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

Tristan Thrush[¶][*], Ryan Jiang[‡], Max Bartolo[§],
Amanpreet Singh[¶], Adina Williams[†], Douwe Kiela[¶], Candace Ross[†][*]
[¶] Hugging Face; [†] Facebook AI Research; [‡] University of Waterloo; [§] University College London
tristan@huggingface.co, ccross@fb.com

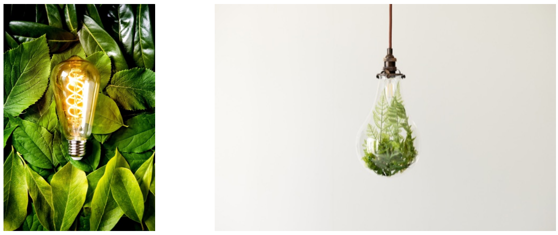
## Abstract

*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 [43, 47, 50], they remain poorly understood [6, 15, 42, 61]. 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 that transformers are often remarkably insensitive to word order [63]. 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.



(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 proves much more difficult for vision and language models. Every model that we tested (UNITER, ViLLA, VinVL, VisualBERT, ViLT, LXMERT, ViLBERT, UniT, 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). Image above is a compilation of assets, including ©Getty Images/Natasha Breen, Maki Nakamura.

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 Chal-

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

lenge [40], 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 [53,54,82]. 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 [9, 19, 31, 36, 43, 47, 50, 68, 81] and RNN-based models [17, 41]. 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 [10], exploitation of language and vision priors [8, 23], single word mismatches [58], hate speech detection [22, 28, 37, 83], memes [35, 67], ablation of one modality to probe the other [18], distracting models with visual similarity between images [29], distracting models with textual similarity between many suitable captions [13], collecting more diverse image-caption pairs beyond the predominately English and North American/Western European datasets [46], and probing for an understanding of verb-argument relationships [26] or specific model failure modes [59, 62]. Many of these stress tests rely only on synthetically generated images, often with minimal visual differences, but no correspondingly minimal textual changes [71]. Other datasets test models with a single caption [66] or a single image [5, 33]. There are also purely visual stress tests with naturalistic images: ImageNet-C/ImageNet-P [27] 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 [57], COCO [44], Visual7W [84] and VQA [2, 23]. 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. [63] 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-occurence 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 [20, 24, 30, 45, 49, 64, 74] or the complex interaction between syntactic and semantic categories [34, 69, 72, 73].

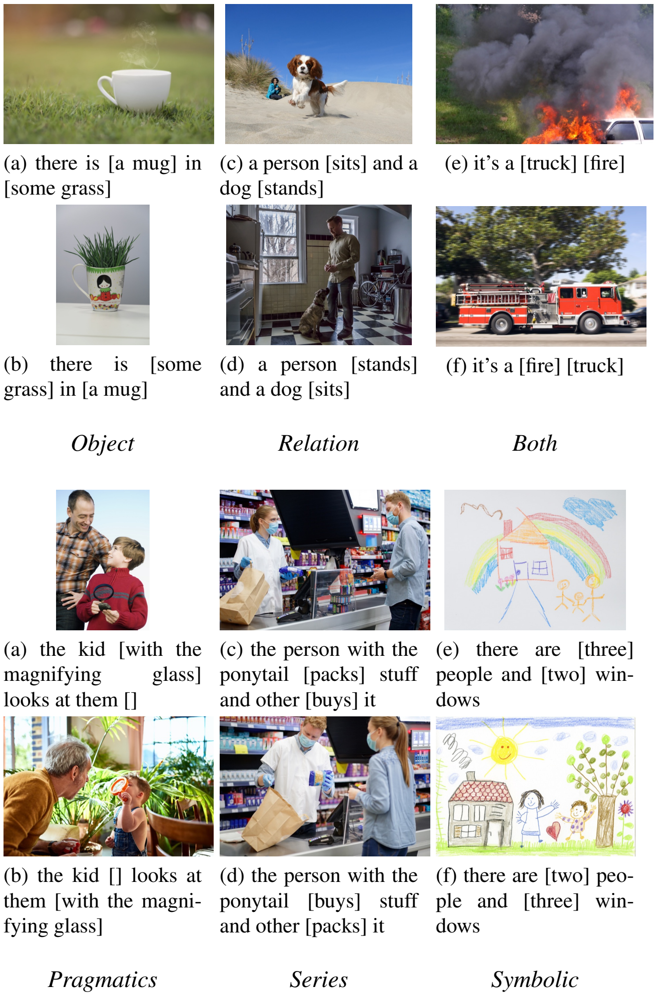
**Winograd schemas.** The Winograd Schema Challenge [40] was named after a coreference resolution problem presented by Terry Winograd [76]. The goal is to correctly resolve (an) ambiguous referent(s) in two English sentences. The sentences have a minor difference that changes how a human resolves the referent. Winograd schema examples are easily handled by humans, and commonsense reasoning is said to be required [3]. 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. [54] 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. [40]. Other approaches use ambiguous pronouns in sentences to probe for gender biases in models [53, 82]. See Kotcijan et al. [38] 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.



