

# Through the Looking Glass: Common Sense Consistency Evaluation of Weird Images

Elisei Rykov<sup>1</sup> Kseniia Petrushina<sup>1,4</sup> Kseniia Titova<sup>1,3</sup>  
Anton Razzhigaev<sup>2</sup> Alexander Panchenko<sup>1,2</sup> Vasily Konovalov<sup>2,4</sup>  
<sup>1</sup>Skoltech <sup>2</sup>AIRI <sup>3</sup>MTS AI  
<sup>4</sup>Moscow Institute of Physics and Technology  
{Elisei.Rykov, Kseniia.Petrushina, A.Panchenko}@skol.tech

## Abstract

Measuring how real images look is a complex task in artificial intelligence research. For example, an image of a boy with a vacuum cleaner in a desert violates common sense. We introduce a novel method, which we call Through the Looking Glass (TLG), to assess image common sense consistency using Large Vision-Language Models (LVLMs) and Transformer-based encoder. By leveraging LVLMs to extract atomic facts from these images, we obtain a mix of accurate facts. We proceed by fine-tuning a compact attention-pooling classifier over encoded atomic facts. Our TLG has achieved a new state-of-the-art performance on the WHOOPS! and WEIRD datasets while leveraging a compact fine-tuning component.<sup>1</sup>

## 1 Introduction

People quickly notice something unusual in images that defy common sense, like Einstein holding a smartphone. We find it odd even though each part seems normal. Our brain’s ability to understand normality goes beyond just identifying objects (Zellers et al., 2019). It involves connecting visual cues with everyday knowledge.

In this work, we propose a visual commonsense model that utilizes the observation that LVLMs may generate contradictory facts when confronted with images defying common sense (Liu et al., 2024b). By leveraging LVLMs to extract atomic facts from these images, we obtain a mix of accurate facts and erroneous hallucinations. Then we fine-tune a compact attention-pooling model over encoded atomic facts.

Our results indicate that using the classifier for basic facts can efficiently spot strange images. Surprisingly, this method outperforms existing more complex techniques.

<sup>1</sup><https://github.com/s-nlp/through-the-looking-glass>

In addition, we introduce a synthesized WEIRD dataset, a dataset of 824 samples of normal and strange images. Using this dataset, we further confirmed the performance of our model.

Our contributions are as follows:

- We present a *new method* called TLG that achieved state-of-the-art performance on the existing dataset of normal and strange images WHOOPS!.
- We present a *new dataset* dubbed WEIRD which is more challenging and nearly four times larger than WHOOPS!.

## 2 Related Work

Recently, commonsense reasoning has attracted substantial interest from the research community, spanning disciplines within NLP and CV, with numerous tasks being introduced.

Guetta et al. (2023) introduced the WHOOPS! benchmark, comprised of purposefully commonsense-defying images created by designers using publicly available image generation tools like Midjourney. They used a supervised approach based on BLIP-2 Flan-T5 (Li et al., 2023a) on multiple scales. The proposed fine-tuned model managed to outperform a random baseline, but still falls significantly short of human performance.

LLMs are capable of producing highly fluent responses to a wide range of user prompts, but they are notorious for hallucinating and making non-factual statements. Manakul et al. (2023b) proposed SelfCheckGPT, a straightforward sampling-based method that enables fact-checking of black-box models with zero resources.

To assess consistency among multiple sampled responses, SelfCheckGPT utilizes several techniques, including BERTScore, an automatic multiple-choice question answering generation (MQAG) framework (Manakul et al., 2023a), and

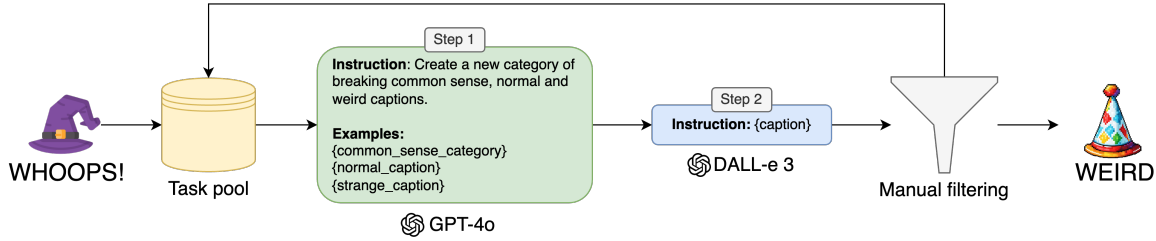


Figure 1: WEIRD dataset generation process. First, we formed a task pool for the few-shot generation of new samples from the WHOOPS! benchmark. Next, we randomly sampled few-shots from the task pool and asked GPT-4o to generate new samples. The samples were then visualized using Dall-E 3 and manually filtered. Good samples were added to the task pool for the next few-shot sampling.

NLI contradiction scores to detect hallucinations in the generated responses. However, the most effective method found was prompting the LLM to verify if the generations are supported by the context or not.

