# Quantifying image realism via language model reasoning

**Kseniia Petrushina**
MIPT
Skoltech
petrushina.ke@phystech.edu

## Abstract

Quantifying the realism of images remains a challenging problem in the field of artificial intelligence. We introduce a novel method to assess image realism using language models and natural language inference. Our approach involves extracting atomic facts from images via multimodal instruct models, computing pairwise entailment scores between these facts, and aggregating these scores to derive a single reality score. This method identifies contradictions within the image, indicating the presence of unusual or implausible scenarios. Unlike traditional fact-checking or deep-fake detection techniques, our focus is on the *weirdness* or impossibility of the depicted scenes, rather than determining the authenticity of the content. Applying this method to a benchmark dataset, we demonstrate its effectiveness in capturing the degree of realism and providing explanatory insights into the nature of visual contradictions. This work advances the understanding of visual realism and commonsense reasoning in AI.

## 1 Introduction

The last decade has witnessed significant advancements in the field of artificial intelligence, particularly in the generation of realistic data across various modalities such as images, text, audio, and video. Thus, with the improvement of generative models, the problem of recognizing realistic images arises.

Fact-checking is a critical task in the era of information overload and rampant misinformation (Yao et al., 2023). The goal of fact-checking is to verify the truthfulness of claims by retrieving and analyzing relevant evidence from various sources. This task is essential for maintaining the integrity of information disseminated through media and social networks. Accurate fact-checking helps prevent the spread of false information, which can have significant social, political, and economic consequences.

In parallel to fact-checking, the detection of deep-fakes and image manipulations has become increasingly important (Zanardelli et al., 2022). The prevalence of easy-to-use image editing tools has led to a surge in the production and dissemination of fake and altered images. These manipulations can range from simple copy-move and splicing attacks to sophisticated deep-fakes that are often indistinguishable from genuine content.

But there is another side to the realism of images, the human perception of the surrounding world and its laws. Quantifying the realism of images involves more than just detecting forgeries or verifying authenticity; it encompasses assessing how closely an image aligns with real-world expectations. The intrinsic challenges arise in designing functions that can reliably differentiate between realistic and unrealistic data (Theis, 2024). L. Theis argues that a good generative model alone is insufficient to solve this problem. Instead, he proposes the concept of a universal critic — a theoretical ideal that can serve as a guide for practical implementations and a tool for analyzing current attempts to measure realism. And there is the need to consider not just the visual fidelity of images but also the contextual and commonsense coherence. An image might appear visually convincing but still fail to align with our understanding of how the world works, thereby diminishing its realism.

Building on the concept of image realism, the challenge extends to evaluating whether the content of an image adheres to commonsense expectations. This can be done on a dataset specifically designed to test visual commonsense (Bitton-Guetta et al., 2023). The WHOOPS! dataset includes images that deliberately defy common sense, such as famous soccer players playing chess instead of competing on a football field. Humans can easily recognize and interpret these unconventional images, but AI models often struggle with this task. This dataset evaluates AI's ability to understand and explain

why certain images are unusual. Tasks include image captioning, cross-modal matching, visual question answering, and explanation generation, where models must identify and articulate why an image defies commonsense expectations.

Necessity of higher-order cognition in visual understanding can be also explained in (Zellers et al., 2018). This study formalizes the task of Visual Commonsense Reasoning, requiring machines not only to answer challenging questions about images but also to provide justifications for their answers.

In the field of NLP, research has been conducted on the numerical evaluation of factual precision in generated long texts (Min et al., 2023). The metric is obtained as follows: the generated statement was divided into many atomic facts, which were checked for compliance with some database and the proportion of reliable facts was calculated.

While traditional fact-checking and deep-fake detection focus on the authenticity of content, our work explores more the concept of *weirdness* in images. The objective is not merely to determine whether an image is real or generated but to assess how plausible the depicted scenario is within the context of the real world. This involves identifying and analyzing contradictions within the image that indicate unusual or impossible scenes. We employ a strategy of selecting atomic facts, but for common sense research we do not compare them with an external source and find contradictions between them.

## 2 Problem statement

Given unknown *real* and *weird* probability distributions

$$\mathrm{P}_{\text{real}}(\mathbf{x}) : \mathbb{R}^{n\times n} \to [0,1] \quad \mathrm{P}_{\text{weird}}(\mathbf{x}) : \mathbb{R}^{n\times n} \to [0,1]$$

and samples from the distributions

$$\mathcal{D}_{\text{r}} = \{\mathbf{x}_{\text{r}}^i \mid \mathbf{x}_{\text{r}}^i \sim \mathrm{P}_{\text{real}}(\mathbf{x})\}_{i=1}^N$$
$$\mathcal{D}_{\text{w}} = \{\mathbf{x}_{\text{w}}^i \mid \mathbf{x}_{\text{w}}^i \sim \mathrm{P}_{\text{weird}}(\mathbf{x})\}_{i=1}^N.$$

We need to find a *reality-check* function

$$\mathrm{f}_{\text{weird}} : \mathbb{R}^{n\times n} \to \mathbb{R}_+$$

that defines the realism score, that is for *real* image $\mathbf{x}_{\text{r}}$ and *weird* image $\mathbf{x}_{\text{w}}$, provided that they are close in a sense of similarity measure $\langle\cdot,\cdot\rangle$:

$$\langle \mathbf{x}_{\text{r}}, \mathbf{x}_{\text{w}} \rangle \le \varepsilon,$$

the following holds true

$$\mathrm{f}_{\text{weird}}(\mathbf{x}_{\text{r}}) < \mathrm{f}_{\text{weird}}(\mathbf{x}_{\text{w}}).$$

