

# Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality

Tristan Thrush<sup>¶\*</sup>, Ryan Jiang<sup>‡</sup>, Max Bartolo<sup>§</sup>,

Amanpreet Singh<sup>¶</sup>, Adina Williams<sup>†</sup>, Douwe Kiela<sup>¶</sup>, Candace Ross<sup>†\*</sup>

<sup>¶</sup> Hugging Face; <sup>†</sup> Facebook AI Research; <sup>‡</sup> University of Waterloo; <sup>§</sup> University College London

tristan@huggingface.co, ccross@fb.com

## Abstract

We present a novel task and dataset for evaluating the ability of vision and language models to conduct visio-linguistic compositional reasoning, which we call Winoground. Given two images and two captions, the goal is to match them correctly—but crucially, both captions contain a completely identical set of words, only in a different order. The dataset was carefully hand-curated by expert annotators and is labeled with a rich set of fine-grained tags to assist in analyzing model performance. We probe a diverse range of state-of-the-art vision and language models and find that, surprisingly, none of them do much better than chance. Evidently, these models are not as skilled at visio-linguistic compositional reasoning as we might have hoped. We perform an extensive analysis to obtain insights into how future work might try to mitigate these models’ shortcomings. We aim for Winoground to serve as a useful evaluation set for advancing the state of the art and driving further progress in the field. The dataset is available at <https://huggingface.co/datasets/facebook/winoground>.

## 1. Introduction

Despite the impressive performance of pretrained vision and language transformers on a wide variety of multimodal tasks [47, 51, 56], they remain poorly understood [8, 19, 46, 47]. One important question is to what extent such models are able to conduct unimodal and multimodal compositional reasoning. For humans, the visual differences between images depicting “the tree is in the shopping cart” and “the shopping cart is in the tree” will be blatantly obvious, even when the words in the captions are identical—but is the same true for machines?

While matching simple images and captions may seem almost too trivial a task, recent work in NLP has shown

\*Equal contribution. TT, AS, and DK conducted most of the work for this paper when they were at Facebook AI Research.



(a) some plants surrounding a lightbulb



(b) a lightbulb surrounding some plants

Figure 1. An example from Winoground. The two sentences contain the same words but in a different order. The task of understanding which image and caption match is trivial for humans but much harder for vision and language models. Every model that we tested (UNITER, ViLLA, VinVL, VisualBERT, ViLT, LXMERT, ViLBERT, UniT, FLAVA, CLIP, VSE++, and VSRN) fails to correctly pair the images and captions, except the large checkpoint of ViLLA by a very thin margin (0.00013 confidence).

that transformers are often remarkably insensitive to word order [70]. Understanding the relationship between text in captions and corresponding visual content is a fundamental goal of computer vision, and the fact that different word orders correspond to wildly different visual depictions should be reflected in the capabilities of our models.

Motivated by this, we propose a novel task, called Winoground, for measuring visio-linguistic compositional reasoning, whereby two images and two captions have to be matched correctly; both captions contain exactly the same set of words, ordered in such a way that each describes primarily one of the images. To perform well on Winoground, models must not only encode text and images well (i.e., be sensitive to the compositional structure present in each modality), but they also must be able to synthesize information across the two modalities.

We draw inspiration from the Winograd Schema Challenge [44], which tests the commonsense capabilities of models. In the challenge, a model is given two sentences

that minimally differ and is tasked with performing coreference resolution. The Winograd twin sentence format has been used for a variety of language-related tasks [59, 60, 91]. In this work, we study the image-grounding of twin sentences with identical but differently ordered words.

Winoground was hand-crafted by expert annotators and is labeled with a rich set of fine-grained tags to assist in analyzing model performance. In efforts to shed better light on what exactly models learn, the NLP community has designed a wide variety of “probing tasks”: specialized, targeted tasks meant specifically for evaluation. The primary purpose of Winoground is to serve as a probing task for vision and language models. See Fig. 1 for an example.

We evaluate a variety of state-of-the-art vision and language (V&L) transformers [12, 23, 35, 40, 47, 51, 56, 68, 76, 90] and RNN-based models [21, 45]. Surprisingly, all of the models rarely—and if so only barely—outperform chance. Our findings indicate that the visio-linguistic compositional reasoning capabilities of these models fall dramatically short of what we might have hoped.

In what follows, we introduce the Winoground task and dataset. We then describe the models we tested and discuss our findings. Next, we conduct an analysis of the performance of different models. We hope that insights from this work will lead to more robust vision and language models.

## 2. Related Work

**Visio-linguistic stress testing.** There are a number of existing multimodal stress tests about correctly understanding implausible scenes [13], exploitation of language and vision priors [11, 27], single word mismatches [64], hate speech detection [26, 32, 41, 92], memes [39, 75], ablation of one modality to probe the other [22], distracting models with visual similarity between images [7, 33], distracting models with textual similarity between many suitable captions [1, 17], collecting more diverse image-caption pairs beyond the predominately English and North American/Western European datasets [50], probing for an understanding of verb-argument relationships [30], counting [53], or specific model failure modes [65, 69]. Many of these stress tests rely only on synthetically generated images, often with minimal visual differences, but no correspondingly minimal textual changes [80]. Other datasets test models with a single caption [74] or a single image [6, 37]. There are also purely visual stress tests with naturalistic images: ImageNet-C/ImageNet-P [31] tests models on perturbations for a variety of image features. Unlike Winoground, these stress tests tend to come from existing datasets that have images and text from typical training domains, such as Conceptual Captions [63], COCO [48], Visual7W [93] and VQA [3, 27]. None of them hold the set of words constant in the captions, which is what allows us to carefully test for compositional reasoning without any biases stemming from

the presence of altogether different words. While it is theoretically possible for unstructured bag of words models to do well on these previous datasets, that is not possible on Winoground.

**Probing.** Measuring what exactly a model knows about word order and linguistic structure has been explored in natural language processing. Sinha et al. [70] found that word order information does not have a large impact on performance when pretraining large transformer language models, across a variety of metrics. This suggests that transformers use high-level word co-occurrence statistics, which gives the illusion of an understanding of word order. Other work in this space has tried to understand what models know about syntax [24, 28, 34, 49, 54, 71, 83] or the complex interaction between syntactic and semantic categories [38, 78, 81, 82].

**Winograd schemas.** The Winograd Schema Challenge [44] was named after a coreference resolution problem presented by Terry Winograd [85]. The goal is to correctly resolve (an) ambiguous referent(s) in two English sentences. The sentences have a minor difference that changes how a human resolves the referent. Winograd schema examples are easily handled by humans, and commonsense reasoning is said to be required [4]. For example, in the sentence “*The city councilmen refused the demonstrators a permit because they [feared/advocated] violence*”, the pronoun *they* can either refer to the councilmen or to the demonstrators depending on which word is chosen. The format has been used in a variety of other tasks and datasets. For instance, Sakaguchi et al. [60] introduce WinoGrande: a large-scale approach to building a Winograd Schema dataset that uses Amazon Mechanical Turk to generate sentences instead of expert annotators like the original work of Levesque et al. [44]. Other approaches use ambiguous pronouns in sentences to probe for gender biases in models [59, 91]. See Kotcijan et al. [42] for an in-depth review. Winoground is the first work to apply these ideas to the vision and language domain, by using twin captions with identical word content and two images that are each associated with one caption over the other.

## 3. Winoground

In this section, we describe how the dataset was constructed and how performance on the task is to be measured.