Regarding multi-modal case, [Jing et al. \(2023\)](#) proposed FAITHSCORE, a reference-free and fine-grained evaluation metric that measures the faithfulness of the generated free-form answers from large vision-language models. The FAITHSCORE uses multistep approach: (1) identify the descriptive content, (2) extract corresponding atomic facts from the identified sentences, and (3) the faithfulness of all atomic facts is verified according to the input image by applying Visual Entailment Model (VEM), which is able to predict whether the image semantically entails the text. Analogously, NLI has been used in textual mode to verify premises and hypotheses and subsequently to detect hallucinations ([Maksimov et al., 2024](#)).

[Rykov et al. \(2025\)](#) proposed an approach, in which LVLM is used to first generate atomic facts from images, resulting in a combination of accurate facts and erroneous hallucinations. The next step involves calculating pairwise entailment scores among these facts and aggregating these values to produce a single reality score.

Our approach is similar to the preceding methods, as we also utilize LVLMs to extract atomic facts from the image. We then train a supervised model to learn the relationships between the derived facts. If the classifier identifies a high contradiction among atomic facts, it indicates that one of the generated atomic facts is likely a hallucination. This often occurs when the LVLMs encounter an unusual image ([Liu et al., 2024b](#)), leading to such inconsistencies in most cases.

	WHOOPS!	WEIRD
# of samples	204	824
# of categories	26	12
# of sub-categories	—	181
Human baseline	92%	82.22%

Table 1: Comparison details between WHOOPS! and WEIRD. WEIRD contains 4 times more samples than WHOOPS!. In addition, WEIRD contains 181 different generated commonsense-breaking categories, which have been grouped into 12 global categories.

### 3 Dataset

This section describes the datasets we used to evaluate our methodology.

#### 3.1 WHOOPS!

To evaluate our methods, we employ the WHOOPS!<sup>2</sup> benchmark, focusing on a subset comprising 102 pairs of weird and normal images. Performance is measured by binary accuracy within this paired dataset, where a random guess would yield 50% accuracy. To assess human performance, three annotators were enlisted to categorize each image as weird or normal, relying on a majority vote for the final determination. Impressively, the human baseline reached 92%, indicating that despite subjectivity, there is a clear consensus on what constitutes weirdness within the specific context of the WHOOPS! benchmark.

#### 3.2 WEIRD

Due to the fact that the WHOOPS! benchmark is relatively small, we generated a larger benchmark for quantifying image realism to validate our methodology – WEIRD<sup>3</sup>.

<sup>2</sup>Weird and Heterogeneous Objects, Phenomena, and Situations

<sup>3</sup>Weird Examples of Images with Real-life Discrepancies

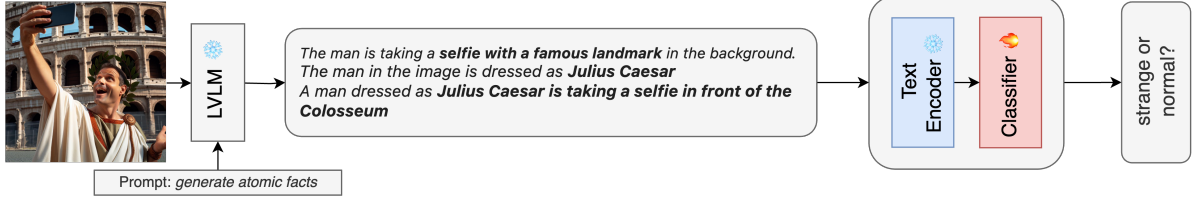


Figure 2: The proposed approach TLG for image commonsense consistency evaluation. Using the LVLm-generated atomic facts about the image, we train a classifier using hidden states from the textual encoder.

The detailed process of WEIRD dataset creation is shown in Figure 1. Like the Self-Instruct (Wang et al., 2023) dataset, WEIRD was generated in an iterative, semi-automatic manner using LLM. Specifically, we used WHOOPS! as an initial task pool with few-shot samples. In each iteration, we randomly sampled 5 pairs of normal and weird situations, along with the commonsense-breaking category. Each few-shot sample contains the breaking commonsense category, a caption of the normal image, and a caption of a strange image. The randomly sampled few-shots were passed to GPT-4o to generate a new category and captions. See the exact prompt used for generation in Appendix I. In the next step, these textual captions were used to generate images with Dall-E 3.

In each iteration, we generated 50 pairs of normal and strange images, resulting in 100 samples after each iteration. We also manually filtered out bad samples. We considered bad samples to be those with inconsistencies between image and caption, or with textual noisy captions. For example, there were many inconsistencies in the captions that mention celebrities. It turned out that Dall-E 3 struggled with the generation of celebrity faces, while some strange captions were based on putting certain celebrities in inappropriate conditions.

In total, we generated 2,000 unique samples of commonsense-breaking situations before the filtering stage. After filtering, only 824 samples remained. To evaluate human performance on WEIRD, we additionally annotated the dataset on the Yandex Tasks<sup>4</sup> crowd-source platform. Each example was annotated by five annotators with overlapping assignments. In order to introduce crowd sources to the task, we added 10 training samples. As a result of the annotation process, Krippendorff’s alpha coefficient of consistency was 0.69 with a human accuracy of 82.22%. WHOOPS! and WEIRD comparison details can be seen in Table 1.

