EViL-Probe
A Composite Benchmark for Extensive Visio-Linguistic Probing

Marie Bexte[1], Andrea Horbach[1, 2], Torsten Zesch[1]
[1]CATALPA, FernUniversität in Hagen, Germany [2]Hildesheim University, Germany

Abstract

Research probing the language comprehension of visio-linguistic models has gained traction due to their remarkable performance on various tasks. We introduce EViL-Probe, a composite benchmark that processes existing probing datasets into a unified format and reorganizes them based on the linguistic categories they probe. On top of the commonly used negative probes, this benchmark introduces positive probes to more rigorously test the robustness of models. Since the language side alone may introduce a bias models could exploit in solving the probes, we estimate the difficulty of the individual subsets with a language-only baseline. Using the benchmark to probe a set of state-of-the-art visio-linguistic models sheds light on how sensitive they are to the different linguistic categories. Results show that the benchmark is challenging for all models we probe, as their performance is around the chance baseline for many of the categories. The only category all models are able to handle relatively well are nouns. Additionally, models that use a Vision Transformer to process the images are also somewhat robust against probes targeting color and image type. Our enrichment of EViL-Probe with positive probes helps further discriminate performance.

Keywords: corpus, evaluation methodologies, neural language representation models

1. Introduction

Visual question answering (Antol et al., 2015), image-text retrieval (Peng et al., 2018) and retrieving image patches that match an expression (Mao et al., 2016) are just some of the visio-linguistic tasks neural models show increasingly impressive performance on. Since the beginning of this development, there has been research into how deep the language understanding of these seemingly competent models actually goes. At times, artifacts in the data are found to be the source of the observed high performance (Kafle et al., 2019). One avenue of putting the models' skills to the test is to design probes, usually composed of two minimally different descriptions of the same image. Models then have to predict whether these descriptions match the image they supposedly describe. Figure 1 shows an example of this: While the description of *two dogs* matches the image, describing them as *two cats* constitutes a mismatch.

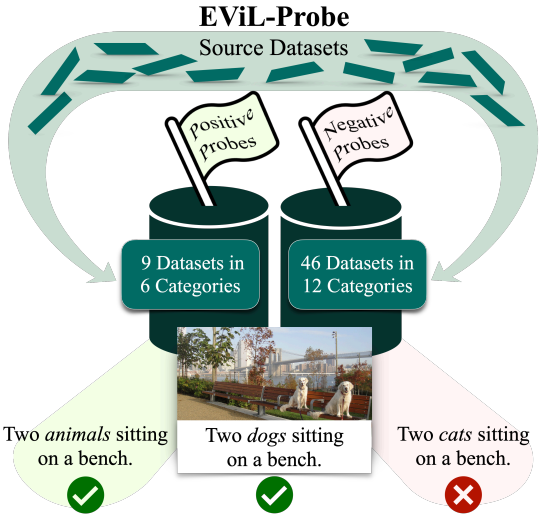

Figure 1: Overview of our probing benchmark.

Gardner et al. (2020) formulate the idea behind this probing as testing a models' decision boundary: While it is unrealistic to have an evaluation dataset that covers the entire space of possible input examples, test examples with just minor differences lead to a more densely populated subspace. This allows to at least probe the decision boundary in this specific area. Such probing datasets have been released to target specific phenomena, including, but not limited to, nouns, verbs or attributes (Shekhar et al., 2017b; Parcalabescu et al., 2022; Zhao et al., 2022; Bexte et al., 2024). However, these datasets are not all of the same format and quality. In addition, models are usually only tested on some of them and not always evaluated with the same metrics. All in all, this leaves potential for a more comprehensive performance estimation.

Our contribution is three-fold: First, we methodically categorize existing probing datasets based on the linguistic categories they cover. From these datasets, we compile the **E**xtensive **Vi**sion-and-**L**anguage Probing resource **EViL-Probe**. It consists of 1.5M probes across 880k images, organized into 55 datasets that cover 18 different probing categories. In contrast to previous benchmarks, EViL-Probe is not limited to what we call *negative* probes, but also includes *positive* probes. In a positive probe, the two descriptions of an im-

LREC-COLING 2024, pages 6682–6700

age are both valid. Such an example is shown in Figure 1, where both *dogs* and *animals* accurately describes the image content, therefore probing robustness against hypernyms.

Second, we apply a language-only baseline to assess the quality of the probes in EViL-Probe. We do this because a critical weakness of probing datasets can be systematic differences in the overall likelihood of sentences. This bears the risk of making probes solvable by not looking at the images. If this were the case, they would not be a true test of visio-linguistic ability.

Third, we probe a set of state-of-the-art visio-linguistic models to assess how robust they are against probes that target the different linguistic categories covered by EViL-Probe.

Our benchmark is publicly available on GitHub[1].

2. Probing Visio-Linguistic Models

We now describe the current state of the art of visio-linguistic models, and then discuss previous work on probing their language understanding.

2.1. Visio-Linguistic Models

In recent years, a plethora of language-and-vision models has emerged, mostly building on the Transformer architecture (Vaswani et al., 2017). Most can be considered as augmenting BERT (Devlin et al., 2019) with the capability to process images in a single-stream (Chen et al., 2020; Li et al., 2019a) and sometimes a dual-stream (Tan and Bansal, 2019; Lu et al., 2019) architecture. Many of the initial models require visual features extracted using a Faster R-CNN (Anderson et al., 2018) as their visual input. More recently, there have been proposals of models where the raw image is directly input into a Vision Transformer (Dosovitskiy et al., 2021). This was motivated by models that use pre-extracted features falling short on inter-object reasoning (Farah et al., 2022; Cho et al., 2022), as it lifts the restriction of having object-centered features. Vision Transformers were also shown to align more closely with human perception than Convolutional Neural Networks (Tuli et al., 2021). CLIP (Radford et al., 2021) is a popular model that embeds image and text into a shared embedding space, where similarity can then be measured using cosine.

As is typical for a transformer architecture, visio-linguistic models are usually pretrained on large amounts of data and then fine-tuned for downstream applications. During pretraining, one task that is often employed and sometimes even the only one used (Radford et al., 2021) is image-text alignment. While some models implement this via contrastive loss (Radford et al., 2021), others apply cross entropy loss to the output of a binary classification head (Tan and Bansal, 2019; Chen et al., 2020). Since this image-text matching is however only a pretraining task, performance on it is usually not reported (Parcalabescu et al., 2021). Instead, models are evaluated on downstream tasks, such as visual question answering (Antol et al., 2015), image-text retrieval (Peng et al., 2018), image captioning (Hossain et al., 2019), and even detecting hateful memes (Kiela et al., 2020). While models continue to achieve more and more impressive performance on these downstream tasks, there is a remaining concern regarding their true integration of the two modalities (Salin et al., 2022). To some extent, they could rely on spurious correlations in the data, such as cows usually appearing on grass or even watermarks being associated with certain image contents (Boreiko et al., 2022).

2.2. Probing

To gain insight into how deep the multi-modal understanding of visio-linguistic models goes, there have been efforts regarding what Gardner et al. (2020) call *contrast sets*. These are sets of minimally different input examples that are designed to probe model decisions in a more densely populated area of the example space. They are usually centered around a decision boundary, which means that the minimally different examples fall into different classes. This is visualized in Figure 1, where two descriptions *match* the image (*two dogs*, *two animals*), while the third is a *mismatch* (*two cats*).

Apart from a probing set, Gardner et al. (2020) provide for the NLVR2 dataset (Suhr et al., 2019), a number of other visio-linguistic probing datasets have been released. FOIL-IT (Shekhar et al., 2017b,a), VALSE (Parcalabescu et al., 2022), VL-Checklist (Zhao et al., 2022) and EqBen (Wang et al., 2023) are some of the more extensive ones that target different linguistic categories. All of these datasets, among others, are part of EViL-Probe.

CREPE (Ma et al., 2023) and ConStruct-VL (Smith et al., 2023) are two further interesting probing datasets that have however not (yet) been publicly released. While the majority of probing datasets is meant for evaluation purposes, some also include training data to increase the robustness of models through fine-tuning (Shekhar et al., 2017b; Cascante-Bonilla et al., 2023).

The existing work on probing generally follows the setup of slightly altering a matching description to derive a mismatched one. Descriptions are thus modified to change the ground truth. Replacing *dogs* with *cats* in Figure 1 is an example of

[1] https://github.com/mariebexte/vl-probing

such a probe, to which we refer as *negative probes*. An interesting avenue that is already pursued in visual question answering is robustness against paraphrasing (Jimenez et al., 2022). They perform ground truth-preserving changes that should thus not affect model predictions either. EViL-Probe contains nine datasets of such nature, which we refer to as *positive probes*. In Figure 1, replacing *dogs* with *animals* is an example of such a probe where both sentences are accurate descriptions of the image. Our positive probes cover six different categories, such as hypernym- or specificity-based alterations.

3. EViL-Probe

Figure 2 (left) gives an overview of the overall setup of our benchmark: Each probe consists of two texts t_1 and t_2 that refer to the same image i. This forms two tuples (i, t_1) and (i, t_2), a setup that allows both ranking-based and classification-based evaluation (Figure 2, right). In the following, we first describe the datasets EViL-Probe builds on and how they were processed to form the benchmark. We then describe the evaluation setup we employ to probe models with it.

EViL-Probe is a compilation of existing probing data. This data is sometimes taken as-is and in other cases rearranged to fit the required form, so that each probe consists of two image-text tuples that share the same image. Table 4 in the Appendix gives an overview of if and how the respective source datasets were processed to fit this desired format.[2] Some of the original datasets also include training and validation data. Since EViL-Probe is meant as an evaluation resource, we only use the testing split of these datasets.

