Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки / специальность: 09.04.01 Информатика и вычислительная техника
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

ДЕТЕКЦИЯ МУЛЬТИМОДАЛЬНЫХ ГАЛЛЮЦИНАЦИЙ НА ОСНОВЕ ВНУТРЕННИХ ПРЕДСТАВЛЕНИЙ И АКТИВАЦИЙ МОДЕЛЕЙ

(магистерская диссертация)

Студент:
Петрушина Ксения Евгеньевна

_____
(подпись студента)

Научный руководитель:
Панченко Александр Иванович, PhD

_____
(подпись научного руководителя)

Консультант *(при наличии)*:
_____
_____
(подпись консультанта)

Москва 2025

Детекция мультимодальных галлюцинаций на основе внутренних представлений и активаций моделей

Петрушина Ксения

Представлено в Московский физико-технический институт
Июнь 18

Аннотация

Данная работа посвящена обнаружению галлюцинаций в мультимодальных моделях при обработке необычных или нарушающих здравый смысл изображений. Мы представляем методы, использующие несовершенства больших визуально-языковых моделей, которые склонны к генерации недостоверных описаний в ответ на визуальные аномалии. Основной подход заключается в применении логического вывода в естественном языке (NLI) к атомарным фактам, извлечённым из описаний изображений, где противоречия служат индикатором визуальной нереалистичности. Дополнительно мы исследуем линейный пробинг скрытых представлений моделей с целью различения реалистичных и нереалистичных изображений. Предлагаемые методы оцениваются на выборках WHOOPS! и WEIRD, а также сравниваются с обучаемым методом, использующим агрегирование представлений фактов с весами, соответствующими их важности в механизме внимания. Полученные результаты показывают, что склонность LVLMs к галлюцинациям может служить полезным свойством для оценки реалистичности изображений, что способствует созданию интерпретируемых инструментов для диагностики мультимодальных галлюцинаций.

Detection of Multimodal Hallucinations Based on Internal Representations and Activations of Models

Kseniia Petrushina

Submitted to the Moscow Institute of Physics and Technology on June 18, 2025

ABSTRACT

The focus of this work is on the detection of hallucinations in multimodal models in presence of unusual or commonsense-defying images. We propose methods that leverage the imperfections of large vision-language models (LVLMs), which tend to generate hallucinations when presented with images that violate everyday knowledge. Our primary method applies natural language inference (NLI) to atomic facts extracted from image descriptions, identifying internal contradictions as a signal of visual abnormality. We further explore linear probing techniques over hidden representations of LVLMs, evaluating their capacity to distinguish between realistic and unrealistic images. Both methods are evaluated on benchmarks such as WHOOPS! and WEIRD. Additionally, we consider a learning-based approach that generalizes contradiction modeling via attention pooling over fact representations. The results demonstrate that hallucination-prone behavior of LVLMs, when carefully analyzed, can serve as a valuable cue for identifying images that lack realism. This work contributes toward developing interpretable tools for multimodal hallucination detection and realism estimation.

Keywords: LVLM, NLI, Probing, WHOOPS!

Research advisor:
Name: Alexander Panchenko
Degree, title: DSc in Computer Science, Associate professor

Contents

Chapter 1
Introduction

Relevance

Measuring how real images look is a complex and important task in artificial intelligence. Despite significant progress in generative modeling, systems still struggle with realism, often producing visually plausible but semantically nonsensical scenes - for example, a child vacuuming in the desert or Einstein using a smartphone. While such images may appear coherent in visual form, they contradict basic commonsense knowledge. Humans detect these inconsistencies effortlessly, yet replicating this ability in artificial systems remains a challenge [18].

Recent works highlight the difficulties that large vision-language models (LVLMs) face when encountering visually abnormal content. These models, such as LLaVA [6], tend to generate hallucinations - textual outputs that misrepresent the visual input - when confronted with images that defy commonsense. Interestingly, these hallucinations can be used as a signal to assess image realism [11].

Benchmarks such as WHOOPS! [1] and WEIRD [12] have been proposed to evaluate model performance on this task. Despite the capabilities of existing multimodal models, their accuracy remains well below human-level performance. This underscores the need for new approaches that integrate commonsense reasoning into realism detection.

Main purpose of the research

The aim of this thesis is to develop and evaluate methods for detecting hallucinations in multimodal systems by analyzing both the textual outputs and internal representations of LVLMs. We propose that hallucinations produced by these models, when describing images that contradict real-world logic, can be repurposed as features for detecting visual inconsistency. The work is focused on two main strategies: the use of natural language inference (NLI) over LVLM-generated atomic facts, and the application of linear probing to the hidden states of LVLMs. Additionally, we consider a supervised method based on attention-pooling classifiers that generalize the NLI pipeline by learning to model relationships between facts.

Scientific novelty

The novelty of this work lies in treating hallucinations as a resource rather than a flaw. The first method introduced in this thesis employs a zero-shot approach that generates atomic facts using LVLMs and applies pairwise NLI to identify contradictions among them. This contradiction signal is aggregated into a single realism score [11].

In the second part, we evaluate a lightweight linear probing classifier that operates on the internal representations of LVLMs, demonstrating that the model's hidden states contain sufficient information to detect visual abnormality [12].

Finally, a supervised model named Through the Looking Glass (TLG) is considered. It generalizes the NLI approach by learning fact-to-fact relationships using attention mechanisms. This method achieves state-of-the-art performance on both WHOOPS! and WEIRD datasets [12].

Statements for defense

- LVLMs generate hallucinations when presented with counter-commonsense images, and these hallucinations can be detected via contradiction analysis over atomic facts.
- The NLI-based aggregation of contradiction scores between facts yields a zero-shot method that outperforms strong zero-shot and instruction-tuned baselines [11].
- Hidden states of LVLMs contain signal relevant for detecting image realism; linear probing on these states can separate realistic and strange images without tuning the model [12].
- The supervised TLG classifier, based on attention-pooling over fact embeddings, achieves superior performance and generalizes to newly generated datasets.

Chapter 2
Author Contribution

The research presented in this thesis is based on two main publications. My contributions to each of them are detailed below.

- I was responsible for the core idea, implementation, and experiments for the NLI-based method of detecting unusual images. I implemented the atomic fact generation pipeline using LLaVA, designed the entailment scoring and aggregation methods, and conducted extensive evaluation on the WHOOPS! dataset. This work was carried out in collaboration with Elisei Rykov and resulted in a paper in press titled ”Don’t Fight Hallucinations, Use Them: Estimating Image Realism using NLI over Atomic Facts” [1].
- I implemented and evaluated the linear probing method, performed an analysis of hidden states across LVLM layers and cross-dataset generalization. Moreover, I ran and analyzed zero-shot LVLM baselines using instruction-tuned models such as InstructBLIP and LLaVA across the WHOOPS! and WEIRD datasets. I also participated in discussions regarding experimental design and contributed to the writing of the experimental section of the paper ”Through the Looking Glass: Common Sense Consistency Evaluation of Weird Images” [2].

[1] https://arxiv.org/abs/2503.15948
[2] https://aclanthology.org/2025.naacl-srw.28

Chapter 3
List of publications

1. 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. The 4th Workshop on Multimodal Fact Checking and Hate Speech Detection co-located with the 39th Annual AAAI Conference on Artificial Intelligence. Philadelphia, PA, USA.

2. Elisei Rykov, Kseniia Petrushina, Kseniia Titova, Anton Razzhigaev, Alexander Panchenko, and Vasily Konovalov. 2025. Through the Looking Glass: Common Sense Consistency Evaluation of Weird Images. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 279–293, Albuquerque, USA. Association for Computational Linguistics.

Chapter 4
Literature Review

Recently, commonsense reasoning has attracted substantial attention across both natural language processing (NLP) and computer vision (CV), leading to the development of numerous benchmark datasets and evaluation frameworks. In this chapter, we provide an overview of existing work related to commonsense reasoning and hallucination detection in both vision-language and text-only models. We begin by discussing the conceptual challenges of defining and evaluating realism in multimodal outputs. Then, we survey benchmark datasets, such as WHOOPS! and PhD, that expose limitations in current models' ability to reason about visual commonsense.

Theoretical Motivation for Realism Evaluation

Theoretical basis for why quantifying realism is challenging is provided in [15]. The author suggests considering not just the visual fidelity of images, but also the contextual and common-sense coherence. An image might appear visually convincing but still fail to align with our understanding of how the world works, thereby diminishing its realism.