<sup>4</sup><https://tasks.yandex.com>

## 4 Visual Commonsense Evaluation Method using Atomic Fact Extraction

The idea of our method dubbed TLG (Through the Looking Glass) is inspired by FactScore (Min et al., 2023): we adopt the principle of atomic facts generation for trustworthiness verification for the image modality. Namely, the common sense evaluation method is based on the classification of atomic facts generated by LVLms using textual encoders. The approach is depicted in Figure 2.

We use LVLms to collect different atomic facts that describe different aspects of the scene in the image. To sample as many different facts as possible, we use the Diverse Beam Search (Vijayakumar et al., 2016). So, given an image  $I$  and an LVLm, we sample  $N$  facts  $F = \{f_1, f_2, \dots, f_N\}$ , where  $F = \text{LVLm}(I)$ .

Next, we use a frozen textual encoder to extract representations  $H$  of the generated atomic facts. Each fact representation is computed as

$$H_i = \text{Encoder}(f_i) \in \mathbb{R}^{N \times T \times d}, \quad (1)$$

where  $T$  – number of tokens,  $d$  – embeddings dimensionality.

Since each encoder output  $H$  is a set of hidden representations for each token and fact, we perform average pooling to extract a single representation  $V$  for each fact. Thus, using the attention masks  $m$  obtained by the encoder tokenizer and the hidden representations  $H$ , we compute a single fact representation by averaging the vectors of its tokens

$$V_i = \frac{\sum_{j=1}^T m_{ij} H_{ij}}{\sum_{j=1}^T m_{ij} + \varepsilon}. \quad (2)$$

Furthermore, we train an attention-based pooling classifier using individual representations  $V$ . This classifier maps each representation to a single value. Then, we convert a set of attention values into probabilities using the softmax function:

$$A = \text{softmax}(W_a V + b_a) \in \mathbb{R}^N. \quad (3)$$

Later, these scores are used to perform a weighted averaging of the set of representations for each fact into a single representation:

$$v_{\text{weighted}} = \frac{\sum_{i=1}^N A_i V_i}{\sum_{i=1}^N A_i} \in \mathbb{R}^d. \quad (4)$$

Finally, we classify the final representation by mapping it to a single common sense violation probability:

$$\text{prob} = \sigma(W_c v_{\text{weighted}} + b_c) \in [0, 1]. \quad (5)$$

## 5 Experimental Setup

To run the experiments, we strictly follow the evaluation setup suggested in WHOOPS! (Guetta et al., 2023). Thus, we evaluate several models using 5-fold cross-validation in a supervised configuration. See the detailed list of checkpoints used for the main approach and baselines in Appendix E.

For fact generation, we set `num_beams` and `num_beam_groups` to 5, and the `diversity_penalty` to 1.0. Regarding penalty, we find this value to be optimal for adding diversity and preserving the model’s ability to follow instructions. For LVLMS, with various backbone architectures, we utilized the following prompt for fact generation: *“Provide a brief, one-sentence descriptive fact about this image”*. To generate atomic facts, we used different LVLMS with different sizes (from 0.5B to 13B) of the LLaVA architecture. Given the generated atomic facts, we encode them using several DeBERTa-v3-large-based encoders.

We also consider the following baselines:

**LVLMS** with the prompt, which was found to be effective in detecting weird images (Liu et al., 2024a): *“<image> Is this unusual? Please explain briefly with a short sentence.”*

**Linear Probing** resemble our approach in that it requires a small learnable component. This baseline involves learning a logistic regression classifier on the hidden representation of LLaVAs at each layer. We consider two setups: (a) using the `<image>` as the sole input (**Image only**), and (b) using `<image>` with a prompt *“Provide a short, one-sentence descriptive fact about this image”* (**+Prompt**), which was used to generate atomic facts.

**CLIP-based** models were evaluated by passing images and measuring the distance from the strange and normal classes in a zero-shot setting. In addition, we fine-tuned CLIP in a cross-validation setting. More details on the hyperparameters and detailed baseline results can be found in the Appendix C.

**LLM** zero-shot baselines were represented by Gemma-2-9B-Instruct and Qwen2.5-7B-Instruct. As input, we passed generated atomic facts about the image and asked the model to determine whether the facts were strange or not using the following prompt: *“Your task is to classify a series of facts as normal or strange. The set of facts is strange if some of the facts contradict common sense. Answer using ‘normal’ or ‘strange’. Do not write anything else”*.

Furthermore, we used two fine-tuned baselines based on BLIP2 (Li et al., 2023b): BLIP2 FlanT5-XL and BLIP2 FlanT5-XXL that were reported in Guetta et al. (2023).

Moreover, we conducted experiments on knowledge transfer between WEIRD and WHOOPS! for fine-tunable methods to explore the generalization ability to another dataset.