### 3.1. Dataset

The Winoground dataset was hand-curated by four expert annotators with extensive experience in vision and language research as well as computational linguistics. Let  $(C_0, I_0)$  and  $(C_1, I_1)$  be two image-caption pairs. An example satisfies the Winoground schema if and only if:

- $(C_0, I_0)$  and  $(C_1, I_1)$  are preferred by the annotator over  $(C_1, I_0)$  and  $(C_0, I_1)$ ; and

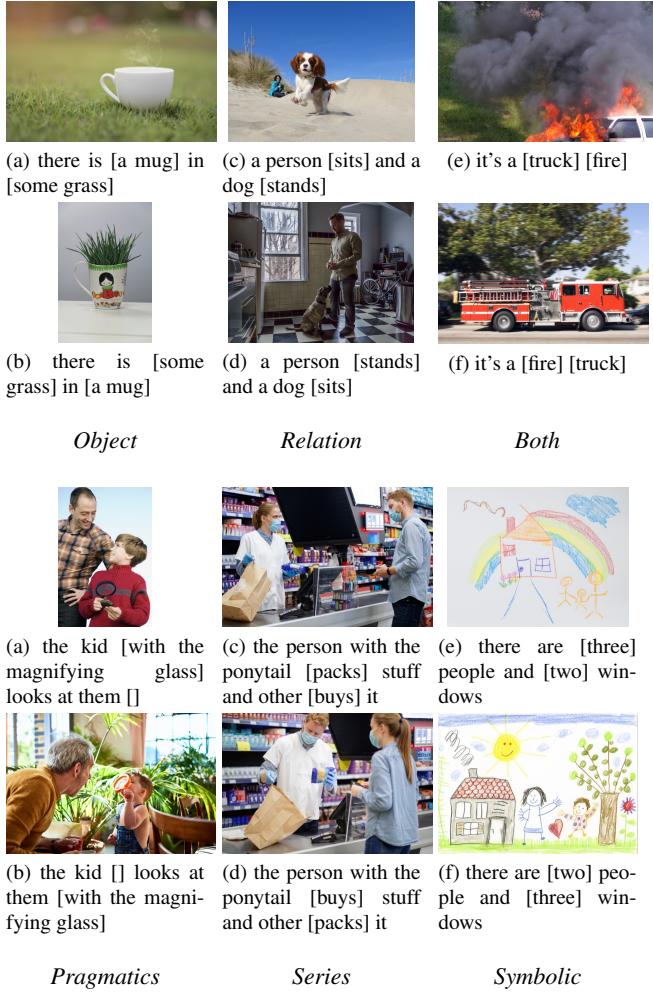


Figure 3. Examples from our dataset for the swap-dependent linguistic tags (top) and visual tags (bottom). The visual examples are additionally tagged with the *Relation* tag, and 1, 2, and 1 main predicates from left to right. The linguistic examples are additionally tagged with 2, 1, and 1 main predicates from left to right.

- $C_0$  and  $C_1$  have the same words and/or morphemes but the order differs.

We have secured a license from Getty Images to distribute images for research purposes. Thus, the expert annotators were given access to the Getty Images API [25], and tasked with jointly creating captions and finding images to compose examples. We encouraged them to be as creative as possible, and to mark each of their examples with fine-grained linguistic tags. If applicable, annotators also marked examples with one or more visual reasoning tags.

The annotators created a total of 70 linguistic tags for the swaps that make caption pairs different. This set of tags can be split into three broad groups: objects, relations, and swaps involving both relations and objects. Object swaps reorder elements such as noun phrases that tend to refer

Category	Tag	Count
Linguistic <sub>swap-dep.</sub>	Object	141
	Relation	233
	Both	26
Linguistic <sub>swap-indep.</sub>	1 Main Pred	293
	2 Main Preds	108
Visual	Symbolic	41
	Series	31
	Pragmatics	24

Table 1. Linguistic and visual tag counts in the Winoground dataset. Every example has a linguistic tag; only examples that contain the visual phenomena have visual tags.

to objects in the real world. Relation swaps reorder elements such as verbs, adjectives, prepositions, and/or adverbs, which tend to take nouns referring to objects as semantic arguments [2]. Swaps of both relations and objects can involve two separate swaps, or can involve a single swap that changes parts of speech (e.g., “it's a [fire] [truck]” vs. “it's a [truck] [fire]”). Examples of each broad tag group can be seen in Fig. 3. For examples for each fine-grained linguistic tag, see Appendix C.

Separately, the annotators tagged examples for how many main predicates were in the captions, which is not dependent on the specific swap happening between the two captions. For example, “left is blue and right is red” has two main predicates and “water is in a bottle” has one main predicate. It turned out that all examples in Winoground have either one main predicate or two.

Finally, examples were tagged from a set of three non-mutually exclusive visual reasoning tags, which are tied in some way to the images in an example, and not necessarily the captions. The “Pragmatics” tag comprises examples where the images need to be interpreted non-literally due to idiomatic uses of language in a caption (e.g. “it starts with Z and ends with A” describing an image of a Zebra) or due to attachment preferences of prepositional phrases in the captions (e.g. “the kid looks at them with the magnifying glass” describing an image of a child looking at someone through a magnifying glass with greater confidence than an image of a child looking at someone while holding a magnifying glass at their side). The “Symbolic” tag represents whether a symbolic depiction of something must be understood to make a correct prediction (e.g., objects in a child’s drawing). Lastly, the “Series” tag is given to examples where both images come from the same photo series on Getty, which typically means that the same people occur in both images, with a similar background and in similar lighting.

See Fig. 3 for representative examples of the tags, and Tab. 1 for tag counts. As noted, Winoground is a probing

dataset and so we prioritize clean, expert annotations over mere size. Our dataset has 1600 image-text pairs in total, with 800 correct and 800 incorrect pairings. These comprise 400 examples, with 800 unique captions and images.

### 3.2. Metrics

Performance on Winoground is computed according to three different metrics that evaluate different aspects of the models’ visio-linguistic reasoning abilities. The first metric is the **text score**, which measures whether a model can select the correct caption, given an image. Given images  $I_0$  and  $I_1$  and captions  $C_0$  and  $C_1$ , the text score for an example  $(C_0, I_0, C_1, I_1)$  is computed according to:

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

where  $s(\cdot)$  is the model’s score for the image/caption pair. This metric tests whether the ground truth caption for a given image in our dataset is scored higher than the alternative caption *and* whether this holds for the other image/caption pair in the example too.

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

$$g(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \quad \text{and } s(C_1, I_1) > s(C_1, I_0) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

This metric tests whether the ground truth image for a given caption is scored higher than the image corresponding to the alternative caption *and* whether this holds vice versa.

Our final metric combines the previous two. In their analysis of the Winograd Schema Challenge, Elazar et al. [20] find that evaluation metrics tend to overestimate model performance by computing scores for the twin sentences individually instead of as a set. So, we also evaluate using the **group score**, where every combination for a given example  $\{(C_0, I_0), (C_0, I_1), (C_1, I_0), (C_1, I_1)\}$  must be correctly scored by the model in order for the example to be considered correct. The group score in our framework is computed according to:

$$h(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } f(C_0, I_0, C_1, I_1) \\ & \quad \text{and } g(C_0, I_0, C_1, I_1) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

## 4. Experimental Setup

We evaluate various configurations of the following multimodal transformers: CLIP [56], FLAVA [68], LXMERT

[76], UniT [35], UNITER [12], VILLA [23], VinVL [90], ViLT [40], VisualBERT [47] and ViLBERT [51]. We also evaluate several configurations of two types of RNN-based models: VSE++ [21] and VSRN [45]. We detail differences between these models and provide a high-level overview in Tab. 2. We also establish a human baseline using crowd-workers, as described in Sec. 4.3.

### 4.1. Vision & Language Transformers

**Image and language embedding.** All transformer models we evaluate use a pretrained BERT tokenizer [16], except CLIP, which uses a Byte-Pair Encoding tokenizer [62] trained from scratch. For the image embedding, five transformers (VisualBERT, ViLBERT, LXMERT, UNITER, VILLA) [12, 23, 47, 51, 76] use region features extracted from the  $f_{\text{c6}}$  layer of a Faster R-CNN [58] trained on Visual Genome [43]. VinVL trains its own feature extractor on a large combined dataset from public sources with a unified object vocabulary [90]. The CLIP, FLAVA, and ViLT that we test all use Vision Transformer (ViT) [18]. In ViT, images are flattened into patches that are linearly projected and combined with a position encoding. UniT [35] alternatively uses a transformer network [79] on top of a convolutional network following Carion et al. [9].

