.1), we study the performance of CLIP models trained on different datasets using a range of publicly available model checkpoints. The performance of CLIP remains below chance level for all tested checkpoints. This might be because all available checkpoints are all trained on rather noisy data, or because the architecture and pretraining objectives of CLIP don't allow it to learn grounded predicate-noun dependencies.

Datasets that are composed of highly descriptive captions seem to be advantageous for the learning of noun-predicate dependencies. Indeed, for datasets such as COCO (Lin et al., 2014) or VQA (Antol et al., 2015), the images are not only strongly associated with the captions or question–answer pairs (as they were crowdsourced specifically for the tasks), but also precise and detailed in nature. In contrast, Conceptual Captions (Sharma et al., 2018) is composed of images with captions that were automatically collected from web pages, and therefore generally rather broad descriptions of the image content.

ViLBERT and models trained in the controlled conditions are only trained using Conceptual Captions, and the resulting performances are around chance level. UNITER and LXMERT perform much worse compared to their original training setups. One main difference for these two models in their original implementation compared to the controlled condition is that they are trained on richer datasets with respect to the language modality, leveraging more descriptive captions.[8]

This observation is coherent with what Hendricks and Nematzadeh (2021) found when studying verb understanding of V&L models: They compare performance of the same model when trained on Conceptual Captions or COCO, and find that the model trained on COCO performs better, despite Conceptual Captions being bigger and closer to the task in terms of image and language distribution.

These results suggest that, when considering multimodal dependencies, having a high quality pretraining dataset with less noise and more de-

[8] The original pretraining datasets for UNITER and LXMERT are also larger in terms of the number of image-text pairs. However, LXMERT is actually trained on much fewer unique images than in the controlled conditions (180K vs. 3.1M). Therefore, we assume that the sheer size is not the driving factor of performance.

scriptive textual data could be more important than having a larger dataset. Highly descriptive textual data is essential to learn precise predicate-noun dependencies.

The role of pretraining objectives While models such as ViLT, Oscar, and VinVL are trained on datasets that are comparable in size and quality to those of LXMERT and UNITER, they still perform substantially worse on the task. One explanation could be that contrary to the other models, UNITER and LXMERT both have multimodal pretraining objectives *in addition* to image-text matching: Visual question answering for LXMERT and word-region alignment for UNITER.[9] This could help the models to establish finer multimodal dependencies. Indeed, ViLT and VinVL show better results than UNITER and LXMERT in the `cropped` task (indicating that their object/predicate recognition performance even surpasses that of the other models), but worse results in the `full` task. Our hypothesis is that the pretraining objectives of UNITER and LXMERT enable them to learn more fine-grained multimodal dependencies than ViLT and VinVL, even though their performance on the `cropped` task is worse.[10]

This gap in performance should not only be due to the training data associated with the additional pretraining objectives, as VinVL also uses data from Visual Question Answering task, but without training on the objective.

The impact of the multimodal pretraining objectives of UNITER and LXMERT can be an additional explanation for the drop in performance of CTRL_UNITER and CTRL_LXMERT, which were only trained using image-text matching as a multimodal pretraining objective. The gap in performance between those controlled models and the original models indicate that using more precise multimodal pretraining objectives and better annotated datasets can greatly improve the learning of multimodal dependencies.

The lack of suitable multimodal pretraining objectives could also offer an explanation for the poor performance of CLIP in our task.

The role of image encoders The authors of ViLT and VinVL motivate their work by suggesting that improved image features are mandatory for improved multimodal reasoning of V&L transformers. Here, we observe that these improved features only translate to better results in the `cropped` task (where ViLT and VinVL perform best). We speculate that the improved image encoders allow for a better understanding of visual entities, but not necessarily of the dependencies between them. In order to obtain more conclusive interpretations regarding the role of image features, we require more targeted experiments which control for other confounding factors present here (such as different pretraining objectives).

The role of model architecture In addition to the lack of suitable pretraining objectives, the worse performance of CLIP compared to the other models could also be due to the fact that it does not support any kind of inter-modal fusion of features within the model (image and text are processed in separate submodules that do not allow for intermodal interaction). This shortcoming of CLIP is also discussed in Kim et al. (2021), where the authors find representations from CLIP to be not useful for the more advanced multimodal reasoning task NLVR2 (Suhr et al., 2019).