(a) there is [a mug] in [some grass]

(c) a person [sits] and a dog [stands]

(e) it's a [truck] [fire]

(b) there is [some grass] in [a mug]

(d) a person [stands] and a dog [sits]

(f) it's a [fire] [truck]

*Object*

*Relation*

*Both*

(a) the kid [with the magnifying glass] looks at them []

(c) the person with the ponytail [packs] stuff and other [buys] it

(e) there are [three] people and [two] windows

(b) the kid [] looks at them [with the magnifying glass]

(d) the person with the ponytail [buys] stuff and other [packs] it

(f) there are [two] people and [three] windows

*Pragmatics*

*Series*

*Symbolic*

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. Image above is a compilation of assets, including ©Getty Images/ Vu Le, Toson Rueangsuksut, Nisian Hughes, Tanja Walter, Douglas Sacha, PBNJ Productions, Glow Images, 10'000 Hours, zoranm, Marlene Ford, Westend61.

| Category | Tag | Count |
|---|---|---|
| $\text{Linguistic}_{\text{swap-dep.}}$ | Object | 141 |
| | Relation | 233 |
| | Both | 26 |
| $\text{Linguistic}_{\text{swap-indep.}}$ | 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.

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
- $C_0$ 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 [21], 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 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 [1]. 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:

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

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:

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

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. [16] 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:

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

## 4. Experimental Setup

We evaluate various configurations of the following multimodal transformers: CLIP [50], LXMERT [68], UniT [31], UNITER [9], VILLA [19], VinVL [81], ViLT [36], VisualBERT [43] and ViLBERT [47]. We also evaluate several configurations of two types of RNN-based models: VSE++ [17] and VSRN [41]. We detail differences between these models and provide a high-level overview in Tab. 2. We also establish a human baseline using crowdworkers, 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 [12], except CLIP, which uses a Byte-Pair Encoding tokenizer [56] trained from scratch. For the image embedding, five transformers (VisualBERT, ViLBERT, LXMERT, UNITER, ViLLA) [9,19,43,47,68] use region features extracted from the `fc6` layer of a Faster R-CNN [52] trained on Visual Genome [39]. VinVL trains its own feature extractor on a large combined dataset from public sources with a unified object vocabulary [81]. The CLIP and ViLT that we test both use Vision Transformer (ViT) [14]. In ViT, images are flattened into patches that are linearly projected and combined with a position encoding. UniT [31] alternatively uses a transformer network [70] on top of a convolutional network following Carion et al. [7].

**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 [9,19,36,43,81]. 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, UniT, LXMERT and ViLBERT [31, 47, 50, 68]. CLIP lacks cross-modal attention. ViLBERT has language-only transformer layers that are then fused by cross-modal transformer layers. LXMERT and

| *Model* | *Datasets* | *# Images, Captions (Millions)* | *Architecture* | *Attention* |
|---|---|---|---|---|
| VinVL [81] | VQA, GQA, VG-QA, COCO, Flickr30k, CC, SBU | 1.89, 4.87 | single-stream | merged |
| UNITER [9] | COCO, VG, CC, SBU | 4.20, 9.58 | single-stream | merged |
| ViLLA [19] | COCO, VG, CC, SBU | 4.20, 9.58 | single-stream | merged |
| VisualBERT [43] | COCO, NVLR2 | 0.30, 0.52 | single-stream | merged |
| ViLT [36] | COCO, VG, SBU, CC | 4.10, 9.85 | single-stream | merged |
| LXMERT [68] | COCO, VG | 0.18, 9.18 | dual-stream | modality-specific, co-attn, merged |
| ViLBERT [47] | CC | 3.30, 3.30 | dual-stream | modality-specific, co-attn, merged |
| UniT [31] | COCO detect., VG detect., VQAv2, SNLI-VE QNLI, MNLI-mm, QQP, SST-2 | 0.69, 1.91 | dual-stream | modality-specific, merged |
| CLIP [50] | – | 400.00, 400.00 | dual-stream | modality-specific |
| VSE++ and VSRN $_{COCO}$ | COCO | 0.11, 0.57 | dual-stream | – |
| VSE++ and VSRN $_{Flickr30k}$ | 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 [44], Visual Genome (VG) [39], Conceptual Captions (CC) [57], SBU Captions [48], Flickr30k [79], VQA 2.0 [23], VCR [80], NLVR2 [66], SNLI-VE [78], QNLI [51], MLNI-mm [75], QQP [32], and SST-2 [65]. CLIP uses their own dataset for pretraining.