**Single-stream vs. dual-stream encoders.** Vision and language transformers are mainly single- or dual-stream models: the embeddings for the image and text modalities are either concatenated and then jointly encoded (single-stream), or encoded by two separate modality-specific encoders with optional cross-modality fusion (dual-stream). Five of our transformers are single-stream [12, 23, 40, 47, 90]. VinVL additionally concatenates object tags, which are the set of objects detected by the X152-C4 model during feature extraction, to the language tokens before encoding. All single-stream models use merged attention, where the language and visual input attend to both themselves and the other modality. The dual-stream transformers we evaluate are CLIP, FLAVA, UniT, LXMERT and ViLBERT [35, 51, 56, 68, 76]. CLIP and the contrastive configuration of FLAVA lack cross-modal attention. ViLBERT has language-only transformer layers that are then fused by cross-modal transformer layers. LXMERT, the ITM configuration of FLAVA, and UniT each use language-only and vision-only layers that are also fused by cross-modal transformer layers, which perform a combo of modality-specific attention and co-attention across modalities.

**Pretraining objectives.** V&L transformers use a number of pretraining objectives including but not limited to masked language modeling, masked region modeling (classification of object classes and regression over image features) and image-text matching. As we are evaluating a model’s ability to determine if an image and a corresponding caption match, we select V&L transformers that are pre-

Model	Datasets	# Images, Captions (Millions)	Architecture	Attention
VinVL [90]	VQA, GQA, VG-QA, COCO, Flickr30k, CC, SBU	1.89, 4.87	single-stream	merged
UNITER [12]	COCO, VG, CC, SBU	4.20, 9.58	single-stream	merged
ViLLA [23]	COCO, VG, CC, SBU	4.20, 9.58	single-stream	merged
VisualBERT [47]	COCO, NLVR2	0.30, 0.52	single-stream	merged
ViLT [40]	COCO, VG, SBU, CC	4.10, 9.85	single-stream	merged
LXMERT [76]	COCO, VG	0.18, 9.18	dual-stream	modality-specific, co-attn, merged
ViLBERT [51]	CC	3.30, 3.30	dual-stream	modality-specific, co-attn, merged
UniT [35]	COCO detect., VG detect., VQAv2, SNLI-VE QNLI, MNLI-mm, QQP, SST-2	0.69, 1.91	dual-stream	modality-specific, merged
FLAVA <i>ITM</i> [68]	COCO, SBU, LN, CC, VG, WIT, CC 12M, RC, YFCC100M	70.00, 70.00	dual-stream	modality-specific, merged
FLAVA <i>Contrastive</i> [68]	COCO, SBU, LN, CC, VG, WIT, CC 12M, RC, YFCC100M	70.00, 70.00	dual-stream	modality-specific
CLIP [56]	—	400.00, 400.00	dual-stream	modality-specific
VSE++ and VSRN <i>coco</i>	COCO	0.11, 0.57	dual-stream	—
VSE++ and VSRN <i>Flickr30k</i>	Flickr30k	0.03, 0.16	dual-stream	—

Table 2. A high-level overview of the differences between the models we evaluate by the pretraining datasets, architecture, and attention mechanisms between the modalities. We omit datasets that were only used to train backbones. We exclude the language embedding from this table as every model uses a pretrained BERT tokenizer, except CLIP, VSE++, and VSRN. The pretraining datasets include COCO [48], Visual Genome (VG) [43], Conceptual Captions (CC) [63], SBU Captions [52], Flickr30k [88], VQA 2.0 [27], VCR [89], NLVR2 [74], SNLI-VE [87], QNLI [57], MNLI-mm [84], QQP [36], Localized Narratives (LN) [55], Wikipedia Image Text (WIT) [73], Conceptual Captions 12M (CC 12M) [10], Red Caps (RC) [15], YFCC100M [77], and SST-2 [72]. CLIP uses their own dataset for pretraining.

trained with an image-text matching classification head or that produce a similarity score between the two modalities<sup>1</sup>.

## 4.2. Multimodal RNNs

To determine whether low performance on Winoground is unique to transformer-based models, we include results for two sequence-based models, which are VSRN [45] and VSE++ [21]. Both VSE++ and VSRN have a loss function that prioritizes minimizing the hardest negative’s score. The hardest negative is the highest-scoring image-caption pair that is not correct. Intuitively, this type of loss function could enable models to get higher scores on Winoground in particular and may be useful in future work. Although we show later in the paper that VSRN and VSE++ do not do well, perhaps due to issues besides the loss function. Both models use a GRU [14] to get language embeddings and a separate pipeline to get image embeddings. Scores for image-caption pairs are found by taking an inner-product of the embeddings. VSE’s image encoder is a linear projection of the embedding from a backbone (either ResNet152 [29] or VGG19 [66]). In VSRN, a ResNet101-based Faster R-CNN with graph convolutions on top is used to get a sequence of features which are fed into a GRU. The GRU’s last hidden state is then used as the image embedding.

## 4.3. Human Performance

We employed crowd workers on the Amazon Mechanical Turk platform to establish a more conservative human baseline than the expert annotator upper bound of a perfect score. Like the models, annotators are shown one image and one caption at a time. Annotators are asked the binary choice question “Does the caption match the image?”. All 1600 combinations of images and captions are labeled by at

<sup>1</sup>UniT is the only model we selected that was not pretrained on image-text matching. To get image-text alignment scores, we finetuned UniT on image-text matching loss using MS-COCO [48]

least ten annotators. We compute the human image-caption score as the ratio of annotators who said the image/caption pair match over the total number of annotators for the pair. More details about the human labelling interface, onboarding criteria, and quality control are provided in Appendix E.

## 5. Results

### 5.1. Compared to humans

As shown in Tab. 3, the models struggle across the board on Winoground, often performing close to or below random chance. Comparatively, as expected, the human performance is high across the full range of linguistic and visual phenomena. For the **text score**, we observe  $\sim 50\%$  absolute difference between humans and the best performing models—UNITER, VILLA VinVL, ViLT, FLAVA, and CLIP—with the remaining models below chance.

The human performance is only slightly lower for the **image score**, whereas all models perform much worse. Even the highest performing model, FLAVA<sub>ITM</sub>, has a  $\sim 70\%$  performance gap compared to humans. This gap is not unique to our dataset: in prior work [21] [56], models also tend to perform significantly better on caption retrieval compared to image retrieval. More investigation is required to pinpoint the reasons: perhaps textual encoders are stronger, or the text modality has different biases.

Lastly, we consider the **group score**. For humans, it is not appreciably lower than their text and image scores. All of the models are below random chance here as well. We report confidence intervals for these results in Appendix A.

### 5.2. Results by Tags

For the swap-dependent linguistic tags, human performance is highest on **object**, followed by the **relation** and then **both**. For the swap-independent linguistic tags, humans do better on examples with two main predicates,

Model	Text	Image	Group
MTurk Human	<b>89.50</b>	<b>88.50</b>	<b>85.50</b>
Random Chance	25.00	25.00	16.67
VinVL	<b>37.75</b>	17.75	14.50
UNITER <sub>large</sub>	<b>38.00</b>	14.00	10.50
UNITER <sub>base</sub>	<b>32.25</b>	13.25	10.00
ViLLA <sub>large</sub>	<b>37.00</b>	13.25	11.00
ViLLA <sub>base</sub>	<b>30.00</b>	12.00	8.00
VisualBERT <sub>base</sub>	15.50	2.50	1.50
ViLT (ViT-B/32)	<b>34.75</b>	14.00	9.25
LXMERT	19.25	7.00	4.00
ViLBERT <sub>base</sub>	23.75	7.25	4.75
UniT <sub>ITM finetuned</sub>	19.50	6.25	4.00
FLAVA <sub>ITM</sub>	<b>32.25</b>	20.50	14.25
FLAVA <sub>Contrastive</sub>	<b>25.25</b>	13.50	9.00
CLIP (ViT-B/32)	<b>30.75</b>	10.50	8.00
VSE++ <sub>COCO</sub> (ResNet)	22.75	8.00	4.00
VSE++ <sub>COCO</sub> (VGG)	18.75	5.50	3.50
VSE++ <sub>Flickr30k</sub> (ResNet)	20.00	5.00	2.75
VSE++ <sub>Flickr30k</sub> (VGG)	19.75	6.25	4.50
VSRN <sub>COCO</sub>	17.50	7.00	3.75
VSRN <sub>Flickr30k</sub>	20.00	5.00	3.50

Table 3. Results on the Winoground dataset across the text, image and group score metrics. Results above random chance in **bold**.

which tend to contain longer and more complicated sentences. The models perform poorly on every category, but they largely show the opposite pattern. They perform better on examples with simpler and shorter sentences which more often have swaps at the morpheme level (see Tab. 4). One exception to the low model performance is that CLIP performs comparably to the humans on the **both** tag text score—the 26 examples with the **both** tag have some of the shortest and least compositional captions in our dataset (e.g. “presenting the watch” vs “watching the present”).