However, there seems to be no major effect of architecture with respect to multimodal fusion in the case of single and dual stream transformers: LXMERT and UNITER have comparable performances, even though one is dual-stream transformer and the other a single-stream transformer.

5.2 Performance for Nouns vs. Predicates

Here, we compare performance for pairs in which the sentences differ with respect to the noun, to sentences with a different predicate. Detailed results for all models are reported in Appendix A.4.3. Overall patterns show a slightly better performance for cases in which the noun was switched, especially in the `cropped` task. This is in line with findings that V&L models are better at grounding nouns than verbs (Hendricks and Nematzadeh, 2021).

5.3 Analysis of individual nouns and predicates

For a given concept (noun or predicate), we consider all pairs that contain this concept in at least one of the two sentences, i.e. cases in which a

[9] ViLT also uses a word-patch alignment objective similar to word-region alignment, but the patches are not based on regions detected by an object detector and therefore the loss can not leverage any semantic labels for the patches during training, making this multimodal objective probably less useful.

[10] Most directly comparable are probably the cases of ViLT and UNITER, which are both trained on the same datasets, but with different pretraining objectives.

model's understanding of a concept is instrumental for making the correct decision.

Figure 3 shows the per-concept accuracies of the best performing model, LXMERT. Appendix A.4.4 shows the per-concept accuracies for all models in their original implementations.

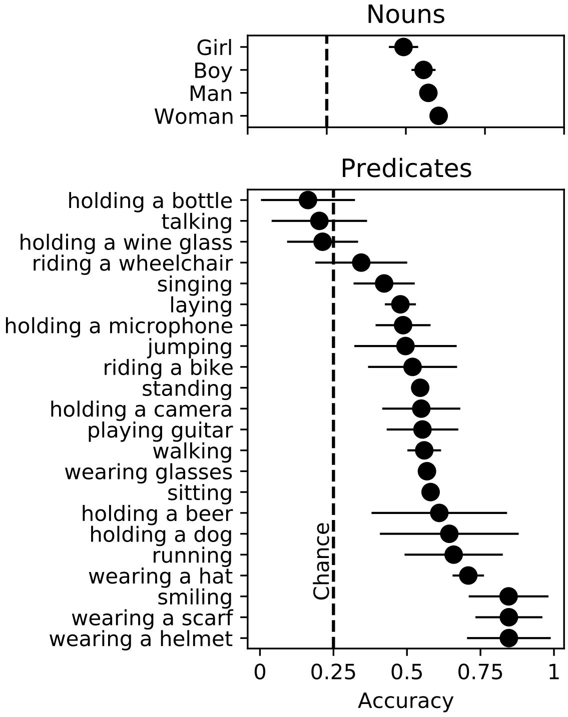


Figure 3: Per-concept accuracies for LXMERT. We display nouns and predicates for which we have at least 10 evaluation triplets. Standard deviation calculated using bootstrapping (100 re-samples).

We observe large variation in accuracy scores of predicates, and less variation for nouns. We could not find any simple reasons that explain the predicates' variability. For example, verbs can have good or bad performances (e.g., "running" vs "talking"), and the same can be said for predicates that are composed of both verb and noun (e.g., "holding a bottle" vs. "wearing a helmet"). That said, factors that may influence model performance on specific nouns or predicates are further discussed in Section 5.4.

Additionally, we observe that for some concepts, the models perform better if the concept is the target, and for others, performance is better if it is the distractor. This is, e.g., the case for the pair "sing" vs "stand", where the models consistently perform better if "sing" is the target predicate. Appendix A.4.5 shows the accuracy for each (target, distractor) concept tuple.

5.4 Confounding Factors

Here we discuss possible factors influencing the models' performances.

Object salience In most of the images in our evaluation set, the target and distractor persons in the image are not of equal size, and not equally salient (sometimes one is more in the foreground than the other). We explore whether there is an effect on the models' decisions by correlating the models' predictions with target and distractor bounding box size and location.

More specifically, we measure the difference in similarity for target and distractor sentence $s(I_1, S_1) - s(I_1, S_2)$ and correlate it with the difference in bounding box size of the target and distractor object. Further, we also correlate it with the difference of distances from the center of the image.

