

# Don’t Learn, Ground: A Case for Natural Language Inference with Visual Grounding

Daniil Ignatev<sup>◊†</sup> and Ayman Santeer<sup>◊</sup> and Albert Gatt<sup>◊</sup> and Denis Paperno<sup>◊</sup>

<sup>◊</sup>Utrecht University

<sup>†</sup>Corresponding author: d.ignitev@uu.nl

## Abstract

We propose a zero-shot method for Natural Language Inference (NLI) that leverages multimodal representations by grounding language in visual contexts. Our approach generates visual representations of premises using text-to-image models and performs inference by comparing these representations with textual hypotheses. We evaluate two inference techniques: cosine similarity and visual question answering. Our method achieves high accuracy without task-specific fine-tuning, demonstrating robustness against textual biases and surface heuristics. Additionally, we design a controlled adversarial dataset to validate the robustness of our approach. Our findings suggest that leveraging visual modality as a meaning representation provides a promising direction for robust natural language understanding.

## 1 Introduction

Language models trained and fine-tuned on various textual tasks exhibit impressive performance, especially with more data. At the same time, the extent to which unimodal language models can truly represent meaning has been criticized (Bender and Koller, 2020; Bisk et al., 2020). Indeed, while models can learn a lot about linguistic *form* in language, it is debatable whether exposure to text alone is sufficient to acquire *functional linguistic competence* — i.e., the ability to use language in real-world situations (Mahowald et al., 2024), which also includes drawing inferences and reasoning with linguistic messages. This arguably requires models to take into account the relationship between language and the world, for instance, through visual or other perceptual channels (but see Pavlick, 2023; Mandelkern and Linzen, 2024, for different positions on this issue). Enhancing language models with

multimodal capabilities has now become common (recent examples include Deitke et al., 2024; Peng et al., 2023; Li et al., 2024; Chen et al., 2025). The fusion of linguistic and visual modalities is often cited as a way to address the classic grounding problem (Harnad, 1990), whereby a natural or artificial agent needs to establish systematic links between symbols (say, in natural language) and elements of perception and experience. Though multimodal models perform well in several downstream tasks, it is yet to be shown whether grounding abilities can yield more robust meaning representations and reasoning abilities, obviating the need to fine-tune models on specific datasets for reasoning with natural language.

Natural Language Inference (NLI; aka Textual Entailment) is a case in point. This task is typically framed in terms of the relationship between pairs of texts: given a premise  $p$  and a hypothesis text  $h$ , the goal is to determine whether  $h$  follows from (is entailed by)  $p$ , contradicts it, or whether the relationship is neutral (Dagan et al., 2006; MacCartney, 2009). This task definition is highly flexible, and it has been argued that many understanding tasks lend themselves to an NLI framing (White et al., 2017; Poliak et al., 2019). Yet such a framing arguably focuses our attention on a very narrow view of inference, one that also makes models highly susceptible to learned biases that arise in text, from frequency effects to spurious correlations (e.g., McCoy et al., 2019). An alternative view can be found in classical, truth-conditional semantic accounts, where entailment is not viewed as a relationship between texts, but as a relationship between the models or situations in which a given text holds true. Thus, if  $p$  entails  $h$ , this is because the "world" in which  $p$  holds is likely to be one in which  $h$  also holds.

In this paper, we consider whether, by exploiting the capabilities of multimodal models to ren-

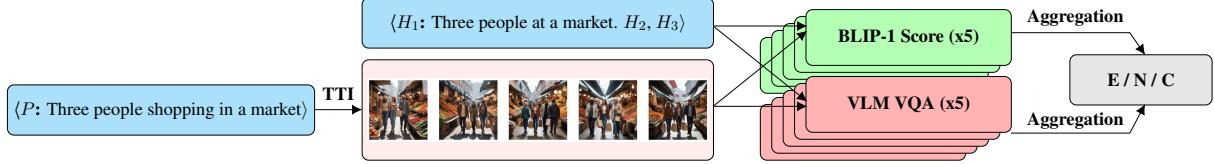


Figure 1: A diagram of the proposed method. NLI is framed in terms of the relationship between a hypothesis ( $h_i$ ) and a visual representation of the situation depicted by a premise ( $p$ ). Visual representations are obtained using a text-to-image (TTI) model. The NLI label is determined based on embedding similarity (e.g. via a model like BLIP) or by directly predicting the label in a VQA setting.

der visually (some of) the possible situations in which  $p$  might hold, we can obtain competitive zero-shot performance on NLI without the need for fine-tuning. In line with the observations above, we are especially interested in whether such an approach can help overcome biases that arise from the exclusively text-based treatment of NLI.

To validate our approach, we make use of data from the SNLI dataset (Bowman et al., 2015) as well as a novel, synthetic adversarial dataset.

Modern approaches to NLI, especially since the introduction of pretrained language models such as BERT (Devlin et al., 2019), often rely on fine-tuning Transformer LMs. Despite their impressive quantitative performance, such NLI models have notable drawbacks. First, successful fine-tuning requires substantial computational resources and large datasets, which typically contain hundreds of thousands of examples. Second, fine-tuned models often perform poorly on new, unseen data due to biases and artifacts present in the training datasets (Nie et al., 2020; Gururangan et al., 2018; McCoy et al., 2019). Attempts to mitigate these biases by expanding fine-tuning datasets exacerbate computational demands while still failing to fully address the issues of bias and out-of-distribution generalization. Finally, as noted earlier, text-based language models addressing semantic tasks like NLI have been criticized on theoretical grounds for capturing a "potentially useful, but incomplete, reflection of [...] actual meaning" (Bender and Koller, 2020). This highlights the need for innovative approaches that leverage the strengths of existing models while reducing the computational burden of fine-tuning and simultaneously exploiting a richer notion of meaning.