Benchmarks for Visual Commonsense Violations

To systematically evaluate such cases, several datasets have been proposed. [1] introduced the WHOOPS! benchmark, a collection of deliberately commonsense-violating images crafted using generative tools like Midjourney. Their approach involved fine-tuning BLIP-2 Flan-T5 [3] using supervised learning. Although their model outperformed a random baseline, it still lagged significantly behind human performance.

Another recent benchmark, PhD [7], focuses on hallucination detection and includes a subset of counter-commonsense (CCS) images. Each image is accompanied by two binary (yes/no) questions designed to probe whether the visual content aligns with commonsense expectations. This setup provides a targeted way to evaluate model inconsistencies in image interpretation.

Hallucination Detection in Text Generation

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. [10] proposed SelfCheckGPT, a straightforward sampling-based method that enables fact-checking of black-box models with zero resources. It uses multiple techniques to assess consistency across generations, including BERTScore, multiple-choice QA generation (MQAG) [9], and NLI-based contradiction detection.

Evaluation of Faithfulness in Multimodal Models

Regarding multi-modal case, [2] 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 [8].

Our Approach Compared to Prior Work

Prior approaches such as FAITHSCORE [2] and SelfCheckGPT [10] rely on consistency between generated content and the original input (image or text context), often using entailment models or self-consistency checks. This work follows a related intuition: hallucinations perform as internal contradictions among atomic facts generated by LVLMs, especially when presented with strange or counter-commonsense images.

The first method explores this idea by using a vision-language model to produce a set of atomic statements describing an image. These statements were then compared pairwise using a pretrained NLI model, and the resulting entailment or contradiction scores were aggregated into a single scalar value, the so-called reality-score, to detect visual inconsistency. This unsupervised NLI-based approach is one of the methods examined in this thesis.

Our second proposed method takes this idea further. While we also begin by extracting atomic facts using an LVLM, we do not rely solely on pairwise NLI scores. Instead, we use hidden representations of the LVLM itself and apply a supervised linear classifier to assess realism. This probing-based method aims to learn implicit realism features encoded across model layers. Notably, strong contradiction signals often emerge when the model encounters visually strange inputs, consistent with observations in benchmarks such as WHOOPS! and WEIRD [12]. Unlike earlier unsupervised approaches, this setup allows us to directly test the expressiveness of the model's internal representations for realism classification.

Chapter 5
Problem Statement

The goal of this thesis is to develop methods for estimating the realism of an image using outputs and internal representations of large vision-language models (LVLMs). Realism in this context refers to an image's consistency with commonsense knowledge, i.e., whether it plausibly reflects the real world.

5.1 Notation and Definitions

Let I denote an input image. We define a *reality-check function* $f_{\text{reality}}(I) \in \mathbb{R}$ that returns a scalar score representing the degree of realism of the image. A higher value indicates a higher likelihood that the image conforms to commonsense.

We denote a set of N atomic textual facts generated from I using an LVLM as:

$$F = \{f_1, f_2, \ldots, f_N\}, \quad f_i \in \mathbb{T}$$

where $\mathbb{T}$ is the set of natural language sentences.

Natural Language Inference (NLI)

For each pair (f_i, f_j) of atomic facts, an NLI model provides a triple of scores:

$$(s_{\text{ent}}, s_{\text{con}}, s_{\text{neu}}) = \text{NLI}(f_i, f_j)$$

These represent the probability of entailment, contradiction, and neutrality between the two facts. The final similarity score is computed as:

$$s_{\text{nli}}(f_i, f_j) = w_{\text{ent}} \cdot s_{\text{ent}} + w_{\text{con}} \cdot s_{\text{con}} + w_{\text{neu}} \cdot s_{\text{neu}}$$

Aggregation Methods

To estimate an overall realism score from the set of pairwise comparisons, the scores are aggregated using one of the following techniques:

- **Minimum:** Select the most contradictory fact pair: $\text{Min}(S_{\text{nli}}) = \min s_{\text{nli}}$
- **Absolute Maximum:** Take the strongest polarity: $\text{AbsMax}(S_{\text{nli}}) = S_{\text{nli}}\,[\arg\max(|S_{\text{nli}}|)]$
- **Clustering:** Apply k-means clustering to $\{s_{\text{nli}}(f_i, f_j)\}$ and take the lower centroid mean.

Linear Probing

We also consider internal representations of LVLMs. For an image I, we extract hidden states $h^{(l)}(I) \in \mathbb{R}^d$ from a specific layer l of the model. A logistic regression classifier $\mathcal{C}(h^{(l)}(I)) \rightarrow \{0, 1\}$ is trained to predict whether the image is realistic or not.

Attention-Based Classifier

An alternative approach encodes each atomic fact f_i as a vector $v_i = \text{Encoder}(f_i)$. These vectors are passed into an attention mechanism:

$$v_{\text{agg}} = \sum_{i=1}^{N} \alpha_i v_i, \quad \alpha = \text{softmax}(W_a V + b_a)$$

$$\text{prob} = \sigma(W_c v_{\text{agg}} + b_c)$$

where $\text{prob} \in [0, 1]$ indicates the estimated probability that the image is realistic.

5.2 Formal Objective

Let $D = \{(I_k, y_k)\}_{k=1}^{K}$ be a dataset of image-label pairs, where $y_k \in \{0, 1\}$ indicates whether the image is realistic (0) or weird (1).

The objective is to construct a classifier based on a realism scoring function $f_{\text{reality}}(I_k)$ such that:

$$\hat{y}_k = \mathbb{I}[f_{\text{reality}}(I_k) < \tau]$$

where τ is a threshold value (possibly tuned on a validation set), and the classifier $\hat{y}_k$ should maximize accuracy over the dataset D.

Thus, the task is formalized as a binary classification problem using image-level realism scores derived from LVLM outputs or internal activations.

Chapter 6
Methodology

In this chapter, we present the methods developed for detecting hallucinations and commonsense violations in images using outputs and internal states of vision-language models. We introduce an unsupervised NLI-based method that quantifies internal contradiction between generated facts, and a supervised probing approach that utilizes hidden representations of LVLMs to classify images as realistic or strange. In addition, we evaluate a variety of baseline models, including vision-language and language-only zero-shot methods, as well as fine-tuned supervised models. As part of the supervised methods, we also include TLG (Through the Looking Glass), a recently proposed model that learns attention-based representations over atomic facts for hallucination detection.

6.1 NLI-Based Detection of Multimodal Hallucinations

We propose a zero-shot method for estimating the realism of an image by analyzing contradictions among textual atomic facts extracted using a Large Vision-Language Model (LVLM). The method is based on the observation that LVLMs may hallucinate factual inconsistencies when describing visually abnormal images.

The method consists of three main stages: (1) atomic fact generation using an LVLM, (2) pairwise natural language inference between facts, and (3) aggregation of contradiction scores into a final realism score. The pipeline of our approach is depicted in Figure 6.1.

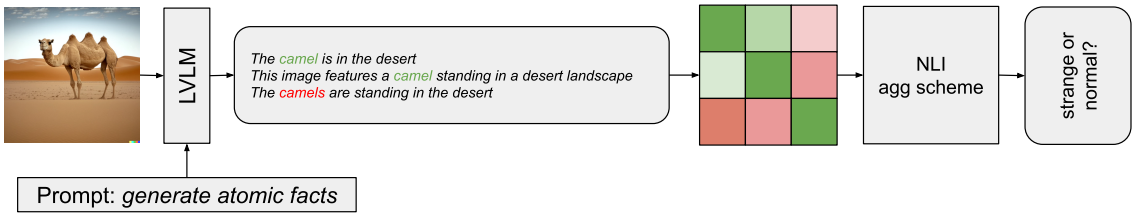

Figure 6.1: Weird images detection pipline. First, we generate five atomic facts that describe the images with LVLM (`llava-v1.6-mistral-7b-hf`). Then, we proceed with the matrix of pairwise NLI scores, where each NLI score is a weighted combination of entailment, neutral, and contradiction scores. The last step is aggregating NLI scores. Then, based on the aggregated score, we decide whether the image is strange or not.

Atomic Fact Generation

Given an input image I, we prompt the LVLM to produce a set of short, descriptive statements that we call *atomic facts* with a prompt P *"Provide a short, one-sentence descriptive fact about this image."*:

$$F = \{f_1, f_2, \ldots, f_N\}, \quad f_i = \text{LVLM}(I, P)$$

This is a camel
The camel is standing on sand
This image features a camel standing on a sandy desert plain

The camel is in the desert
The camels are standing in the desert
This is a digitally manipulated image of a camel with two heads

Figure 6.2: A pair of images from the WHOOPS! dataset with corresponding generated atomic facts. The normal image is on the left, and the unusual image is on the right. All the facts associated with the normal image are consistent and accurately describe the image. However, in the case of the weird image, LVLM hallucinates and generates untruthful facts.