For LXMERT, we find no significant correlation (Bounding box size: Pearson $r = -0.03, p = 0.16$, bounding box distance from center: Pearson $r = 0.03, p = 0.14$). Correlation scores for other models can be found in Appendix A.4.6. While there are statistically significant correlations for some models, these are small and of varying direction. The largest correlations are found for CLIP (Bounding box size: Pearson $r = 0.14, p < 0.01$, bounding box distance from center: Pearson $r = -0.24, p < 0.01$), indicating that the performance of CLIP could be affected, to some extent, by object salience.

Concept recognizability We also correlate the models' similarity judgments differences to differences in concept recognizability, which we operationalize by taking the object or attribute confidence score for a given concept in an image from a Faster R-CNN (Ren et al., 2015) trained on VisualGenome.[11]

For most models, we find a small positive correlation (see Appendix A.4.6), indicating that the models' similarity judgments are affected by the varying degree to which the concepts are recognizable in the image.

Linguistic biases Another aspect, already mentioned earlier, is that models' performance could be affected by linguistic biases in the training data, such as the frequency and co-occurrence of words and phrases.

[11] If there are multiple objects/attributes with the corresponding label in the image, we take the maximum confidence.

To explore this possible effect, we correlate the difference in similarity for target and distractor sentence with the difference of target and distractor sentence perplexity. We calculate the perplexity for each sentence using a single-modality BERT model (`bert-base-uncased`), that was fine-tuned for 3 epochs on the textual data of Conceptual Captions.

For LXMERT, we find no significant correlation (Pearson $r = -0.01, p = 0.48$). For the other models, we find very small positive correlations (see Appendix A.4.6). We conclude that the models do not rely only on shallow heuristics of the training data in the textual modality.

6 Conclusion

This work examines whether state-of-the-art V&L models learn multimodal syntactic dependencies, by focusing on a case study on simple predicate-noun dependencies. Our controlled experiments and analyses on a range of recent models reveal that their capability track such dependencies is variable, with some models (e.g., LXMERT and UNITER) show performance above chance level and others (e.g., CLIP) performing even below chance.

In contrast to the recent trend in the field focused on increasing pretraining data and using simple general-purpose pretraining objectives (Brown et al., 2020; Devlin et al., 2019), here we observe that best performance is achieved, rather, with high-quality pretraining data, and more fine-grained pretraining objectives.

More specifically, our results suggest that multimodal pretraining objectives have a major impact on the model's learning of grounded predicate-noun dependencies. Models that include more targeted objectives such as visual question answering and word region alignment in addition to the general image-text matching objective show better performance. In addition, having highly descriptive pretraining datasets seems to help with learning fine-grained multimodal dependencies. In comparison, models trained on larger, web-scraped datasets do not perform well.

In the future, the proposed highly-controlled evaluation protocol can be used to conduct more targeted studies regarding the role of model architecture, pretraining objectives, as well as training data quality and quantity in order to build V&L models that are better at learning grounded predicate-noun dependencies, and possibly also other, move advanced multimodal reasoning tasks.

7 Limitations

While our analyses revealed patterns that seems to explain observed variability in the models' performance, the role of some architectural choices such as image encoding techniques remains ambiguous. A better understanding of all factors influencing the learning of grounded predicate-noun dependencies could be achieved by training sets of models on comparable conditions and by varying only one factor at a time (as done for example in Bugliarello et al., 2021, regarding the role of model architecture).

The range of concepts evaluated is rather small and therefore not representative for the understanding of grounded predicate-noun dependencies in general. More targeted data collection will be necessary in order to obtain more large-scale evaluation datasets. Additionally, our zero-shot evaluation paradigm introduces a possible mismatch between training and evaluation: Models are trained using pairs of images and descriptions where the descriptions often describe *all* salient parts of the image, whereas in our evaluation set the descriptions focus on only *one* aspect/person in the image.

In the `cropped` condition, the images are not are not representative of the typical photographic framing of image-text corpora, which could deteriorate our results. That said, random cropping is a frequent data augmentation technique in computer vision research, where it has been successfully applied to improve generalization performance (Krizhevsky et al., 2012).

Further, some scenes, actions and cultures are disproportionally represented in our evaluation dataset. As proposed in (Liu et al., 2021), it is important to pursue further work on more diverse datasets.