## 6 Results

The results of our experiments on both WHOOPS! and WEIRD datasets are presented in Table 2. The proprietary GPT-4o model has been included as a baseline to illustrate the complexity of benchmarks for proprietary systems and to demonstrate the performance gap between human-generated and proprietary systems. It is not directly comparable to other open-source methods. The results of the linear probing baselines can be found in the Appendix B. For the TLG method and LLM-based baselines, we used facts produced by LLaVA 1.6 Mistral 7B; see the Appendix F for more details. The total number of parameters is calculated as the sum of all parameters in the method. As LLMs and text encoders use pre-generated atomic facts, we report their parameters together with the LVLMS parameters. See also Appendix D for more details of the generated facts.

**TLG** achieves an accuracy of 73.54% on WHOOPS! and 87.57% on WEIRD, demonstrating the state-of-the-art performance both datasets.

**BLIP2 FlanT5 vs. TLG** Next, we compare our best-performing approach to the baselines from Guetta et al. (2023). TLG outperforms the

Method	# Total	Mode	WHOOPS!	WEIRD
Humans	–	–	92.00	82.22
BLIP2 FlanT5-XL	3.94B	fine-tuned	60.00	71.47
BLIP2 FlanT5-XXL	12.4B		73.00	72.31
BLIP2 FlanT5-XXL	12.4B	zero-shot	50.00	63.84
nanoLLaVA Qwen1.5 0.5B	1.05B		66.66	70.90
LLaVA 1.6 Mistral 7B	7.57B		56.86	61.18
LLaVA 1.6 Vicuna 7B	7.06B		65.68	76.54
LLaVA 1.6 Vicuna 13B	13.4B		56.37	58.36
InstructBLIP Vicuna 7B	7B		61.27	69.41
InstructBLIP Vicuna 13B	13B		62.24	66.58
Qwen2.5 7B Instruct	15.18B	zero-shot	67.65	66.46
Gemma2-9B	16.57B		73.04	82.92
LP - LLaVA	13B	fine-tuned	73.50	85.26
CLIP	0.65B	–	60.78	81.57
TLG (Ours)	8B	fine-tuned	<b>73.54</b>	<b>87.57</b>
GPT-4o	–	zero-shot	79.90	81.64

Table 2: The results of different approaches on WHOOPS! and WEIRD datasets. Both benchmarks are balanced and accuracy is the evaluation metric. Fine-tuned methods are displayed at the top, while zero-shot methods are presented in the middle. The best linear probing results for all configurations along with our method are displayed at the bottom.

original fine-tuned approach (BLIP2-FLAN-T5-XXL). This suggests that the task of detecting anomalous images should be tackled by fine-tuning a compact classifier on either textual representations or images, rather than adapting an entire LVLM for this purpose.

**Linear Probing and CLIP vs. TLG** The results of our baselines, which were conducted using Linear Probing and CLIP, are detailed in the Appendices B, C. For the LLaVA models, hidden states of the Vicuna 13B achieved the second-best accuracy on both datasets, with 73.50% on WHOOPS! with prompt and 85.26% on WEIRD in image-only mode. Since WHOOPS! is a smaller dataset, evaluating methods with cross-validation results in high variance, making the ranking of methods less stable. However, the strong performance on WEIRD supports the effectiveness of this approach.

As for the CLIP baseline, OpenAI/CLIP excelled with an accuracy of 60.78% in zero-shot mode for WHOOPS!. On the other hand, on the WEIRD dataset, SigLIP outperformed other models, achieving an accuracy of 81.57% in fine-tuning mode.

**LLM** Qwen2.5-7B-Instruct achieved a relatively high score of 67.65% on WHOOPS! and 66.46% on WEIRD. However, it falls behind Gemma2-9B-Instruct with a score of 73.04% on WHOOPS! and 82.92% on WEIRD. Although LLMs show strong performance, they require more computing resources than TLG.

**GPT-4o** performance illustrates the complexity of the benchmarks for proprietary systems and demonstrates the performance gap between human-generated content and proprietary systems (it should not be directly compared with other open-source methods). The results are rather surprising; GPT-4o outperforms all the methods mentioned here on the WHOOPS! dataset (Guetta et al., 2023). However, it lags significantly behind all the considered baselines and our method on the newly generated WEIRD dataset.

Method	#	Accuracy
<b>WEIRD→WHOOPS!</b>		
BLIP-XL	4B	70.59
BLIP-XXL	12B	72.06
LP (+Prompt)	13B	72.06
LP (Image only)	13B	<b>75.00</b>
TLG (Ours)	8B	<u>74.02</u>
<b>WHOOPS!→WEIRD</b>		
BLIP-XL	4B	72.11
BLIP-XXL	12B	75.06
LP (+Prompt)	13B	74.69
LP (Image only)	13B	<u>79.61</u>
TLG (Ours)	8B	<b>83.05</b>

Table 3: Knowledge transfer between datasets. WEIRD→WHOOPS! means that the approach has been fine-tuned on the WEIRD dataset and tested on the WHOOPS! dataset.