We organize the many subdatasets into the linguistic categories they probe. As described, a probe consists of two descriptions of the same image that are minimally different with regard to the tested aspect. In the existing datasets, one of these descriptions matches the image, while the other does not (*negative probe*). EViL-Probe also contains *positive probes*, where both descriptions match the image. Note that the examples in the positive probes are therefore all of the same class. Thus, they are solvable with perfect accuracy by a model that classifies every description as matching its respective image. Since EViL-Probe does however also include negative probes, these serve as a control on models exhibiting such a bias. Table 1 gives an overview of the different categories we probe, grouped into negative (top) and positive (bottom) probes.

[2]We ensure that the two texts in a probe are never identical and that no duplicate probes are present.

3.1. Negative Probes

To ensure that models can generally reject mismatched descriptions, we compile two datasets with pairs of a matching and a **random** mismatched description. One is derived from MS COCO (Lin et al., 2014), the other bases on Flickr30k (Young et al., 2014)[3]. Both of these source datasets contain high-quality crowdsourced descriptions of photographs.

We incorporate the probes from Winoground (Thrush et al., 2022), where **word order** changes semantics. An example of such a description pair would be *a person underneath lights* and *lights underneath a person*. Diwan et al. (2022) did however find that some probes in this dataset are ambivalent, because *both* descriptions match the image[4]. We exclude these cases.

VL-Checklist (Zhao et al., 2022) contains two subsets of **attribute** probes that are based on Visual Genome (VG; Krishna et al. (2017)) and VAW Pham et al. (2021). These are each further split into the attribute types *action*, *material*, *size* and *state*. From these probes, we create one combined probing set for each of the four attribute types. In addition to this, we include the *attribute* probes from EQ-Kubric (Wang et al., 2023) and ARO (Yuksekgonul et al., 2023).

To probe **color** understanding, we use the VL-Checklist *color* probes and isolate the probes in EQ-SD (Wang et al., 2023) that target color. EQ-SD can be separated into multiple subsets that target different categories. We further derive probes that vary the described **image type** and test **noun** understanding. Since manipulating the objects mentioned in a description is a straightforward way of creating probes, there are many other datasets that target nouns: We aggregate the *object* probes from VL-Checklist into three subsets that are based on different source datasets: HAKE (Li et al., 2019b), SWiG (Pratt et al., 2020), and VG.[5] We also use the probes from the FOIL-IT dataset (Shekhar et al., 2017b) and the *noun* examples from Nikolaus et al. (2022). Further, we combine the respective *subject* and *object* subsets from SVO Probes (Hendricks and Nematzadeh, 2021) and ComVG (Jiang et al., 2022) into two additional noun probing subsets.

Just as we do for nouns, we create three **verb** probing subsets from the VL-Checklist *action* probes based on the three different source

[3]We use the Karpathy and Fei-Fei (2017) test splits.

[4]E.g. *The taller (shorter) person hugs the shorter (taller) person* both match an image of a taller and a shorter person hugging.

[5]A substantial amount of the Visual Genome probes has texts that do not have any overlap, e.g. *blue cotton tee shirt* paired with *potato salad*. We exclude such examples.

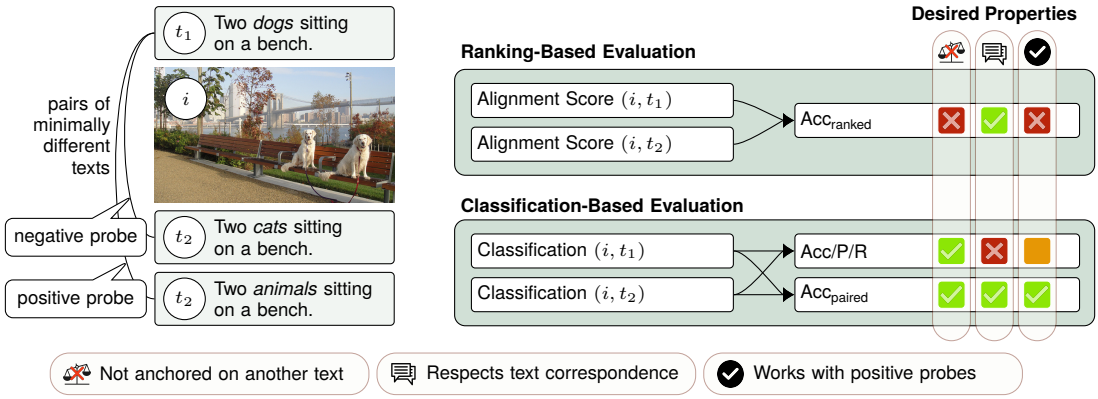

Figure 2: Overview of the structure of EViL-Probe (left) and the properties of different evaluation possibilities (right). Our preferred metric paired accuracy is non-relative, takes the correspondence between the two texts in a probe into account and can accommodate positive probes.

Category	# Datasets	# Images	# Triples	**Example:** Text 1 (matches respective image)	**Example:** Text 2
Negative Probes					
Attribute	6	29289	72426	The lace gown and the large painting.	The large gown and the lace painting.
Color	2	36656	75102	silver sculpture	bright yellow sculpture
Image Type	1	414	414	An oil painting of car.	A pencil sketch of car.
Negation	3	1590	2647	There are people in the water.	There are no people in the water.
Noun	8	36656	493829	An animal sits in a meadow.	A girl sits in a meadow.
Number	7	8554	11318	There are 5 players.	There are 6 players.
Random	2	6000	30010	A bunch of bananas are sitting on the stand.	The airplane has begun its ascent in the skies.
Semantic Role	1	1028	1028	A woman bites her shoe.	A shoe bites a woman.
Spatial Relation	5	19721	56776	The bowl is on the plate.	The plate is on the bowl.
Verb	7	144322	272557	A woman is riding a horse.	A woman is feeding a horse.
Video-based	3	487048	487048	Spread the dough out in the pan.	Sprinkle garlic powder on the crust.
Word Order	1	708	708	A person underneath lights.	Lights underneath a person.
Positive Probes					
Hypernyms	1	633	1232	A car smashed into a tree.	A vehicle smashed into a tree.
Paraphrase	2	534	534	Flat on the bottom and pointy on top.	Flat bottom and pointed top.
Perspective	3	1499	59580	A professional baseball player in a game.	He is playing baseball.
Specificity	1	369	369	A photo of dog wearing a scarf.	A photo of dog.
Word Order	1	91	127	Someone bakes the dough before it is eaten.	Before the dough is eaten someone bakes it.
Slang	1	417	608	There are two people and three windows.	There are two peeps and three windows.

Table 1: Data in EViL-Probe. For statistics on individual datasets see Table 5 in the Appendix.

datasets. We also incorporate the *verb* subsets from SVO Probes, ComVG and Nikolaus et al. (2022). In addition, we use the *action replacement* examples from the VALSE (Parcalabescu et al., 2022) benchmark. As a special case of verb understanding, we include the VALSE *actant swap* probes to probe **semantic role** comprehension.

To probe **negation** understanding, we include the *existence* and *coreference*[6][7] examples from the VALSE benchmark.

VALSE also includes *relation* probes, which we use as part of our probes targeting **spatial relations**. On top of these, we also include the *relation* probes from ARO (Yuksekgonul et al., 2023), EQ-Kubric (Wang et al., 2023) and VL-Checklist. Further, we add the probes from the Visual Spatial Reasoning dataset (Liu et al., 2023).

Another category we base off of VALSE are **number** probes. We use their *plurals*, *adversarial*, *balanced* and *small numbers* counting probes. In addition, we incorporate the the *standard* and *hard* splits of Parcalabescu et al. (2021) and the *counting* examples from EQ-Kubric. Note that these number-based probes follow the Gricean maxim of quantity (Grice, 1975): In a strictly logical sense, an image showing three girls sitting on a bench would still be correctly described as *two girls sitting on a bench*. However, as the sentence bears the implicature of *not more than two girls*, such a description would be considered uncooperative.

A special group of examples are formed by the EQ-AG, EQ-GEBC, and EQ-YouCook2 subsets of EqBen (Wang et al., 2023). These are all based on **videos** and derive probes from different frames of

[6]We merge the *standard* and *hard* split.

[7]These are of the form *Statement. Question? Yes/No*. This format is intended to require resolution between question and statement to determine whether the yes/no assessment is correct. However, our experiments (see Appendix A.5) show that they do not really require resolution. Therefore, we also include a version where they are reduced to just *Question? Yes/No*.

the same video. This makes them challenging due to the small degree of change between frames.

3.2. Positive Probes

The image descriptions in many of the standard datasets are what Hodosh et al. (2013) call *conceptual*: face-value descriptions of what is shown in an image. Cafagna et al. (2023) augment the object-centered image descriptions from MS COCO (Lin et al., 2014) with alternate descriptions. These either describe the *scene* in general, the *actions* people are taking, or possible *rationales* for why people are taking these actions. We adopt this data as three subsets that probe **perspective**. Cafagna et al. (2023) perform a second round of annotation with plausibility scores. We only use descriptions that have a plausibility score of at least 4 out of 5.

The EQ-SD subset of EqBen (Wang et al., 2023) has examples where one sentence is a substring of the other. Combined with the images depicting what is described in the longer sentence, these form an additional set of probes. They test whether both descriptions are accepted by a model, irrespective of their level of **specificity**.