Our framework, which is depicted in Figure 1 and elaborated in Section 3 below, builds on the idea that language understanding should rely on grounding language in the world, which neural

models with perceptual interfaces are naturally adapted for. If a model can represent situations in which  $p$  is true and assess the truth of  $h$  in these situations, this should suffice for inference-making. For the sake of argument, we restrict ourselves to visual grounding with static images, the type of grounding for which pre-trained computational models are the most mature. While experimental work has long highlighted the potential of exploiting grounding for inference (cf. Young et al., 2014), it is the recent progress in pre-trained grounded models that makes our approach to inference technically possible. Since our approach is zero-shot, it also does not depend on large-scale exposure to training data. Furthermore, the approach aligns with how entailment is characterized in semantic theory (Winter, 2016), as a relation between the truth of the premise and the hypothesis across possible situations (represented as models).

Our study makes the following contributions:

- We propose a novel method for zero-shot NLI;
- We show that this method attains high accuracy, comparable to that of text-only baselines;
- We validate the robustness of our method against surface biases inherent in fine-tuned NLI models;
- We release a new adversarial NLI dataset based on string overlap bias.

## 2 Related work

**Natural Language Inference** NLI is typically addressed as a unimodal task involving the relationship between textual premises and hypotheses. Following an early phase which considered the project from a logical perspective, for example in the FraCaS framework (Cooper et al., 1996), NLI

came to be defined in probabilistic terms as a result of developments in the PASCAL RTE challenge (Dagan et al., 2006), which also gave rise to the current three-way classification between entailment, contradiction, and neutral (Giampiccolo et al., 2008). Large-scale datasets developed to address this challenge include SICK (Bentivogli et al., 2016), SNLI (Bowman et al., 2015), and MultiNLI (Williams et al., 2018). The present paper uses a subset of SNLI; the full dataset consists of 570k premise-hypothesis pairs, with premises sampled from an image captioning dataset (and which are hence linked to images) (Young et al., 2014) and hypotheses which were crowdsourced. With the advent of large-scale pretraining, the field has witnessed a steady increase in NLI performance on standard benchmarks. For example, a fine-tuned version of DeBERTa (He et al., 2020) achieves around 90% accuracy on SNLI, while RoBERTa (Zhuang et al., 2021) achieves 90.8%, which goes up to 92.8% with the addition of a self-explaining layer. More recently, large language models can be deployed in a zero-shot or few-shot fashion (Brown et al., 2020), although their accuracy in this case remains below that of models fine-tuned on the task. For example, on SNLI, Mistral-7B (Jiang et al., 2023) has a reported accuracy of around 90%, while SOTA performance is achieved with a few-shot version of T5 (Raffel et al., 2020) further trained with synthetic data, reaching 94.7% (Banerjee et al., 2024).

**Bias and shortcut learning** Performance on NLI benchmarks, however, is subject to shortcut learning (Geirhos et al., 2020), sensitivity to data permutations, and an inability to handle paraphrases with similar meanings (Schluter and Varab, 2018). For the specific case of SNLI, Gururangan et al. (2018) found that in a significant proportion of samples, models can guess the correct label based on surface heuristics such as the presence of negation in the hypothesis. They identified a ‘hard’ subset of SNLI where such surface heuristics are not present. In this paper, we use the hard subset in Section 3.2. Subsequent efforts have been dedicated to developing reliable evaluation methodologies to detect when models are not actually solving the inference task, with techniques ranging from stress testing (Naik et al., 2018) to adversarial examples (Beklinkov et al., 2019; Nie et al., 2020). For example, the HANS dataset (McCoy et al., 2019) contains adversarial examples designed to ensure failure

for models that rely on surface heuristics such as lexical overlap. In this paper, we develop similar adversarial examples to compare the susceptibility of unimodal and visually-grounded models to string similarity heuristics (Section 3.3).

**Visually grounded inference** A separate line of work investigates the role of visual grounding in reasoning. This includes tasks involving reasoning about the differences between pairs of images (Suhr et al., 2019; Ventura et al., 2024); Visual Question Answering (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019), where a model needs to answer a question based on an image, with some types of questions requiring models to go beyond object labeling (for example, counting; e.g. Acharya et al., 2019); and visual commonsense reasoning (Zellers et al., 2019; Park et al., 2020). Sun et al. (2023) show that SOTA zero-shot performance on these tasks can be achieved by taking into account fine-grained visual and textual features. Conceptually closer to the original definition of the NLI task, Vu et al. (2018) created a grounded version of SNLI by linking the premises to their original images in Flickr30k (Young et al., 2014). They found that the inclusion of visual information sometimes led to a change in the gold label for a premise-hypothesis pair, but also showed that models do not benefit significantly from the inclusion of images, relative to only using textual premises. Later approaches benefited more from grounding, but not by a large margin (Kiela et al., 2019; De et al., 2023). In a different vein, Suzuki et al. (2019) propose a logical formalism to represent both images and texts, in order to model entailment relationships between them. Xie et al. (2019) presented a new visual-textual entailment task (VTE) and developed the SNLI-VE dataset, where the entailment relationship is defined purely between an image (which replaces the textual premise) and a hypothesis. As with Vu et al. (2018), it was observed that the grounding of image-hypothesis pairs sometimes results in a change of label compared to the original, text-only pairs in SNLI (Do et al., 2021). Lastly, Reijtenbach et al. (2025) showed that generated images can be used for visual entailment just as effectively as the original images. This is similar in spirit to our approach; however, the goal of the present paper is to solve the original NLI task via a visual rendition of a *textual* premise, and to address whether this can help models overcome a reliance on surface heuristics. To that end, we

propose two approaches that we describe in the following section.