We also evaluate performance for the visual reasoning tags as shown in Tab. 5. Models and humans are particularly good at the **symbolic** examples, but the models are poor comparatively. On the **pragmatics** tag, humans have the lowest performance. Ten crowdworkers probably didn’t capture slight pragmatics preferences that our expert linguist annotators agreed on. One example that the crowdworkers failed is Fig. 3(a): “the kid [with the magnifying glass] looks at them []”. All ten annotators said that “the kid with the magnifying glass looks at them” was acceptable for both images, but captured the correct preference for the second caption. This reveals a limitation in how the task was presented to humans: our hypothesis is that if we gave humans both images and both captions at the same time, or if significantly more human annotators gave their

judgements, then the human scores would be substantially higher. Finally, models do worst on the **series** tag where most get a 0% group score, which indicates that they are always choosing one image over the other regardless of the caption (or vice versa).

## 6. Discussion

Despite the fact that every model struggled on Winoground compared to humans, we hope to gain further insights by analyzing which aspects of these models could contribute to their performance differences.

### 6.1. Capabilities of Encoders

**Richer features.** UNITER, VILLA, VinVL, ViLT, FLAVA, and CLIP are the only models that get above random chance performance in Tab. 3, and only for the text score. We hypothesize that these models perform better than others due to their richer features (unimodal features for CLIP and FLAVA<sub>Contrastive</sub>, multimodal features for the others). A potential explanation could be the large-scale pretraining used by CLIP and FLAVA, the large training dataset used to train the object detector for VinVL, or the ViT approach for image features used by ViLT, FLAVA, and CLIP that encodes every portion of the image.

**Common failure modes.** We highlight again that most of the models fail with 0% group score on the *same image series* tag. One explanation is that the models’ visual encoders might be too weak to correctly discriminate between substantially similar images. This could cause the models to fall back on their unimodal priors, picking one caption or image over the other in the majority of the four potential caption-image pairings.

**Heat maps.** We show a heatmap in Fig. 4 of the word-region alignment between ViLT’s vision and language features as a visualization for a model with some of the better performance on our dataset. ViLLA and UNITER are also trained with word-region alignment and we provide their heatmaps in Appendix D.

**Complicated captions.** The above-chance models do worse on examples with longer captions, possibly due to weak language encoding abilities. As shown in Tab. 6, caption length and lower model performance significantly correlate for the best models, even though the correlation is reversed for humans. The examples with the shortest captions are also the least compositional; they are primarily the examples where the parts of speech change between swapped words, or where there is a morpheme-level swap. Finally, we show in Tab. 6 correlations between caption perplexity<sup>2</sup> and model scores. We found that there is typically a weak correlation between models assigning an image-caption pair a higher score and a caption having low perplexity.

<sup>2</sup>We used the standard size GPT2 checkpoint from Hugging Face transformers to get perplexity [86].

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

Table 4. The results by linguistic tag. Results above chance are in **bold**.

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

Table 5. The results by visual tag. Results above chance are in **bold**.

## 6.2. By Architecture & Type of Attention

As shown in Tabs. 3 to 5, both single-stream and dual-stream models perform significantly worse than humans on the text, image and group scores. We find at least one single-stream model and at least one dual-stream model are above chance for most of our experiments, suggesting there is not a distinct performance difference by architecture. Although, six single-stream model checkpoints do above chance overall, compared to only the very large dual-stream models (CLIP and FLAVA). CLIP and FLAVA were trained on an order of magnitude more data than the other models. Across all types of attention, models struggled compared to humans. But neither of the two models using co-attention, in conjunction with single-modality and/or merged attention, performed above chance.

## 6.3. By Multimodal Pretraining Dataset Size

We find highly significant correlations between the size of the multimodal pretraining dataset and the scores, if we remove CLIP and FLAVA as outliers. Tab. 7 shows these correlations, and Appendix B has graphs showing each model's score versus the pretraining data size. The unimodal training data (for image backbones or pre-initialized text encoders) is not included in these calculations.

## 7. Conclusion

We introduced a novel task and dataset, Winoground, aimed at measuring visio-linguistic compositional reasoning in state of the art vision and language models. We demonstrate that models fall short, in most cases performing no better than chance. Our findings highlight that there

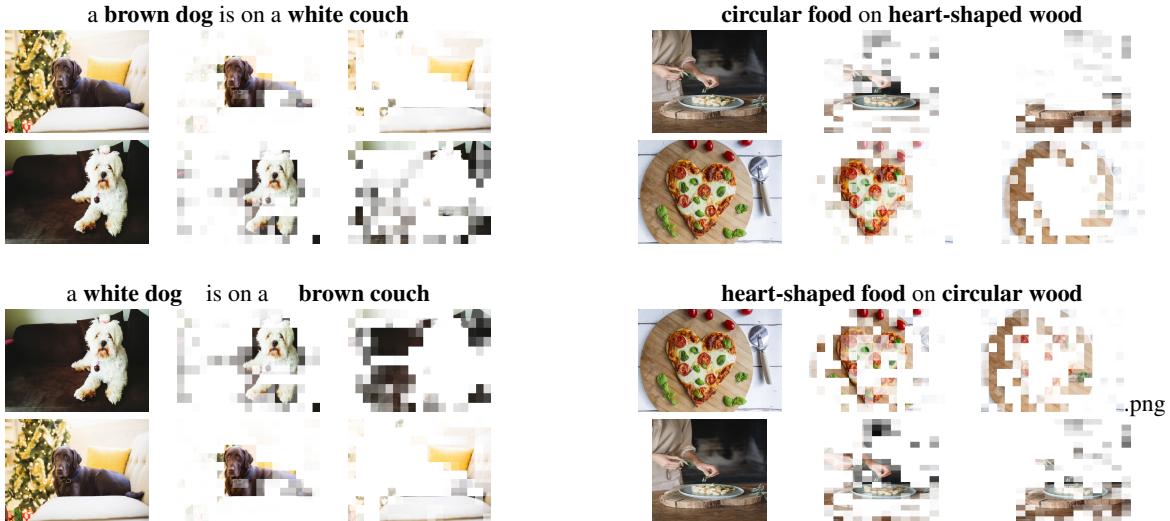


Figure 4. Word-region alignment scores between the image and text features for ViLT [40] on examples from Winoground. In this case study, ViLT appears to disregard the information from adjectives. E.g., the heatmaps highlight the brown dog just as strongly regardless of whether the text was “brown dog” or “white dog”.

