Don't learn, Ground. Image Generation for Grounded NLI

Anonymous ACL submission

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 with textual hypotheses. The pipeline achieves an accuracy comparable to text-based NLI classifiers while offering additional transparency. Our findings suggest that grounding language in vision is a viable and effective strategy for advancing 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). Achieving *functional linguistic competence* requires models to take into account the relationship between language and the world, e.g., 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., 2025; Peng et al., 2024; Li et al., 2025; Chen et al., 2025).

Natural Language Inference (NLI; aka Textual Entailment) is a case in point. This task is typically framed in terms of the relationship between a premise text p and a hypothesis text h: whether h follows from (is entailed by) p, contradicts it, or whether the relationship is neutral (Dagan et al., 2006; MacCartney, 2009). This narrow framing of inference makes NLI classifiers susceptible to learned biases that arise in text (e.g., McCoy et al., 2019).

Modern approaches to NLI often focus on fine-tuning pretrained language models. Despite their impressive quantitative performance, such NLI models have notable drawbacks. First, successful fine-tuning requires substantial computational resources and large datasets. Second, fine-tuned models often scale poorly to unseen data due to biases present in the training datasets (Nie et al., 2020; Gururangan et al., 2018; McCoy et al., 2019; Zgreabăn et al., 2025).

We introduce a framework that operationalizes a truth-conditional view of entailment by grounding the premise in a concrete situation: we first render a visual scene consistent with the premise, and then ask a vision-language model to judge whether the hypothesis holds in that scene. This design is in line with recent evidence that generated assistive images can boost chain-of-thought (CoT) reasoning (Zhou et al., 2025). As our pipeline is fully zero-shot, it avoids task-specific fine-tuning and lets us probe whether grounding can reduce reliance on textual heuristics. Our contributions are as follows:

- We evaluate a zero-shot, visually grounded NLI pipeline that uses generated images as intermediate representations.
- On SNLI *easy*/*hard* subsets, we show that grounded inference can approach strong text-only baselines when paired with high-fidelity image generators (Table 1).
- We quantify hypothesis-side shortcut behavior via a premise-ablation control (RoBERTa (abl.)) and compare sensitivity to dataset artifacts across text-only and grounded settings.
- We benchmark multiple image generators and show that a smaller, faster model (FLUX.1-schnell, Labs, 2024) can yield competitive grounded NLI accuracy, suggesting a practical route to scaling the method.

2 Related work

Natural Language Inference Large-scale datasets developed to study NLI include SICK (Bentivogli et al., 2016), SNLI (Bowman et al., 2015), and MultiNLI (Williams et al., 2018). With the advent of large-scale pretraining, the field has witnessed a steady increase in NLI performance on standard benchmarks. For example, RoBERTa (Zhuang et al., 2021) achieves 90.8% accuracy on SNLI. 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).

Visually grounded inference A separate line of work investigates the role of visual grounding in reasoning. This includes Visual Question Answering (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019; Acharya et al., 2019) and visual commonsense reasoning (Zellers et al., 2019; Park et al., 2020). Importantly, Zhou et al. (2025) showed that generated intermediary images (e.g., blueprints) can be used to guide CoT reasoning on complex tasks; we adapt a similar approach to NLI.

Conceptually close 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. Later approaches benefited more, 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 for entailment involving both images and texts.

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 can result in a change of label compared to the original, text-only pairs in SNLI (Do et al., 2021; Kayser et al., 2021). Reijtenbach et al. (2025) showed that images generated from SNLI premises are applicable for VTE. While our approach is similar, we do not consider images independently, but rather as intermediary representations. Their study also focused on the applicability of generated data, while we consider the *pro* and *contra* in more detail.

3 Experiments

Our primary question is whether zero-shot NLI can be achieved by means of a visual representation of the text. To that end, we compare our implementation of this pipeline with text-only baselines. In addition, we consider two questions:

1) Whether grounded NLI helps bypass some of the recognized problems in text-centric NLI, such as hypothesis-side heuristics. The latter are keywords and grammatical features in hypotheses (e.g., negation) that are spuriously correlated with NLI classes in SNLI (Gururangan et al., 2018; Geirhos et al., 2020).