Diwan et al. (2022) augment Winoground (Thrush et al., 2022) through **hypernym** replacement, the introduction of **slang**, and creating **paraphrases**[8]. In performing *rule-based reordering* they further create semantics-preserving changes in **word order**. We include all of these probes as additional subsets of EViL-Probe. We do not include their *synonym* substitutions due to a relatively high number of pairs where semantics change substantially. Examples of this are changes of *car* to *cable car* or *fall* to *precipitate* when it is a *person* falling.

3.3. Evaluation

Figure 2 (right) shows different possibilities of evaluating performance on our probes and the properties of these metrics. Each probe consists of a pair of two image-text tuples (i, t_1) and (i, t_2) that describe the same image i. Processing these yields two separate outputs, one for each tuple. All visiolinguistic models we probe output the probability of an input image and text matching. These probabilities can either be taken as-is (ranking-based evaluation), or be mapped to the classes *match* (p > .5) and *mismatch* (p <= .5) (classification-based evaluation).

Ranking-Based Evaluation A metric that is often used to evaluate probing is ranking-based accuracy (**Acc$_{\text{ranked}}$**). Here, the alignment score a_1 between an image i and a matching text t_1 is compared to the alignment score a_2 of i and a mismatched text t_2. If $a_1 > a_2$, i.e. the matching description is found to have higher alignment with the image, this counts as the model having made the correct decision.

This is however a relative evaluation that compares the scores of the two texts. It does therefore only permit the assessment of whether one description aligns better *than another*.

Classification-Based Evaluation To avoid the need for a reference text to evaluate against, we evaluate probes in a binary classification setting. This requires a separate binary decision regarding the alignment of each image-text tuple. Such binary decisions can be evaluated using accuracy, precision, recall and f-score. This evaluation would however not take into account the paired nature of the two texts in a probe.

In the spirit of testing the model's decision boundary, it is of interest whether the model classifies *both* texts in a probe correctly. In this regard, Nikolaus et al. (2022) and Thrush et al. (2022) calculate image-based accuracy scores. These require a model to be correct on *all* texts an image is paired with. We argue for this same mode of evaluation and focus on paired accuracy (**Acc$_{\text{paired}}$**). Figure 3 demonstrates how this metric can uncover differences between models that achieve the same performance when evaluated based on individual image-text tuples.

A crucial requirement for a metric that evaluates EViL-Probe is the ability to give useful results for both negative and positive probes. In a positive probe, both texts match the image. This means that all examples are of the same class *match*, which prohibits calculation of precision or recall. Accuracy does not require the definition of a target class and can therefore be calculated for both negative and positive probes alike. Therefore, we report both standard (tuple-based) accuracy and paired accuracy in our experiments. Do note that while the random baseline of standard accuracy lies at .5, it is .25 for paired accuracy.

To summarize, as is visualized in Figure 2 (right), paired accuracy satisfies three crucial requirements: First, it is based on classification. It does not anchor the alignment estimate on another text, which is the case for ranking-based accuracy. Second, it respects the correspondence between the two texts in a probe. Traditional classification evaluation measures such as precision or recall do not take this into account. Third, it yields useful results even for positive probes, i.e. when both texts match the image.

[8] We incorporate their *backtranslation* and *diverse-paraphrase* probes.

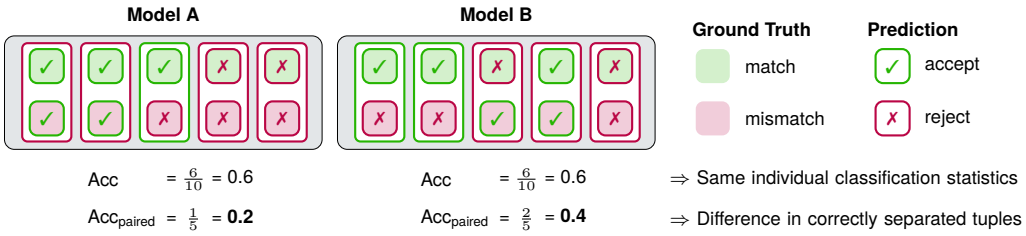

Figure 3: Each square represents a tuple of an image and a description. The respective upper and lower tuple share the same image, but differ in its description of it. While the description in the upper tuple matches the image, the one in the lower tuple does not. The two hypothetical models both correctly recognize three matches and three mismatches. They therefore have identical accuracy, precision and recall. However, calculating accuracy on the basis of pairs of tuples (white rectangles) reveals that model B is superior in separating the matching from the mismatched description of an image.

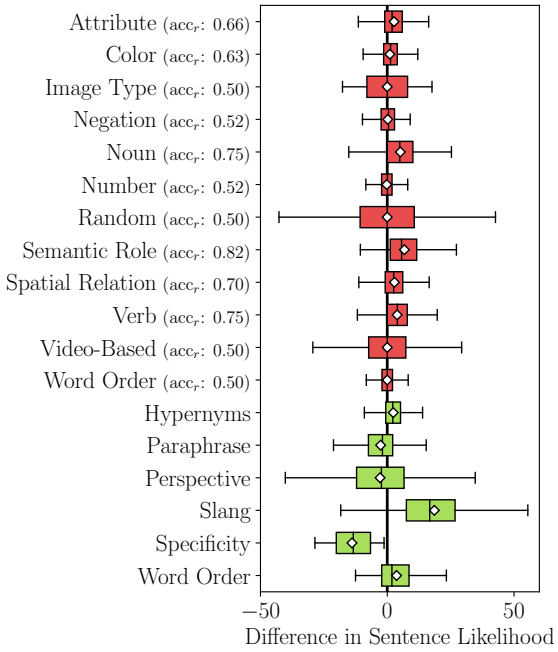

Figure 4: Language-only baseline that shows the differences in likelihood of the first and second sentence in a probe. For negative probes (■), we give the ranking-based accuracy this baseline would achieve in parentheses. Since both texts match the image in the positive probes (■), we cannot calculate the ranking-based accuracy for these.

4. Benchmark Analysis

We now estimate the linguistic bias present in the different datasets of EViL-Probe. Afterward, we discuss results of probing a set of models.

4.1. Language-Only Baseline

To assess the quality of the probes in EViL-Probe, we determine to what extent there may be a language bias in them. Such a bias could systematically make one of the two texts in a probe more likely than the other. If for example the matching sentence is *A green banana* and the probe *A blue banana*, it may be determined from the language alone that the former is more likely a match. Note that in all probes, the first text always *matches* the respective image, while the second one is a *mismatch* in negative and a *match* in positive probes.

To determine the language-only baseline, we need a way of assigning probabilities to texts. This should preferably be done using a language model that is similar to how language is processed in visio-linguistic models. These models usually rely on BERT. This is why we use the approach of Salazar et al. (2020), who obtain pseudo log-likelihood scores from BERT models. They apply sequential token masking to the inputs and determine the probabilities of the correct tokens[9].

We determine the difference in likelihood for each probe by subtracting the likelihood of the second text from that of the first. Results are displayed in Figure 4[10]. For the negative probes, we also show the ranking-based accuracy this language-only approach would achieve (in parentheses).

Irrespective of whether probes are positive or negative, the desire is for them to be unbiased, i.e. close to the zero line (vertical black line in Figure 4). While deviations vary, the average difference in likelihood is close to this line for all but two categories: *specificity* and *slang*. This indicates that BERT prefers the less specific and (presumably more familiar) sentences that do not contain slang.

Some of the datasets with negative probes are cleverly designed to achieve a ranking-based accuracy of .5. In these, each text occurs both as a matching description and a mismatched one. Datasets for which this is not the case display a

[9] https://github.com/awslabs/mlm-scoring. We use the *bert-base-en-cased* model.

[10] See Figures 6 and 7 in the Appendix for results on the individual datasets.

Model	Architecture	Visual Input	#Images
LXMERT	2-stream	Faster R-CNN	0.2M
UNITER (large)	1-stream	Faster R-CNN	4M
VILLA (large)	1-stream	Faster R-CNN	4M
SOHO	end2end	ResNet	0.2M
ALBEF (large)	end2end	ViT-B/16	14M
TCL (base)	end2end	ViT-B/16	4M
BLIP (base_{129M})	end2end	ViT-B/16	129M

Table 2: Overview of the models we probe.

	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
Negative Probes							
- Acc	.54	.62	.61	.57	.63	.62	.63
- $\text{Acc}_{\text{paired}}$	.27	.26	.24	.20	.29	.27	.29
Positive Probes							
- Acc	.47	.80	.83	.82	.54	.52	.75
- $\text{Acc}_{\text{paired}}$	.25	.70	.75	.73	.41	.37	.65

Table 3: Macro-averaged performance of models.

slight bias towards the correct sentences, which is most pronounced for the *semantic role* category. This might be due to the fact that a large proportion of the role reversals are nonsensical. An example is *A woman is sucking a candy cane*, which is turned into *A candy cane is sucking a woman*. To examine the general quality of the language in the probes, we give absolute likelihood scores for the different datasets in Appendix A.2.

4.2. Probing Models with EViL-Probe

We now probe a set of existing pretrained models with our benchmark.

4.2.1. Models

We need to obtain matching probabilities for input image-text pairs and want to evaluate existing pretrained models without further fine-tuning them. Therefore, we can only consider models that were pretrained on image-text matching. We use the output of the respective pretraining head and apply softmax to obtain the matching probability of image-text pairs. Models are provided by the respective authors and summarized in Table 2.