Model	Perplexity		Caption Length	
	Corr.	p-value	Corr.	p-value
MTurk Human	0.05	0.07	<b>0.20</b>	<b>0.00</b>
VinVL	<b>-0.05</b>	<b>0.04</b>	<b>-0.20</b>	<b>0.00</b>
UNITER <sub>large</sub>	-0.01	0.57	<b>-0.16</b>	<b>0.00</b>
UNITER <sub>base</sub>	-0.03	0.22	<b>-0.14</b>	<b>0.00</b>
ViLLa <sub>large</sub>	-0.02	0.39	<b>-0.12</b>	<b>0.01</b>
ViLLa <sub>base</sub>	-0.04	0.13	<b>-0.11</b>	<b>0.03</b>
VisualBERT <sub>base</sub>	-0.04	0.15	-0.06	0.22
ViLT (ViT-B/32)	-0.04	0.16	<b>-0.16</b>	<b>0.00</b>
LXMERT	-0.04	0.12	<b>-0.11</b>	<b>0.02</b>
ViLBERT <sub>base</sub>	-0.04	0.11	<b>-0.14</b>	<b>0.00</b>
UniIT <sub>ITM finetuned</sub>	-0.01	0.73	-0.02	0.73
FLAVA <sub>ITM</sub>	-0.03	0.22	<b>-0.23</b>	<b>0.00</b>
FLAVA <sub>Contrastive</sub>	<b>-0.06</b>	<b>0.01</b>	<b>-0.19</b>	<b>0.00</b>
CLIP (ViT-B/32)	-0.04	0.09	<b>-0.22</b>	<b>0.00</b>
VSE++ <sub>COCO</sub> (ResNet)	<b>-0.05</b>	<b>0.04</b>	0.01	0.90
VSE++ <sub>COCO</sub> (VGG)	-0.04	0.08	0.03	0.56
VSE++ <sub>Flickr30k</sub> (ResNet)	-0.02	0.43	0.02	0.67
VSE++ <sub>Flickr30k</sub> (VGG)	0.01	0.74	<b>-0.10</b>	<b>0.04</b>
VSRN <sub>COCO</sub>	<b>-0.07</b>	<b>0.01</b>	-0.05	0.36
VSRN <sub>Flickr30k</sub>	-0.02	0.32	-0.05	0.29

Table 6. (left) The correlation between model image-caption scores and the caption perplexity from GPT2. (right) The correlation between the model group scores and the caption length.

is more work to be done. Particularly, the field could investigate possible strengths of single-stream models, the compilation of more pretraining data, improving image-encoding capabilities, and pretraining objectives that emphasize similar but wrong images. We hope that our task and dataset will help guide research in this important direction.

Pretraining Modality	Score	Corr.	p-value
Image	Text	<b>0.84</b>	<b>0.00</b>
	Image	<b>0.76</b>	<b>0.00</b>
	Group	<b>0.75</b>	<b>0.00</b>
Caption	Text	<b>0.77</b>	<b>0.00</b>
	Image	<b>0.75</b>	<b>0.00</b>
	Group	<b>0.71</b>	<b>0.00</b>

Table 7. Correlations between the number of pretraining images and captions and the model text, image, and group scores. CLIP and FLAVA are excluded as outliers.

**Broader Impact & Limitations.** Winoground is English-only and translation to other languages may be non-trivial [50]. Expert curation is time-consuming and our dataset is limited in size. Multimodal datasets containing images of people require thoughtful consideration of how people are represented (see [5] for a detailed analysis of the stereotypes present in many multimodal datasets). We used gender underspecified human denoting terms (e.g., person, child) to avoid issues with inferring gender identity from images [61]. Our annotators disproportionately come from the USA and the same could be true for our crowdworkers.

**Getty Acknowledgement.** Images in the paper are a compilation of assets, including ©Getty Images/Natasha Breen, Maki Nakamura, Jessica Peterson, Kundanlall Sharma, lacaosa, Alberto Bogo, Vu Le, Toson Rueangsuk-sut, Nisian Hughes, Tanja Walter, Douglas Sacha, PBNJ Productions, Glow Images, 10’000 Hours, zoranm, Marlene Ford, Westend61.

## References

- [1] Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *ACL*, 2020. [2](#)
- [2] Daniel Altshuler, Terence Parsons, and Roger Schwarzschild. *A Course in Semantics*. MIT Press, 2019. [3](#)
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. [2](#)
- [4] David Bender. Establishing a human baseline for the winograd schema challenge. In *Modern Artificial Intelligence and Cognitive Science*, 2015. [2](#)
- [5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kambemwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. In *arXiv preprint arXiv:2110.01963*, 2021. [8](#)
- [6] Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of GQA. In *NAACL: Human Language Technologies*, 2021. [2](#)
- [7] Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. Covr: A test-bed for visually grounded compositional generalization with real images. In *EMNLP*, 2021. [2](#)
- [8] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *ECCV*, 2020. [1](#)
- [9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. [4](#)
- [10] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. [5](#)
- [11] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *arXiv preprint arXiv:1704.07121*, 2017. [2](#)
- [12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. [2, 4, 5](#)
- [13] Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. Context models and out-of-context objects. In *Pattern Recognition Letters*, 2012. [2](#)
- [14] Junyoung Chung, Caglar Gulcehr, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS*, 2014. [5](#)
- [15] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people. In *NeurIPS Datasets and Benchmarks*, 2021. [5](#)
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL: Human Language Technologies*, 2019. [4](#)
- [17] Nan Ding, Sebastian Goodman, Fei Sha, and Radu Soricut. Understanding image and text simultaneously: a dual vision-language machine comprehension task. In *arXiv preprint arXiv:1612.07833*, 2016. [2](#)
- [18] 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. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. [4](#)
- [19] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuhang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. In *arXiv preprint arXiv:2111.02387*, 2021. [1](#)
- [20] Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. Back to square one: Artifact detection, training and commonsense disentanglement in the winograd schema. In *EMNLP*, 2021. [4](#)
- [21] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. [2, 4, 5](#)
- [22] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *EMNLP*, 2021. [2](#)
- [23] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. [2, 4, 5](#)
- [24] Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. SyntaxGym: An online platform for targeted evaluation of language models. In *ACL: System Demonstrations*, 2020. [2](#)
- [25] <https://www.gettyimages.com/>. [3](#)
- [26] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *ICCV*, 2020. [2](#)
- [27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. [2, 5](#)
- [28] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *NAACL: Human Language Technologies*, 2018. [2](#)
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. [5](#)
- [30] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *ACL-IJCNLP*, 2021. [2](#)
- [31] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. [2](#)
- [32] Homa HosseiniMardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detec-

- tion of cyberbullying incidents on the instagram social network. In *arXiv preprint arXiv:1503.03909*, 2015. 2
- [33] Hexiang Hu, Ishan Misra, and Laurens van der Maaten. Evaluating text-to-image matching using binary image selection (bison). In *ICCV*, 2019. 2
- [34] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In *ACL*, 2020. 2
- [35] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *arXiv preprint arXiv:2102.10772*, 2021. 2, 4, 5
- [36] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. First quora dataset release: Question pairs, 2017. 5
- [37] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2
- [38] Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. Verb argument structure alternations in word and sentence embeddings. In *SCiL*, 2019. 2
- [39] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *arXiv preprint arXiv:2005.04790*, 2020. 2
- [40] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2, 4, 5, 8
- [41] Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, Tristan Thrush, and Scott A Hale. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *arXiv preprint arXiv:2108.05921*, 2021. 2
- [42] Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. A review of winograd schema challenge datasets and approaches. In *arXiv preprint arXiv:2004.13831*, 2020. 2
- [43] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv preprint arXiv:1602.07332*, 2016. 4, 5
- [44] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Conference on the Principles of Knowledge Representation and Reasoning*, 2012. 1, 2
- [45] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019. 2, 4, 5
- [46] Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. In *arXiv preprint arXiv:2012.08673*, 2020. 1
- [47] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *arXiv preprint arXiv:1908.03557*, 2019. 1, 2, 4, 5
- [48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5
- [49] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. In *TACL*, 2015. 2
- [50] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *EMNLP*, 2021. 2, 8
- [51] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 2019. 1, 2, 4, 5
- [52] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 5
- [53] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. In *ACL*, 2021. 2
- [54] Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau, and Adina Williams. Sometimes we want ungrammatical translations. In *Findings of the Association for Computational Linguistics: EMNLP*, 2021. 2
- [55] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 5
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 4, 5
- [57] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *arXiv preprint arXiv:1606.05250*, 2016. 5
- [58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 4
- [59] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *arXiv preprint arXiv:1804.09301*, 2018. 2
- [60] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, 2020. 2
- [61] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. In *ACM: Human Computer Interaction*, 2019. 8
- [62] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *arXiv preprint arXiv:1508.07909*, 2015. 4
- [63] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, im-