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 specifically select V&L transformers that are pretrained with an image-text matching classification head or that produce a similarity score between the two modalities[1].

### 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 [41] and VSE++ [17]. 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 [11] 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 [25] or VGG19 [60]). 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 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 observed 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 ∼50% absolute difference between humans and the best performing models—UNITER, VILLA VinVL, ViLT and CLIP—with the remaining models at or below chance performance.

The human performance is only slightly lower for the **image score**, whereas all models perform much worse. Even the highest performing model, VinVL, has a ∼70% performance gap compared to humans. This gap is not unique to our dataset: in prior work [17] [50], 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

[1] 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 [44]

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

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, 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 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, multimodal features for the others). A potential explanation could be the large-scale pretraining used by CLIP, the large training dataset used to train the object detector for VinVL, or the ViT approach for image features used by ViLT and CLIP that encodes every portion of the image.

**Common failure modes.** We highlight again that nearly all 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 best 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,

| | *Object* | | | *Relation* | | | *Both* | | | *1 Main Pred* | | | *2 Main Preds* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Model* | Text | Image | Group | Text | Image | Group | Text | Image | Group | Text | Image | Group | Text | Image | Group |
| MTurk Human | **92.20** | **90.78** | **88.65** | **89.27** | **90.56** | **86.70** | **76.92** | **57.69** | **57.69** | **87.33** | **85.62** | **82.53** | **95.37** | **96.30** | **93.52** |
| VinVL | **36.88** | 17.73 | 14.18 | **37.77** | 17.60 | 14.16 | **42.31** | 19.23 | **19.23** | **39.38** | 21.23 | **17.47** | **33.33** | 8.33 | 6.48 |
| UNITER$_{large}$ | **39.01** | 12.77 | 9.93 | **36.05** | 14.16 | 9.87 | **50.00** | 19.23 | **19.23** | **40.07** | 16.44 | 13.36 | **32.41** | 7.41 | 2.78 |
| UNITER$_{base}$ | **34.04** | 11.35 | 9.22 | **30.04** | 14.16 | 10.30 | **42.31** | 15.38 | 11.54 | **35.27** | 14.73 | 11.99 | 24.07 | 9.26 | 4.63 |
| ViLLA$_{large}$ | **36.88** | 14.89 | 11.35 | **37.34** | 12.88 | 11.16 | **34.62** | 7.69 | 7.69 | **39.73** | 17.12 | 14.38 | **29.63** | 2.78 | 1.85 |
| ViLLA$_{base}$ | **33.33** | 15.60 | 9.93 | **27.04** | 9.01 | 6.01 | **38.46** | 19.23 | 15.38 | **33.22** | 14.04 | 10.27 | 21.30 | 6.48 | 1.85 |
| VisualBERT$_{base}$ | 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) | **31.91** | 15.60 | 9.22 | **36.91** | 11.59 | 8.15 | **30.77** | **26.92** | **19.23** | **35.27** | 17.12 | 11.64 | **33.33** | 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$_{base}$ | **29.08** | 10.64 | 7.09 | 19.31 | 3.00 | 1.72 | **34.62** | **26.92** | **19.23** | 23.97 | 8.90 | 5.82 | 23.15 | 2.78 | 1.85 |
| UniT$_{ITM finetuned}$ | 17.73 | 5.67 | 2.13 | 18.03 | 4.72 | 3.43 | **42.31** | 23.08 | **19.23** | 21.58 | 6.85 | 4.11 | 13.89 | 4.63 | 3.70 |
| CLIP (ViT-B/32) | **34.75** | 7.80 | 6.38 | 22.75 | 8.58 | 5.58 | **80.77** | **42.31** | **38.46** | **35.27** | 13.01 | 10.27 | 18.52 | 3.70 | 1.85 |
| VSE++$_{COCO}$ (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++$_{COCO}$ (VGG) | 17.73 | 2.13 | 2.13 | 18.45 | 7.30 | 3.86 | **26.92** | 7.69 | 7.69 | 18.49 | 4.79 | 2.74 | 19.44 | 7.41 | 5.56 |
| VSE++$_{Flickr30k}$ (ResNet) | 20.57 | 6.38 | 3.55 | 18.88 | 4.29 | 2.15 | **26.92** | 3.85 | 3.85 | 21.58 | 6.51 | 3.42 | 15.74 | 0.93 | 0.93 |
| VSE++$_{Flickr30k}$ (VGG) | 17.73 | 4.96 | 2.84 | 19.74 | 6.87 | 5.15 | **30.77** | 7.69 | 7.69 | 20.55 | 6.16 | 4.79 | 17.59 | 6.48 | 3.70 |
| VSRN$_{COCO}$ | 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$_{Flickr30k}$ | 16.31 | 4.96 | 2.13 | 21.03 | 4.29 | 3.86 | **30.77** | 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**.