The first group of models we test are those that rely on pre-extracted image region features, either in a one-stream or a two-stream architecture. We also include the more recent models that do not have this restriction and instead take the entire image as input (end2end in Table 2). Wherever multiple pretrained model versions are available, we use the largest one. The rationale behind this is that our aim is less so a between-model comparison. Rather, we want to gain an overall picture of how challenging EViL-Probe is for the current visio-linguistic models.[11]

VILLA is an improvement of UNITER that introduces adversarial training through permutations in the embedding space. UNITER, VILLA and LXMERT (Tan and Bansal, 2019) all use random mismatched examples in their pretraining of image-text matching. ALBEF (Li et al., 2021) makes an effort to sample hard negatives with shared semantics, but different fine-grained details. TCL (Yang et al., 2022) adds to ALBEF by introducing intra-modal self-supervision. To make the most of noisy web-scraped data, BLIP (Li et al., 2022) performs bootstrapping on captions by generating and then filtering them. Note that due to the nature of the source datasets and the rather extensive datasets used to pretrain models, some of them may have already seen some of the probing images during pretraining.

In our experiments, we focus on pair-wise accuracy as the metric that fulfills all three of our requirements (see Figure 2). We also include standard accuracy as an alternative metric that gives useful results for both negative and positive probes. Wherever we aggregate results of multiple datasets, we take the macro average.

4.2.2. Negative vs. Positive Probes

Table 3 shows aggregated results for negative and positive probes.[12]

Comparing standard and paired accuracy, models show different relative performances. While TCL outperforms LXMERT when it comes to standard accuracy, the two have the same level of paired accuracy. Similarly, while SOHO outperforms LXMERT when it comes to standard accuracy, it is the other way around when evaluated with paired accuracy.

The highest paired accuracy is achieved by ALBEF and BLIP, two models that use a Vision Transformer to process images. Looking at the positive probes can further differentiate the performance of these two models: While they perform equally on the negative probes, BLIP outperforms ALBEF by a large margin on the positive probes. As we remarked when we introduced them, performance on positive probes is to be seen as an extension of results on the negative probes. While UNITER, VILLA and SOHO perform well on the positive probes, this is linked to a general tendency of these models to accept a high proportion of texts. On the negative probes, VILLA and SOHO achieve the lowest paired accuracy.

Models in Table 3 are ordered in chronological

[11] For results that include different model sizes see Appendix A.3.

[12] See Appendix A.4 for additional metrics.

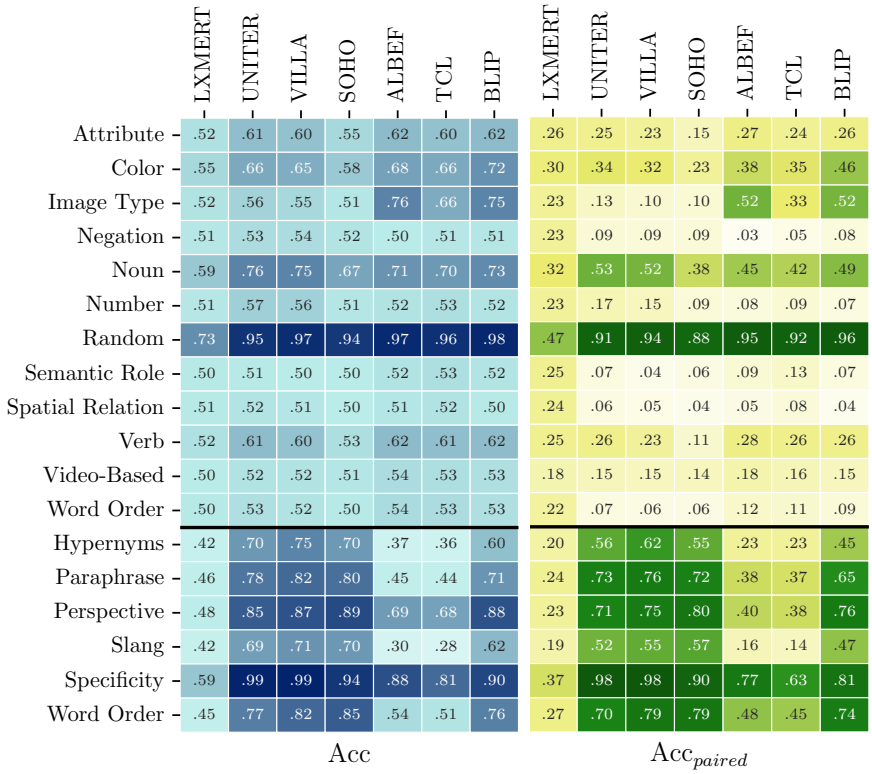

Figure 5: Category-wise performance of models on negative (top) and positive (bottom) probes. Results are macro-averaged over all datasets that belong to the respective category.

order of emergence. Newer models do not perform drastically better than older ones. Do also note that while the newest model (BLIP) gives an overall relatively good performance, this model is trained with a substantially higher number of images than previous models (see Table 2).

We now take a closer look at how the different linguistic categories play into the observed results.

4.2.3. Category-Based Results

Figure 5 shows score breakdowns for the individual categories. LXMERT shows poor performance across all probing categories. While this model is superior on some categories, its performance on these is around the chance baseline.

Results for the *random* probes confirm that all models other than LXMERT are able to reliably reject mismatched descriptions. Another category many of the models are able to handle relatively well are *nouns*. Interestingly, UNITER and VILLA are performing best here. Both of these models use pre-extracted image features to process the visual inputs. These features were obtained from a Faster R-CNN that was designed for object detection - this focus on objects seems to reflect in the superior performance of models using these features on probes that target nouns, i.e. objects.

Image type and *color* are two other categories for which at least ALBEF and BLIP show some discriminatory ability.

When it comes to the positive probes, all models (except LXMERT) perform relatively well on *specificity* and show the greatest difficulty regarding *slang* and *hypernym* probes. While it may be understandable that models are simply not familiar with slang, they will be familiar with the more general terms used in the hypernym probes. The superior performance of BLIP over ALBEF on the positive probes we saw in Table 3 is revealed to be consistent across all categories.

5. Conclusion and Future Work

We present EViL-Probe, an extensive composite benchmark of visio-linguistic probes. It is designed to reveal a detailed picture of the capabilities of visio-linguistic models when it comes to comprehending different linguistic categories. Our benchmark not only contains the standard negative probes, but augments these with additional sets of positive probes. In our experiments, these help distinguish further between models that demonstrate equal performance on the negative probes.

Overall, all models struggle with the majority of the negative probes. Even the best-performing models only scores higher than the random baseline for *random*, *noun*, *image type* and *color*

probes. For the eight other categories, performance is at or below chance level.

Future efforts may derive a smaller, manually validated version of EViL-Probe, filtering out cases where the gold standard labels taken from the original datasets may be questionable. Another exciting direction for future work are more interpretable evaluations of image-text alignment. This might be pursued through identifying the parts of the description that constitute the mismatch. The FOIL-IT (Shekhar et al., 2017a,b) benchmark already includes this as a visionary follow-up task to detecting a mismatched description, even going as far as to correct the misalignment.

Limitations

One limitation of our benchmark is that it is unbalanced, as some aspects (such as probing noun comprehension) have received greater attention in previous work. While an ideal setting would have an equal amount of probes for all categories, we did not want to artificially down-scale EViL-Probe by reducing the number of examples for all categories to the size of the category with the least amount of probes. To prevent larger subsets from diluting the effects of smaller ones we use macro averaging when aggregating results.

An issue that is passed on from the source datasets of EViL-Probe is that some of them include images some models have already seen during pretraining. While the steadily growing number of images used to pretrain visio-linguistic models make this increasingly hard to prevent, a perfect zero-shot setting would consist of images that are entirely novel to the models.

A further limitation arises from the fact that many of the images were collected from Flickr[13] (e.g. MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017)). Flickr gets over half of its traffic from the USA, UK, Germany and France[14], making it predominantly a reflection of Western culture.

Ethics Statement

Composite benchmarks should make sure that all included datasets are free of ethical concerns. To the best of our knowledge all source datasets of EViL-Probe fulfill this requirement. In case new issues should surface, we would drop the respective dataset from the benchmark.

[13] https://www.flickr.com

[14] https://www.similarweb.com/website/flickr.com

Acknowledgements

This work was partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany.

6. Bibliographical References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086. ISSN: 2575-7075.

Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. 2022. Sparse Visual Counterfactual Explanations in Image Space. In *Pattern Recognition*, Lecture Notes in Computer Science, pages 133–148, Cham. Springer International Publishing.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *CoRR*, abs/2202.04053.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality. ArXiv:2211.00768 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers

for image recognition at scale. In *International Conference on Learning Representations*.

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

H. P. Grice. 1975. Logic and Conversation. In *Speech Acts*, pages 41–58. Brill. Section: Speech Acts.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

M. Hodosh, P. Young, and J. Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys*, 51(6):118:1–118:36.

Kushal Kafle, Robik Shrestha, and Christopher Kanan. 2019. Challenges and Prospects in Vision and Language Research. *Frontiers in Artificial Intelligence*, 2.

Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual Spatial Reasoning. ArXiv:2205.00363 [cs].

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921.

Yuxin Peng, Xin Huang, and Yunzhen Zhao. 2018. An Overview of Cross-Media Retrieval: Concepts, Methodologies, Benchmarks, and Challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2372–2385. Conference Name: IEEE Transactions on Circuits and Systems for Video Technology.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257.

James Seale Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, and Leonid Karlinsky. 2023. ConStructVL: Data-free continual structured VL concepts learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14994–15004.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Tom Griffiths. 2021. Are convolutional neural networks or transformers more like human vision? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

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.

Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. 2022. GEB+: A Benchmark for Generic Event Boundary Captioning, Grounding and Retrieval. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 709–725, Cham. Springer Nature Switzerland.