### 3 Experiments

We report two sets of experiments, designed to satisfy two requirements: (i) the approach requires no prior task-specific training; (ii) it is visually grounded. Our overall pipeline is depicted in Figure 1. Given an NLI instance  $\langle h, p, L \rangle$ , where  $L$  is the label, our method consists of the following two steps.

1. **Visual representation:** we rely on a text-to-image generation model to create a sample of visual representations  $V_p$  from premise  $p$ .
2. **Image-grounded inference:** We compare two different techniques to infer  $L$  from  $h$  and  $V_p$ :

- **Cosine Similarity Score (CSS):** determines  $L$  based on the similarity between embeddings of  $V_p$  and  $h$ . We assume that high scores indicate entailment, low scores indicate contradiction, and moderate scores indicate neutrality. Given that SNLI premises are paired with at least one hypothesis of each label,  $L$  can be approximated through ordering these hypotheses by similarity: the most similar  $h$  is an entailment, the least similar a contradiction, and the intermediate one is neutral.
- **Visual Question Answering (VQA):** using a multimodal generative model, generates the most likely  $L$  based jointly on  $V_p$  and  $h$ .

Additionally, in each experiment, we attempted to estimate the impact of text-to-image bias and the extent to which it can be mitigated. We did so by generating 5 images per premise and inferring the labels from each one of those; we then aggregated the resulting labels using the following techniques: (i) *Majority vote* selects the most common label, picking randomly in case of a tie; (ii) *Average Value* assigns each prediction a numeric value (1 for Entailment, 0 for Neutral, -1 for Contradiction), and the 5 values are averaged; (iii) *Oracle-Guided*: if any of the five predictions matches the gold label, this label is selected. Note that the oracle-guided method explores the upper performance bound of either method.

#### 3.1 Experiment 1

**Data** We evaluate the proposed approach using V-SNLI, Vu et al. (2018)'s grounded version of SNLI.

For the purpose of this experiment, we select the first 100 premises and their respective hypotheses from the SNLI training split.

**Visual Representation** To generate visual representations  $V_p$ , we employ stable-diffusion-xl-base (Podell et al., 2023). Five images are generated for each individual premise.

**Inference** The two inference methods are implemented as follows. (i) For CSS, we use BLIP-1 (Li et al., 2022) to estimate the alignment between each premise and its respective hypotheses. (ii) For VQA, we utilize gpt-4-vision-preview<sup>1</sup> (OpenAI, 2023). However, due to the deprecation of that model, we later repeated the same experiment using gpt-4o-2024-05-13 to ensure the reproducibility of our tests. We query the VQA models with an image and a textual prompt; the latter includes all three hypotheses and instructs the model to produce NLI labels for all three at once. Exposing the models to all three hypotheses ensures consistency with CSS inference, which considers the hypotheses jointly rather than independently. The full prompt is provided in Appendix A.

**Baselines** We use three zero-shot baselines, both symbolic and neural, to test the validity of visual grounding for NLI. Much like the CSS approach, our baselines rank hypotheses based on their similarity to  $p$  and infer the labels accordingly: entailment is the most similar, contradiction is the least similar, and neutral is in between. However, they only consider textual features and differ in the feature extraction method:

**BLEU:** We compute BLEU (Papineni et al., 2002; Post, 2018) between  $p$  and  $h$  and consider it as a measure of similarity between them. Good performance based on BLEU would suggest that n-gram overlap heuristics contribute to solving the NLI task.

**NSP:** This baseline takes advantage of the next sentence prediction capabilities of BERT-like models. Devlin et al. 2019 demonstrate that NSP pre-training enhances the performance of BERT on

<sup>1</sup>Experiment 1 was first conducted in March and April 2024. Aside from proprietary models, we tested several alternatives: LLaVa-NeXT (Liu et al., 2024) and InstructBLIP (Dai et al., 2023). However, we found that they struggled to consistently follow instructions given the complexity of the task, making them less suitable for our demonstration.

Task	Method	Aggregation	Score
NLI	CSS	Oracle	79.1%
		Average Val.	69.0%
		Maj. Class	69.4%
	VQA <sub>GPT4V</sub>	Oracle	81.0%
		Average Val.	77.0%
		Maj. Class	74.3%
	VQA <sub>GPT4o</sub>	Oracle	80.0%
		Average Val.	74.3%
		Maj. Class	73.0%
Ent.	CSS	Oracle	78.6%
		Average Val.	68.7%
		Maj. Class	66.9%
	VQA <sub>GPT4V</sub>	Oracle	95.1%
		Average Val.	87.3%
		Maj. Class	90.2%
	VQA <sub>GPT4o</sub>	Oracle	92.0%
		Average Val.	81.1%
		Maj. Class	85.1%
Contr.	CSS	Oracle	87.9%
		Average Val.	82.6%
		Maj. Class	82.8%
	VQA <sub>GPT4V</sub>	Oracle	98.9%
		Average Val.	92.9%
		Maj. Class	94.9%
	VQA <sub>GPT4o</sub>	Oracle	98.0%
		Average Val.	94.0%
		Maj. Class	96.0%
Neut.	CSS	Oracle	70.5%
		Average Val.	56.0%
		Maj. Class	57.9%
	VQA <sub>GPT4V</sub>	Oracle	47.9%
		Average Val.	50.0%
		Maj. Class	36.7%
	VQA <sub>GPT4o</sub>	Oracle	49.5%
		Average Val.	47.5%
		Maj. Class	37.4%

Table 1: Accuracy scores for experiment 1 (**5 images** per premise) overall (NLI) and per-class.