- age alt-text dataset for automatic image captioning. In *ACL*, 2018. 2, 5
- [64] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. "foil it! find one mismatch between image and language caption". In *ACL*, 2017. 2
- [65] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020. 2
- [66] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for largescale image recognition. In *CVPR*, 2015. 5
- [67] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining. In *arXiv preprint arXiv:2004.08744*, 2020. 1
- [68] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 2, 4, 5
- [69] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 2
- [70] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *EMNLP*, 2021. 1, 2
- [71] Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. UnNatural Language Inference. In *ACL-IJCNLP*, 2021. 2
- [72] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013. 5
- [73] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *arXiv preprint arXiv:2103.01913*, 2021. 5
- [74] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017. 2, 5
- [75] Shardul Suryawanshi and Bharathi Raja Chakravarthi. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, 2021. 2
- [76] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2020. 2, 4, 5
- [77] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. In *Communications of the ACM*, 2016. 5
- [78] Tristan Thrush, Ethan Wilcox, and Roger Levy. Investigating novel verb learning in BERT: Selectional preference classes and alternation-based syntactic generalization. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2020. 2
- [79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4
- [80] Ramakrishna Vedantam, Arthur Szlam, Maximillian Nickel, Ari Morcos, and Brenden M Lake. Curi: A benchmark for productive concept learning under uncertainty. In *ICML*, 2021. 2
- [81] Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *EMNLP-IJCNLP*, 2019. 2
- [82] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. In *TACL*, 2020. 2
- [83] Adina Williams, Andrew Drozdov, and Samuel R. Bowman. Do latent tree learning models identify meaningful structure in sentences? In *TACL*, 2018. 2
- [84] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *arXiv preprint arXiv:1704.05426*, 2017. 5
- [85] Terry Winograd. Understanding natural language. In *Cognitive psychology*, 1972. 2
- [86] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, 2020. 6
- [87] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. In *arXiv preprint arXiv:1811.10582*, 2018. 5
- [88] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, 2014. 5
- [89] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 5
- [90] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 2, 4, 5
- [91] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *arXiv preprint arXiv:1804.06876*, 2018. 2
- [92] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, 2016. 2

- [93] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. [2](#)

## A. Confidence Intervals

We provide confidence intervals for the overall model results on Winoground. We divided the dataset into 4 groups of equal size to get 4 scores for each model and score-type, and used Student's t-distribution to compute the confidence intervals.

<i>Model</i>	<i>Text</i>	<i>Image</i>	<i>Group</i>
MTurk Human	<b>89.50</b> [80.83,98.17]	<b>88.50</b> [79.00,98.00]	<b>85.50</b> [73.84,97.16]
VinVL	<b>37.75</b> [28.71,46.79]	17.75 [11.21,24.29]	14.50 [6.65,22.35]
UNITER <sub>large</sub>	<b>38.00</b> [33.32,42.68]	14.00 [6.77,21.23]	10.50 [8.45,12.55]
UNITER <sub>base</sub>	<b>32.25</b> [25.84,38.66]	13.25 [7.68,18.82]	10.00 [7.75,12.25]
ViLLA <sub>large</sub>	<b>37.00</b> [31.05,42.95]	13.25 [7.83,18.67]	11.00 [7.10,14.90]
ViLLA <sub>base</sub>	<b>30.00</b> [25.32,34.68]	12.00 [8.33,15.67]	8.00 [5.75,10.25]
VisualBERT <sub>base</sub>	15.50 [9.34,21.66]	2.50 [0.00,6.29]	1.50 [0.00,3.55]
ViLT (ViT-B/32)	<b>34.75</b> [29.03,40.47]	14.00 [8.49,19.51]	9.25 [6.53,11.97]
LXMERT	19.25 [16.53,21.97]	7.00 [3.10,10.90]	4.00 [2.70,5.30]
ViLBERT <sub>base</sub>	23.75 [18.03,29.47]	7.25 [3.97,10.53]	4.75 [1.47,8.03]
UniT <sub>ITMFinetuned</sub>	19.50 [14.73,24.27]	6.25 [0.53,11.97]	4.00 [2.70,5.30]
FLAVA <sub>ITM</sub>	<b>32.25</b> [20.04,44.46]	20.50 [14.34,26.66]	14.25 [8.53,19.97]
FLAVA <sub>Contrastive</sub>	<b>25.25</b> [19.99,30.51]	13.50 [8.55,18.45]	9.00 [5.10,12.90]
CLIP (ViT-B/32)	<b>30.75</b> [25.03,36.47]	10.50 [6.29,14.71]	8.00 [4.56,11.44]
VSE++ <sub>COCO</sub> (ResNet)	22.75 [19.22,26.28]	8.00 [6.70,9.30]	4.00 [1.40,6.60]
VSE++ <sub>COCO</sub> (VGG)	18.75 [17.23,20.27]	5.50 [3.45,7.55]	3.50 [2.58,4.42]
VSE++ <sub>Flickr30k</sub> (ResNet)	20.00 [12.77,27.23]	5.00 [0.89,9.11]	2.75 [0.75,4.75]
VSE++ <sub>Flickr30k</sub> (VGG)	19.75 [14.49,25.01]	6.25 [2.27,10.23]	4.50 [2.91,6.09]
VSRN <sub>COCO</sub>	17.50 [9.54,25.46]	7.00 [1.19,12.81]	3.75 [0.00,8.50]
VSRN <sub>Flickr30k</sub>	20.00 [13.25,26.75]	5.00 [2.09,7.91]	3.50 [2.58,4.42]

Table 1. 95% confidence intervals for the aggregate results on Winoground. Results above chance are shown in **bold**.

## B. Impact of Pretraining Data Size and Model Performance

Correlations between pretraining data size and model performance are highly significant in every case and the numbers are shown in the main paper. We show plots in the figures below. Most of the single-stream models perform slightly above chance on the text score. CLIP and FLAVA are the only dual-stream models which perform above chance, and they have drastically more training data than all other models.

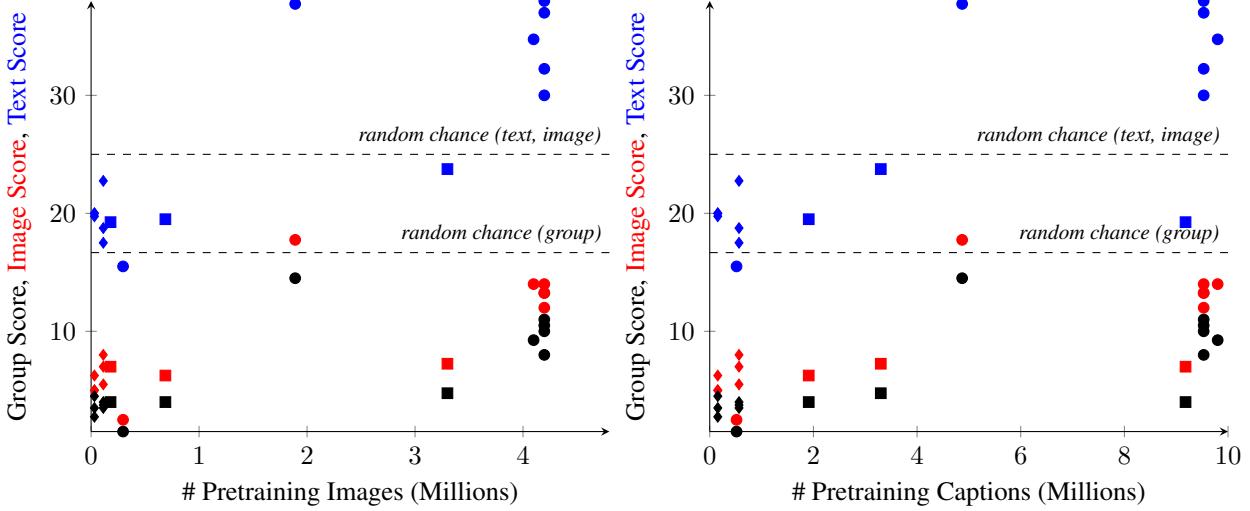


Figure 1. Graphs of the model performance on Winoground for each model by the number of pretraining images (left) and pretraining captions (right).  $\diamond$  = dual-stream RNNs,  $\square$  = dual-stream transformers,  $\circ$  = single-stream transformers. CLIP and FLAVA are removed as outliers. Backbone pretraining data is not included.