7. Language Resource References

Antol, Stanislaw and Agrawal, Aishwarya and Lu, Jiasen and Mitchell, Margaret and Batra, Dhruv and Zitnick, C. Lawrence and Parikh, Devi. 2015. *VQA: Visual Question Answering*. ISSN: 2380-7504.

Bexte, Marie and Horbach, Andrea and Zesch, Torsten. 2024. *Rainbow - A Benchmark for Systematic Testing of How Sensitive Visio-Linguistic Models are to Color Naming*. Association for Computational Linguistics.

Cafagna, Michele and van Deemter, Kees and Gatt, Albert. 2023. *HL Dataset: Grounding*

High-Level Linguistic Concepts in Vision. arXiv. ArXiv:2302.12189 [cs].

Cascante-Bonilla, Paola and Shehada, Khaled and Smith, James Seale and Doveh, Sivan and Kim, Donghyun and Panda, Rameswar and Varol, Gül and Oliva, Aude and Ordonez, Vicente and Feris, Rogerio and Karlinsky, Leonid. 2023. *Going Beyond Nouns With Vision & Language Models Using Synthetic Data*. arXiv. ArXiv:2303.17590 [cs].

Changpinyo, Soravit and Sharma, Piyush and Ding, Nan and Soricut, Radu. 2021. *Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts*.

Chen, Yen-Chun and Li, Linjie and Yu, Licheng and El Kholy, Ahmed and Ahmed, Faisal and Gan, Zhe and Cheng, Yu and Liu, Jingjing. 2020. *UNITER: UNiversal Image-TExt Representation Learning*. Springer International Publishing, Lecture Notes in Computer Science.

Gan, Zhe and Chen, Yen-Chun and Li, Linjie and Zhu, Chen and Cheng, Yu and Liu, Jingjing. 2020. *Large-Scale Adversarial Training for Vision-and-Language Representation Learning*. Curran Associates, Inc.

Gardner, Matt and Artzi, Yoav and Basmov, Victoria and Berant, Jonathan and Bogin, Ben and Chen, Sihao and Dasigi, Pradeep and Dua, Dheeru and Elazar, Yanai and Gottumukkala, Ananth and Gupta, Nitish and Hajishirzi, Hannaneh and Ilharco, Gabriel and Khashabi, Daniel and Lin, Kevin and Liu, Jiangming and Liu, Nelson F. and Mulcaire, Phoebe and Ning, Qiang and Singh, Sameer and Smith, Noah A. and Subramanian, Sanjay and Tsarfaty, Reut and Wallace, Eric and Zhang, Ally and Zhou, Ben. 2020. *Evaluating Models' Local Decision Boundaries via Contrast Sets*. Association for Computational Linguistics.

Hendricks, Lisa Anne and Nematzadeh, Aida. 2021. *Probing Image-Language Transformers for Verb Understanding*.

Huang, Zhicheng and Zeng, Zhaoyang and Huang, Yupan and Liu, Bei and Fu, Dongmei and Fu, Jianlong. 2021. *Seeing Out of the Box: End-to-End Pre-Training for Vision-Language Representation Learning*.

Ji, Jingwei and Krishna, Ranjay and Fei-Fei, Li and Niebles, Juan Carlos. 2020. *Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs*. IEEE.

Jiang, Kenan and He, Xuehai and Xu, Ruize and Wang, Xin Eric. 2022. *ComCLIP: Training-Free Compositional Image and Text Matching*. arXiv. ArXiv:2211.13854 [cs].

Jimenez, Carlos E. and Russakovsky, Olga and Narasimhan, Karthik. 2022. *CARETS: A Consistency And Robustness Evaluative Test Suite for VQA*. Association for Computational Linguistics.

Kiela, Douwe and Firooz, Hamed and Mohan, Aravind and Goswami, Vedanuj and Singh, Amanpreet and Ringshia, Pratik and Testuggine, Davide. 2020. *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes*. Curran Associates, Inc.

Krishna, Ranjay and Zhu, Yuke and Groth, Oliver and Johnson, Justin and Hata, Kenji and Kravitz, Joshua and Chen, Stephanie and Kalantidis, Yannis and Li, Li-Jia and Shamma, David A. and Bernstein, Michael S. and Fei-Fei, Li. 2017. *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*.

Kuznetsova, Alina and Rom, Hassan and Alldrin, Neil and Uijlings, Jasper and Krasin, Ivan and Pont-Tuset, Jordi and Kamali, Shahab and Popov, Stefan and Malloci, Matteo and Kolesnikov, Alexander and Duerig, Tom and Ferrari, Vittorio. 2020. *The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale*. ArXiv:1811.00982 [cs].

Li, Junnan and Li, Dongxu and Xiong, Caiming and Hoi, Steven. 2022. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. PMLR. ISSN: 2640-3498.

Li, Junnan and Selvaraju, Ramprasaath and Gotmare, Akhilesh and Joty, Shafiq and Xiong, Caiming and Hoi, Steven Chu Hong. 2021. *Align before Fuse: Vision and Language Representation Learning with Momentum Distillation*. Curran Associates, Inc.

Li, Liunian Harold and Yatskar, Mark and Yin, Da and Hsieh, Cho-Jui and Chang, Kai-Wei. 2019a. *VisualBERT: A Simple and Performant Baseline for Vision and Language*.

Li, Yong-Lu and Xu, Liang and Liu, Xinpeng and Huang, Xijie and Xu, Yue and Chen, Mingyang and Ma, Ze and Wang, Shiyi and Fang, Hao-Shu and Lu, Cewu. 2019b. *HAKE: Human Activity Knowledge Engine*. arXiv. ArXiv:1904.06539 [cs].

Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C. Lawrence. 2014. *Microsoft COCO: Common Objects in Context*. Springer International Publishing, Lecture Notes in Computer Science.

Lu, Jiasen and Batra, Dhruv and Parikh, Devi and Lee, Stefan. 2019. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. Curran Associates, Inc.

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

Nikolaus, Mitja and Salin, Emmanuelle and Ayache, Stephane and Fourtassi, Abdellah and Favre, Benoit. 2022. *Do Vision-and-Language Transformers Learn Grounded Predicate-Noun Dependencies?* Association for Computational Linguistics.

Ordonez, Vicente and Kulkarni, Girish and Berg, Tamara. 2011. *Im2Text: Describing Images Using 1 Million Captioned Photographs*. Curran Associates, Inc.

Parcalabescu, Letitia and Cafagna, Michele and Muradjan, Lilitta and Frank, Anette and Calixto, Iacer and Gatt, Albert. 2022. *VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena*. Association for Computational Linguistics.

Parcalabescu, Letitia and Gatt, Albert and Frank, Anette and Calixto, Iacer. 2021. *Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks*. Association for Computational Linguistics.

Pham, Khoi and Kafle, Kushal and Lin, Zhe and Ding, Zhihong and Cohen, Scott and Tran, Quan and Shrivastava, Abhinav. 2021. *Learning To Predict Visual Attributes in the Wild*.

Pratt, Sarah and Yatskar, Mark and Weihs, Luca and Farhadi, Ali and Kembhavi, Aniruddha. 2020. *Grounded Situation Recognition*. Springer International Publishing, Lecture Notes in Computer Science.

Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and Krueger, Gretchen and Sutskever, Ilya. 2021. *Learning Transferable Visual Models From Natural Language Supervision*. PMLR.

Schuhmann, Christoph and Vencu, Richard and Beaumont, Romain and Kaczmarczyk, Robert and Mullis, Clayton and Katta, Aarush and Coombes, Theo and Jitsev, Jenia and Komatsuzaki, Aran. 2021. *LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs*. NeurIPS Workshop Datacentric AI, online (online), 14 Dec 2021 - 14 Dec 2021.

Sharma, Piyush and Ding, Nan and Goodman, Sebastian and Soricut, Radu. 2018. *Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning*. Association for Computational Linguistics.

Shekhar, Ravi and Pezzelle, Sandro and Herbelot, Aurélie and Nabi, Moin and Sangineto, Enver and Bernardi, Raffaella. 2017a. *Vision and Language Integration: Moving beyond Objects*.

Shekhar, Ravi and Pezzelle, Sandro and Klimovich, Yauhen and Herbelot, Aurélie and Nabi, Moin and Sangineto, Enver and Bernardi, Raffaella. 2017b. *FOIL it! Find One mismatch between Image and Language caption*. Association for Computational Linguistics.

Suhr, Alane and Zhou, Stephanie and Zhang, Ally and Zhang, Iris and Bai, Huajun and Artzi, Yoav. 2019. *A Corpus for Reasoning about Natural Language Grounded in Photographs*. Association for Computational Linguistics.

Tan, Hao and Bansal, Mohit. 2019. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. Association for Computational Linguistics.

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

Wang, Tan and Lin, Kevin and Li, Linjie and Lin, Chung-Ching and Yang, Zhengyuan and Zhang, Hanwang and Liu, Zicheng and Wang, Lijuan. 2023. *Equivariant Similarity for Vision-Language Foundation Models*. arXiv. ArXiv:2303.14465 [cs].

Yang, Jinyu and Duan, Jiali and Tran, Son and Xu, Yi and Chanda, Sampath and Chen, Liqun and Zeng, Belinda and Chilimbi, Trishul and Huang, Junzhou. 2022. *Vision-Language Pre-Training With Triple Contrastive Learning*.

Young, Peter and Lai, Alice and Hodosh, Micah and Hockenmaier, Julia. 2014. *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions*.