To promote diversity among the generated facts, we use Diverse Beam Search with parameters: `num_beams` $= 5$, `num_beam_groups` $= 5$, and `diversity_penalty` $= 1.0$. The model used for fact generation is LLaVA v1.6 Mistral 7B [4].

Pairwise Natural Language Inference (NLI)

Each pair of atomic facts (f_i, f_j) is evaluated using an NLI model. The model outputs three probabilities:

$$(s_{\text{ent}}, s_{\text{con}}, s_{\text{neu}}) = \text{NLI}(f_i, f_j)$$

representing entailment, contradiction, and neutrality. These scores are combined into a single scalar:

$$s_{\text{nli}}(f_i, f_j) = w_{\text{ent}} \cdot s_{\text{ent}} + w_{\text{con}} \cdot s_{\text{con}} + w_{\text{neu}} \cdot s_{\text{neu}}$$

In practice, the neutrality weight is fixed to zero ($w_{\text{neu}} = 0$), and the other weights are tuned using cross-validation on the WHOOPS! dataset.

Aggregation of Contradiction Scores

After aggregation and weighting, we calculate a sum of weighted scores for each pair of facts $s_{\text{nli}}(f_i, f_j)$ and $s_{\text{nli}}(f_j, f_i)$. We hypothesize that such a summation strategy will amplify negative contradictions and vice versa. In addition, we propose three strategies for aggregating the NLI score.

$$S_{\text{nli}} = \{s_{\text{nli}}(f_i, f_j) \mid i, j \in \{1, \ldots, N\},\ i \neq j\} \tag{6.1}$$

Minimum (min) For a given list of scores, we simply select the lowest value as the metric. We assume that the lowest value could represent the contradictory of the whole set of facts.

$$\text{Min}(S_{\text{nli}}) = \min(S_{\text{nli}}) \tag{6.2}$$

Absolute maximum (absmax) We transform all values from the list of scores to their absolute values, then select the index of the largest absolute value and return the value from the original list to preserve the sign of the original value. So, if some set of facts has a relatively strong contradiction, we choose it as a weird image and vice versa:

$$\text{AbsMax}(S_{\text{nli}}) = S_{\text{nli}}\left[\arg\max\left(|S_{\text{nli}}|\right)\right] \tag{6.3}$$

Clustering (clust) We run the k-means clustering algorithm on the set of individual scores to split them into 2 clusters and then select the centroid with the lowest value as the metric. The scheme is shown in Algorithm 6.1. The choice of 2 clusters corresponds to the binary classification task. The idea is similar to the min method, but instead of the lowest value over all, we select an average of the values from the lowest cluster. We expect that contradictory facts from the weird images will have lower cluster centers than a related one.

Algorithm 1 NLI Interval Clustering Aggregation

Input: S_{nli} – List of weighted and aggregated NLI scores
Output: centroids – The minimum centroid NLI score

```
kmeans ← KMeans(clusters=2)
centroids ← kmeans.fit(S_nli)
return min(centroids)
```

Classification

The final realism score is compared against a threshold τ to produce a binary decision:

$$\hat{y} = \mathbb{I}[f_{\text{reality}}(I) < \tau]$$

6.2 Theoretical Guarantees for Reality-Check Aggregations

Let I be an input image and

$$F = \{f_1, \ldots, f_N\}$$

the set of atomic facts generated by the LVLM for I. Define for each pair (i, j), $i \neq j$,

$$s_{ij} \;=\; s_{\text{nli}}(f_i, f_j) \;\in \mathbb{R}.$$

We then aggregate $\{s_{ij}\}_{i \neq j}$ into a single score $f_{\text{reality}}(I)$ via one of three strategies:

$$\text{min}, \quad \text{absmax}, \quad \text{or} \quad \text{clust},$$

and compare against a threshold τ. We now give sufficient conditions under which $f_{\text{reality}}(I) < \tau$ whenever I is *strange* (i.e. defies common sense).

Assumptions

- For normal images all facts are logically consistent: $s_{\text{nli}}(f_i, f_j) \geq 0$.
- For weird images there exists at least one fact pair with strong contradictions: $s_{\text{nli}}(f_i, f_j) < 0$.

- Generated facts cover the main realism constraints.
- NLI reliably assigns negative values for contradictions and positive for entailment.

Lemma 6.2.1 (Normal images)**.** *If input image is normal, then each aggregation would lead to positive score.*

By Lemma 6.2.1, we have that for any *real* image I, the reality-check function satisfies:

$$f_{\text{reality}}(I_{\text{r}}) \geq 0.$$

In the absence of additional constraints or side information, the minimal score for real images approaches zero. Consequently, the natural choice for the threshold is:

$$\tau = 0.$$

This choice maximally separates real images from weird ones, ensuring that any negative score is classified as strange. Setting $\tau = 0$ is both optimal and justified under the assumption that normal images produce non-negative scores.

Lemma 6.2.2 (Min aggregation)**.** *If there exists a pair* (i, j) *such that*

$$s_{nli}(f_i, f_j) < 0,$$

then then the method predicts strange.

Proof.

$$\min_{s \in S_{\text{nli}}} s \leq s_{ij} = s_{\text{nli}}(f_i, f_j) < 0 \quad \Longrightarrow \quad f^{\text{min}}_{\text{reality}}(I) < \tau$$

□

From now on we will address $s_{\text{nli}}(f_i, f_j)$ as s_{ij}.

Lemma 6.2.3 (AbsMax aggregation)**.** *If there exists an element* $s_{ij} < 0$ *such that*

$$|s_{ij}| \geq \max_{s \in S_{nli}} |s|$$

then the method predicts strange.

Proof.

$$\text{AbsMax}(S_{\text{nli}}) = \text{sign} s_{ij} |s_{ij}| < 0 \quad \Longrightarrow \quad f^{\text{absmax}}_{\text{reality}}(I) < \tau$$

□

Lemma 6.2.4 (Clustering aggregation)**.** *If there exists a cluster* S_1 *such that* $\sum_{s \in S_1} s < 0$ *then the method predicts strange.*

Proof.

$$\text{Clust}(S_{\text{nli}}) = \min(\mu_1, \mu_2) \leq \mu_1 = \frac{1}{|S_1|} \sum_{s \in S_1} s < 0 \quad \Longrightarrow \quad f^{\text{clust}}_{\text{reality}}(I) < \tau$$

□

Theorem 6.2.5 (Aggregation Hierarchy)**.** *If an image I is strange, the following implication chain holds:*

$$f^{absmax}_{reality}(I) < 0 \implies f^{clust}_{reality}(I) < 0 \implies f^{min}_{reality}(I) < 0.$$

Proof. Assume that $f^{\text{absmax}}_{\text{reality}}(I) < 0$. By definition of AbsMax, this implies that there exists an element $s_{min} < 0$ such that:

$$|s_{min}| \geq \max_{s \in S_{\text{nli}}} |s|.$$

Now, consider the two clusters produced by k-means, with centroids denoted as μ_1 and μ_2, where we assume without loss of generality that:

$$\mu_1 < \mu_2.$$

Suppose, for the sake of contradiction, that:

$$f^{\text{clust}}_{\text{reality}}(I) \geq 0 \implies \mu_1 \geq 0.$$

Since $\mu_1 < \mu_2$, the element s_{min} must belong to the cluster represented by μ_1, i.e., $s_{min} \in S_1$. We analyze the difference:

$$\mu_1 - s_{min} < \max_{s \in S_1} |s_{min} - s| < \min_{t \in S_2} |s_{min} - t|.$$

This inequality guarantees that s_{min} is strictly closer to points in S_1 than to any point in S_2. Now, consider any $t \in S_2$. We would have:

$$|t - \mu_1| \leq \mu_1 - s_{min}.$$

However, by the AbsMax condition, the magnitude of s_{min} dominates all other points, implying:

$$t - \mu_1 \leq |s_{min}| - \mu_1 \leq |s_{min}| + \mu_1 = \mu_1 - s_{min}.$$

This is a contradiction since $t \in S_2$ would not satisfy the inequality. Hence, our assumption that $f^{\text{clust}}_{\text{reality}}(I) \geq 0$ must be false, and we conclude:

$$f^{\text{absmax}}_{\text{reality}}(I) < 0 \implies f^{\text{clust}}_{\text{reality}}(I) < 0.$$

Finally, we move to the second part:

$$\mu_1 < 0 \implies \exists\, s_{ij} < 0.$$

Since $\min(S_{\text{nli}})$ selects the smallest score, it follows immediately that:

$$f^{\text{clust}}_{\text{reality}}(I) < 0 \implies f^{\text{min}}_{\text{reality}}(I) < 0.$$

Thus, the implication chain is complete:

$$f^{\text{absmax}}_{\text{reality}}(I) < 0 \implies f^{\text{clust}}_{\text{reality}}(I) < 0 \implies f^{\text{min}}_{\text{reality}}(I) < 0.$$

□

6.3 Linear Probing of LVLM Hidden States

In this approach, we examine whether the hidden states of large vision-language models (LVLMs) encode sufficient information to distinguish realistic from commonsense-violating images. This technique involves extracting intermediate representations from a frozen LVLM and training a lightweight classifier on top of these features.