## 3 Proposed method

We divide *reality-check* function f computation into three steps:

### 3.1 Extracting atomic facts

Using multi-modal model $\mathrm{f}_{\text{cap}} : \mathbb{R}^{n\times n} \to \mathrm{T}^{m\times L}$ we obtain $m$ sequences of language tokens of length $L$, which describe the details about the image:

$$\mathrm{f}_{\text{cap}}(\mathbf{x}) = \mathrm{F}_{\text{A}} = \{[t_1^i, \ldots, t_L^i] \mid i \in \overline{1,m}\}$$

### 3.2 Pairwise natural language inference

For each ordered pair of facts $(f_i, f_j) \in \mathrm{F}_{\text{A}} \times \mathrm{F}_{\text{A}}$ we calculate entailment score via $\mathrm{f}_{\text{nli}} : \mathrm{T}^L \times \mathrm{T}^L \to [-1,1]$. The results are presented in the form of a matrix

$$S_{ij} = \mathrm{f}_{\text{nli}}(f_i, f_j).$$

### 3.3 Aggregating pairwise scores

We take the sum of matrix elements, if both pairs $(f_i, f_j)$ and $(f_j, f_i)$ are contradictory and average it by the number of pairs:

$$\mathrm{f}_{\text{agg}}(S) = -\frac{1}{m^2}\sum_{i<j}(S_{ij} + S_{ji})\mathbb{I}[S_{ij}, S_{ji} < 0]$$

### 3.4 Resulting metric

So, the final formula is

$$\mathrm{f}_{\text{weird}} = \mathrm{f}_{\text{agg}} \circ \mathrm{f}_{\text{nli}} \circ \mathrm{f}_{\text{cap}}$$

**Hypothesis 1** *Resulting reality scores* $R = \{f_{weird}(\boldsymbol{x})\}$ *will correlate with probability densities* $P = \{\mathrm{P}_{real}(\boldsymbol{x})\}$*:*

$$r_s = \rho_{\mathrm{R}(f_{weird}(\boldsymbol{x})),\mathrm{R}(\mathrm{P}_{real}(\boldsymbol{x}))} \ge 0.5$$

## 4 Computational experiments

### 4.1 Dataset

Dataset consists of 102 paired *real* and *weird* images from WHOOPS! benchmark. Examples are presented in 1.

### 4.2 Models

We tested four models for obtaining atomic facts from the images: LLaVa (Liu et al., 2023), BLIP (Li et al., 2022), GPT-2 (Radford et al., 2019), GPT-4o (OpenAI, 2023).

For solving the natural language inference task we chose sileod (Sileo, 2022), MoritzLaurer (Laurer et al., 2022) and t5-true (Honovich et al., 2022).

Figure 1: Examples of *real* and *weird* images.

### 4.3 Results

The comparison results of applying various models from the previous section are presented in Table 1.

Reality scores for the whole dataset obtained using *LLaVa* model for captioning and *sileod* model for contradiction detection are presented in Figure 2. Kolmogorov–Smirnov test shows that the distributions are different with one-sided alternative with p-value 4e-5.

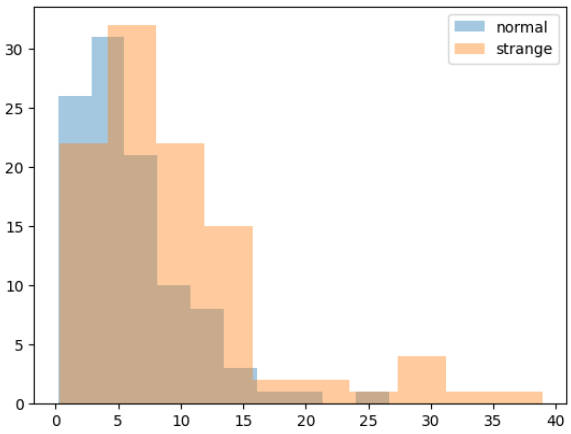


Figure 2: Scores distribution.

| | sileod | MoritzLaurer | t5-true |
|---|---|---|---|
| LLaVa | **0.68** | 0.42 | 0.63 |
| BLIP | 0.53 | **0.68** | 0.53 |
| GPT-2 | 0.37 | 0.32 | 0.37 |
| GPT-4o | 0.63 | **0.68** | 0.37 |

Table 1: Accuracy of various methods.

## 5 Conclusion

In conclusion, the proposed metric for quantifying image realism effectively differentiates between real and weird images by assessing their contextual and commonsense coherence. This approach not only advances the detection of image manipulations and deep-fakes but also enhances AI's ability to align with human perception, ensuring more reliable and nuanced evaluations of visual content. By focusing on how believable the scenarios in images are, we create smarter AI systems that can better understand and interpret the real world.

## References

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. *arXiv preprint*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep

Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

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, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Sewon Min, Kalpesh Krishna, Xinxi 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.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Damien Sileo. 2022. tasknet, multitask interface between Trainer and datasets.

L. Theis. 2024. What makes an image realistic? In *Proceedings of the 41st International Conference on Machine Learning*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Marcello Zanardelli, Fabrizio Guerrini, Riccardo Leonardi, and Nicola Adami. 2022. Image forgery detection: a survey of recent deep-learning approaches. *Multimedia Tools Appl.*, 82(12):17521–17566.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724.