Mert Yuksekgonul and Federico Bianchi and Pratyusha Kalluri and Dan Jurafsky and James Zou. 2023. *When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?*

Zhao, Tiancheng and Zhang, Tianqi and Zhu, Mingwei and Shen, Haozhan and Lee, Kyusong and Lu, Xiaopeng and Yin, Jianwei. 2022. *VL-CheckList: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations*. arXiv. ArXiv:2207.00221 [cs].

Zhou, Luowei and Xu, Chenliang and Corso, Jason. 2018. *Towards Automatic Learning of Procedures From Web Instructional Videos*. Number: 1.

A. Appendix

For more insight into the properties of EViL-Probe and the performance models achieve on it, we include some additional statistics that may be of interest.

A.1. Source Datasets

Section 3 only describes which datasets contribute to which of the linguistic categories in EViL-Probe. Table 4 gives details on how these different source datasets were processed to fit the desired format. The result are probes that consist of two descriptions of the same image.

When it comes to statistics on the number of probes per dataset, Section 3 only gives values that are aggregated to category-level. This is why we include more extensive overviews of the individual subdatasets of EViL-Probe. In Table 5, we give statistics on dataset level. Table 6 shows exemplary probes of the different datasets and the average likelihood of the texts in these datasets, determined using our language-only baseline.

These values can serve as an estimate of the overall quality of the language in the probes. Lower likelihoods may hint at more unusual, perhaps ungrammatical or implausible texts. The two sets of random probes are based on crowdsourced high-quality captions, which is reflected in higher likelihoods for the texts in these probes. The majority of the other subsets falls into the same range as these, with just some exceptions where the average likelihood is lower. It is lowest for the SWiG-based probes, which is in tune with our manual inspection of these: They are frequently ungrammatical and use uncommon terms, such as describing *water* as *h2o* or a *horse* as an *equus caballus*.

A.2. Language-Only Baseline

In Section 4.1, we report our language-only baseline. These results are however aggregated to category-level. Figures 6 and 7 therefore give the per-dataset differences in likelihood. This reveals FOIL-IT and some of the datasets from VL-Checklist to be solved more easily by our language-only baseline.

A.3. Influence of Model Size

In Section 4.2, we limit results to one version per model (the respective largest available one) for the sake of brevity. Table 8 includes results for different model sizes. Perhaps unsurprisingly, albeit minor in some places, larger models tend to perform better. The only exception to this is BLIP, where the *large* model is outperformed by the *base* model trained with 129M images. Details on the different

models are shown in Table 7. We also give information on which image datasets went into the pretraining of the respective models, for transparency on how these may overlap with the image data in EViL-Probe (which can be gathered from Table 5).

A.4. Probing Results

As an extension to the results in Table 3, Table 8 gives more detailed performance statistics of the different models, including separate results for matching and mismatched texts.

Results for the negative probes reveal that UNITER, VILLA and SOHO have some imbalance in their ability to classify matching and mismatched texts. They show high recall for matching texts, but low recall for mismatched ones, because they tend to accept texts rather than reject them. These numbers are much more balanced for the models that use a Vision Transformer to process the images (ALBEF, TCL and BLIP).

A.5. Analysis: Coreference Probes

In compiling a benchmark such as EViL-Probe, source datasets should not be included blindly, and we felt it necessary to delve deeper into the coreference probes of the VALSE benchmark (Parcalabescu et al., 2022). These are of the form *Statement. Question. Yes/No*, supposedly requiring coreference resolution between statement and question to determine whether they match the respective image. Upon manual inspection of these probes, we did however note examples such as *A book that is on the edge of a desk. is this a color photo? yes*, which can be answered without considering the statement. We thus include a derived version of the coreference probes that reduces them to just *Question? Yes/No*. Table 9 shows how this adaptation affects performance. Focusing on our target metric of paired accuracy, we see that dropping the question tends to increase performance. This seems to confirm that the question, and therefore coreference resolution between question and statement, is not necessary to solve the probes. On the contrary, the questions seem to introduce noise that makes solving the probes harder. This finding is why we include the coreference probes as part of our *negation* probes rather than dedicating a *coreference* category to them.

Dataset	Initial Example Structure	Adapted Structure	Comments
ARO (Yuksekgonul et al., 2023)	(i, t, f), t matches i, but f does not	taken as-is	
FOIL-IT (Shekhar et al., 2017b)	(i, t, f), t matches i, but f does not	taken as-is	
Predicate-noun (Nikolaus et al., 2022)	(i, t, f), t matches i, but f does not	taken as-is	
VALSE (Parcalabescu et al., 2022)	(i, t, f), t matches i, but f does not	taken as-is	
Visual Spatial Reasoning (Liu et al., 2023)	(i, t, f), t matches i, but f does not	taken as-is	we use the cleaned (validated) data provided by Liu et al. (2023)
VL-Checklist (Zhao et al., 2022)	(i, t, f), t matches i, but f does not	taken as-is	
EqBen (Wang et al., 2023)	(i_1, i_2, t_1, t_2), t_1 (t_2) matches i_1 (i_2), but not i_2 (i_1)	(i_1, t_1, t_2), (i_2, t_2, t_1)	
Winoground (Thrush et al., 2022)	(i_1, i_2, t_1, t_2), t_1 (t_2) matches i_1 (i_2), but not i_2 (i_1)	(i_1, t_1, t_2), (i_2, t_2, t_1)	
Compositional Visual Genome (Jiang et al., 2022)	(i_1, i_2, t_1, t_2), t_1 (t_2) matches i_1 (i_2), but not i_2 (i_1)	(i_1, t_1, t_2), (i_2, t_2, t_1)	t_2 is SVO triple; we only use those where the corresponding sentence can be looked up from the remaining dataset
SVO Probes (Hendricks and Nematzadeh, 2021)	(i_1, i_2, t_1, t_2), t_1 (t_2) matches i_1 (i_2), but not i_2 (i_1)	(i_1, t_1, t_2), (i_2, t_2, t_1)	t_2 is SVO triple, we only use those where the corresponding sentence can be looked up from the remaining dataset
Counting probe (Parcalabescu et al., 2021)	(i, t, f_1, f_2, f_3), t matches i, but f_j do not	(i, t, f_1), (i, t, f_2), (i, t, f_3)	we use the declarative statements
Winoground (augmented) Diwan et al. (2022)	$(i, t, f_1, ..., f_n)$, t and all f_j match i, f_j are paraphrases of t	(i, t, f_j) for all f_j	
High Level Dataset (Cafagna et al., 2023)	$(i, t_1, t_2, ..., t_5, f_1, f_2, f_3)$, all t_j and f_j match i, but with different focus	(i, t_j, f_1), (i, t_j, f_2), (i, t_j, f_3) for all t_j	
Flickr30k (Young et al., 2014)	$(i, t_1, t_2, ..., t_5)$, all t_j match i	(i, t_j, f_j) for all t_j, with a random mismatched description f_j drawn from other examples	all descriptions appear once as match and once as mismatch
MS COCO (Lin et al., 2014)	$(i, t_1, t_2, ..., t_5)$, all t_j match i	(i, t_j, f_j) for all t_j, with a random mismatched description f_j drawn from other examples	all descriptions appear once as match and once as mismatch

Table 4: Source datasets of EViL-Probe and how they were processed to fit its format. Each probe in EViL-Probe is made up of two descriptions of the same image.