8 Ethics Statement

The proposed evaluation set relies on subjective annotations of perceived gender. Attempting to classify gender based on physical appearance is an ill-posed problem (e.g., due to limitations of object detectors, biases of the human annotators). Additionally, the annotations only consider *binary* gender classes (woman/man, girl/boy). Algorithms that perform such classifications are neither ideal nor desirable, as they perpetuate harmful stereotypes and exclude non-binary gender identities (e.g., Dev et al., 2021; Hamidi et al., 2018; Blodgett et al., 2020; Bender et al., 2021). We explored whether it would be possible to use other classes, but we did not find many examples that would allow for an evaluation of sensitivity to predicate-noun dependencies in a controlled fashion. As our image selection is very constrained (we require a visual distractor, and a counter-example image with reverse properties), we found only sufficient examples for the categories used in the paper. We initially started a bottom-up data exploration in which we considered all labels present in the Open Images dataset, but found only very few examples for a few other categories (generally less than 5 examples after manual filtering). This might be due to the focus of the annotations in Open Images, future work could explore the use of other datasets that are focused on other types of annotations, the main challenge being the requirement for sufficiently large datasets in order to find matching examples and counter-examples. Future efforts should be dedicated to creating datasets that aim at more inclusive annotations.

We acknowledge the severity of these issues, and emphasize that our work does not promote applications of gender classification in downstream tasks, but only uses it as a basis for analysis of existing models.

9 Acknowledgements

We thank the anonymous reviewers for their useful comments and feedback.

Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), ANR-21-CE28-0005-01 (MACOMIC), AMX-19-IET-009 (Archimedes Institute) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). This work was performed using HPC resources from GENCI–IDRIS (Grant 2021-[101693]).

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. pages 4971–4980.

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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. pages 2425–2433.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. 2021. COVR: A Test-Bed for Visually Grounded Compositional Generalization with Real Images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9824–9846, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models. In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 565–580, Cham. Springer International Publishing.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. pages 1017–1025.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: Learning Universal Image-Text Representations. page 14.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. 2020. Cops-Ref: A New Dataset and Task on Compositional Referring Expression Comprehension. pages 10086–10095.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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. pages 6904–6913.

Saurabh Gupta and Jitendra Malik. 2015. Visual Semantic Role Labeling. Technical report. Publication Title: arXiv e-prints ADS Bibcode: 2015arXiv150504474G Type: article.

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. pages 771–787.

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.

Hexiang Hu, Ishan Misra, and Laurens van der Maaten. 2019. Evaluating Text-to-Image Matching using Binary Image Selection (BISON). In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1887–1890, Seoul, Korea (South). IEEE.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. pages 6700–6709.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

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*, pages 5583–5594. PMLR. ISSN: 2640-3498.

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.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4. *International Journal of Computer Vision*, 128(7):1956–1981.

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:1908.03557 [cs]*. ArXiv: 1908.03557.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. What Does BERT with Vision Look At? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

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.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual Relationship Detection with Language Priors. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 852–869, Cham. Springer International Publishing.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. pages 11–20.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 200–210, Online. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded Situation Recognition. *arXiv:2003.12058 [cs]*. ArXiv: 2003.12058.

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. *Proceedings of the 38th International Conference on Machine Learning*, page 16.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing Common Human Visual Actions in Images. *arXiv:1506.02203 [cs]*. ArXiv: 1506.02203.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective. *AAAI*.

Joni O. Salminen, Hind A. Al-Merekhi, Partha Dey, and Bernard J. Jansen. 2018. Inter-Rater Agreement for Social Computing Studies. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 80–87.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! Find One mismatch between Image and Language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations.

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.

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.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. *arXiv:2204.03162 [cs]*. ArXiv: 2204.03162.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5534–5542, Las Vegas, NV, USA. IEEE.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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. pages 5579–5588.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. pages 4995–5004.

A Appendix

A.1 Image Pre-Filtering

We consider labels that occur at least 100 times in the dataset.

As some labels are similar and sometimes used interchangeably by the annotators, we create groups of synonyms for some labels and treat labels within a group as identical in the following. The groups of synonyms can be found in Table 3.

Further, we verify that the bounding boxes of the target and distractor objects are big enough (at least 20% width and 20% height of the image) and that the bounding box sizes of target and distractor objects don’t differ by more than a factor of 2.