2) How suitable are different models for image generation for the task of NLI grounding.

Data We evaluate the proposed approach using text and images from V-SNLI, (Vu et al., 2018). Due to the large size of the corpus, we restrict our analysis to a limited, manually analyzable subset. Specifically, to test our method's robustness against heuristics, we draw on Gururangan et al. (2018)'s distinction between *hard* and *easy* subsets of SNLI data. We sample items both subsets leaving only those where all three NLI labels are present; as the *easy* set is larger, we obtain 1080 *easy* hypotheses related to 360 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 images, we use four text-to-image (TTI) models: Stable Diffusion XL (Podell et al., 2024), DALL-E 3 (Betker et al., 2024), FLUX.1-schnell (Labs, 2024), and Qwen Image (Wu et al., 2025). We produce three images from each model and report the average of three inference passes. In addition, we make use of images generated by Reijtenbach et al. (2025) using a fine-tuned version of Stable Diffusion 1.5 (Rombach et al., 2022); their dataset includes 1 image per premise p. As a sanity check, we also use the original Flickr images from V-SNLI.

Our choice of TTI models allows for a comparison between parameter-heavy, large language

Method		Easy	Hard
RoBERTa		97.6%	83.2% (-14.4)
RoBERTa (abl.)		52.3%	28.4%(-23.9)
gpt-4o-2024-05-13		92.1%	80.6% (-11.5)
Qwen-VL-72B-Instruct		83.4%	80.2% (-3.2)
Qwen-VL-72B-Instruct+CoT		90.6%	85.1% (-5.5)
Images+Qwen-VL	DSG		
Flickr	-	84.1%	73.3% (-10.8)
Qwen-Image	0.949	92.6%	82.1% (-10.5)
DALL-E 3	0.914	87.5%	79.8% (-7.7)
FLUX.1-schnell	0.884	89.0%	78.5% (-10.5)
SDXL	0.883	90.0%	78.4% (-11.4)
SD1.5	0.623	79.6%	60.7% (-18.9)

Table 1: Percentage accuracy of NLI methods on easy and hard data. All scores except Flickr and SD1.5 are the average of 3 runs. In parentheses: accuracy delta between the subsets. DSG (Cho et al., 2024) scores are reported for image generators (see Section 4)

model (LLM)-enhanced models (Qwen Image, DALL-E 3) and lighter models with weaker text encoding (FLUX.1-schnell, Stable Diffusion).

Inference For visual inference, we utilize `Qwen2.5-VL-72B-Instruct` (Bai et al., 2025), an open-weight MLLM that offers a solid approximation of the current state of the art. In addition, we considered LLaVa-NeXT (Liu et al., 2024) as an alternative backbone, but left it out due to a high rate of formatting errors (8%). The model is queried with an image and a textual prompt that includes three hypotheses and instructs the model to produce NLI labels for all three (see Appendix A).

Baseline We employed four baselines to represent zero-shot and fine-tuned NLI classifiers. We use `gpt-4o-2024-05-13` (Hurst et al., 2024) as a strong zero-shot text-only baseline. We also run `Qwen2.5-VL-72B-Instruct` on textual premises and hypotheses, with and without CoT. The prompt is adjusted to accommodate texts. Lastly, we test fine-tuned RoBERTa (Zhuang et al., 2021) from `sentence-transformers` (Reimers and Gurevych, 2019).

Even though RoBERTa can compete with state-of-the-art models on SNLI, it is known to be prone to bias as a consequence of fine-tuning (Gururangan et al., 2018). To control for how much this affects RoBERTa, we additionally report an ablation: we substitute textual premises with uninformative text ("something is happening") that cannot entail or contradict most hypotheses and report the results. If these are notably above random, this may indicate that the model is using textual shortcuts.

3.1 Results

We report the results in Table 1. Overall, grounded zero-shot NLI attains reasonably high accuracy, showing that the method is generally applicable. With high-fidelity image generators, grounded inference approaches the performance of strong baselines: Qwen-Image+VL reaches 92.6% accuracy on the *easy* subset and 82.1% on the *hard* subset, comparable to strong textual models in the zero-shot setting.