| | *Symbolic* | | | *Pragmatics* | | | *Same Image Series* | | |
|---|---|---|---|---|---|---|---|---|---|
| *Model* | Text | Image | Group | Text | Image | Group | Text | Image | Group |
| MTurk Human | **96.43** | **92.86** | **92.86** | **58.82** | **41.18** | **41.18** | **95.65** | **91.30** | **91.30** |
| VinVL | 25.00 | 17.86 | 14.29 | **29.41** | 5.88 | 5.88 | **34.78** | 17.39 | 13.04 |
| UNITER$_{large}$ | **39.29** | **28.57** | **17.86** | **35.29** | 0.00 | 0.00 | 4.35 | 8.70 | 0.00 |
| UNITER$_{base}$ | **46.43** | 14.29 | 14.29 | **29.41** | 17.65 | 11.76 | 8.70 | 8.70 | 0.00 |
| ViLLA$_{large}$ | **39.29** | 14.29 | 10.71 | 17.65 | 0.00 | 0.00 | 17.39 | 4.35 | 0.00 |
| ViLLA$_{base}$ | **42.86** | 17.86 | 14.29 | **29.41** | 5.88 | 5.88 | 13.04 | 8.70 | 4.35 |
| VisualBERT$_{base}$ | **28.57** | 0.00 | 0.00 | 5.88 | 0.00 | 0.00 | 13.04 | 0.00 | 0.00 |
| ViLT (ViT-B/32) | **28.57** | 17.86 | 10.71 | **35.29** | 0.00 | 0.00 | **26.09** | 0.00 | 0.00 |
| LXMERT | **28.57** | 3.57 | 3.57 | 17.65 | 5.88 | 0.00 | 8.70 | 4.35 | 0.00 |
| ViLBERT$_{base}$ | **28.57** | 10.71 | 7.14 | **29.41** | 5.88 | 5.88 | 13.04 | 0.00 | 0.00 |
| UniT$_{ITM finetuned}$ | 14.29 | 10.71 | 7.14 | 17.65 | 5.88 | 5.88 | 21.74 | 4.35 | 4.35 |
| CLIP (ViT-B/32) | **39.29** | 3.57 | 3.57 | **35.29** | 5.88 | 5.88 | 8.70 | 0.00 | 0.00 |
| VSE++$_{COCO}$ (ResNet) | **32.14** | 10.71 | 10.71 | 23.53 | 11.76 | 0.00 | 13.04 | 4.35 | 4.35 |
| VSE++$_{COCO}$ (VGG) | 17.86 | 14.29 | 7.14 | 17.65 | 0.00 | 0.00 | 13.04 | 4.35 | 4.35 |
| VSE++$_{Flickr30k}$ (ResNet) | 21.43 | 3.57 | 0.00 | 23.53 | 0.00 | 0.00 | 17.39 | 4.35 | 0.00 |
| VSE++$_{Flickr30k}$ (VGG) | **28.57** | 10.71 | 10.71 | 11.76 | 0.00 | 0.00 | 13.04 | 4.35 | 0.00 |
| VSRN$_{COCO}$ | 7.14 | 3.57 | 0.00 | 11.76 | 0.00 | 0.00 | 13.04 | 0.00 | 0.00 |
| VSRN$_{Flickr30k}$ | 21.43 | 3.57 | 3.57 | **35.29** | 11.76 | 5.88 | 8.70 | 4.35 | 4.35 |

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

we show in Tab. 6 correlations between caption perplexity[2] 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.