NLI. In our experiment, hypotheses are ranked based on how probable they are as the next sentence after the premise, according to the BERT NSP classifier.

**BERTcess:** We extract `CLS` embeddings for both  $p$  and  $h$  using BERT-base and compute the cosine similarity between them to rank hypotheses.

### 3.1.1 Results

Results of Experiment 1 are summarized in Tables 1 & 2. They demonstrate that image-based methods can attain meaningful classification accuracy. Furthermore, they show that visual information is primarily effective at distinguishing entailment and

contradiction. It is also evident that image-based inference struggles to handle neutral instances well, resulting in a sizeable accuracy drop. Aggregating predictions based on multiple (diverse) images leads to better accuracy on that class (compare Tables 1 & 2).

### 3.2 Experiment 2: Biased Data

An important argument in favor of the method we propose is that we expect it to be robust against surface heuristics compared to task-specific fine-tuning. Nevertheless, one of our individual pipelines, VQA, could potentially be affected by annotation artifacts in hypotheses and use NLI-relevant heuristics, such as negation, to shortcut on image analysis. This could have impacted the results of the first experiment, in which we did not account for possible biases of this kind. The second experiment is designed to address the question of hypothesis-level bias.

**Data** In experiment 2, we draw on Gururangan et al. (2018)'s distinction between "hard" and "easy" subsets of SNLI data (see Section 2). We sample 300 hypotheses related to 100 premises from both subsets. When more than three hypotheses were associated with one premise, we discarded the surplus hypotheses. This resulted in 276 "easy" hypotheses related to 92 premises and 285 "hard" hypotheses related to 95 premises. In what follows, we refer to these as the "easy" and "hard" subsets.

**Visual Representation** To generate the images  $V_p$  from premises, we made use of two alternative text-to-image (TTI) models: Stable Diffusion and DALL-E 3 (Betker et al., 2024), a state-of-the-art image generation model at the time these experiments were conducted (September 2024). Our intuition was that leveraging DALL-E images would improve inference accuracy, since this model is less prone to concept bleeding<sup>2</sup> and other potentially misleading TTI generation errors (see Section 4). As in the first experiment, we additionally considered the results of aggregated inference based on 5 images per premise.

**Inference** The inference procedure largely followed the first experiment for both the CSS and the VQA approaches. Since

<sup>2</sup>Concept bleeding occurs when, given a predicate that applies to a specific argument, a text-to-image model renders the image such that the property denoted by the predicate applies to other arguments as well (Podell et al., 2023).

Task	CSS	VQA-1	VQA-2	BLEU	NSP	BERTcss
NLI	69.0%	74.6%	74.3%	39.0%	47.0%	42.0%
Entailment	69.7%	90.1%	86.1%	44.9%	50.0%	47.9%
Contradiction	80.6%	95.0%	97.9%	43.8%	56.2%	41.0%
Neutral	57.0%	37.7%	39.3%	28.8%	36.9%	38.2%

Table 2: Performance of CSS and VQA against baselines on 100 premises (**1 image** per premise), overall (NLI) and per-class.

gpt4-vision-preview was no longer available, VQA experiments were conducted with gpt-4o-2024-05-13.

**Baseline** In the second experiment, our objective was to determine whether multimodal or unimodal inference methods handle bias more efficiently. Our main text-only baseline is NLI fine-tuned RoBERTa (Zhuang et al., 2021), a part of the sentence-transformers library (Reimers and Gurevych, 2019). RoBERTa was trained and fine-tuned on a known set of data, and it is certain that it has not been exposed to the test split of SNLI. At the same time, RoBERTa performs comparably to state-of-the-art models on the NLI task. As an additional sanity check baseline, we once again use BLEU, which only considers surface overlap.

### 3.2.1 Results

We report the results in Table 3. One trend that we observe is that both RoBERTa and the VQA model show considerably lower accuracy on the hard compared to the easy subset, suggesting that both models might rely on surface heuristics related to the hypothesis text  $h$ . In contrast, CSS shows a much smaller gap in performance between the two subsets; however, it also demonstrates lower overall accuracy. Regarding specific classes, distinguishing contradictions remains relatively easy for both VQA and CSS (to the extent that VQA surpasses RoBERTa on the "hard" set), but the Neutral class continues to be challenging to handle.

The results also reveal another tendency: the BLEU baseline is *more* accurate on the hard subset, compared to the easy one. This suggests that while the hard subset mitigates the utility of heuristic biases on the hypothesis text, it may be more susceptible to another heuristic, namely the lexical overlap between  $p$  and  $h$ , an issue we return to in Section 3.3.

Some evidence for the impact of visual grounding comes from the use of multiple images. The results of aggregation over several images, provided in Table 4, show that averaging and oracle aggre-

Task	Method	Easy	Hard
NLI	BLEU	39.1%	44.6% (+5.5)
	RoBERTa	98.9%	83.1% (-15.8)
	GPT-4O	96.3%	85.6% (-10.7)
	CSS-DALL-E	70.3%	69.1% (-1.2)
	CSS-SD	71.7%	65.2% (-6.5)
	VQA-DALL-E	90.6%	74.7% (-15.9)
	VQA-SD	89.5%	70.2% (-19.3)
Ent.	BLEU	41.0%	48.2% (+7.2)
	RoBERTa	98.9%	81.6% (-17.3)
	GPT-4O	95.8%	86.8% (-9)
	CSS-DALL-E	64.2%	63.1% (-1.1)
	CSS-SD	65.2%	57.0% (-8.2)
	VQA-DALL-E	93.7%	78.1% (-15.6)
	VQA-SD	87.4%	73.7% (-13.7)
Contr.	BLEU	46.2%	52.5% (+6.3)
	RoBERTa	97.8%	86.1% (-11.7)
	GPT-4O	98.9%	92.1% (-6.8)
	CSS-DALL-E	87.1%	82.2% (-4.9)
	CSS-SD	89.2%	82.2% (-7)
	VQA-DALL-E	97.8%	93.1% (-4.7)
	VQA-SD	97.8%	91.1% (-6.7)
Neut.	BLEU	29.5%	27.1% (-2.4)
	RoBERTa	100.0%	81.4% (-18.6)
	GPT-4O	94.3%	74.3% (-20)
	CSS-DALL-E	59.1%	60.0% (+0.9)
	CSS-SD	60.2%	54.2% (-6)
	VQA-DALL-E	79.5%	42.8% (-36.7)
	VQA-SD	82.9%	34.3% (-48.6)