Concerning robustness to surface heuristics, the results do not provide definite evidence that grounding eliminates sensitivity to dataset artifacts. All models, independent of grounding, exhibit lower performance on the *hard* subset. Fine-tuned RoBERTa shows a larger-than-average accuracy drop (14.4%), suggesting strong reliance on hypothesis-side heuristics; this is supported by the ablation experiment, in which RoBERTa remains above chance on the *easy* subset (52.3% vs. 33%). At the same time, GPT-4o and image-conditioned models show comparable *easy*/*hard* deltas (≈11%), indicating that the *easy*/*hard* split reflects general dataset difficulty rather than hypothesis-only reasoning alone.

Nevertheless, the results indicate that visual representations consistently influence inference. Performance varies systematically with the source and fidelity of the images: higher-quality generators (Qwen-Image) yield stronger results than lower-fidelity ones (SD1.5), despite identical textual inputs. Notably, using original Flickr images results in lower accuracy than most generated images on both subsets, suggesting that visual properties such as resolution and salience affect grounded inference. Further evidence comes from error overlap analysis (Table 2). The limited overlap between errors made by textual models and those made by image-assisted inference suggests that grounded models are not driven by the same hypothesis-level features alone. Remarkably, this is more pronounced on the *hard* subset, where the ablated RoBERTa performs below chance level.

Finally, the comparison across image generators reveals that LLM-backed diffusion models produce images that support more accurate grounded inference; particularly, Qwen-Image achieves the strongest overall performance. At the same time, FLUX.1-schnell performs competitively despite its smaller size and reduced computational requirements, indicating that effective grounded NLI does

Method	Easy	Hard
RoBERTa	26% (23)	35.4% (48)
gpt-4o-2024-05-13	20.9% (86)	37.5% (55)
Qwen-VL-72B-Instruct	26.5% (162)	42.8% (56)
Qwen-VL-72B-Instruct+CoT	19.6% (102)	34.9% (43)

Table 2: Baselines: relative error overlap with Qwen-Image + Qwen-VL; in parentheses: total error count.

not necessarily require the most resource-intensive image generation models. With FLUX.1-schnell requiring the fewest diffusion steps (16 vs. ≈ 50), it could have sufficient throughput for real-time production of assistive images.

4 Discussion

The fact that `Qwen2.5-VL-72B-Instruct` yields higher accuracy on artificial data than on Flickr images, contrary to expectations, may challenge its adequacy for our task. We consider the explanation that the low resolution of Flickr images (500×500) may be too coarse for the model. Indeed, similarly sized SD1.5 images (512×512) also lead to low accuracy; by contrast, 1024×1024 images systematically yield higher accuracy. Our analysis of predictions further reinforces this view: we find that Flickr+VL predicts the Neutral label disproportionately often: its ratio in *hard* and *easy* subsets is 36.8% and 40.2%. In contrast, the proportions of Neutral in the gold labels are 24.6% and 31.5% respectively. For SD1.5+VL, the ratio is also above 33%. We supply the confusion matrices in Appendix B. We see this as an indication that low resolution prevents the model from distinguishing the supporting details, although the VL model appears adequate overall. Still, the resolution issue may have implications for further work with V-SNLI and its descendants, as it could explain why these datasets continue to be challenging for the most modern VL models (Pitta et al., 2025); likewise, artificial VTE data production (Reijtenbach et al., 2025) has to account for different resolutions being easy or difficult for vision models.

Generally, we distinguish two types of errors pertaining mostly to text-to-image: factual errors (1) and neutrality errors (2). The former include incorrect rendering of object counts, concept bleeding (Podell et al., 2024), etc. We estimate the impact of these errors on our pipeline using two measures. First, the average score on Davidsonian Scene Graphs (Cho et al., 2024), i.e., the proportion of correctly rendered atomic statements from 0 to 1 calculated over a balanced set of hard and easy items (see Appendix D). The score is computed with `Qwen-2.5-VL-Instruct` and verified manually with $\alpha \approx 0.7$ (Castro, 2017). Second, the proportion of Contradiction labels. As factual errors often lead to the Entailment label flipping, the latter score can be a coarse proxy. Our conclusions are as follows: we find that faithfulness and contradiction ratio are negatively correlated (Pearson's $r = -0.97$, $n = 10$, $p < 0.05$); more importantly, we also observe that the impact of factual errors is substantial in SD1.5, but much less so in the modern diffusion models with $\approx 5\%$ inaccuracies for Qwen-Image (see Table 1). This calls for caution when using TTI models, but also shows that some of them are fit for the task.