## 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 models do above chance overall, compared to only one dual-stream model (CLIP). CLIP was trained on an order of magnitude more data than the other models. Across all types of attention, models struggled compared to humans. But, compared to the random baseline, models that use merged attention (VinVL, VILLA, UNITER and ViLT) and modality-specific attention (CLIP) performed above chance on the full Winoground dataset; none of the remaining models, which all use 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 as an outlier. Tab. 7 shows these correlations, and Appendix B has graphs showing each model's score versus the pretraining data size. The training data

[2] We used the standard size GPT2 checkpoint from Hugging Face transformers to get perplexity [77].

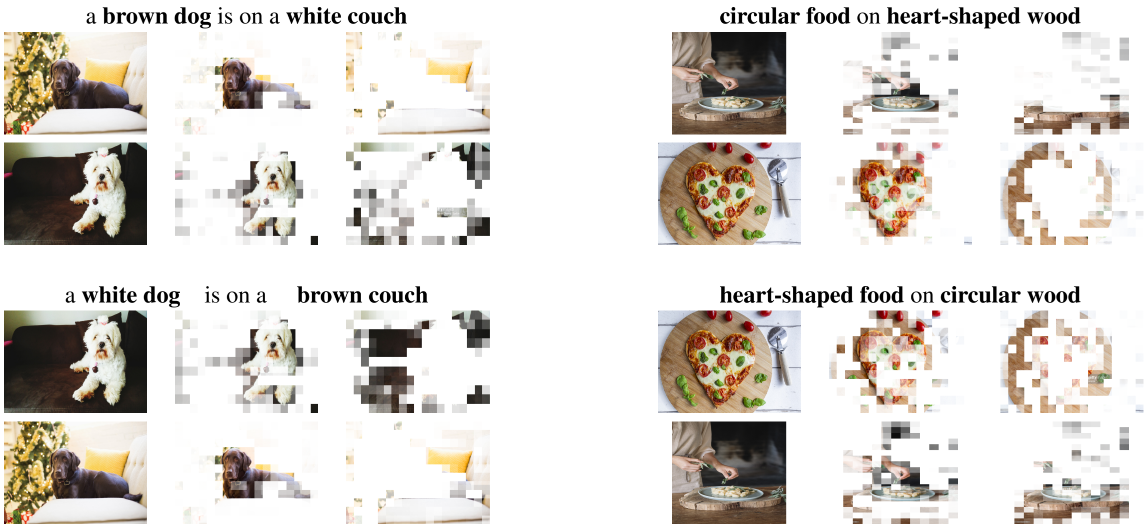

a **brown dog** is on a **white couch**
**circular food** on **heart-shaped wood**
a **white dog** is on a **brown couch**
**heart-shaped food** on **circular wood**

Figure 4. Word-region alignment scores between the image and text features for ViLT [36] 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". Image above is a compilation of assets, including ©Getty Images/Jessica Peterson, Kundanlall Sharma, lacaosa, Alberto Bogo

for unimodal components (e.g. an image backbone or pre-initialized unimodal language model embeddings) is not included in these calculations.

| | *Perplexity* | | *Caption Length* | |
|---|---|---|---|---|
| *Model* | Corr. | p-value | Corr. | p-value |
| MTurk Human | 0.05 | 0.07 | **0.20** | **0.00** |
| VinVL | **-0.05** | **0.04** | **-0.20** | **0.00** |
| UNITER$_{large}$ | -0.01 | 0.57 | **-0.16** | **0.00** |
| UNITER$_{base}$ | -0.03 | 0.22 | **-0.14** | **0.00** |
| ViLLA$_{large}$ | -0.02 | 0.39 | **-0.12** | **0.01** |
| ViLLA$_{base}$ | -0.04 | 0.13 | **-0.11** | **0.03** |
| VisualBERT$_{base}$ | -0.04 | 0.15 | -0.06 | 0.22 |
| ViLT (ViT-B/32) | -0.04 | 0.16 | **-0.16** | **0.00** |
| LXMERT | -0.04 | 0.12 | **-0.11** | **0.02** |
| ViLBERT$_{base}$ | -0.04 | 0.11 | **-0.14** | **0.00** |
| UniT$_{ITM finetuned}$ | -0.01 | 0.73 | -0.02 | 0.73 |
| CLIP (ViT-B/32) | -0.04 | 0.09 | **-0.22** | **0.00** |
| VSE++$_{COCO}$ (ResNet) | **-0.05** | **0.04** | 0.01 | 0.90 |
| VSE++$_{COCO}$ (VGG) | -0.04 | 0.08 | 0.03 | 0.56 |
| VSE++$_{Flickr30k}$ (ResNet) | -0.02 | 0.43 | 0.02 | 0.67 |
| VSE++$_{Flickr30k}$ (VGG) | 0.01 | 0.74 | **-0.10** | **0.04** |
| VSRN$_{COCO}$ | **-0.07** | **0.01** | -0.05 | 0.36 |
| VSRN$_{Flickr30k}$ | -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.