Finally, we ensure that there is at least one counter-example for each triplet before starting the manual image selection phase.

“Table”, “Desk”, “Coffee table”
“Mug”, “Coffee cup”
“Glasses”, “Sunglasses”, “Goggles”
“Sun hat”, “Fedora”, “Cowboy hat”, “Sombrero”
“Bicycle helmet”, “Football helmet”
“High heels”, “Sandal”, “Boot”
“Racket”, “Tennis racket”, “Table tennis racket”
“Crown”, “Tiara”
“Handbag”, “Briefcase”
“Cart”, “Golf cart”
“Tree”, “Palm tree”
“Football”, “Volleyball (Ball)”, “Rugby ball”, “Cricket ball”, Tennis ball”

Table 3: Groups of label synonyms. Each line corresponds to one group.

A.2 Dataset statistics

Figure 4 shows the number of triplets for each noun and predicate. For a given noun or predicate, we count all pairs that contain this concept in at least one of the two sentences, i.e. cases in which correct understanding of a concept is useful for making the correct decisions.

A.3 Details on V&L Models

A.3.1 Pretraining Datasets

Table 4 details the multimodal datasets used for V&L models. Dataset sizes as reported in the corresponding papers. Note that these sizes are also affected by the fact that some models leverage validation sets for pretraining, while others constrain the data to the training sets. Also, different approaches for dataset overlap detection have been applied. The pretraining data size for CLIP is reportedly 400M image-text pairs.

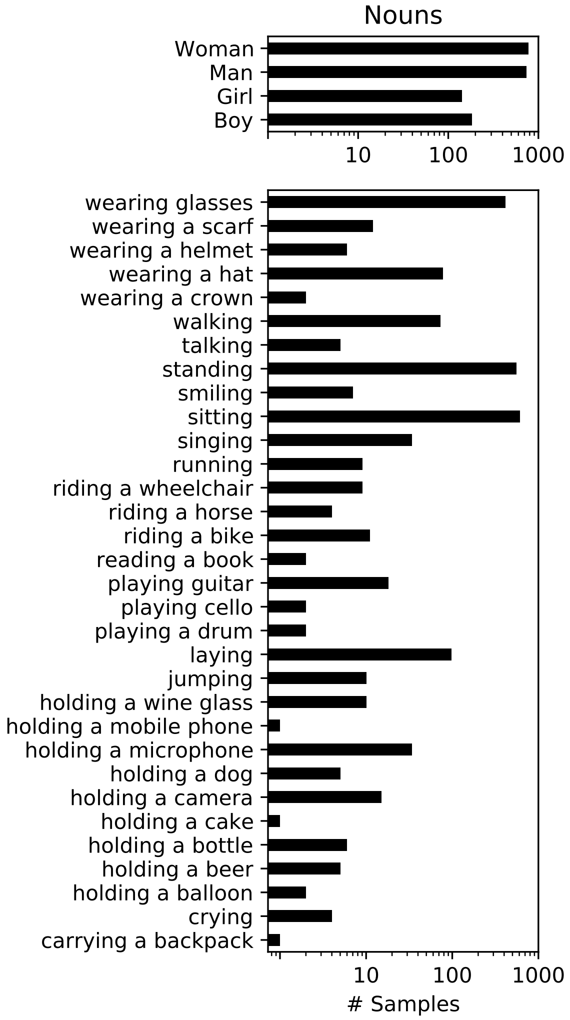


Figure 4: Number of triplets per concept. Note that the x-axis is on logarithmic scale.

A.3.2 Number of parameters

Table 5 compares the number of trainable parameters for each model that was tested in their original implementations.

A.4 Additional Analyses

A.4.1 Results for CLIP with varying pretraining data

Table 6 presents the accuracy scores of multiple publicly available checkpoints for CLIP trained on different training data.