The method is motivated by findings that deep models often learn structured representations across layers, where information relevant to downstream tasks may emerge even without fine-tuning.

Feature Extraction

We consider two setups: (a) using the <image> as the sole input (**Image only**), and (b) using <image> the with a prompt *"Provide a short, one-sentence descriptive fact about this image"* (**+Prompt**), which was used to generate atomic facts. Given an input image I, we pass it through an LVLM such as LLaVA-1.6 Vicuna 13B with a corresponding prompt and collect the hidden states at each decoder layer:

$$h^{(l)}(I, P) \in \mathbb{R}^d, \quad l \in \{1, 2, \ldots, L\}$$

where d is the hidden size of the decoder layer. The representation $h_l(I)$ is pooled from last token. We collect features from each layer l for every sample in the dataset.

Classifier Training

We train a logistic regression classifier $\mathcal{C}_l$ for each layer l using the corresponding features:

$$\mathcal{C}_l(h_l(I, P)) \rightarrow \{0, 1\}$$

to predict whether an image is realistic or weird. All LVLM parameters remain frozen during training.

Cross-Validation with Layer Selection

To ensure robust evaluation and fair model selection, we perform 5-fold cross-validation. For each fold, we split the training set into training and validation subsets. We train separate classifiers for each layer and select the best-performing layer l^* based on validation accuracy. This selection is done once and used for all test folds. We trained a logistic regression with L2 regularization, with a maximum of 100 iterations and a tolerance of 0.1 on standardized hidden states.

Algorithm 2 Cross-validated layer selection and evaluation for linear probing

Require: Hidden state features $H \in \mathbb{R}^{L \times N}$, labels $y \in \{0, 1\}^N$, number of folds K
Initialize validation accuracy map $\texttt{val\_acc}[l] \leftarrow []$ for $l = 1 \dots L$
for each fold i in $1 \dots K$ **do**
    Split H and y into train and test sets
    Further split train set into training and validation sets
    for each layer l **do**
        Train logistic regression $\mathcal{C}_l$ on training set
        Compute validation accuracy of $\mathcal{C}_l$ and store in $\texttt{val\_acc}[l]$
    end for
end for
Select best layer $l^* = \arg\max_l \mathbb{E}[\texttt{val\_acc}[l]]$
Initialize $\texttt{test\_acc} \leftarrow []$
for each fold i in $1 \dots K$ **do**
    Train $\mathcal{C}_{l^*}$ on train set of fold i
    Evaluate on test set of fold i and store accuracy in $\texttt{test\_acc}$
end for
return Mean and std of $\texttt{test\_acc}$

6.4 Through the Looking Glass

In addition to the NLI-based and probing approaches, we include a third method based on supervised learning over textual representations of atomic facts. The approach, referred to as TLG (Through the Looking Glass), was developed by the co-authors of [12] and represents a generalization of the NLI-based method by removing the need for explicit pairwise comparison.

Motivation

While the NLI-based approach detects contradictions between facts using human-defined logic, TLG seeks to learn such relationships directly from data. The method uses a lightweight attention-based classifier trained over embeddings of atomic facts generated by an LVLM. It allows the model to learn which facts contribute most to realism estimation, without relying on predefined inference mechanisms.

Architecture

Given an input image I, a set of atomic facts $\mathcal{F} = \{f_1, \dots, f_N\}$ is generated using an LVLM as described previously.

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

$$H_i = \text{Encoder}(f_i) \in \mathbb{R}^{N \times T \times d}, \tag{6.4}$$

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

$$V_i = \frac{\sum_{j=1}^{T} m_{ij} H_{ij}}{\sum_{j=1}^{T} m_{ij} + \varepsilon}. \tag{6.5}$$

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:

$$A = \text{softmax}(W_a V + b_a) \in \mathbb{R}^N. \tag{6.6}$$

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

$$v_{\text{weighted}} = \frac{\sum_{i=1}^{N} A_i V_i}{\sum_{i=1}^{N} A_i} \in \mathbb{R}^d. \tag{6.7}$$

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

$$\text{prob} = \sigma(W_c v_{\text{weighted}} + b_c) \in [0, 1]. \tag{6.8}$$

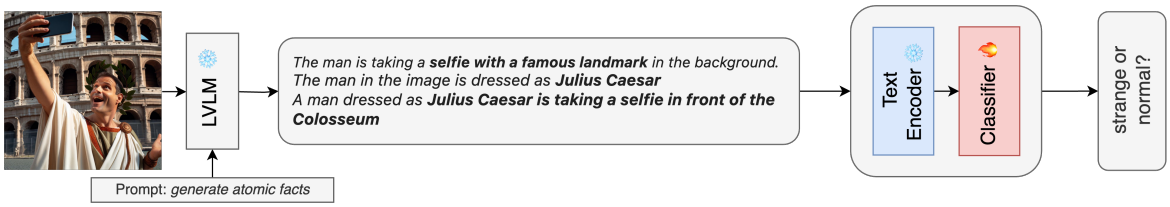

Figure 6.3: Architecture of the TLG method: atomic facts are encoded into embeddings, attention pooling is applied, and a classifier is trained to predict realism. Adapted from [12].

Training and Evaluation

The classifier is trained in a supervised setting using binary labels from datasets such as WHOOPS! and WEIRD. The training objective is binary cross-entropy loss between the predicted realism score and the ground-truth label.

TLG is evaluated in a 5-fold cross-validation setup. It achieves state-of-the-art performance on both datasets, significantly outperforming zero-shot and fine-tuned LVLM baselines. The method also demonstrates strong generalization capabilities in cross-dataset evaluation, indicating that it effectively captures commonsense inconsistency patterns across image domains.

Position in This Work

While TLG is not the main contribution of this thesis, we include it to provide a broader view of the design space for realism detection. It serves as a supervised generalization of the contradiction-based approach and provides a high-performance upper bound against which simpler zero-shot and probing-based methods can be compared.

6.5 Baselines

To assess the effectiveness of our proposed approaches, we compare them against a range of baseline models. These include both zero-shot and supervised methods, covering vision-language and language-only architectures.

LVLM

We include a zero-shot baseline using a vision-language model with a prompt designed to elicit commonsense realism judgments. Following [5], we use the prompt:

> *"<image> Is this unusual? Please explain briefly with a short sentence."*

This setup uses the model's instruction-following capabilities to assess the unusualness of the image directly.

LLM

We additionally evaluate language-only models as zero-shot classifiers. We use Gemma-2-9B-Instruct [13] and Qwen2.5-7B-Instruct [17]. As input, we provide the set of atomic facts generated from the image and prompt the model to classify the realism of the fact set based on internal consistency. The prompt is as follows:

> *"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."*

This approach tests the model's ability to detect contradictions among factual descriptions without visual input.

BLIP-2 Fine-Tuned Models

Furthermore, we include two supervised baselines based on BLIP-2 [3], which were fine-tuned on the WHOOPS! dataset as reported in [1]. Specifically, we evaluate the performance of BLIP-2 FlanT5-XL and BLIP-2 FlanT5-XXL models. These serve as strong supervised references for visual commonsense violation detection.

Chapter 7
Computational Experiments

In this chapter, we present the experimental evaluation of the proposed methods across multiple benchmarks for commonsense-based realism detection. All experiments are performed on the WHOOPS! [1] and WEIRD [12] datasets. We evaluate both zero-shot and supervised methods, including our proposed NLI-based approach, linear probing (LP), and the supervised TLG classifier.

7.1 Data

We evaluate our proposed methods on two benchmark datasets designed for commonsense-based realism detection: WHOOPS! [1] and WEIRD [12]. These datasets serve complementary roles: WHOOPS! provides controlled, synthetic abnormalities, while WEIRD introduces more nuanced, real-world hallucinations for robustness testing.