Category Dataset	Text Source	Image Source	# Images	# Triples
Negative Probes				
Attribute				
- ARO-attribute	ARO	Visual Genome (Krishna et al., 2017)	4517	28053
- EQ-Kubric-attribute	Wang et al. (2023)	EqBen (Wang et al., 2023)	4000	4000
- VL-Checklist-attribute-action	Zhao et al. (2022)	Visual Genome ()VisualGenome	3321	4854
- VL-Checklist-attribute-material	Zhao et al. (2022)	Visual Genome (Krishna et al., 2017)	9435	14207
- VL-Checklist-attribute-size	Zhao et al. (2022)	Visual Genome (Krishna et al., 2017)	11453	15921
- VL-Checklist-attribute-state	Zhao et al. (2022)	Visual Genome (Krishna et al., 2017)	4551	5391
Color				
- EQ-SD-color	Wang et al. (2023)	EqBen (Wang et al., 2023)	426	426
- VL-Checklist-attribute-color	Zhao et al. (2022)	Visual Genome (Krishna et al., 2017)	36230	74676
Random				
- Flickr30k-Random	Young et al. (2014)	Flickr30k (Young et al., 2014)	1000	5000
- MS COCO-Random	Lin et al. (2014)	MS COCO 2014 (test) (Lin et al., 2014)	5000	25010
Image Type				
- EQ-SD-image-type	Wang et al. (2023)	EqBen (Wang et al., 2023)	414	414
Negation				
- VALSE-coreference	Parcalabescu et al. (2022)	MS COCO (train/val2014) (Lin et al., 2014)	1057	1057
- VALSE-coreference-q-only	Parcalabescu et al. (2022)	MS COCO (train/val2014) (Lin et al., 2014)	1057	1057
- VALSE-existence	Parcalabescu et al. (2022)	Visual Genome (Krishna et al., 2017)	533	533
Noun				
- ComVG-noun	Jiang et al. (2022)	Visual Genome (Krishna et al., 2017)	473	2040
- EQ-SD-noun	Wang et al. (2023)	EqBen (Wang et al., 2023)	1448	1448
- FOIL-IT-noun	Shekhar et al. (2017b)	MS COCO (val2014) (Lin et al., 2014)	32150	99456
- Predicate-Noun-subject	Nikolaus et al. (2022)	Open Images (Kuznetsova et al., 2020)	901	1098
- SVO-Probes-noun	Hendricks and Nematzadeh (2021)	SVO Probes (Hendricks and Nematzadeh, 2021)	3562	21285
- VL-Checklist-noun-HAKE	Zhao et al. (2022)	HAKE (Li et al., 2019b)	72207	143441
- VL-Checklist-noun-SWiG	Zhao et al. (2022)	SWiG (Pratt et al., 2020)	23139	109688
- VL-Checklist-noun-VG	Zhao et al. (2022)	Visual Genome (Krishna et al., 2017)	9963	115373
Number				
- Counting-probe-hard	Parcalabescu et al. (2021)	Visual Genome (Krishna et al., 2017)	1348	1348
- Counting-probe-standard	Parcalabescu et al. (2021)	Visual Genome (Krishna et al., 2017)	3610	3610
- EQ-Kubric-counting	Wang et al. (2023)	EqBen (Wang et al., 2023)	4000	4000
- VALSE-counting-adversarial	Parcalabescu et al. (2022)	Visual Genome (Krishna et al., 2017)	756	756
- VALSE-counting-balanced	Parcalabescu et al. (2022)	Visual Genome (Krishna et al., 2017)	991	991
- VALSE-counting-small-numbers	Parcalabescu et al. (2022)	Visual Genome (Krishna et al., 2017)	1000	1000
- VALSE-plurals	Parcalabescu et al. (2022)	MS COCO (val2017) (Lin et al., 2014)	939	1000
Semantic Role				
- VALSE-actant-swap	Parcalabescu et al. (2022)	SWiG (Pratt et al., 2020)	1028	1028
Spatial relations				
- ARO-relation	Yuksekgonul et al. (2023)	Visual Genome (Krishna et al., 2017)	5316	22921
- EQ-Kubric-Location	Wang et al. (2023)	EqBen (Wang et al., 2023)	4000	4000
- VALSE-Relations	Parcalabescu et al. (2022)	MS COCO (val2017) (Lin et al., 2014)	546	614
- VL-Checklist-relation-spatial	Zhao et al. (2022)	Visual Genome (Krishna et al., 2017)	8686	26183
- Visual-Spatial-Reasoning	Liu et al. (2023)	MS COCO (train/val2017) (Lin et al., 2014)	1179	3058
Verb				
- ComVG-verb	Jiang et al. (2022)	Visual Genome (Krishna et al., 2017)	374	648
- Predicate-Noun-object	Nikolaus et al. (2022)	Open Images (Kuznetsova et al., 2020)	1136	1486
- SVO-Probes-verb	Hendricks and Nematzadeh (2021)	SVO Probes (Hendricks and Nematzadeh, 2021)	9088	30502
- VALSE-action-replacement	Parcalabescu et al. (2022)	SWiG (Pratt et al., 2020)	779	779
- VL-Checklist-relation-action-HAKE	Zhao et al. (2022)	HAKE (Li et al., 2019b)	104815	201312
- VL-Checklist-relation-action-SWiG	Zhao et al. (2022)	SWiG (Pratt et al., 2020)	25147	25147
- VL-Checklist-relation-action-VG	Zhao et al. (2022)	Visual Genome (Krishna et al., 2017)	5281	12683
Video-based				
- EQ-AG	Wang et al. (2023)	Action Genome (Ji et al., 2020)	391744	391744
- EQ-GEBC	Wang et al. (2023)	GEBC (Wang et al., 2022)	3624	3624
- EQ-YouCook2	Wang et al. (2023)	YouCook2 (Zhou et al., 2018)	91680	91680
Word Order				
- Winoground	Thrush et al. (2022)	Winoground (Thrush et al., 2022)	708	708
Positive Probes				
Hypernyms				
- Winoground-hypernyms	Diwan et al. (2022)	Winoground (Thrush et al., 2022)	633	1232
Paraphrase				
- Winoground-backtranslation	Diwan et al. (2022)	Winoground (Thrush et al., 2022)	534	534
- Winoground-diverseparaphrase	Diwan et al. (2022)	Winoground (Thrush et al., 2022)	534	534
Perspective				
- High-Level-action	Cafagna et al. (2023)	MS COCO (train2014) (Lin et al., 2014)	1496	19969
- High-Level-rationale	Cafagna et al. (2023)	MS COCO (train2014) (Lin et al., 2014)	1486	19203
- High-Level-scene	Cafagna et al. (2023)	MS COCO 2014 (train2014) (Lin et al., 2014)	1497	20468
Slang				
- Winoground-slang	Diwan et al. (2022)	Winoground (Thrush et al., 2022)	417	608
Specificity				
- EQ-SD-specificity	Wang et al. (2023)	EqBen (Wang et al., 2023)	369	369
Word Order				
- Winoground-rulebased	Diwan et al. (2022)	Winoground (Thrush et al., 2022)	91	127

Table 5: Overview of the probing datasets in EViL-Probe and the linguistic categories they target.

Category Dataset	ø Likelihood Text 1	 Text 2	Example Probe Text 1 (Matches respective image)	 Text 2
Negative Probes				
Attribute				
- ARO-attribute	-36.50	-41.17	The metal floor and the striped shirt	The striped floor and the metal shirt
- EQ-Kubric-attribute	-61.43	-61.43	The black hat is above the green turtle toy	[...] grey hat [...] above the green turtle toy
- VL-Checklist-attribute-action	-28.52	-30.73	Hair on reading man	Hair on drinking man
- VL-Checklist-attribute-material	-24.28	-27.52	Tiled floor	Cement floor
- VL-Checklist-attribute-size	-25.51	-25.49	Long counter	Short counter
- VL-Checklist-attribute-state	-25.52	-26.70	Dry floor	Wet floor
Color				
- EQ-SD-color	-37.49	-37.49	A painting of black car	A painting of red car
- VL-Checklist-attribute-color	-25.85	-26.89	Brown flooring	Purple flooring
Random				
- Flickr30k-Random	-28.09	-28.09	A Boston terrier is running in the grass	Two nuns stand talking to a person
- MS COCO-Random	-27.35	-27.35	Girl blowing out the candle on an ice-cream	A Giraffe standing outside [...]
Image Type				
- EQ-SD-image-type	-28.27	-28.27	An oil painting of train	A pencil sketch of train
Negation				
- VALSE-coreference	-50.88	-51.24	A woman [...] is gasping. is she inside? yes.	A woman [...] is gasping. is she inside? no.
- VALSE-coreference-q-only	-25.21	-25.65	Is she inside? yes.	Is she inside? no.
- VALSE-existence	-16.03	-15.50	There are no pets pictured	There are pets pictured
Noun				
- ComVG-noun	-20.59	-21.48	A man is sitting at a table	A man is sitting on a bridge
- EQ-SD-noun	-30.68	-30.68	A photo of camel in the zoo	A photo of cattle in the zoo
- FOIL-IT-noun	-27.02	-34.75	A woman in a room with a cat	A woman in a room with a dog
- Predicate-Noun-subject	-21.37	-21.37	A woman is holding a bottle	A man is holding a bottle
- SVO-Probes-noun	-19.70	-19.89	A businessman walks down a beach	A businessman walks down a busy street
- VL-Checklist-noun-HAKE	-33.91	-36.94	Person ride horse	Person ride mouse
- VL-Checklist-noun-SWiG	-58.20	-63.03	Man pats man using manus at a place	Man pats man using dozer at a place
- VL-Checklist-noun-VG	-28.69	-35.69	Train has wheels	Snail has wheels
Number				
- Counting-probe-hard	-21.13	-20.42	There are 4 windows and doors	There are 6 windows and doors
- Counting-probe-standard	-20.98	-20.17	There are 4 pizzas	There are 5 pizzas
- EQ-Kubric-counting	-38.09	-38.09	1 brown bull is in the scene	There are 4 brown bulls in the scene
- VALSE-counting-adversarial	-25.81	-27.24	There are exactly 4 cows shown	There is exactly 1 cow shown
- VALSE-counting-balanced	-26.30	-26.13	There are exactly 5 buses	There are exactly 0 buses
- VALSE-counting-small-numbers	-26.54	-26.41	There is exactly 1 hot dog	There are exactly 2 hot dogs
- VALSE-plurals	-30.17	-30.09	A single man and woman are in a kitchen	A number of men and woman are in a kitchen
Semantic Role				
- VALSE-actant-swap	-16.07	-22.82	A man displays a certificate	A certificate displays a man
Spatial relations				
- ARO-relation	-28.24	-30.02	The truck is in front of the tree	The tree is in front of the truck
- EQ-Kubric-Location	-56.20	-56.20	[...] is behind the blue gloves	[...] is located on the top of the blue gloves
- VALSE-Relations	-25.96	-28.63	An elephant is standing in a dirt field	An elephant is standing beside a dirt field
- VL-Checklist-relation-spatial	-28.35	-32.89	Scissors on table	Scissors under table
- Visual-Spatial-Reasoning	-23.87	-23.75	The bus is behind the horse	The bus is in front of the horse
Verb				
- ComVG-verb	-21.23	-21.77	A man is standing on the water	A man is drinking the water
- Predicate-Noun-object	-21.74	-21.74	A woman is wearing glasses	A woman is standing
- SVO-Probes-verb	-20.12	-20.23	Girl stand near tree	A girl sits in a tree
- VALSE-action-replacement	-16.35	-19.68	The people package the food	The people eat the food
- VL-Checklist-relation-action-HAKE	-34.43	-38.41	Person sit on chair	Person break chair
- VL-Checklist-relation-action-SWiG	-54.74	-60.62	The elephants stampede in srubland	The elephants mopping in scrubland
- VL-Checklist-relation-action-VG	-29.38	-38.23	Man wearing a shirt	Man holding shirt
Video-based				
- EQ-AG	-25.21	-25.21	The person is sitting on the chair [...]	The person is not contacting the chair [...]
- EQ-GEBC	-46.64	-46.64	Man [...] standing in the pool	[...] move his head inside the water [...]
- EQ-YouCook2	-36.06	-36.06	Rinse the rice	Add vinegar and mix together
Word Order				
- Winoground	-31.61	-31.61	A car smashed into a tree	A tree smashed into a car
Positive Probes				
Hypernyms				
- Winoground-hypernyms	-32.20	-34.57	A car smashed into a tree	A vehicle smashed into a tree
Paraphrase				
- Winoground-backtranslation	-33.40	-30.79	A tree smashed into a car	Tree crashed into car
- Winoground-diverseparaphrase	-33.40	-30.79	A brown dow is on a white couch	A brown dog sitting on a white sofa
Perspective				
- High-Level-action	-27.08	-25.33	A man in glasses is wearing a tie	The person is posing for a photo
- High-Level-rationale	-27.06	-25.91	A man in glasses is wearing a tie	He's working and took a professional photo
- High-Level-scene	-27.05	-21.62	A man in glasses is wearing a tie	In an office
Slang				
- Winoground-slang	-32.44	-51.01	A young person kisses and old person	A young bod kisses and old bod
Specificity				
- EQ-SD-specificity	-41.22	-27.32	A photo of dog wearing a scarf	A photo of dog
Word Order				
- Winoground-rulebased	-37.76	-41.44	They're enjoying cold water on a hot day	On a hot day they're enjoying cold water