Model	CC	COCO	SBU	VG	QA	F30K	OI	Total size	
								# images	# image-text pairs
LXMERT		✓		✓	✓			0.18M	9.18M
UNITER	✓	✓	✓	✓				4.16M	9.59M
ViLBERT	✓							3.10M	3.10M
ViLT	✓	✓	✓	✓				4.05M	9.85M
Oscar	✓	✓	✓		✓	✓		4.10M	6.50M
VinVL	✓	✓	✓		✓	✓	✓	5.65M	8.85M

Table 4: Pretraining datasets of the tested models: CC (Conceptual Captions; Sharma et al., 2018), COCO (Lin et al., 2014), SBU captions (Ordonez et al., 2011), VG (Krishna et al., 2017, Visusal Genome;), F30K (Flickr 30K; Young et al., 2014), OI (Open Images), and QA (including VQA2.0 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), and VG-QA (Zhu et al., 2016)).

Model	# Parameters
LXMERT	228,051,752
UNITER	112,938,887
ViLBERT	250,044,029
ViLT	111,596,546
Oscar	111,062,018
VinVL	111,686,973
CLIP	151,277,313

Table 5: Number of parameters for each model.

Visual Encoder	Dataset	Accuracy
RN101	YFCC-15M	0.18
RN101	400M	0.21
RN50	cc12m	0.18
RN50	400M	0.20
RN50	YFCC-15M	0.17
ViT-B-32	aion2b_e16	0.21
ViT-B-32	laion400m_e31	0.20
ViT-B-32	laion400m_e32	0.19
ViT-B-32	400M	0.20
ViT-L-14	400M (336px)	0.20

Table 6: Accuracy (Full) of CLIP models with varying visual encoders and pretraining data.

A.4.2 Controlling for linguistic robustness

As the sentences used in our evaluation dataset are built from a template, they do not vary in syntax. We verify that results obtained do not depend on the exact template chosen.

We vary the original templates by using the definite article ("the") at the beginning of sentences, and using verbs in simple present instead of present progressive tense (e.g., "the woman sits." or "the man holds a camera.").

Results with these alternative sentences are show in Table 7. We find that overall result patterns are highly similar to those with the original sentences in Table 1.

Model	Accuracy	
	Full	Cropped
LXMERT	0.55	0.70
UNITER	0.54	0.66
ViLBERT	0.26	0.67
ViLT	0.34	0.72
Oscar	0.32	0.65
VinVL	0.30	0.74
CLIP	0.21	0.58

Table 7: Results with alternative sentences.

A.4.3 Switching noun vs. switching predicate

Table 8 presents the accuracy for cases in which target and distractor sentence differ with respect to the predicate, or noun. We report scores for all models in their original implementations.

A.4.4 Analysis of individual nouns and predicates for all models

Figure 5 shows the accuracies for split up for the different predicates and nouns for all models. For more details, refer to Section 5.3.

A.4.5 Accuracies for (target, distractor) tuples

Figure 6 shows the accuracy for target-distractor tuples for all models.

A.4.6 Confounding Factors

In Table 9 we show the correlation scores for several confounding factors as described in Section 5.4.

	Full		Cropped	
Model	Noun	Predicate	Noun	Predicate
LXMERT	0.60	0.55	0.78	0.62
UNITER	0.60	0.50	0.76	0.56
ViLBERT	0.27	0.28	0.74	0.59
ViLT	0.44	0.37	0.80	0.72
Oscar	0.36	0.30	0.75	0.62
VinVL	0.33	0.28	0.83	0.71
CLIP	0.21	0.19	0.69	0.52

Table 8: Accuracy of models for cases in which the distractor sentence contains a different noun, or a different predicate.

Model	Bounding box size	Distance from center	Perplexity	Object detector confidence
LXMERT	-0.03 (p=0.12)	0.03 (p=0.08)	-0.01 (p=0.48)	0.30 (p=0.00)
UNITER	-0.09 (p=0.00)	0.09 (p=0.00)	0.05 (p=0.01)	0.26 (p=0.00)
ViLBERT	0.11 (p=0.00)	-0.16 (p=0.00)	0.05 (p=0.02)	0.22 (p=0.00)
ViLT	-0.01 (p=0.73)	0.03 (p=0.13)	0.05 (p=0.02)	0.26 (p=0.00)
Oscar	0.12 (p=0.00)	-0.17 (p=0.00)	0.06 (p=0.00)	0.15 (p=0.00)
VinVL	0.12 (p=0.00)	-0.12 (p=0.00)	0.05 (p=0.01)	0.04 (p=0.05)
CLIP	0.14 (p=0.00)	-0.24 (p=0.00)	0.08 (p=0.00)	0.17 (p=0.00)

Table 9: Correlations between difference in similarity and various factors related to targets and distractors: difference in bounding box size, distance from the image center of the bounding boxes, perplexity, and confidence scores of the bounding box as calculated using a Faster R-CNN object detector model.