*The child is vacuuming the floor* **0.60**  
*This is a photo of a child vacuuming the floor* **0.12**  
*A child vacuuming a wooden floor* **-0.28**



*The man is using a vacuum cleaner on the beach* **2.38**  
*This image features a man vacuuming the beach* **1.65**  
*The vacuum cleaner is silver* **-0.25**

Figure 3: A pair of images from WHOOPS! with corresponding generated atomic facts. The normal image is on the left, and the unusual image is on the right.

**Knowledge Transfer** To measure the knowledge transfer ability, we fine-tuned a model on one dataset and tested it on another. The results are shown in Table 3.

For WHOOPS!, the linear probing baseline with image-only input on 13B Vicuna backbone with WEIRD calibration outperforms other approaches with an accuracy of 75%. However, the TLG approach with *deberta-v3-large-tasksource-nli* is a second best method with an accuracy of 74.02%. As for WEIRD, TLG trained on WHOOPS! is the best performing approach - 83.05%. Linear probing in image-only mode on 13B Vicuna with a score of 79.61% accuracy. Unlike the previous setting with WEIRD training and WHOOPS! testing, there is a large gap between the best performing approach and the second. This probably indicates that our approach is robust to a small training set, while linear probing requires a larger amount of data for calibration.

**TLG Attention Scores Analysis** Since TLG is based on a learning classifier that includes part of assigning an attention weight to each fact, we interpreted the meaning of these scores. The example of the score distribution for images is shown in Figure 3. In fact, TLG assigns higher attention weights to facts that violate common sense. In this example, the fact “*The vacuum cleaner is silver and purple*” has a lower score than the more inconsistent fact “*The man is using a vacuum cleaner on the beach*”. As a result, TLG gives higher scores to more strange facts, meaning that TLG could also be used as a pure text reality ranker, rating the realism of text facts.

## 7 Conclusion

In this work, we propose a straightforward yet effective approach to visual common sense recognition. Our method exploits an imperfection in LVLMs, causing them to generate hallucinations when presented with unrealistic or strange images. The method entails transitioning to a text modality and addressing the problem from this perspective. Our three-step process involves generating atomic facts, encoding atomic facts with Transformer-based text encoder, and training classifier based on attention-pooling to detect strange images.

Despite the shift in modality, our approach outperforms previous baselines and other supervised methods applied in the image domain, including CLIP-based image encoders and linear probing of LVLMs.

In addition, we developed a methodology to synthesize strange images. Using this methodology, we created WEIRD, a dataset consisting of 824 images that include both strange and normal visuals, which we have made openly available. Surprisingly, our TLG method outperformed the proprietary GPT-4o on our newly generated WEIRD benchmark.

## Limitations

First, we acknowledge that we did not consider all possible open LVLMs that became available recently, such as Qwen2.5-VL. Also, among the proprietary systems, we only evaluated GPT-4o. However, we believe that our choice of both proprietary and open models was representative of the state-of-the-art.

Second, although we tested several prompts for zero-shot baselines and selected the best one, more prompt engineering work could lead to better performance.

## Ethics Statement

We have carefully curated the generated WEIRD dataset, and we have not encountered any inappropriate or offensive content within it.

## References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#). *CoRR*, abs/2309.16609.
- 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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#).
- Nitzan Bitton Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. [Breaking common sense: Whoops! A vision-and-language benchmark of synthetic and compositional images](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2616–2627. IEEE.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. [FAITHSCORE: evaluating hallucinations in large vision-language models](#). *CoRR*, abs/2311.01477.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.
- Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024b. [Phd: A prompted visual hallucination evaluation dataset](#). *CoRR*, abs/2403.11116.
- Ivan Maksimov, Vasily Kononov, and Andrei Gliniskii. 2024. [DeepPavlov at SemEval-2024 task 6: Detection of hallucinations and overgeneration mistakes with an ensemble of transformer-based models](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 274–278, Mexico City, Mexico. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023a. [MQAG: multiple-choice question answering and generation for assessing information consistency in summarization](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 39–53. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023b. [Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*,

- pages 9004–9017. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- 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 and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.
- Elisei Rykov, Kseniia Petrushina, Kseniia Titova, Alexander Panchenko, and Vasily Konovalov. 2025. [Don't fight hallucinations, use them: Estimating image realism using nli over atomic facts](#). *Preprint*, arXiv:2503.15948.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. [LAION-5B: an open large-scale dataset for training next generation image-text models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Damien Sileo. 2024. [tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.
- Gemma Team. 2024. [Gemma](#).
- Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. [Diverse beam search: Decoding diverse solutions from neural sequence models](#). *CoRR*, abs/1610.02424.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khoshdel, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. 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. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [From recognition to cognition: Visual commonsense reasoning](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.