Neutrality errors occur when the image inadvertently overspecifies initially ambiguous details, an issue that complicated turning SNLI into a VTE dataset (Do et al., 2021; Kayser et al., 2021). These errors often align with commonsense inference as opposed to logical inference; consider the premise "The dog about to catch a frisbee", inadvertently pictured as an outside scene, and the neutral hypothesis "The dog is outside". Human annotators label it as Entailment in 2 out of 5 cases, resorting to common sense inference due to limited instructions (Kalouli et al., 2023). The transparency offered by our method is beneficial here, as it can help identify the premise-hypothesis pairs prone to mislabeling due to commonsense reasoning.

5 Conclusion

This paper explored the use of assistive images as intermediate representations for human-interpretable zero-shot Natural Language Inference. Our results show that visually grounded NLI is feasible when premise information can be rendered visually, achieving accuracy comparable to strong text-only baselines. By grounding inference in concrete visual situations, the proposed pipeline aligns with a truth-conditional view of entailment and offers increased transparency.

Our comparison of text-to-image models shows that while none of them is error-free, they can support accurate NLI judgments, and the impact of factual rendering errors appears limited for strong models. Meanwhile, grounded inference does not require the most resource-intensive models: systems like FLUX.1-schnell yield competitive results, suggesting that image-assisted inference may be viable beyond the small scale.

6 Limitations

Several limitations of the present study should be noted. First, text-to-image generation models are known to exhibit social, demographic, and representational biases, and they may fail to visualize certain statements due to safety filters or training artifacts. Although we encountered only 8 such cases in our experimental data, these issues remain a general concern for image-assisted reasoning methods and warrant careful consideration.

Second, while we evaluated a diverse set of image generators and a strong VL model, our coverage is not exhaustive. Other architectures, prompting strategies, or vision–language models may yield different results, and a more systematic exploration of the design space could provide further insights.

Finally, scalability remains a practical consideration. Generating high-resolution images incurs considerable computational cost, which may limit throughput in large-scale applications. Since we propose our solution for interpretably diagnosing the problems in the existing data and interpretable inference, this problem is not as critical, as this implies a limited scale. In addition, our results indicate that cheaper and faster generators can still support accurate grounded inference.

In the present work, we used generative AI tools for spelling and grammar checks.

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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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.

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, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † 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)*, pages 8718–8735.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. `https://github.com/pln-fing-udelar/fast-krippendorff`.

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

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*.

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.

Arkadipta De, Maunendra Sankar Desarkar, and Asif Ekbal. 2023. Towards improvement of grounded cross-lingual natural language inference with visio-textual 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, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104.

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.

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.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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].

Aikaterini-Lida Kalouli, Hai Hu, Alexander F. Webb, Lawrence S. Moss, and Valeria de Paiva. 2023. Curing the SICK and Other NLI Maladies. *Computational Linguistics*, 49(1):199–243.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *IEEE/CVF International Conference on Computer Vision*, pages 1224–1234. IEEE.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *ArXiv*, abs/1909.02950.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*. Accepted by TMLR.

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.

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.

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.

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, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *International Conference on Learning Representations (ICLR)*. ICLR 2024 poster.

Elena Pitta, Tom Kouwenhoven, and Tessa Verhoef. 2025. Probing vision-language understanding through the visual entailment task: promises and pitfalls. In *Proceedings of the 2nd LUHME Workshop*, pages 74–83, Bologna, Italy.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*. ICLR 2024 spotlight.

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. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

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.

Hoa Trong Vu, Claudio Greco, Aliia 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.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

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.

Mădălina Zgreabăn, Tejaswini Deoskar, and Lasha Abzianidze. 2025. Merge: Minimal expression-replacement generalization test for natural language inference. *arXiv preprint arXiv:2510.24295*.