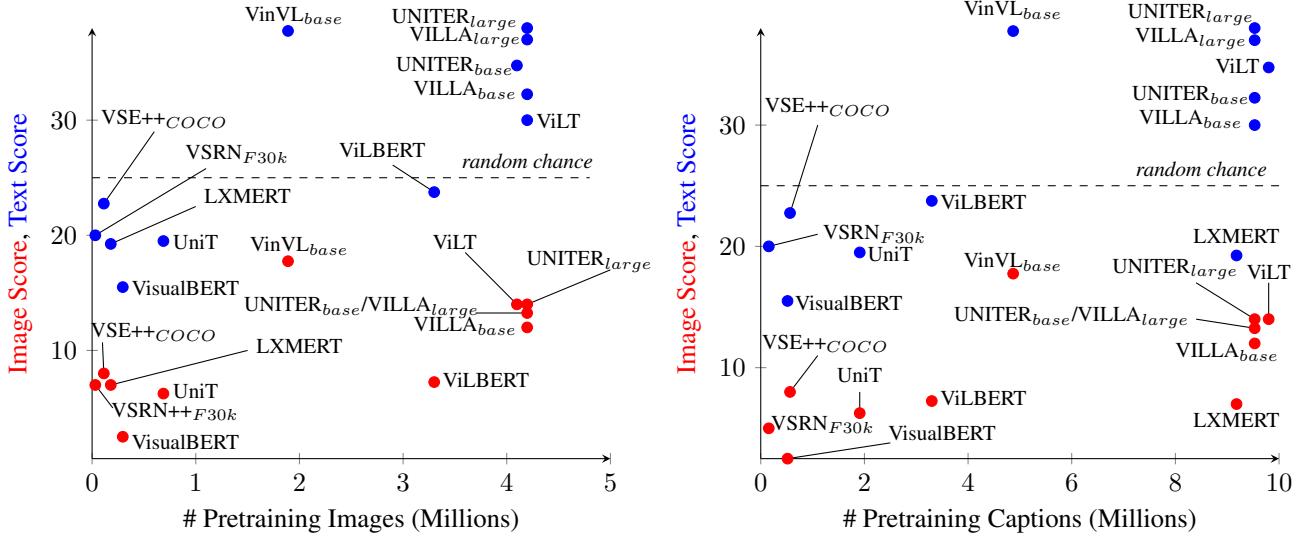


Figure 2. Graphs of the model performance on Winoground for each model by the number of pretraining images (left) and pretraining captions (right). This is a finer-grained version of Tab. 1, with model names instead of grouping by architecture; we again exclude CLIP and FLAVA as their pretraining dataset sizes are outliers. We only show the best VSE++ and VSRN configurations and do not show group scores due to clutter issues.

## C. Linguistic Tag Breakdown

This section reports every different swap-dependent linguistic tag that our annotators gave examples. Many of these fine-grained linguistic tags are used for multiple examples, although some tags are only used once in the dataset.

Tag	Fine-Grained Tag	Example
Object	Noun Phrase, Determiner-Numerical	[a person] carrying [more than one flotation device]
	Noun Phrase	[a person] holding up [books]
	Determiner-Numerical, Noun Phrase	[a lightbulb] surrounding [some plants]
	Noun Phrase, Determiner-Possessive	[a deer's nose] is resting on [a child's hand]
	Noun Phrase, Adjective-Color	aerial view of a green tree in [the brown freshly turned soil] next to [a green field]
	Pronoun, Noun Phrase	[the person] wears a hat but [it] doesn't
	Determiner-Numerical Phrase	[one] is in a boat and [almost everyone] is swimming
	Pronoun, Verb-Intransitive	[it] ran away while [they] pursued
	Noun	more [bicycles] than [cars]
Relation	Adjective-Age	[an older] person blocking [a younger] person
	Scope, Preposition	racing [over] it []
	Verb-Intransitive, Verb-Transitive Phrase	a kid [threw a basketball] then [jumped]
	Verb-Intransitive, Adjective-Manner	the younger person is [making noise] while the other is [silent]
	Negation, Noun Phrase, Preposition Phrase	a person [with long braids] is exercising in front of a person [without braids]
	Scope, Preposition, Verb-Intransitive	[out]1[swam]2 the person in the red swimcap []2[]1
	Noun Phrase, Adjective-Animate	the one on the left is [sad] and the other is [happy]
	Adjective-Size	the [taller] person hugs the [shorter] person
	Determiner-Possessive	the [person's] leg is on the [dog's] torso
	Adjective-Texture	[smooth] shoes are on a [soft] floor
	Adjective-Color	painting the [white] wall [red]
	Scope	[getting] a horse [] wet
	Preposition Phrase	flat [at the bottom] and pointy [on top]
	Relative Clause, Scope	the person [who is wearing a crown] is kissing a frog []
	Adjective-Height	a [taller] person wearing blue standing next to a [shorter] person
	Verb-Intransitive Phrase, Preposition	the gesture of the person [sitting down] is supporting the understanding of the person [standing up]
	Verb-Intransitive, Determiner-Numerical	some people are [standing] but more are [sitting]
	Determiner-Numerical	[one]1 person[]2 wearing [two]1 scarf[]2
	Adjective-Weight	the larger ball is [lighter] and the smaller one is [heavier]
	Verb-Intransitive, Noun	the dog is [standing] and the person is [swimming]
	Verb-Intransitive Phrase, Adverb-Animate	the person on the left is [crying sadly] while the one on the right is [smiling happily]
	Scope, Relative Clause	a fencer [who is wearing black pants] having a point scored against them by another fencer [] using a wheelchair
	Adjective-Speed	the train is [still] while the person is [moving fast]
	Adverb-Temporal	a person is drinking [now] and eating [later]
	Adverb-Spatial	the car is sitting [upside down] while the person is standing [rightside up]
	Adjective-Shape	the [round] table has a [square] base
Both	Noun, Adjective-Color	Young person playing baseball with a [blue] bat and [green] ball
	Verb-Transitive	the person with the ponytail [buys] stuff and other [packs] it
	Scope, Verb-Transitive	[] gears for [moving] something
	Scope, Preposition Phrase	[] child in [front facing] row of yellow rubber ducks
	Adjective-Temperature	a [hot] drink on a [cold] day
	Adjective-Temporal	the [first] vowel is E and the [last] consonant is N
	Scope, Conjunction	a person spraying water on [someone else]1 [and]2 a person on a bike []2[]1
	Scope, Conjunction Phrase	A child [] riding a bike [and an adult]
	Preposition Phrase, Scope	someone [with an apple] is hurt by a tree []
	Adjective-Manner Phrase	two people wearing clothes of [different] colors are on [the same] side of the tennis net
	Verb-Intransitive	a person [stands] and a dog [sits]
	Adjective-Animate	[toy] cat with [real] baby
	Adverb-Spatial Phrase	the sailboat sails [close] but the beach is [far away]
	Scope, Adjective-Texture	A [] small animal with [curled] hair
	Adverb-Animate	someone talks on the phone [angrily] while another person sits [happily]
Altered POS	Adjective-Manner	[poor] [unfortunate] people
	Verb-Transitive Phrase	they [drank water] then they [worked out]
	Adjective-Color (3-way swap)	The [red]→[yellow] book is above the [yellow]→[blue] book and below the [blue]→[red] book
	Scope, Adjective-Manner	[] living things [drinking]
	Preposition	seat numbers increasing from [right] to [left]
	Verb-Intransitive Phrase	a cat is [stretching] and a person is [lying down]
	Sentence	[the coffee is poured] before [it is ground]
	Adjective-Speed Phrase, Verb-Intransitive	the person with green legs is running [quite slowly] and the red legged one runs [faster]
	Adjective-Spatial	A [left] hand pulls a glove onto a [right] hand
Noun, Adjective-Size	Negation, Scope	The [un]caged bird has an []opened cage door
	Verb-Transitive Phrase, Verb-Intransitive, Preposition Phrase	the dog [bite]1s []2 what someone would normally [wear]1 [as a hat]2
	Altered POS	[watch]ing the [present]
Both	Verb-Transitive, Noun	someone []1 on [the ground]2 [is]1 spraying water towards [a vehicle]2
	Scope, Altered POS, Verb-Intransitive, Verb-Transitive	[walking]1 someone []1 [cut]2 [lines]2 into green plants
	Noun, Adjective-Size	the [person]1 is too [big]2 for the [small]2 [door]1
	Noun, Verb-Intransitive	a [dog sitting] on a couch with a [person lying] on the floor
	Scope, Noun, Preposition	[]1 a person [near]1 [water]2 using a []2 lasso
	Noun, Preposition Phrase, Scope	a person wearing a [bear]1 mask []2 in blue on the left hand side of a person wearing a [panda]1 mask [with glasses]2 in pink
	Scope, Preposition Phrase, Adjective-Color	[darker]1 things []2 become [light]1 [in stripes]2
	Altered POS, Determiner-Numerical	[one] ear that some [donkey] is whispering a secret into

Table 2. Examples showcasing the full linguistic (swap-dependent) tag breakdown.