## A Performance on WEIRD with Standard Deviation

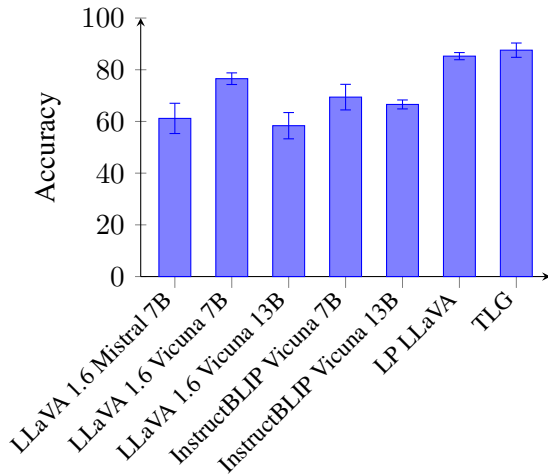


Figure 4: Accuracy with standard deviation for different setups

## B Linear Probing Baseline

We collect hidden states by passing the image with corresponding to the setup (**Image only**, **+Prompt**) prompt through LLaVA decoder. The results are presented in Table 4.

We trained a logistic regression with L2 regularization, with a maximum of 100 iterations and a tolerance of 0.1 on standardized hidden states.

Model	Image only	+Prompt
<b>WHOOPS!</b>		
LLaVA 1.6 Mistral 7B	67.63	67.13
LLaVA 1.6 Vicuna 7B	73.01	72.02
LLaVA 1.6 Vicuna 13B	69.06	<b>73.50</b>
<b>WEIRD</b>		
LLaVA 1.6 Mistral 7B	78.13	81.82
LLaVA 1.6 Vicuna 7B	84.65	83.91
LLaVA 1.6 Vicuna 13B	<b>85.26</b>	84.02

Table 4: Linear probing baseline results on WHOOPS! and WEIRD.

## C CLIP Baseline

We fine-tuned the model for 5 epochs with batch size 1 using AdamW optimizer with learning rate  $1e-3$ . Other hyperparameters are the same as in the HuggingFace trainer.

The detailed results for WHOOPS! and WEIRD are given in Table 5. An interesting result is that SigLIP is more accurate than the standard CLIP-based models of OpenAI and LAION.

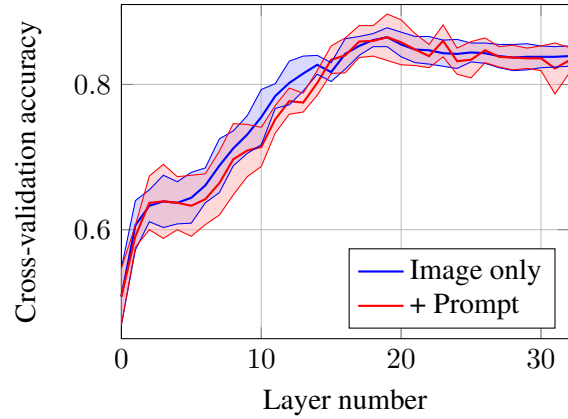


Figure 5: Cross-validation accuracy depending on the LLaVA 1.6 Vicuna 13B index layer for linear probing on the WEIRD dataset. Layers containing the most relevant information are in the middle of the decoder.

Model	#	zero-shot	fine-tuned
<b>WHOOPS!</b>			
OpenAI/CLIP	0.15B	<b>60.78</b>	56.86
Google/SigLIP	0.88B	50.49	73.01
LAION/CLIP	0.43B	53.92	54.39
<b>WEIRD</b>			
OpenAI/CLIP	0.15B	56.15	65.65
Google/SigLIP	0.88B	48.87	<b>81.57</b>
LAION/CLIP	0.43B	57.34	74.86

Table 5: CLIP results on WHOOPS! and WEIRD.

## D Analysis of the Generated Facts

Category	Keywords
common	<i>common</i>
	<i>usual</i>
	<i>normal</i>
	<i>natural</i> <i>real</i>
weird	<i>unusual</i>
	<i>strange</i>
	<i>playful</i>
	<i>creative</i>
	<i>unreal</i> <i>weird</i>
real (as not generated)	<i>real</i>
	<i>realistic</i>
	<i>photo</i>
digital	<i>digital</i>
	<i>generated</i>
	<i>3D</i>
	<i>fantastic</i>
	<i>rendering</i> <i>artistic</i>

Table 6: List of keywords with corresponding categories to analyze generated atomic facts.

LLaVA Backbone	Type	Length	ROUGE	Cosine Similarity	Marker words			
					common	weird	real	digital
<b>WHOOPS!</b>								
Mistral-7B	normal	61.80	45.46	79.65	9	1	33	37
	strange	64.34	46.28	79.57	5	12	19	68
Qwen-0.5B	normal	140.15	45.02	83.19	55	4	20	8
	strange	144.01	45.07	83.36	46	26	17	17
Vicuna-7B	normal	99.57	64.71	88.27	8	0	54	42
	strange	103.63	63.75	87.88	5	4	25	66
Vicuna-13B	normal	86.69	64.24	88.24	8	0	21	37
	strange	92.88	64.64	88.13	4	8	15	58
<b>WEIRD</b>								
Mistral-7B	normal	72.94	52.43	72.94	24	1	95	201
	strange	77.81	51.37	77.81	31	57	79	270
Qwen-0.5B	normal	129.17	54.67	68.46	170	24	35	36
	strange	131.84	54.70	68.40	184	130	24	69
Vicuna-7B	normal	74.39	60.09	68.41	6	1	146	213
	strange	79.35	60.32	68.55	3	16	130	262
Vicuna-13B	normal	67.13	58.04	69.36	10	0	108	242
	strange	69.82	59.08	69.46	3	19	106	291

Table 7: Metrics for generated atomic facts on the WHOOPS! and WEIRD datasets are computed separately for each of the four models, assessing them on both normal and strange images. ROUGE and Cosine Similarity metrics evaluate the similarity of facts derived from a single image, while marker words denote the presence of at least one characteristic marker term in the group of facts. From these results, we can conclude that the facts generated by *llava-v1.6-mistral-7b* are of the finest quality in atomicity — they are the briefest and exhibit the greatest semantic independence.