Yiyang Zhou, Haoqin Tu, Zijun Wang, Zeyu Wang, Niklas Muennighoff, Fan Nie, Chaorui Deng, Shen Yan, Haoqi Fan, Yejin Choi, and 1 others. 2025. When visualizing is the first step to reasoning: Mira, a benchmark for visual chain-of-thought. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

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 Prompts

Our final prompts result from systematic varying of the initial prompt across several dimensions. We picked the combination yielding the highest accuracy overall. The dimensions are as follows:

– terms 'premise' and 'hypothesis' kept/replaced with 'statement';

– terms 'entailment', 'contradiction', 'neither'

kept/replaced with 'accurate', 'contradicting', 'neither';
– provided/omitted explanations of NLI classes;
– provided/omitted detailed explanations of the Neutral class.

A.1 Vision-language prompt

Question: With respect to the objects in the image, is the statement in square brackets a) entailment, b) contradiction, c) neutral?
Answer 'entailment' if the statements accurately describe the objects in the image.
Answer 'contradiction' if the statement contradicts the image.
Answer 'neutral' if the statement isn't a contradiction, but adds unverifiable details (hidden/off-screen facts, motifs, intentions, identities/relationships etc.).
Put your answers into angled brackets.
Statement 1: []
Statement 2: []
Statement 3: []
Answer 1 (entailment/contradiction/neutral): <...>
Answer 2 (entailment/contradiction/neutral): <...>
Answer 3 (entailment/contradiction/neutral): <...>

A.2 Language prompt

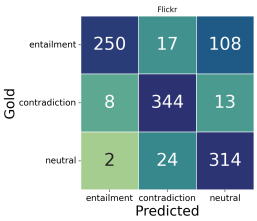
Question: With respect to the premise, is the statement in square brackets a) entailment, b) contradiction, c) neutral?
Answer 'entailment' if the statements accurately describe the premise.
Answer 'contradiction' if the statement contradicts the premise.
Answer 'neutral' if the statement isn't a contradiction, but adds unverifiable details (hidden/off-screen facts, motifs, intentions, identities/relationships etc.).
Put your answers into angled brackets.
Statement 1: []
Statement 2: []
Statement 3: []
Premise:
Answer 1 (entailment/contradiction/neutral): <...>
Answer 2 (entailment/contradiction/neutral): <...>
Answer 3 (entailment/contradiction/neutral): <...>
(CoT-only:) Think step by step.

A.3 Text-to-image

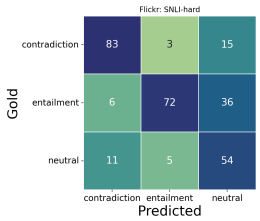
For text-to-image generation, we use no special prompt and supply the premise directly. The only exception is DALL-E 3, for which we use the officially recommended prompt (Betker et al., 2024).

I NEED to test how the tool works with extremely simple prompts. DO NOT add any detail, DO NOT alter the prompt, just use it AS-IS: .

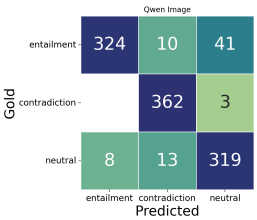
B Confusion Matrices



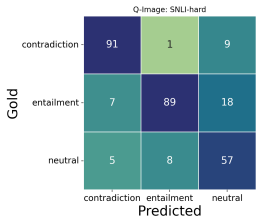
(a) Flickr + Qwen2.5-VL-72B-Instruct: *easy*



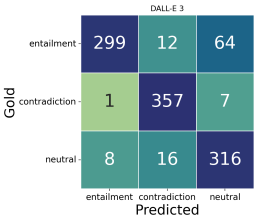
(b) Flickr + Qwen2.5-VL-72B-Instruct: *hard*



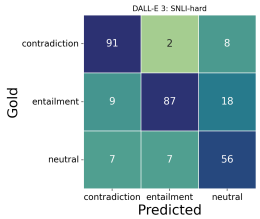
(a) Qwen-Image + Qwen2.5-VL-72B-Instruct: *easy*



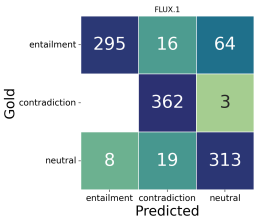
(b) Qwen-Image + Qwen2.5-VL-72B-Instruct: *hard*