## D. Heatmaps for the Word-Region Alignment Models

We provide heatmaps for models that were trained with a word-region alignment objective: UNITER, ViLLA and ViLT. See the main text for ViLT heatmaps.

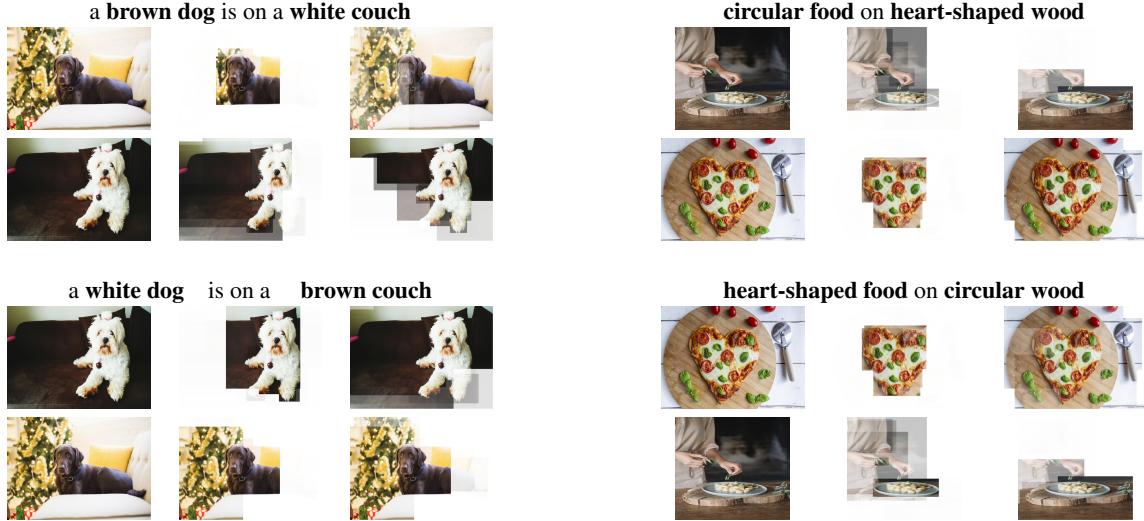


Figure 3. Word-region alignment scores between the image and text features for ViLLA<sub>base</sub> on examples from Winoground.

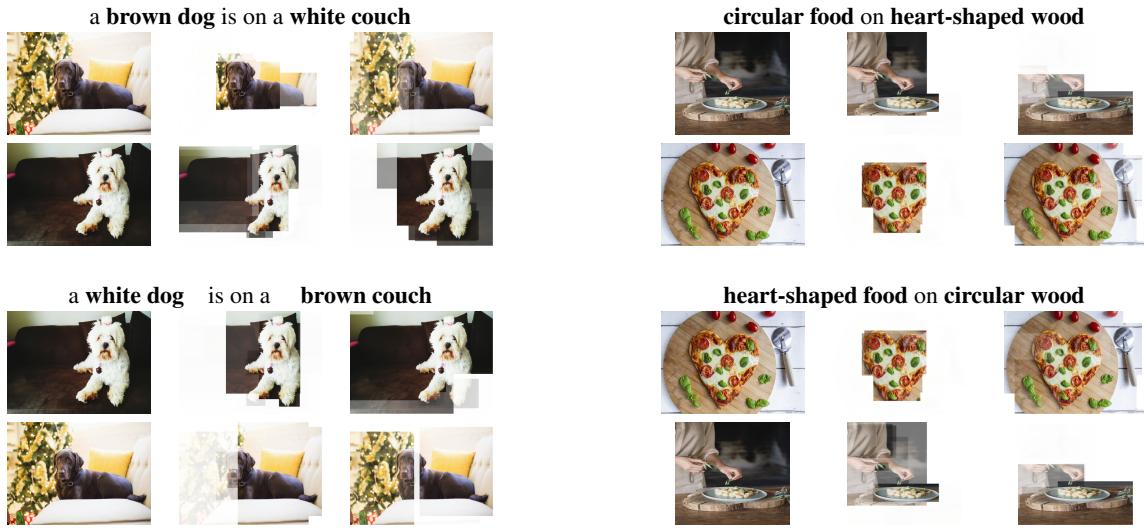


Figure 4. Word-region alignment scores between the image and text features for UNITER<sub>base</sub> on examples from Winoground.

## E. Mechanical Turk Interface

In order to participate, crowdworkers needed to satisfy several criteria: be an English speaker, have 98% previous HIT approval, have completed 1000 previous HITs, and pass the onboarding test. The onboarding test used the same interface as the actual task. It consisted of ten image-caption match questions, with images and captions that are independent from the actual Winoground dataset. If they made one mistake, a pop-up asked them if they were sure, and they would be allowed to select whether there was a match or not again. If they made any additional mistakes during onboarding, they were disqualified.

The screenshot shows the Amazon Mechanical Turk validation interface. At the top, it displays the Amazon Mechanical Turk logo, the worker's name 'Noah Turk', the number of HITs completed '16', the reward '\$0.35', and the time elapsed '1:47 of 60 Min'. There is also a 'Return' button. Below this, the main task area is titled 'Validate examples'. It contains instructions: 'Select whether the image matches the caption. Pay close attention to the word order.' An image of a person kite surfing is shown, with a watermark for 'gettyimages Ben Welsh' and the ID '523073400'. The caption provided is 'a kite is lifting up a person'. Below the image, there is a question 'DOES THE CAPTION MATCH THE IMAGE?' followed by two radio buttons: 'Yes' and 'No'. A blue 'Submit' button is at the bottom. At the very bottom of the page, there are links for 'Report this HIT' and 'Why Report'.

Figure 5. The Amazon Mechanical Turk validation interface.

## F. Ethical Considerations

A key consideration while designing Winoground centered on how the expert annotators would describe the people contained in the images. We avoided using gendered terms (*e.g.* using “person” in place of “woman” or “man”) in our captions and did not include any swaps between pairs of captions based on gender, race or ethnicity (*e.g.* “[*the man*] hands a water to [*the woman*]”). We recognize that, barring direct access to the people in the images, we would be merely making a guess at a person’s identity based on our own cultural norms and experiences.

In addition, we encouraged the expert annotators to find images that represent a variety of people across the dimensions of perceived race, gender, disability, *etc.*. We gathered the Getty Images metadata (title and short alt text-like description) and searched them for specific words as a rough proxy for gender representation. The relevant words are either words referring to women (*e.g.* girl, her), words referring to men (*e.g.* boy, him) or words that are gender-neutral (*e.g.* them, themselves). Using the Getty Images metadata corresponding to the 800 images in Winoground, 371 images have corresponding metadata that contained at least one word from the lists we created. Using this metadata for these 371 images, we estimate that 152 images only contain women, 123 images only contain men, 22 images only contain people without gender descriptors, and the remaining 74 images contain people described by multiple genders. This serves only as a rough estimate as much of the metadata contain words referring to people that are inherently non-gendered (*e.g.* scuba diver, friend, *etc.*) and because the relevant gendered words we found are themselves subject to the assumptions of those who wrote the titles and captions.