Table 3: Overall and per-class percentage accuracy for experiment 2 on SNLI easy and hard subsets (**1 image** per premise). In parentheses: change in accuracy between easy and hard subsets.

gation yield more accurate predictions compared to those based on a single image. However, we also observe that TTI does not consistently yield diverse images for a given premise, which limits the effectiveness of aggregation; we return to this point in our error analysis (Section 4). Additionally, we note the influence of the text-to-image generation procedure: in the VQA setting, generating images with DALL-E leads to noticeably better accuracy compared to Stable Diffusion.

Task	Method	Aggregation	Score
NLI	VQA-DALL-E	Oracle	83.1%
		Average Val.	74.4%
		Maj. Class	74.4%
	VQA-SD	Oracle	78.2%
		Average Val.	70.2%
		Maj. Class	71.6%
Ent.	VQA-DALL-E	Oracle	89.5%
		Average Val.	76.3%
		Maj. Class	79.8%
	VQA-SD	Oracle	85.1%
		Average Val.	71.0%
		Maj. Class	77.2%
Contr.	VQA-DALL-E	Oracle	96.0%
		Average Val.	93.0%
		Maj. Class	93.0%
	VQA-SD	Oracle	93.0%
		Average Val.	91.0%
		Maj. Class	91.0%
Neut.	VQA-DALL-E	Oracle	54.3%
		Average Val.	44.3%
		Maj. Class	38.6%
	VQA-SD	Oracle	45.7%
		Average Val.	38.6%
		Maj. Class	34.3%

Table 4: Overall and per-class percentage accuracy on SNLI-hard in Experiment 2. (**5 images** per premise)

### 3.2.2 Quantifying hypothesis-side bias

Data from SNLI-easy and SNLI-hard can help us quantify how much different models depend on hypothesis-level heuristics, disregarding the premise. This question is not fully answered by comparing the performances on the two subsets, because models may still rely on information in an informative premise, even when the hypothesis alone can be leveraged to solve the task.

To mitigate this, we conduct a separate experiment where we replace all premises with an uninformative statement, "Something is happening." A model that can correctly predict the original  $L$  from this altered  $p$  and the original  $h$  must be exploiting hypothesis-side heuristics. These are known to be present in the easy subset and absent in the hard subset. Therefore, with  $p$ -s replaced, the delta between a model's accuracy on "easy" vs. "hard" accurately reflects how often our models make predictions based on hypotheses alone.

We compare VQA-DALL-E, CSS-DALL-E, and RoBERTa based on these criteria. Since the predictions of VQA-DALL-E and CSS-DALL-E can vary substantially depending on the image, we report

Method	Easy	Hard
Random	33.33%	33.33% (0%)
RoBERTa	51.4%	28.1% (-23.3%)
CSS-DALL-E	42.0%	29.8% (-12.2%)
VQA-DALL-E	45.5%	37.0% ( <b>-8.5%</b> )

Table 5: Accuracy on uninformative premises (averaged over **5 images** for VQA); deltas (in parentheses) estimate the degree of hypothesis side bias.

the average accuracy derived from five images.

As shown in Table 5, RoBERTa exhibits a wide delta of 23.3% and performs below random on hard hypotheses, indicating a substantial hypothesis-side bias. In contrast, VQA-DALL-E and CSS-DALL-E demonstrate narrower deltas than the text-based model: 8.5% and 12.2%, respectively. VQA-DALL-E's and CSS-DALL-E's accuracy on hard hypotheses remains close to random, indicating some reliance on hypothesis properties but no strong bias.

Qualitatively, both methods tend to classify most examples as contradictions. Manual inspection suggests that VQA pipeline only relies on the heuristic involving the mention of thoughts, intentions, and other non-visual predicates in the hypothesis (e.g., "Boys **trying to escape** an incoming storm"), which leads it to classify such examples as neutral. This behavior may be influenced by the prompt, where this type of predicates is explicitly mentioned as a criterion for the Neutral class: see Appendix A. RoBERTa, on the other hand, appears to consider additional properties of the hypothesis, such as the occurrence of modifiers.

### 3.3 Experiment 3: Adversarial Data

Experiment 2 revealed a potential source of bias in the hard SNLI subset, beyond the hypothesis-side heuristics. The higher accuracy of the BLEU baseline on the hard subset suggests the presence of considerable textual overlap between  $p$  and  $h$ , which could provide an additional shortcut for language models to exploit. Our third experiment was designed to address this question and investigate whether our methods are susceptible to this bias.

**Data** Inspired by previous work (e.g., McCoy et al., 2019), we designed a synthetic dataset intended to mitigate overlap-based heuristics, while ensuring that premises could be rendered accurately as images. By creating a controlled synthetic dataset, we aimed to achieve two goals: first, address both lexical and sequence overlap as NLI

heuristics; second, analyze on a small scale how frequently incorrect text-to-image generation leads to errors.