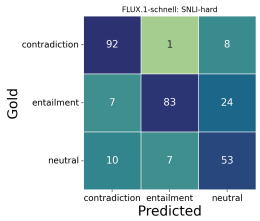
(a) DALL-E 3 + Qwen2.5-VL-72B-Instruct: *easy*



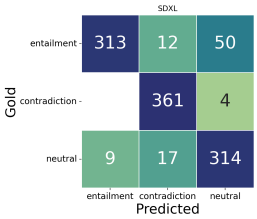
(b) DALL-E 3 + Qwen2.5-VL-72B-Instruct: *hard*



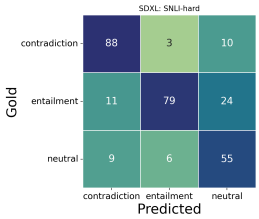
(a) FLUX.1-schnell + Qwen2.5-VL-72B-Instruct: *easy*



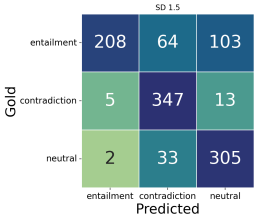
(b) FLUX.1-schnell + Qwen2.5-VL-72B-Instruct: *hard*



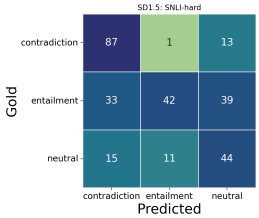
(a) SDXL + Qwen2.5-VL-72B-Instruct: *easy*



(b) SDXL + Qwen2.5-VL-72B-Instruct: *hard*



(a) SD1.5 + Qwen2.5-VL-72B-Instruct: *easy*



(b) SD1.5 + Qwen2.5-VL-72B-Instruct: *hard*

C Misclassified neutrals

Method	Agr. cases
Qwen-Image	8/10 (80%)
DALL-E 3	10/11 (90%)
FLUX.1-schnell	8/11 (73%)
SDXL	11/13 (85%)

Table 3: Proportion of neutral instances misclassified in all 3 inference runs, where TTI+VL pipeline agrees with 1+ SNLI annotators. We consider these instances diagnostic of common sense inference.

D Davidsonian Scene Graphs

TTI evaluation with DSGs was proposed in Cho et al. (2024). The pipeline consists of three steps: 1) breaking the source text down into atomic statements (DSG tuples); 2) rewriting the tuples as questions; 3) answering the questions based on the generated image. Steps 1 and 2 can be performed with a text-only LLM, while 3 requires a VQA model. The answers can then be manually verified. To ensure that easy and hard statements are equally represented, we compute the DSG score on a balanced set of 95 hard and 92 easy premises. Below, we provide the used prompts (same as in the original paper) as well as an evaluation example.

Tuple generation: Task: given input prompts, describe each scene with skill-specific tuples. Do not generate same tuples again. Do not generate tuples that are not explicitly described in the prompts.
output format: id | tuple

Question generation: Task: given input prompts and skill-specific tuples, re-write tuple each in natural language question.
output format: id | question

Question answering: Task: Answer all questions by number with 'yes' or 'no' based on the image; reply using tuple syntax. e.g., [(1, yes), (2, no)].

Example: Text: A dog jumps over the pole.
Tuples:
1 | entity - whole (dog)
2 | entity - whole (pole)
3 | action - (dog, jump)
4 | relation - spatial (dog, pole, over)
Questions:
1 | Is there a dog?
2 | Is there a pole?
3 | Is the dog jumping?
4 | Is the dog jumping over the pole?

Figure 7: Image: FLUX.1-schnell

Answers:
[(1, yes), (2, yes), (3, yes), (4, yes)] → score=1.0

E Generation comparison

Premise: A dog jumps over the pole.

(a) Original Flickr image

(b) Qwen Image

(a) FLUX.1-schnell

(b) SDXL

(a) DALL-E 3

(b) SD 1.5

Premise: Tho babies, one in red, asleep in their highchairs.

(a) Original Flickr image

(b) Qwen Image

(a) FLUX.1-schnell

(b) SDXL

(a) DALL-E 3

(b) SD 1.5