

# Детекция мультимодальных галлюцинаций на основе внутренних представлений и активаций моделей

Ксения Петрушина

Кафедра интеллектуальных систем ФПМИ МФТИ

Специализация: Интеллектуальный анализ данных

Направление: 09.04.01 Информатика и вычислительная техника.

Сколковский институт науки и технологий

Научный руководитель: д.к.н. Панченко А. И.

2025

# Предмет исследования

## Проблема

По мере того как изображения, создаваемые искусственным интеллектом, становятся все более убедительными, отличить реализм от вымысла становится все сложнее.

## Цель

Разработать численную меру реализма для обнаружения несоответствий контексту и здравому смыслу в визуальном контенте.

## Задачи

1. Разработать метод для получения оценки реализма изображения при помощи визуально-языковой модели.
2. Проверить метод на выборке реальных и *странных* изображений.
3. Проанализировать объяснение *странных* изображений.

## Постановка задачи

Предложить функцию проверки реализма изображения  
 $f_{\text{reality}} : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ , которая присваивает *reality score* изображениям следующим образом:

- ▶  $f_{\text{reality}}(I_r) > f_{\text{reality}}(I_w)$  для **нормальных** изображений  $I_r$  и **странных** изображений  $I_w$ .
- ▶ Учитывая порог  $\tau$ , функция  $f_{\text{reality}}$  может быть использована для классификации изображений:

Нормальное:  $f_{\text{reality}}(I_r) \geq \tau$ , Странное:  $f_{\text{reality}}(I_w) < \tau$ .

# Методы

## Метод на основе NLI



### 1. Выделение атомарных фактов

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

### 2. Попарное логическое следствие между фактами

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

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

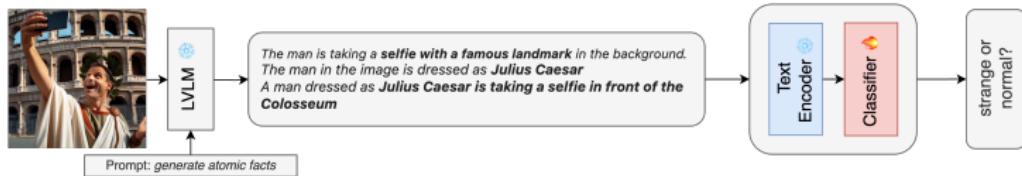
### 3. Агрегация

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

Методы агрегации: min, abs max, clustering.

# Методы

## Метод на основе механизма внимания (TLG)



### 1. Векторное представление фактов

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

### 2. Вычисление весов внимания

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

### 3. Получение предсказания

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

# Методы

## Линейный пробинг (LP)

- Передача изображения и запроса в модель.  $P$  “Provide a short, one-sentence descriptive fact about this image“

$$\text{LVLM}(I, P)$$

- Получение внутренних состояний модели

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

- Обучение логистической регрессии

$$\mathcal{C}_l(h_l(I, P)) \in [0, 1]$$

# Вычислительный эксперимент

## Данные



	WHOOPS!	WEIRD
# число изображений	204	824
# число категорий	26	12
# число подкатегорий	—	181
Классификация человеком	92%	82.22%

Таблица: Характеристики рассматриваемых выборок

# Результаты

Метод	# Всего	Режим	WHOOPS!	WEIRD
Человек	–	–	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	–	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	zero-shot	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
NLI	7B	zero-shot	72.55	60.00
LP - LLaVA	13B	fine-tuned	73.50	85.26
TLG	8B	fine-tuned	<b>73.54</b>	<b>87.57</b>
GPT-4o	–	zero-shot	79.90	81.64

# Результаты

Модель	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

Таблица: Линейный пробинг на WHOOPS! и WEIRD по моделям и вариантам входного запроса.

# Результаты

Method	#	Accuracy
<b>WEIRD → WHOOPS!</b>		
LP (+Prompt)	13B	72.06
LP (Image only)	13B	<b>75.00</b>
TLG	8B	74.02
<b>WHOOPS! → WEIRD</b>		
LP (+Prompt)	13B	74.69
LP (Image only)	13B	79.61
TLG	8B	<b>83.05</b>

Таблица: Перенос знаний между выборками. WEIRD → WHOOPS! значит что метод был обучен на выборке WEIRD и протестирован на WHOOPS!.