WHOOPS! Dataset

The WHOOPS! dataset consists of 204 images, evenly split between realistic and weird categories. Each image is accompanied by short descriptive captions and annotations reflecting commonsense plausibility. WHOOPS! is constructed to expose visual contradictions -images that, while photo-realistic, defy logical reasoning or physical constraints. Examples include objects with implausible scale, vehicles in unnatural environments, and human interactions with physically impossible poses.

Images in WHOOPS! are grouped into 26 commonsense-breaking categories, covering various violation types such as:

- **Physical Implausibility:** Objects with impossible configurations (e.g., a car floating in mid-air).
- **Semantic Inconsistency:** Actions that contradict realistic behavior (e.g., a dog reading a book).
- **Contextual Errors:** Objects misplaced in environments where they are not typically found (e.g., a snowplow on a beach).

The dataset provides a controlled setting for evaluating the ability of LVLMs to identify visual hallucinations via logical contradiction. Human accuracy on WHOOPS! is reported at 92%, establishing a high standard for model-based hallucination detection.

WEIRD Dataset

The WEIRD dataset, introduced in [12], is a large-scale benchmark designed to extend realism evaluation beyond the limitations of WHOOPS!. It contains 824 image pairs representing a broad spectrum of commonsense violations. In contrast to WHOOPS!, which primarily focuses on synthetic abnormalities, WEIRD introduces more nuanced and diverse violations of commonsense realism, with a total of 181 sub-categories grouped into 12 global categories.

Generation Process. The WEIRD dataset was generated in a semi-automatic manner, inspired by the Self-Instruct framework [16]. The creation process began with a task pool of image-caption pairs sampled from WHOOPS!. For each iteration, five pairs of normal and weird images were randomly selected, along with their commonsense-breaking categories. These samples were fed into GPT-4o, which generated new categories and captions for both normal and weird scenarios. The new textual descriptions were then visualized using DALL-E 3.

In each generation cycle, 50 new pairs were created, totaling 100 samples per iteration. Generated images underwent a manual filtering process to remove low-quality or inconsistent samples, particularly focusing on mismatches between captions and visual content. For instance, images involving celebrity likenesses were frequently discarded due to generation inaccuracies.

In total, 2,000 samples were generated, of which 824 high-quality samples remained after filtering. This rigorous curation ensured that the dataset reflected genuine commonsense violations while maintaining visual coherence.

Annotation Process. The dataset was annotated through the Yandex Tasks crowd-sourcing platform, where each image was evaluated by five independent annotators. To familiarize annotators with the task, 10 training samples were provided before the main annotation phase. The overall annotation agreement, measured by Krippendorff's alpha, was reported at 0.69, with a final human accuracy of 82.22%. This high inter-annotator agreement highlights the dataset's consistency and the clarity of commonsense violations represented in the images.

Comparison with WHOOPS!. Compared to WHOOPS!, WEIRD is approximately four times larger and incorporates more granular categories of weirdness. While WHOOPS! captures visually distinct contradictions, WEIRD emphasizes subtle violations that require deeper commonsense reasoning, such as abstract scene logic errors and artistic contradictions. The larger scale and broader category set of WEIRD make it a robust benchmark for evaluating the hallucination detection capabilities of vision-language models (VLMs).

7.2 Ablation Study: NLI Model and Aggregation Method

To explore the impact of different NLI models and aggregation techniques on performance, we conduct an ablation study using a subset of the WHOOPS! dataset. This analysis provides insights into how various models and aggregation methods influence the accuracy of the NLI-based hallucination detection approach. The results are shown in Table 7.1, where we evaluate models such as 'nli-deberta-v3-large', 'nli-deberta-v3-base', and 'nli-deberta-v3-small' under three aggregation strategies: min, absmax, and clust.

Model	#	**Method**		
		min	absmax	clust
nli-deberta-v3-large	304M	<u>63.73</u>	62.75	**72.55**
nli-deberta-v3-base	86M	<u>60.78</u>	55.39	**61.76**
nli-deberta-v3-small	47M	**61.27**	58.33	<u>60.78</u>

Table 7.1: A comparison of various NLI models for distinct aggregation techniques for subset with 5 facts is provided. Accuracy as the evaluation metric.

7.3 Comparison of Methods

Table 7.2 compares all evaluated methods on both datasets in terms of classification accuracy. Supervised models are trained with 5-fold cross-validation. Zero-shot baselines include open-ended prompting with instruction-tuned LVLMs such as InstructBLIP and LLaVA.

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
NLI	7B	zero-shot	72.55	60.00
LP - LLaVA	13B	fine-tuned	73.50	85.26
TLG	8B	fine-tuned	**73.54**	**87.57**
GPT-4o	–	zero-shot	79.90	81.64

Table 7.2: 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.

Among the zero-shot baselines, InstructBLIP and LLaVA perform better than random, but significantly worse than trained models, with accuracy around 50-55% on both datasets. This illustrates the difficulty of detecting commonsense inconsistencies without explicit training or relational modeling, as usually LVLMs classify the images as weird due to the stylistic features and not its content.

The proposed NLI-based method achieves 73.04% accuracy on WHOOPS!, which is on par with supervised models like TLG and significantly higher than zero-shot baselines. This result confirms the viability of contradiction detection as a proxy for realism estimation, particularly on synthetic images where contradictions are more evident and fact generation is stable. However, on WEIRD, the NLI method drops to 60.0%, indicating reduced robustness in scenarios with subtler or more diverse abnormalities. This drop is attributed to the fact generation pipeline, which may struggle with the broader visual and semantic variance in WEIRD images.

Despite this, the NLI method retains important advantages: it requires no labels, supports direct interpretability, and provides meaningful diagnostic signals (i.e., contradictions) that can be analyzed or visualized. It thus remains a useful tool, particularly in low-resource or zero-shot settings.

The TLG classifier achieves the highest overall accuracy on both datasets, reaching over 83% in WHOOPS→WEIRD transfer and 74% in WEIRD→WHOOPS!. Its performance confirms the effectiveness of learning attention-based interactions over fact embeddings. However, this comes with the cost of supervision and increased model complexity.

Linear probing performs surprisingly well, achieving results close to or exceeding TLG in certain settings. For example, when trained on WEIRD and evaluated on WHOOPS!, LP (Image only, LLaVA Vicuna 13B) reaches 75.00% accuracy, surpassing all TLG variants. This supports the conclusion that internal representations in large LVLMs encode realism-related signals that are robust across domains.

In summary, TLG offers the highest in-domain performance; LP demonstrates stronger cross-domain generalization and simplicity; the NLI method excels in interpretability and zero-shot applicability.

7.4 Ablation Study: Linear Probing Models and Inputs

Table 7.3 summarizes results across different linear probing configurations, including variations in model backbone, input modality (**Image only** or **+Prompt**), and parameter size. Larger models such as LLaVA Vicuna 13B yield the best overall performance. Including prompt information in the encoder input also improves accuracy on WHOOPS!, though performance gains on WEIRD are less consistent.

Model	Image only	+Prompt
WHOOPS!		
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	**73.50**
WEIRD		
LLaVA 1.6 Mistral 7B	78.13	81.82
LLaVA 1.6 Vicuna 7B	84.65	83.91
LLaVA 1.6 Vicuna 13B	**85.26**	84.02

Table 7.3: Linear probing results on WHOOPS! and WEIRD across models and input variants.

7.5 Cross-Dataset Generalization

The table 7.4 evaluates the generalization ability of Linear probing and TLG methods by training on one dataset and testing on another. The direction WEIRD→WHOOPS! evaluates whether models trained on the more diverse and visually complex WEIRD images can generalize to the simpler, more synthetically structured WHOOPS! dataset. Conversely, WHOOPS!→WEIRD tests whether models trained on simpler examples can detect realism violations in the more diverse WEIRD images.

Linear Probing shows strong performance across all configurations. In particular, LP (image-only) with LLaVA Vicuna 13B achieves 75.00% on WEIRD→WHOOPS!, outperforming all TLG configurations. This result suggests that internal LVLM representations encode rich commonsense information that generalizes across domains, even when textual prompt is not used.

LP with both image and prompt inputs tends to perform slightly better than image-only in the WHOOPS!→WEIRD direction. The addition of textual context appears helpful in scenarios where the visual content is less semantically rich (e.g., WHOOPS!), but may introduce noise or redundancy in more complex images (e.g., WEIRD).