Sample premises were generated according to the following template: *The [noun1] who [transitive\_verb] the [noun2] [intransitive\_verb]*. (e.g., *The girl who greets the dog laughs*). From such premises, an entailed hypothesis was generated based on the template *The [noun1] [intransitive\_verb]*. (e.g., *The girl laughs*); similarly, non-entailment examples were generated based on the template *The [noun2] [intransitive\_verb]* (e.g., *The dog laughs*). In order to reduce ambiguity, we did not generate neutral statements. We expected the hypotheses to be adversarial concerning both lexical overlap and subsequence, since both of them share an equal number of words with the premise, but *[noun2]* also directly precedes *[intransitive\_verb]*, prompting a heuristic-influenced model to likely recognize it as the verb’s subject. Nouns and verbs were chosen from a small, manually selected set that maximized the chances of accurate visual representation: *[intransitive\_verb]* values denoted actions with clear facial expressions (e.g., *laughs*); the two noun values consisted of a human- and an animal-denoting noun. Initial experiments with two human-denoting nouns with distinct visual features (e.g., *policeman* and *mechanic*) showed that Stable Diffusion and, to a lesser extent, DALL-E were prone to introducing concept bleeding, whereby both characters would share the same facial expression described by *[intransitive\_verb]*. In total, we generated 100 premises, each paired with both an entailed and a non-entailed hypothesis, resulting in a total of 200 pairs<sup>3</sup>.

**Baseline** Once again, we include fine-tuned RoBERTa as a strong comparison. The BLEU-based baseline used in experiment 2 is not applicable here, as the lexical overlap between hypotheses and premises is identical for both entailments and non-entailments.

**Inference** The pipelines for the compared models remained the same as in Experiment 2, utilizing DALL-E-3, BLIP-1, and GPT4o. The only adjustment was that both models operated with a set of two labels (“entailment” and “non-entailment”) in-

Task	Method	Adversarial
Ent. vs Non-Ent.	RoBERTa	65.5%
	CSS-DALL-E	56.0%
	VQA-DALL-E	85.0%
	VQA-SD	74.0%

Table 6: Percentage accuracy on adversarial data, Experiment 3 (**1 image** per premise).

stead of the three standard NLI labels. The VQA prompt was updated accordingly.

### 3.3.1 Results

VQA achieves an accuracy of 85% on the adversarial data using DALL-E images, outperforming the fine-tuned RoBERTa model (65.5%), see Table 6. In contrast, CSS yields a lower accuracy of 56%, indicating a noticeable word overlap bias. A possible explanation for this is that BLIP-1 focuses on matching tokens with image regions while losing syntactic information. We discuss this issue in more detail in Section 4. We conducted a manual analysis of the generated images whereby we found that up to 15% of these suffered from concept bleeding. This issue largely overlaps with the erroneous predictions made by VQA: out of 30 errors, 14 were due to factual inaccuracies in the images, 11 due to concept bleeding, and only 5 due to incorrect image-to-text inference. These observations further highlight that improvements in TTI generation will enhance the performance of our method.

## 4 Error analysis

We distinguish two general types of errors in our pipeline: those that stem from visualizing NLI premises (1) and those from the subsequent inference (2).

### 4.1 Text-to-image: Factual Errors

Text-to-image models have some widely known limitations, such as factual inaccuracies, concept bleeding, and incorrect object counts (Podell et al., 2023), which we also encountered in practice for some of the SNLI examples. When this occurred, the inference often led to incorrect results. One such example is illustrated in Figure 2a: “One man sits inside and plays the banjo; there are trees behind him outside” is depicted as an outdoor scene. Despite this, the fact that DALL-E is less prone to such issues compared to SD-XL suggests that

<sup>3</sup>We will make our data available on GitHub after the anonymity period.

more advanced text-to-image models may become increasingly robust in this respect.

## 4.2 Text-to-image: Neutrality Errors

When constructing hypotheses for the neutral class, SNLI annotators employed a limited set of strategies (Gururangan et al., 2018). One common strategy observed in our subsets of SNLI involves introducing an object that is semantically fitting but not mentioned in the premise. For instance, consider the premise "Three men standing on grass **by the water** looking at something on a table" and the hypothesis "The three men are **by the lake**". While the textual premise can remain agnostic about specific details, such as the type of water, translating this premise into the visual domain forces the model to depict a specific landscape, such as a seashore, a lake-shore or a riverbank (as illustrated by Figure 2b). Consequently, neither of our methods can reliably identify the neutral class. The same issue affects many other neutral instances as well.

The multi-image approach described in Section 3 could potentially address this issue: assuming that the problematic object appears in some images but not in others, aggregating the inference results could lead to the neutral class being selected as a middle ground. However, in our evaluation attempts, even with varying random seeds and temperature values, the model-generated images either consistently omitted or consistently included the problematic visual objects ("lake", etc.). As a result, normalizing predictions over sets of five images, whether through averaging or majority voting, did not yield the expected improvement in scores.

## 4.3 CSS: Object overlap

Powered by BLIP, CSS is prone to a multimodal overlap bias: it assigns higher similarity scores to descriptions that include more words with visual referents in the image, regardless of their semantic correctness. Pezzelle, 2023 observes a similar tendency for CLIP. Likewise, models of this kind are also unable to fully leverage compositional information due to how they handle text and image encoding (Kamath et al., 2023). An example of this can be found in Figure 2c, where for  $p$  "A boat worker securing a line", CSS favors the neutral  $h$ , "The boat worker works hard to establish the line", and not the entailment, "A worker is doing something to a boat". The "line" object, present in the



Figure 2: Examples of incorrectly classified premise representations.

neutral hypothesis and in the images, contributes to a higher overlap.