| Pretraining Modality | Score | Corr. | p-value |
|---|---|---|---|
| Image | Text | **0.84** | **0.00** |
| | Image | **0.76** | **0.00** |
| | Group | **0.75** | **0.00** |
| Caption | Text | **0.77** | **0.00** |
| | Image | **0.75** | **0.00** |
| | Group | **0.71** | **0.00** |

Table 7. Correlations between the number of pretraining images and captions and the model text, image, and group scores. CLIP is excluded as an outlier.

## 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 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.

**Broader Impact & Limitations.** Winoground is English-only and translation to other languages may be nontrivial [46]. 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 [4] 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 [55]. Our annotators disproportionately come from the USA and the same could be true for our crowdworkers. See Appendix F for our ethics statement.

## References

[1] Daniel Altshuler, Terence Parsons, and Roger Schwarzschild. *A Course in Semantics*. MIT Press, 2019. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2

[3] David Bender. Establishing a human baseline for the winograd schema challenge. In *Modern Artificial Intelligence and Cognitive Science*, 2015. 2

[4] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. In *arXiv preprint arXiv:2110.01963*, 2021. 9

[5] 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

[6] 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

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4

[8] 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

[9] 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

[10] Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. Context models and out-of-context objects. In *Pattern Recognition Letters*, 2012. 2

[11] Junyoung Chung, Caglar Gulcehr, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS*, 2014. 5

[12] 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

[13] 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

[14] 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

[15] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] https://www.gettyimages.com/. 3

[22] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *ICCV*, 2020. 2

[23] 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

[24] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *NAACL: Human Language Technologies*, 2018. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[26] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *ACL-IJCNLP*, 2021. 2

[27] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 2

[28] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. In *arXiv preprint arXiv:1503.03909*, 2015. 2

[29] Hexiang Hu, Ishan Misra, and Laurens van der Maaten. Evaluating text-to-image matching using binary image selection (bison). In *ICCV*, 2019. 2

[30] 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

[31] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *arXiv preprint arXiv:2102.10772*, 2021. 2, 4, 5

[32] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. First quora dataset release: Question pairs, 2017. 5

[33] 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

[34] Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. Verb argument structure alternations in word and sentence embeddings. In *SCiL*, 2019. 2

[35] 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

[36] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2, 4, 5, 8

[37] 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

[38] 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

[39] 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

[40] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Conference on the Principles of Knowledge Representation and Reasoning*, 2012. 2

[41] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019. 2, 4, 5

[42] 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

[43] 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

[44] 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

[45] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. In *TACL*, 2015. 2

[46] 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

[47] 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

[48] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 5

[49] 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

[50] 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

[51] 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

[52] 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

[53] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *arXiv preprint arXiv:1804.09301*, 2018. 2

[54] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, 2020. 2

[55] 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. 9

[56] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *arXiv preprint arXiv:1508.07909*, 2015. 4

[57] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2, 5

[58] 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

[59] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020. 2

[60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for largescale image recognition. In *CVPR*, 2015. 5

[61] 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

[62] 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

[63] 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

[64] Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. UnNatural Language Inference. In *ACL-IJCNLP*, 2021. 2

[65] 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

[66] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017. 2, 5

[67] 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

[68] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2020. 2, 4, 5

[69] 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

[70] 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

[71] 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

[72] 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

[73] 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

[74] Adina Williams, Andrew Drozdov, and Samuel R. Bowman. Do latent tree learning models identify meaningful structure in sentences? In *TACL*, 2018. 2

[75] 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

[76] Terry Winograd. Understanding natural language. In *Cognitive psychology*, 1972. 2

[77] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric 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. 7

[78] 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

[79] 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

[80] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 5

[81] 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

[82] 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

[83] 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

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