Effect of Model Size For LP, model size has a notable impact on generalization. The 13B variant of LLaVA Vicuna consistently outperforms 7B models, indicating that deeper and wider networks

Method	#Params	Model	WEIRD→WHOOPS!	WHOOPS!→WEIRD
LP (+Prompt)	7B	LLaVA Mistral	70.10	66.58
	7B	LLaVA Vicuna	70.59	76.04
	13B	LLaVA Vicuna	72.06	74.69
LP (Image only)	7B	LLaVA Mistral	67.65	63.88
	7B	LLaVA Vicuna	70.59	70.64
	13B	LLaVA Vicuna	**75.00**	<u>79.61</u>
TLG (Ours)	8B	LLaVA-Mistral + deberta-v3-large-tasksource-nli	<u>74.02</u>	**83.05**
	8B	LLaVA-Mistral + nli-deberta-v3-large	71.08	74.82
	8B	LLaVA-Mistral + deberta-v3-large	70.59	73.10

Table 7.4: Knowledge transfer across datasets. WEIRD→WHOOPS! means training on WEIRD and testing on WHOOPS!, and vice versa. Best result in bold, second-best underlined.

encode more transferable features. This effect is especially visible in the image-only setting, where the 13B model reaches 75.00% and 79.61% in WEIRD→WHOOPS! and WHOOPS!→WEIRD respectively.

TLG Among all models, TLG achieves the best cross-dataset transfer in the WHOOPS!→WEIRD direction, reaching 83.05% accuracy. This suggests that TLG is capable of learning generalizable fact-level consistency patterns even from a simpler dataset. Notably, the best TLG result uses a fact encoder fine-tuned on NLI and task-related data (`deberta-v3-large-tasksource-nli`), highlighting the importance of pretraining alignment in the textual encoder.

In the opposite direction (WEIRD→WHOOPS!), the best TLG accuracy is 74.02%, which remains competitive, but slightly lower than the best LP result.

Chapter 8
Analysis

In this chapter, we conduct an investigation into the behavior of our proposed methods, with a focus on understanding the internal mechanisms driving their predictions. We begin with a qualitative analysis of contradiction heatmaps produced by the NLI-based method, highlighting the alignment between fact-level contradictions and image-level weirdness. We then examine the statistical association between hallucination markers and model predictions. Finally, we analyze the effectiveness of linear probing across different layers of the LVLM, revealing insights about where realism-related features are encoded.

8.1 Pairwise NLI Scores of Atomic Facts with Corresponding Images

To analyze how the NLI-based method identifies inconsistencies, we visualize pairwise scores between atomic facts as weighted contradiction matrices. These matrices reflect the weighted sum of entailment, contradiction, and neutrality probabilities as previously described in Section **??**. In this setting:

- **Negative values** indicate contradiction between two facts (with more negative values denoting stronger contradictions),
- **Positive values** indicate entailment or mutual consistency.

Figures 8.1–8.4 display four such matrices for two normal and two weird images. In the first normal example (Figure 8.1), the image shows a snow plow clearing a snowy street. The generated atomic facts are mutually consistent. The matrix shows predominantly positive values across all fact pairs, indicating no strong contradiction signals.

In contrast, the visually weird image in Figure 8.2 depicts a large yellow truck in the desert. Here, the LVLM struggles to settle on a coherent interpretation, producing conflicting facts like *"a large yellow bulldozer"* and *"a yellow school bus"*. These contradictions appear in the matrix as highly negative values along specific fact pair entries, reflecting the model's internal confusion due to the object's ambiguous nature and environment.

The second normal image (Figure 8.3) features a man sleeping in a bed. All generated facts relate to typical objects and scenarios and are semantically coherent. The resulting matrix again shows mostly positive values, reflecting consistent entailment between facts.

In the corresponding weird image (Figure 8.4), the man is sleeping on a rock. The LVLM hallucinates a fact stating *"The rock appears to be the man's head"*, which contradicts more literal facts like *"The man is sleeping on a rock"*. This contradiction is evident as a sharply negative entry in the matrix, while the remaining facts show only weak entailment or neutral scores.

Overall, the presence of sharply negative values localized within the matrix signals conflicting factual interpretations, which correspond with low realism scores. This confirms that the model is able to surface internal contradictions induced by counter-commonsense visual content, validating the design of the NLI-based pipeline.

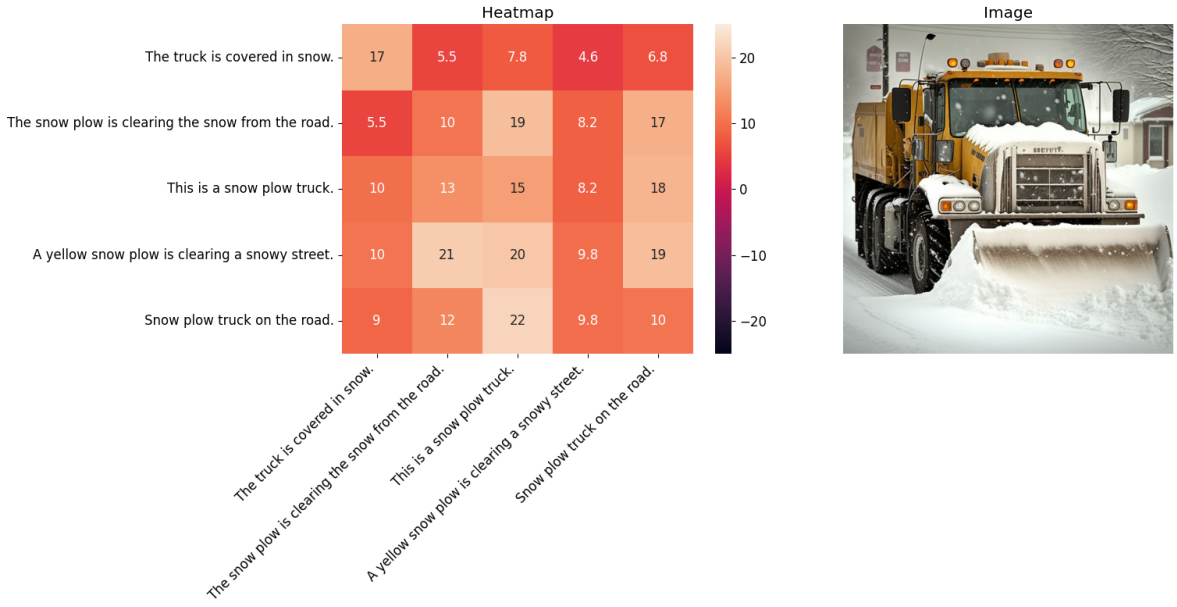

Figure 8.1: A snow plow driving down a snowy street (normal image).

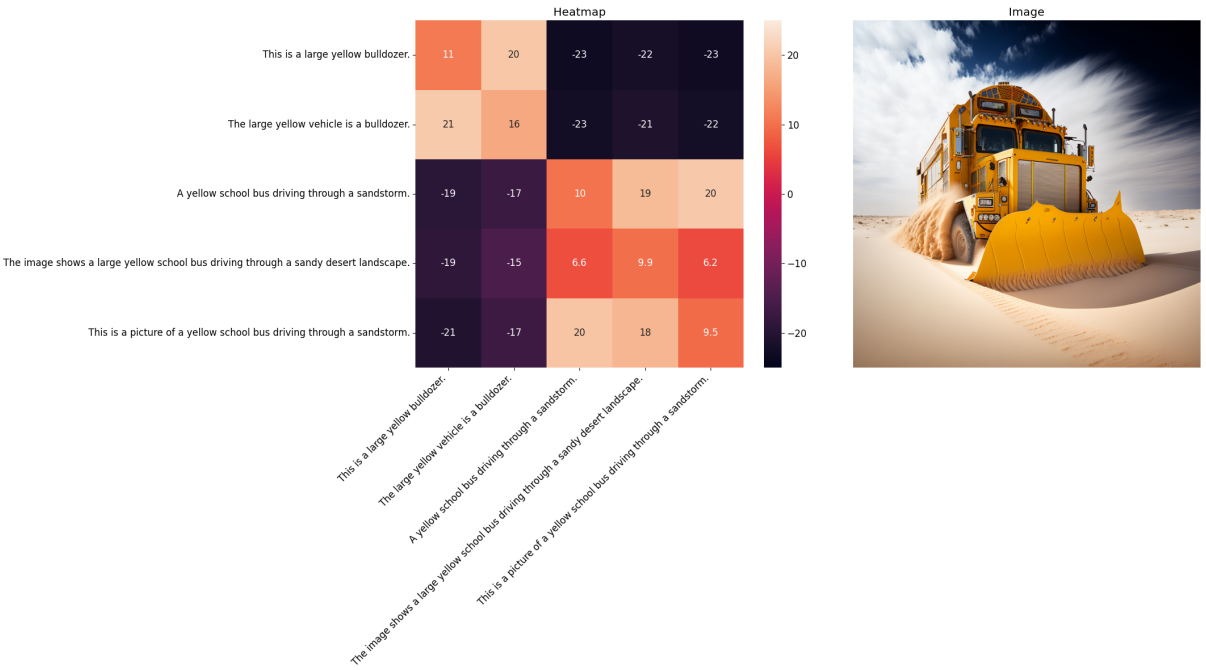

Figure 8.2: A large yellow truck driving through the sand (weird image).