## 5 Conclusion

In this paper, we present the first implementation of an NLI classifier based purely on visual grounding. Our method is zero-shot and exhibits high accuracy, surpassing the few-shot results reported for NLI in Brown et al. (2020). Unlike state-of-the-art task-specific fine-tuning, our methods (VQA and, to a lesser extent, CSS) are cognitively and theoretically grounded and avoid known superficial biases. They are also robust against injecting irrelevant biases from training data (although the specific implementation of CSS tested here exhibits overlap effects). While the standard approach of fine-tuning models on adversarial data can mitigate biases (and potentially introduce new ones), we demonstrated how these biases can be mitigated by *avoiding* fine-tuning. Due to the use of images as an intermediate representation, our method offers additional interpretability. For example, in our error analysis, we were able to track errors down to specific components of our system, such as text-to-image generation or image-based inference.

We tested our method on examples describing everyday scenes as represented by SNLI. With necessary modifications, our approach can be extended to other reasoning tasks where relevant information can be visualized, such as spacial reasoning. At the same time, as is the case with visual grounding per se, dealing with abstract entities remains challenging (Beinborn et al., 2018).

A notable limitation of our approach is that it

works much better with entailments and contradictions than with neutral hypotheses. We believe that aggregation of predictions from multiple diverse images can help bridge this gap; our preliminary results in this directions are cautiously promising.

Many of the errors we observed are linked to limitations in the existing image generation models. Therefore, one can expect that the same basic method will demonstrate significant improvements over time, given the rapid pace at which image generation models are evolving. Furthermore, we find it promising to eventually extend this approach to more diverse types of grounding, such as audio or video.

We argued in this paper that grounding can provide a radically different approach to NLI. However, our general hypothesis suggests that grounding offers a qualitatively better handling of meaning in general. Can the essence of our method extend to other tasks? One could envision, for instance, story generation with interleaved generated illustrations that also help support the narrative coherence, or visualization as an aid for understanding and generating metaphorical language.

In the context of NLI, our approach reduces training compute but adds more steps at inference time, including image generation. Interestingly, the current practice of text-only models also trends toward increased test-time compute, sometimes to a much greater extent than our approach (OpenAI, 2024).

Lastly, our approach might have further theoretically interesting properties. First, while we tested our method on predicting categorical NLI labels, relying on multiple images and possibly on the continuous nature of image-hypothesis match can address the probabilistic nature of NLI, recognized since the first competitions on the task (Glickman et al., 2005). Second, our work can contribute to old theoretical debates. Take the example of two views on semantics (Lepore, 1983), model theoretic (based on reference) vs. structural (based on formal relations between expressions that support semantic inference). Grounded modeling of meaning like in our method corresponds to model theoretic semantics while language modeling based approaches correspond to structural semantics. The two might therefore not only be theoretically complementary but also computationally implemented differently using state-of-the-art AI models.

## Acknowledgments

## References

- Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. [TallyQA: Answering complex counting questions](#). In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii. Association for the Advancement of Artificial Intelligence. ArXiv: 1810.12440 ISSN: 2159-5399.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. [VQA: Visual Question Answering](#). In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*, pages 2425–2433, Santiago, Chile. IEEE. ArXiv: 1505.00468v1.
- Sourav Banerjee, Anush Mahajan, Ayushi Agarwal, and Eishkaran Singh. 2024. [First Train to Generate, then Generate to Train: UnitedSynT5 for Few-Shot NLI](#). ArXiv:2412.09263 [cs] version: 2.
- Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. [Multimodal grounding for language processing](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush. 2019. [Don’t Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. 2016. [SICK Through the](#)

**SemEval Glasses.** *Language Resources and Evaluation*, 50(1):95–124.

James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † Jun-tangZhuang, † JoyceLee, † YufeiGuo, † Wesam-Manassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. 2024. Improving image generation with better captions.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. **Experience Grounds Language.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP’20)*, volume 2004.10151, pages 8718–8735, Online. Association for Computational Linguistics. ArXiv: 2004.10151.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xiee, Xingkai Yu, and Chong Ruan. 2025. Janus-Pro: Unified Multi-modal Understanding and Generation with Data and Model Scaling.

Robin Cooper, Dick Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. **FRA-CAS: A Framework for Computational Semantics.** (LRE 62-051 Deliverable D16).

Ido Dagan, Oren Glickman, Bernardo Magnini, and Ramat Gan. 2006. The PASCAL Recognising Textual Entailment.pdf. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F D’Alche-Buc, editors, *Machine Learning Challenges*, pages 177–190. Springer, Berlin and Heidelberg.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang,

Boyang Li, Pascale Fung, and Steven Hoi. 2023. **Instructblip: Towards general-purpose vision-language models with instruction tuning.**

Arkadipta De, Maunendra Sankar Desarkar, and Asif Ekbal. 2023. Towards improvement of grounded cross-lingual natural language inference with visiotextual attention. *Nat. Lang. Process. J.*, 4:100023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, Yen-Sung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Fayyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. **Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models.** ArXiv:2409.17146 [cs].

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

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2021. **e-SNLI-VE: Corrected Visual-Textual Entailment with Natural Language Explanations.** ArXiv:2004.03744 [cs].

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. **Shortcut learning in deep neural networks.** *Nature Machine Intelligence*, 2(11):665–673. Number: 11 Publisher: Nature Publishing Group.