We measured Cosine Similarity of the generated facts by using *all-MiniLM-L6-v2*<sup>5</sup> embedder. We also calculated ROUGE (Lin, 2004) metric for lexical similarity. We calculate the metric values pairwise for each unique pair of facts and then averaging the results. There is no significant difference in lexical/semantic similarity (as measured by ROUGE and Cosine Similarity) between strange and normal images within the same LLaVA. However, a significant difference can be observed when comparing similarity between different LLaVAs. In Table 7 we provide metrics on generated atomic facts. We noticed that there are several groups of different marker words that all LVLMs tend to generate. Table 6 shows the exact list of marker words for each observed group.

**nanoLLaVA 1.5B** generates significantly different facts from all other LLaVA models in terms of used vocabulary. By analyzing occurring marker words, it becomes evident that nanoLLaVA-1.5 more frequently employs words from the common and weird sets, indicating a greater tendency to comment on the plausibility of images and use evaluative terms. Conversely, it uses words from the real and digital sets less often. The facts of nanoLLaVA-1.5 are significantly longer than others.

**LLaVA 1.6 Mistral 7B vs LLaVA 1.6 Vicuna 7B** The difference between facts generated by these two is quite noticeable. The Mistral-based LLaVA generates the shorter responses, and judging by the ROUGE metric, these responses are less similar to each other. In terms of the atomicity of the generated facts, the facts produced by Mistral can be considered more qualitative. However, the presence of digital markers can be misleading for the model.

**LLaVA 1.6 Vicuna 7B vs 13B** The metrics of both Vicuna-based models are similar; however, the generations from 13B are shorter on average. We also notice that the facts generated for strange images are generally longer than those for truthful ones.

<sup>5</sup><https://hf.co/sentence-transformers/all-MiniLM-L6-v2>

## E Checkpoints

For generating atomic facts we leverage the following LVLMs:

- `llava-v1.6-mistral-7b-hf`: a 7B LVLM with based on a Mistral (Jiang et al., 2023);
- `nanoLLaVA-1.5`: a 2B LVLM based on a Qwen1.5-0.5B (Bai et al., 2023);
- `llava-v1.6-vicuna-7b-hf`: a 7B LVLM based on a Vicuna (Chiang et al., 2023);
- `llava-v1.6-vicuna-13b-hf`: a 13B LVLM based on a Vicuna.

The following encoders were used for our main approach:

- `deberta-v3-large`: an original DeBERTa without fine-tuning;
- `nli-deberta-v3-large`: DeBERTa fine-tuned by Sentence Transformer (Reimers and Gurevych, 2019) on NLI datasets. Specifically, the model was fine-tuned on the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets.
- `deberta-v3-large-tasksource-nli`: a multi-task text encoder based on DeBERTa-v3-large fine-tuned on 600 tasksource tasks, outperforming every publicly available text encoder of comparable size in an external evaluation (Sileo, 2024).

As for the CLIP-based baseline, the following models were utilized:

- `clip-vit-base-patch32`: a pre-trained CLIP model published by OpenAI with 0.15B parameters (Radford et al., 2021).
- `siglip-so400m-patch14-384`: a novel image encoder with 0.88B parameters trained by Google. This encoder inherits the CLIP architecture but features a better loss function (Zhai et al., 2023).
- `CLIP-ViT-L-14-laion2B-s32B-b82K`: a pre-trained CLIP encoder with 0.43B parameters, trained on the LAION-2B dataset (Schuhmann et al., 2022).

For the LLM zero-shot baseline, these LLMs were used:

- `Qwen2.5-7B-Instruct`: a 7B instruction-tuned LLM trained by Qwen (Yang et al., 2024).
- `Gemma-2-9b-it`: a 9B instruction-tuned LLM trained by Google (Team, 2024).

## F TLG Evaluation Details

Detailed results of TLG evaluation are given in Table 8. A distinct pattern emerges: DeBERTa models fine-tuned on the tasksource collection outperform methods that rely on alternative text encoders, largely due to their enhanced encoding capabilities. This superiority can be attributed to extensive fine-tuning on a diverse range of knowledge-intensive tasks sourced from the tasksource repository. Using tasksource DeBERTa, the best performance was achieved with Mistral-7B backbone, while the poorest performance was observed with the smallest Qwen-0.5B model, and Vicuna fell in the middle.