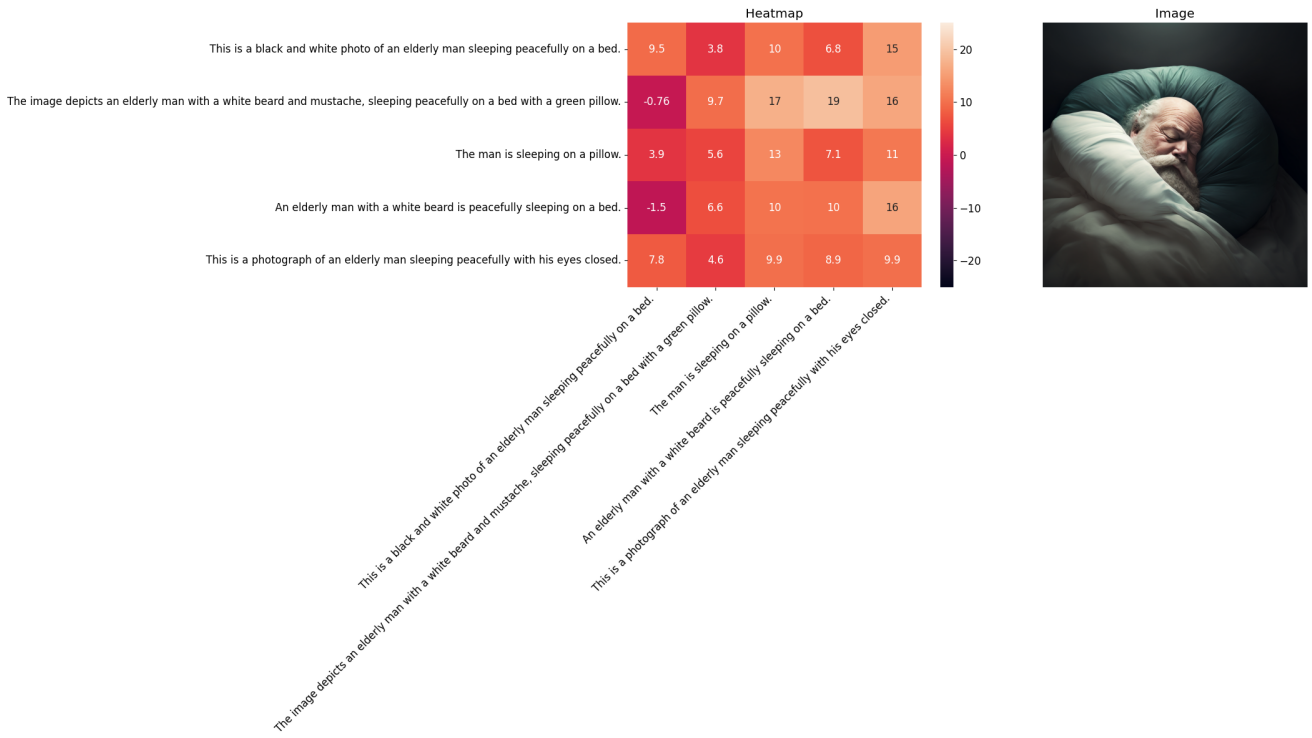

Figure 8.3: A man sleeping in a bed with a pillow (normal image).

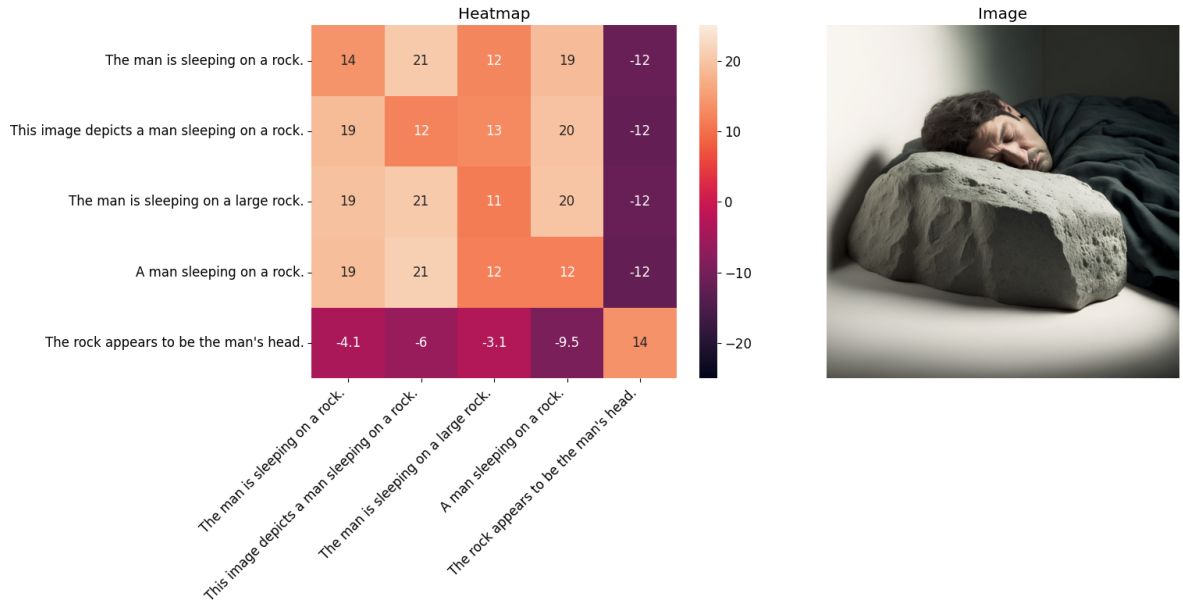

Figure 8.4: A man sleeping on a rock (weird image).

8.2 Underlying Phenomena of NLI-based Predictions

We analyze how the presence of certain phenomena affects the prediction of our approach. digital indicates the presence of marker words in the generated facts: *"digital"*, *"artistic"*, *"rendering"*. hallucination indicates the presence of hallucinations of an object, feature, or relationship, manually determined from the generated facts.

The relation between the model's decision and the existence of hallucinations or marker words is presented in Table 8.1. Our analysis reveals that hallucinations serve as a significant

indicator for our model to categorize images as weird. Interestingly, we also discovered that, in certain instances, the LVLM introduces distinctive marker words specifically to denote unusual images.

Condition	Probability
$\mathbb{P}(\text{weird} \mid \text{digital})$	0.76
$\mathbb{P}(\text{weird} \mid \text{hallucination})$	0.81
$\mathbb{P}(\text{weird} \mid \text{hallucination\&digital})$	**0.93**

Table 8.1: The conditional probability of model prediction being weird given the presence of marker words or hallucinations. For example, $\mathbb{P}(\text{weird} \mid \text{hallucination}) = 0.81$ indicates the probability that the model predicts an image as weird when hallucination is present in the atomic facts.

These findings suggest that in a zero-shot setup, the model implicitly relies on hallucination-related linguistic patterns to make realism judgments. The presence of such words or hallucinated concepts reliably increases the likelihood of the image being classified as weird.

8.3 Layer-Wise Performance in Linear Probing

To analyze where realism-related information is encoded in the LVLM, we examined the validation accuracy of linear probing classifiers trained on hidden states from different decoder layers. The results are visualized in Figure 8.5.