Danilo Giampiccolo, H. Dang, B. Magnini, Ido Dagan, Elena Cabrio, and W. Dolan. 2008. **The Fourth PASCAL Recognizing Textual Entailment Challenge.**

- Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. Web based probabilistic textual entailment. In *Proceedings of the 1st Pascal Challenge Workshop*, pages 33–36.
- Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering.** In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR’17)*, pages 6904–6913. ArXiv: 1612.00837 ISSN: 15731405.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. **Annotation artifacts in natural language inference data.** In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Drew A. Hudson and Christopher D. Manning. 2019. **GQA: A new dataset for real-world visual reasoning and compositional question answering.** *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:6693–6702. ArXiv: 1902.09506v3 ISBN: 9781728132938.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. **Mistral 7B.** ArXiv:2310.06825 [cs].
- Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. **Text encoders bottleneck compositionality in contrastive vision-language models.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.
- Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. **Supervised multi-modal bitransformers for classifying images and text.** *ArXiv*, abs/1909.02950.
- Ernest Lepore. 1983. What model theoretic semantics cannot do? *Synthese*, pages 167–187.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. **LLaVA-OneVision: Easy Visual Task Transfer.** ArXiv:2408.03326 [cs].
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. **BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation.** In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. **Llava-next: Improved reasoning, ocr, and world knowledge.**
- Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University.
- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. **Dissociating language and thought in large language models.** ArXiv:2301.06627.
- Matthew Mandelkern and Tal Linzen. 2024. **Do Language Models’ Words Refer?** *Computational Linguistics*, 50(3):1191–1200.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. **Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. **Stress test evaluation for natural language inference**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. **Adversarial NLI: A new benchmark for natural language understanding**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- OpenAI. 2023. **Gpt-4v(ision) system card**.
- OpenAI. 2024. Learning to reason with llms. Accessed 03.02.2025.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. 2002. **BLEU : a Method for Automatic Evaluation of Machine Translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.
- Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. **VisualCOMET: Reasoning About the Dynamic Context of a Still Image**. In *Proceedings of the European Conference on Computer Vision*, pages 508–524, Berlin and Heidelberg. Springer. ArXiv: 2004.10796 ISSN: 16113349.
- Ellie Pavlick. 2023. **Symbols and grounding in large language models**. *Philosophical Transactions of the Royal Society A*.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. **Kosmos-2: Grounding Multi-modal Large Language Models to the World**. ArXiv:2306.14824.
- Sandro Pezzelle. 2023. **Dealing with semantic underspecification in multimodal nlp**.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. **Sdxl: Improving latent diffusion models for high-resolution image synthesis**. *arXiv*.
- Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2019. **Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation**. pages 337–340.
- Matt Post. 2018. **A call for clarity in reporting BLEU scores**. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. **Exploring the limits of transfer learning with a unified text-to-text transformer**. *Journal of Machine Learning Research*, 21:1–67. ArXiv: 1910.10683.
- Rob Reijtenbach, Suzan Verberne, and Gijs Wijnholds. 2025. **Dataset creation for visual entailment using generative AI**. In *Proceedings of the 5th Workshop on Natural Logic Meets Machine Learning (NALOMA)*, pages 8–17, Bochum, Germany. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. **Sentencebert: Sentence embeddings using siamese bert-networks**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Natalie Schluter and Daniel Varab. 2018. **When data permutations are pathological: the case of neural natural language inference**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4935–4939, Brussels, Belgium. Association for Computational Linguistics.
- Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. **A Corpus for Reasoning about Natural Language Grounded in Photographs**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*, pages 6418–6428. ArXiv: 1811.00491.
- Rui Sun, Zhecan Wang, Haoxuan You, Noel Codella, Kai-Wei Chang, and Shih-Fu Chang.

2023. UniFine: A Unified and Fine-grained Approach for Zero-shot Vision-Language Understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 778–793, Toronto, Canada. Association for Computational Linguistics.
- Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6883–6893, Torino, Italia. ELRA and ICCL.
- Riko Suzuki, Hitomi Yanaka, Masashi Yoshikawa, Koji Mineshima, and Daisuke Bekki. 2019. Multimodal Logical Inference System for Visual-Textual Entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 386–392, Florence, Italy. Association for Computational Linguistics.
- Mor Ventura, Michael Toker, Nitay Calderon, Zorik Gekhman, Yonatan Bitton, and Roi Reichart. 2024. Nl-eye: Abductive nli for images.
- Hoa Trong Vu, Claudio Greco, Alia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. 2018. Grounded textual entailment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2354–2368, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- A. S. White, P. Rastogi, K. Duh, and B. Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing (IJCNLP’17)*, pages 996–1005, Taipei, Taiwan. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Yoad Winter. 2016. *Elements of formal semantics: An introduction to the mathematical theory of meaning in natural language*. Edinburgh University Press.
- Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. *arXiv*, 1901.06706. ArXiv: 1901.06706v1.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’19)*, pages 6713–6724. ArXiv: 1811.10830 ISSN: 10636919.
- Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

## A Appendix VQA Prompt

In line with recommendations by Sun et al. (2024), we replace the task-specific labels (entailment/contradiction/neutral) with natural language labels (accurate/contradicting/neither).

Question: With respect to the objects in the image, is the statement in square brackets a) accurate, b) contradicting, c) neither?

Answer ‘accurate’ if the statement accurately describes the objects in the image.

Answer ‘contradicting’ if objects miss from the image or if the description is incorrect.

Answer ‘neither’ if the statement is factually correct, but makes claims that don’t follow from the image (intentions, social relations etc.).

Put your answers into angle brackets.

Statement 1: [...]

Statement 2: [...]

Statement 3: [...]

Answer 1 (accurate/contradicting/neither): <...>

Answer 2 (accurate/contradicting/neither): <...>

Answer 3 (accurate/contradicting/neither): <...>