Table 6: Exemplary probes and the average likelihood of the texts in the individual subdatasets.

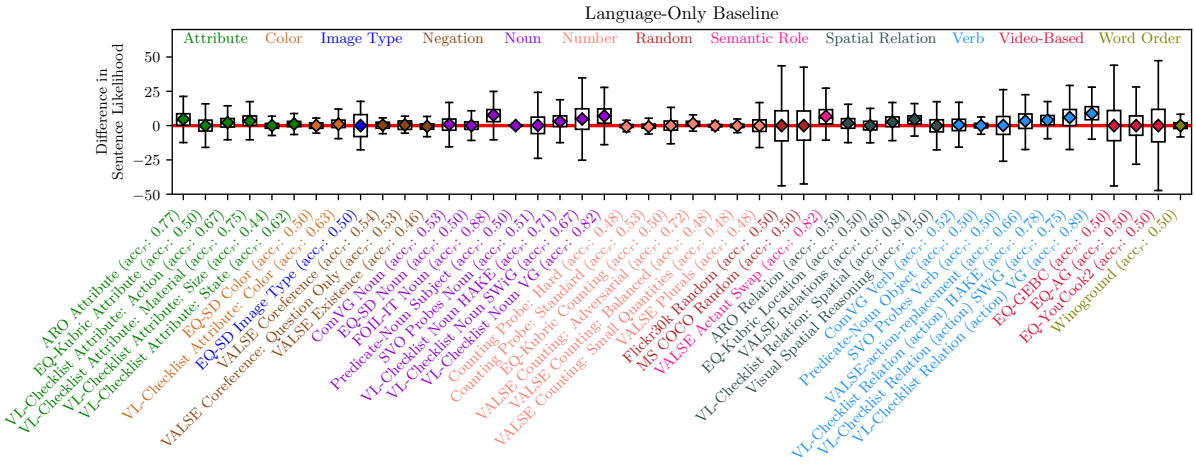

Figure 6: Language-only baseline for the negative probes. The ranking-based accuracy this approach would achieve is given in parentheses.

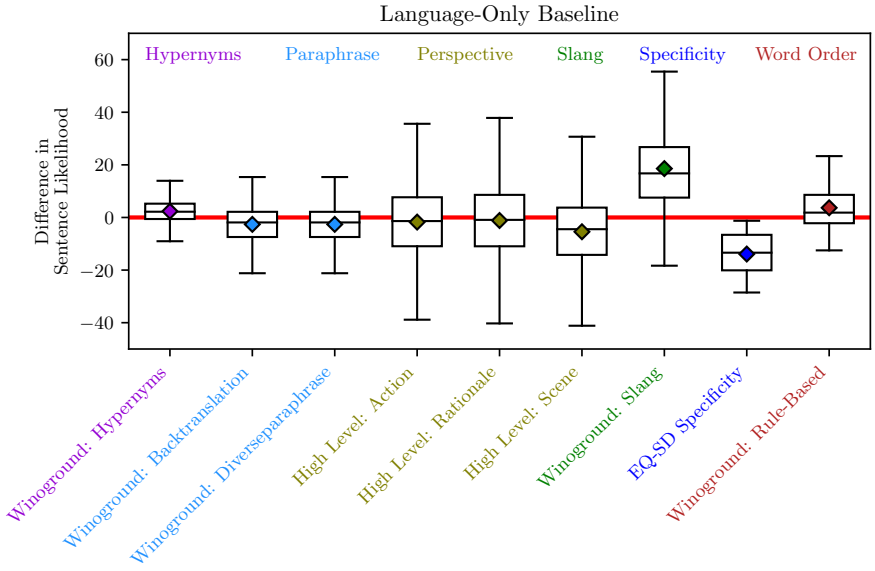

Figure 7: Language-only baseline for the positive probes.

Model		Architecture	Visual Input	Datasets in Pretraining	#Images
LXMERT (Tan and Bansal, 2019)		2-stream	Faster R-CNN	COCO, VG	0.2M
UNITER (Chen et al., 2020)	base	1-stream	Faster R-CNN	CC, SBU, COCO, VG	4M
	large	1-stream	Faster R-CNN	CC, SBU, COCO, VG	4M
VILLA (Gan et al., 2020)	base	1-stream	Faster R-CNN	CC, SBU, COCO, VG	4M
	large	1-stream	Faster R-CNN	CC, SBU, COCO, VG	4M
SOHO (Huang et al., 2021)		end2end	ResNet	COCO, VG	0.2M
ALBEF (Li et al., 2021)	base	end2end	ViT-B/16	CC, SBU, COCO, VG	4M
	large	end2end	ViT-B/16	CC, SBU, COCO, VG, CC12M	14M
TCL (Yang et al., 2022)	base	end2end	ViT-B/16	CC, SBU, COCO, VG	4M
BLIP (Li et al., 2022)	base_{14M}	end2end	ViT-B/16	CC, SBU, COCO, VG, CC12M	14M
	base_{129M}	end2end	ViT-B/16	CC, SBU, COCO, VG, CC12M, LAION	129M
	large	end2end	ViT-L/16	CC, SBU, COCO, VG, CC12M, LAION	129M

Datasets are: Conceptual Captions (CC, Sharma et al. (2018)), SBU Captions (SBU; Ordonez et al. (2011)), MS COCO 2014 (COCO; Lin et al. (2014)), Visual Genome (VG, Krishna et al. (2017)), Conceptual 12M (CC12M, Changpinyo et al. (2021)), LAION (Schuhmann et al., 2021). Faster R-CNN (Anderson et al., 2018) is pretrained on Visual Genome, ResNet (He et al., 2016) and all Visual Transformer (ViT, Dosovitskiy et al. (2021)) models are pretrained on ImageNet (Deng et al., 2009).

Table 7: Overview of the models we test with EViL-Probe.

	LXMERT	UNITER-B	UNITER-L	VILLA-B	VILLA-L	SOHO	ALBEF-B	ALBEF-L	TCL-B	BLIP-B$_{14M}$	BLIP-B$_{129M}$	BLIP-L
Negative Probes												
- Acc$_{ranked}$	.56	.74	.76	.76	.78	.64	.72	.74	.73	.75	.76	.76
- % Accept	.46	.78	.70	.80	.74	.78	.39	.48	.45	.39	.63	.61
- Acc	.54	.60	.62	.59	.61	.57	.61	.63	.62	.63	.63	.64
- Precision$_{match}$	.56	.58	.60	.58	.59	.56	.65	.64	.64	.67	.62	.63
- Recall$_{match}$	.50	.88	.82	.89	.85	.85	.50	.61	.57	.52	.76	.74
- Precision$_{mismatch}$	.54	.71	.71	.72	.72	.63	.63	.66	.65	.63	.68	.68
- Recall$_{mismatch}$	.59	.32	.41	.29	.37	.30	.73	.65	.67	.73	.50	.53
- Acc$_{paired}$	.27	.22	.26	.20	.24	.20	.26	.29	.27	.28	.29	.30
Positive Probes												
- Acc (= % Accept)	.47	.86	.80	.88	.83	.82	.44	.54	.52	.54	.75	.72
- Acc$_{paired}$	.25	.79	.70	.81	.75	.73	.28	.41	.37	.39	.65	.62

Table 8: Detailed results for all tested models, split into positive and negative probes. As positive probes consist entirely of matching image-text pairs, not all metrics can be calculated for these examples.

	LXMERT	UNITER-B	UNITER-L	VILLA-B	VILLA-L	SOHO	ALBEF-B	ALBEF-L	TCL-B	BLIP-B$_{14M}$	BLIP-B$_{129M}$	BLIP-L
- Acc$_{ranked}$	.01	.03	.00	.00	.00	.00	-.04	.01	.02	.00	-.05	-.02
- Acc$_{paired}$	-.05	.03	.05	.03	.03	.09	-.02	.01	.04	.00	.07	.09

Table 9: Shift in performance when reducing the *Statement. Question? Yes/No* descriptions in the VALSE *coreference* probes to just *Question? Yes/No*. Overall, dropping the statement tends to increase performance. This indicates that the task is solvable without them, and that they potentially even introduce noise.