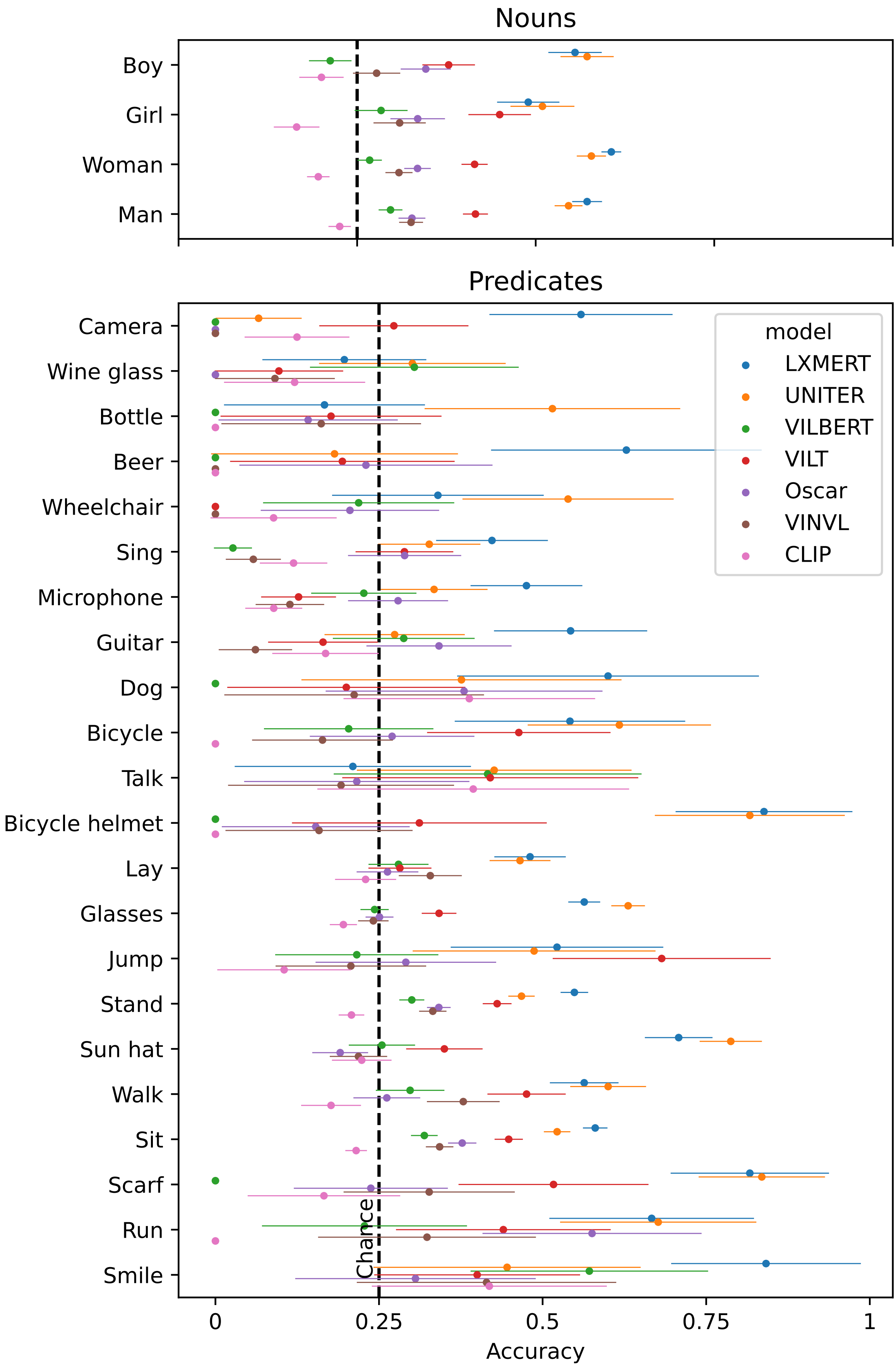


Figure 5: Per-concept accuracies for all models. We display nouns and predicates for which we have at least 10 evaluation triplets. Standard deviation calculated using bootstrapping (100 re-samples).