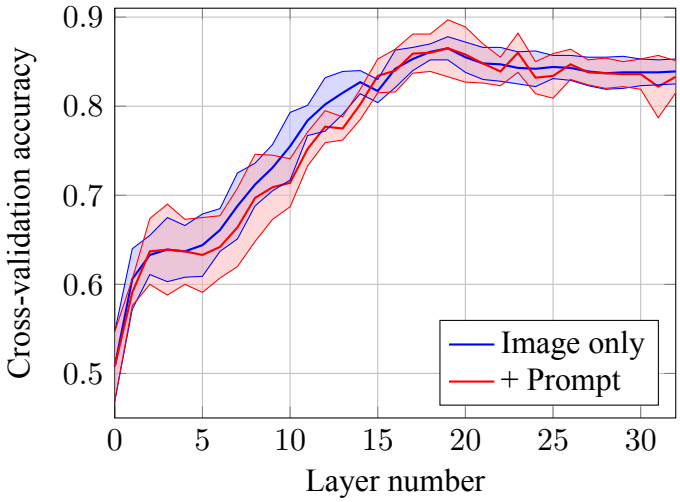

Figure 8.5: Cross-validation accuracy depending on the LLaVA 1.6 Vicuna 13B index layer for linear probing on the WEIRD dataset. Layers containing the most relevant information are in the middle of the decoder.

As shown in the plot, performance typically peaks in the middle layers of the decoder stack and drops toward both the early and final layers. This pattern is consistent with prior findings in transformer-based models [14], where middle layers tend to encode higher-level semantic or task-relevant features, while early layers focus on low-level structure and final layers shift toward task-specific output formatting.

The fact that realism information peaks mid-network implies that LVLMs develop an internal understanding of scene plausibility, which is encoded before final output decoding layers. This supports the central thesis hypothesis that internal activations carry sufficient signal for hallucination detection, even without direct supervision or fine-tuning.

Chapter 9
Discussion and Conclusion

In this chapter, we reflect on the methods proposed for detecting multimodal hallucinations and evaluating image realism based on commonsense understanding. The discussion addresses the strengths and limitations of the approaches explored, their performance across datasets, and their broader implications in the context of vision-language model (VLM) research.

9.1 Summary of Results

We introduced and evaluated several methods for estimating the realism of images that challenge commonsense reasoning. Our primary contribution is the development of a zero-shot classifier based on Natural Language Inference (NLI) over atomic facts generated by LVLMs. This method leverages hallucinations themselves as signals: the assumption is that when LVLMs process images that contradict common sense, they are more likely to generate contradictory or nonsensical statements. We use an NLI model to detect such contradictions among the generated facts and aggregate contradiction scores into a realism score.

This approach achieves competitive performance with fine-tuned models. On the WHOOPS! dataset, the NLI-based method with clustering-based aggregation reaches 72.55% accuracy, outperforming other zero-shot baselines and approaching the performance of the fine-tuned BLIP2 FlanT5-XXL model. Our extended experiments further show that the choice of NLI model and aggregation strategy plays a crucial role, with clustering outperforming simpler methods such as min and absmax.

We also introduced a linear probing (LP) baseline that classifies image realism based on hidden states from a frozen LVLM. This approach performs especially well on the larger WEIRD dataset, achieving 85.26% accuracy when trained in image-only mode on the LLaVA Vicuna 13B backbone. LP is computationally efficient and interpretable, with the probing layers exhibiting performance peaks in the middle layers - an effect observed in prior works on probing transformer representations.

Finally, we compared both methods to TLG (Through the Looking Glass), a fine-tuned attention-based classifier trained on LVLM-generated atomic facts. TLG achieves the highest performance overall, with 87.57% accuracy on WEIRD and 73.54% on WHOOPS!. However, the margin between TLG and LP narrows on WEIRD, where LP benefits from larger data and model capacity. Meanwhile, the NLI-based method, while slightly lower in raw accuracy, remains unique in its complete zero-shot setup and interpretability.

9.2 Position in Global Research Landscape

Our study contributes to a growing body of work aiming to understand and detect hallucinations in multimodal systems. While most existing efforts focus on suppressing hallucinations, our approach embraces them as diagnostic signals. This perspective aligns with recent trends in using self-consistency or contradiction as implicit indicators of factuality in LLM outputs.

Compared to supervised baselines and large proprietary systems (e.g., GPT-4o), our methods offer a transparent and open-source alternative. The NLI-based approach is particularly noteworthy as it relies on small, reusable components and achieves strong performance without requiring any model fine-tuning. This shows that multimodal commonsense reasoning does not always necessitate complex architectures if linguistic inconsistency is properly leveraged.

9.3 Comparative Analysis

The NLI-based method is distinguished by its simplicity, interpretability, and zero-shot capabilities. It offers a clear rationale: realistic images yield consistent facts; weird images yield contradictions. Heatmap analyses confirm that contradiction hotspots frequently emerge around hallucinated elements or unexpected object combinations.

Linear probing, while relying on supervised training, remains lightweight and generalizes surprisingly well across datasets. It performs well even in zero-shot transfer scenarios, particularly from WEIRD to WHOOPS!, and its accuracy scales with model size. Furthermore, our analysis shows that the most informative features for LP lie in the middle layers, which is consistent with previous research [14].

TLG offers the best accuracy by learning to weight and pool atomic facts through an attention mechanism. Its performance gap over LP is most visible on the smaller WHOOPS! dataset, suggesting better data efficiency and robustness to limited training signals. However, LP outperforms TLG in some transfer setups, particularly when using image-only features on the larger WEIRD dataset.

9.4 Limitations and Future Work

The NLI method, while competitive, is sensitive to the quality and diversity of generated facts. Its performance depends on the behavior of the underlying LVLM, and may degrade if the hallucinations are too subtle or if the facts are redundant. Additionally, it assumes that contradiction between facts is a sufficient proxy for unreality, which might not always hold in abstract or artistic scenarios.

Linear probing requires access to large hidden state matrices and benefits from high-capacity backbones. While efficient at inference, its training still requires careful layer selection and standardization, which could be challenging in production settings.

The supervised TLG method achieves strong accuracy but at the cost of training a dedicated classifier. It may also be more brittle to domain shift if the training data is not representative.

In future work, we plan to shift the focus towards detecting and highlighting hallucination spans in textual outputs. Building on the NLI-based contradiction analysis explored in this thesis, we aim to develop methods for partitioning text into equivalent, neutral, and contradictory segments.

Acknowledgements

I would like to express my gratitude to Elisei Rykov for his close collaboration and valuable contributions throughout the research process. I am also grateful to Vasily Konovalov for our insightful discussions and his support.

Bibliography

[1] Guetta, N. B., Bitton, Y., Hessel, J., Schmidt, L., Elovici, Y., Stanovsky, G., and Schwartz, R. 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* (2023), IEEE, pp. 2616–2627.

[2] Jing, L., Li, R., Chen, Y., Jia, M., and Du, X. FAITHSCORE: evaluating hallucinations in large vision-language models. *CoRR abs/2311.01477* (2023).

[3] Li, J., Li, D., Savarese, S., and Hoi, S. C. H. 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* (2023), A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202 of *Proceedings of Machine Learning Research*, PMLR, pp. 19730–19742.

[4] Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 26296–26306.

[5] Liu, H., Li, C., Li, Y., and Lee, Y. J. 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* (2024), IEEE, pp. 26286–26296.

[6] Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (2023), A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds.

[7] Liu, J., Fu, Y., Xie, R., Xie, R., Sun, X., Lian, F., Kang, Z., and Li, X. Phd: A prompted visual hallucination evaluation dataset. *CoRR abs/2403.11116* (2024).

[8] Maksimov, I., Konovalov, V., and Glinskii, A. 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)* (Mexico City, Mexico, June 2024), A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds., Association for Computational Linguistics, pp. 274–278.

[9] Manakul, P., Liusie, A., and Gales, M. J. F. 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* (2023), J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, Eds., Association for Computational Linguistics, pp. 39–53.

[10] Manakul, P., Liusie, A., and Gales, M. J. F. 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* (2023), H. Bouamor, J. Pino, and K. Bali, Eds., Association for Computational Linguistics, pp. 9004–9017.

[11] Rykov, E., Petrushina, K., Titova, K., Panchenko, A., and Konovalov, V. Don't fight hallucinations, use them: Estimating image realism using nli over atomic facts, 2025.

[12] Rykov, E., Petrushina, K., Titova, K., Razzhigaev, A., Panchenko, A., and Konovalov, V. Through the looking glass: Common sense consistency evaluation of weird images, 2025.

[13] Team, G. Gemma.

[14] Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), A. Korhonen, D. Traum, and L. Màrquez, Eds., Association for Computational Linguistics, pp. 4593–4601.

[15] Theis, L. What makes an image realistic?, 2024.

[16] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. 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)* (Toronto, Canada, July 2023), A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Association for Computational Linguistics, pp. 13484–13508.

[17] Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report. *CoRR abs/2412.15115* (2024).

[18] Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. 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* (2019), Computer Vision Foundation / IEEE, pp. 6720–6731.