The results, averaged over five folds, for the evaluated text encoders paired with various LLaVAs on both benchmarks are presented in Table 8. The highest performance for both benchmarks was achieved by generating facts using LLaVA 1.6 Mistral 7B in conjunction with *deberta-v3-large-tasksource-nli* as the text encoder. Thus, we used facts produced by LLaVA 1.6 Mistral 7B in our other approaches and baselines.

Text Encoder	LLaVA Backbone			
	Mistral-7B	Vicuna-7B	Vicuna-13B	Qwen-0.5B
<b>WEIRD Cross-Validation</b>				
deberta-v3-large-tasksource-nli	<b>87.57</b>	80.51	81.37	77.11
nli-deberta-v3-large	77.97	74.00	77.11	74.57
deberta-v3-large	59.92	63.86	63.59	63.29
<b>WHOOPS! Cross-Validation</b>				
deberta-v3-large-tasksource-nli	<b>73.54</b>	69.15	64.72	64.68
nli-deberta-v3-large	64.60	63.61	66.59	65.15
deberta-v3-large	49.49	50.48	47.57	53.93

Table 8: The results of our approach with various LVLMs and text encoders for both benchmarks, WHOOPS! and WEIRD, are presented. Accuracy, averaged over five folds, serves as the performance metric. For both benchmarks, LLaVa 1.6 Mistral-7B paired with *deberta-v3-large-tasksource-nli* demonstrates the best outcome. A clear trend emerges: tasksource DeBERTa outperforms all others, partly due to its superior encoding capabilities. This trend is clearer for the WEIRD dataset due to its larger size.

**G Examples of Strange Images From WEIRD**



**H Examples of Normal Images From WEIRD**



## I Prompt for WEIRD Samples Generation Using GPT-4o

---

Your task is to generate a new COMMONSENSE\_CATEGORY, EXPLANATION, NORMAL\_CAPTION, STRANGE\_CAPTION using the presented ones from the EXAMPLES.

COMMONSENSE\_CATEGORY is the category of common sense disturbance, so follow this information when creating your own captions, as they must disturb common sense in the same category.

Use presented COMMONSENSE\_CATEGORIES only as an example, because your task is to generate a new one.

After generating a new COMMONSENSE\_CATEGORY, generate 1 new pair based on this category.

Each pair should start with EXPLANATION. EXPLANATION is a description of an inconsistent situation.

You should create EXPLANATION first.

Next, based on EXPLANATION, generate NORMAL\_CAPTION and a STRANGE\_CAPTION.

NORMAL\_CAPTION describes an image that is suitable for common sense, it does not contradict facts about the world, etc.

On the other hand, STRANGE\_CAPTION contradicts common sense. Also, captions can represent past time, so a caption about something that happened a long time ago is not strange.

Do not generate something that is too hard to understand or imagine.

Make the captions as specific and descriptive as possible. Describe all the details.

Generate only 1 pair of EXPLANATION, NORMAL\_CAPTION and a STRANGE\_CAPTION.

EXAMPLES:

COMMONSENSE\_CATEGORY: Tool Misapplication

EXPLANATION: A whisk is a kitchen tool specifically designed for mixing ingredients together smoothly or incorporating air into a mixture, such as when making whipped cream or beating eggs. Its structure, consisting of multiple loops of wire, is not intended for hammering nails into wood. Using a whisk to hammer nails is not only ineffective but is likely to damage the whisk and offer no benefit, as its delicate wires are neither strong nor solid enough to drive nails.

NORMAL\_CAPTION: A whisk being used to beat eggs in a bowl

STRANGE\_CAPTION: A whisk being used to hammer nails into a wooden plank

COMMONSENSE\_CATEGORY: Impossible interaction

EXPLANATION: Cats are known for their playful and curious nature, but they do not have the physical ability to solve complex math problems, as they lack the understanding and cognitive functions necessary for such tasks.

NORMAL\_CAPTION: a cat playing with a ball of yarn on the floor

STRANGE\_CAPTION: A cat solving a complex math equation on a blackboard.

COMMONSENSE\_CATEGORY: Untypical behavior

EXPLANATION: Octopuses are sea creatures that live underwater and are adapted to life in the ocean. However, seeing an octopus wearing clothes, something made specifically for humans to provide warmth and protection, is highly unusual and outside the realms of normal behavior or biological needs.

NORMAL\_CAPTION: An octopus swimming in the ocean.

STRANGE\_CAPTION: An octopus wearing a suit and tie.

COMMONSENSE\_CATEGORY: Inappropriate Object Utility

EXPLANATION: Hairdryers are designed to dry hair by blowing warm air. Using a hairdryer to open a locked door is incorrect and impractical, as hairdryers do not have the functionality or mechanism to open locks.

NORMAL\_CAPTION: A person drying their hair with a hairdryer in front of a mirror.

STRANGE\_CAPTION: A person using a hairdryer to open a locked door.

---

Figure 6: Example of prompt used for synthetic samples generation for WEIRD benchmark. In total, 5 random categories from the task pool were taken on each step of generation. The model is expected to generate a new common sense category, a new explanation and a pair of caption. Further, captions are used for image generation.