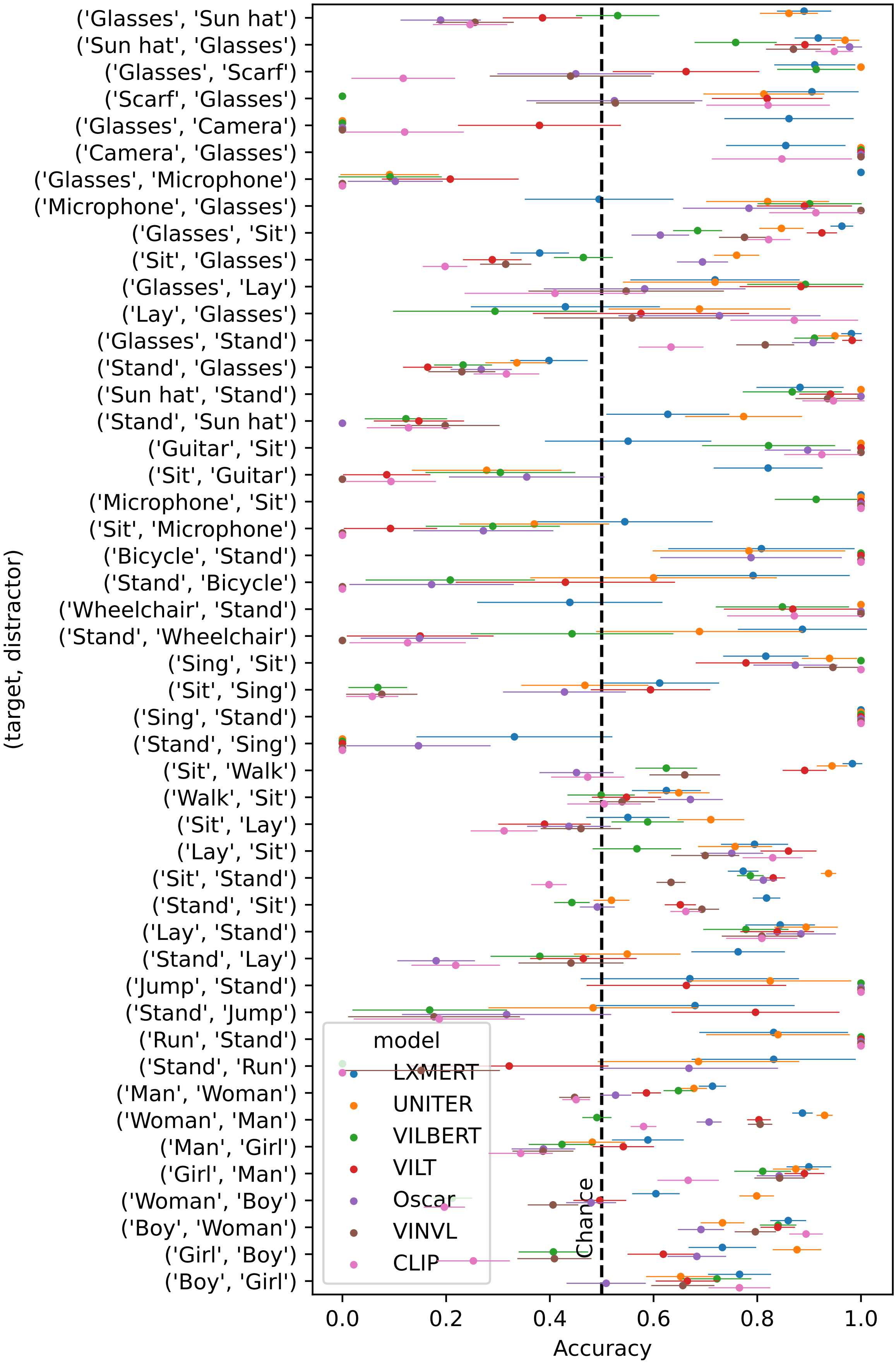


Figure 6: Accuracy for (target, distractor) tuples for all models. We display tuples for which we have at least 10 evaluation triplets. Note that chance performance is at 0.5, because we report per-triplet (and not per-pair) accuracy.