# Анализ

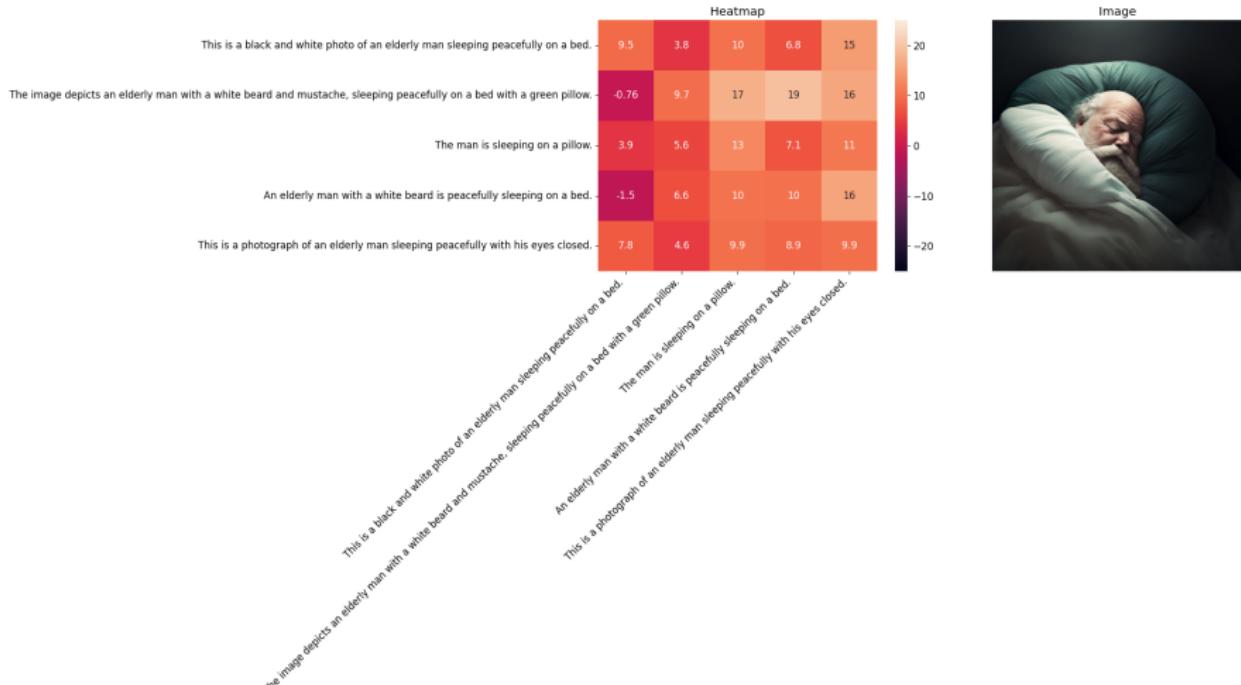


Рис.: Применение NLI метода к нормальному изображению

# Анализ

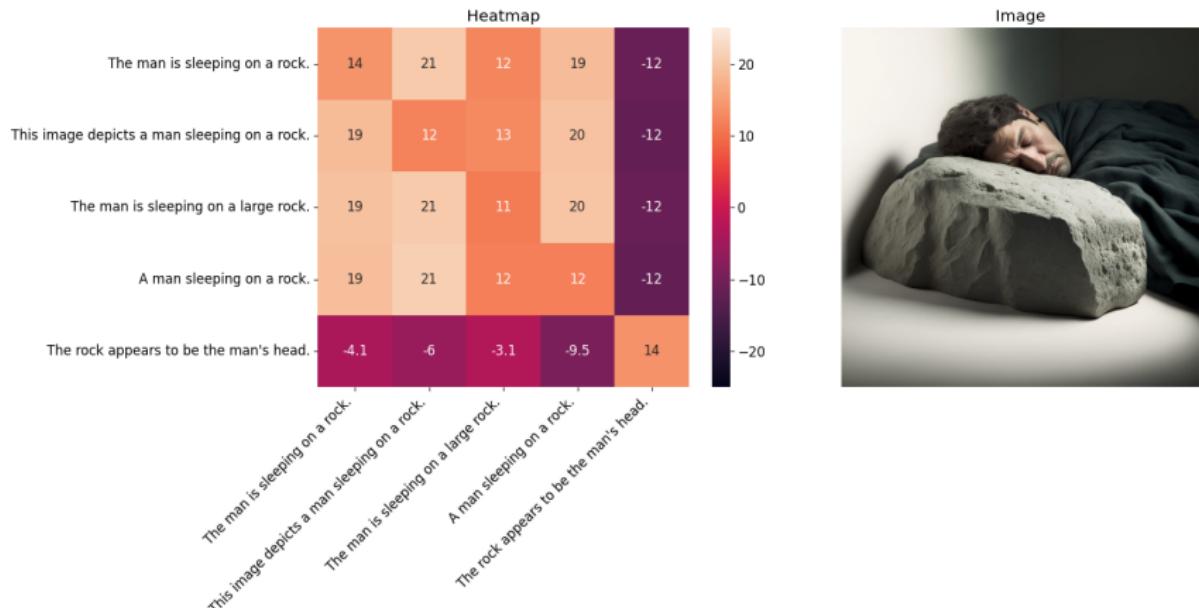


Рис.: Применение NLI метода к странному изображению

# Анализ

Вероятность	Значение
$\mathbb{P}(\text{weird} \mid \text{digital})$	0.76
$\mathbb{P}(\text{weird} \mid \text{hallucination})$	0.81
$\mathbb{P}(\text{weird} \mid \text{hallucination} \ \& \ \text{digital})$	0.93

Таблица: Условная вероятность модели предсказать *странные* при условии наличия галлюцинаций или маркерных слов в фактах.

Анализ таблицы сопряженности  $\chi^2$ -теста

Предсказание модели		Галлюцинация		Всего
		Нет	Да	
	Нормальное	78	10	88
	Странное	74	42	116
	Всего	152	52	204

$\phi$ -коэффициент = 0.27, р-значение=10<sup>-3</sup>

## Анализ

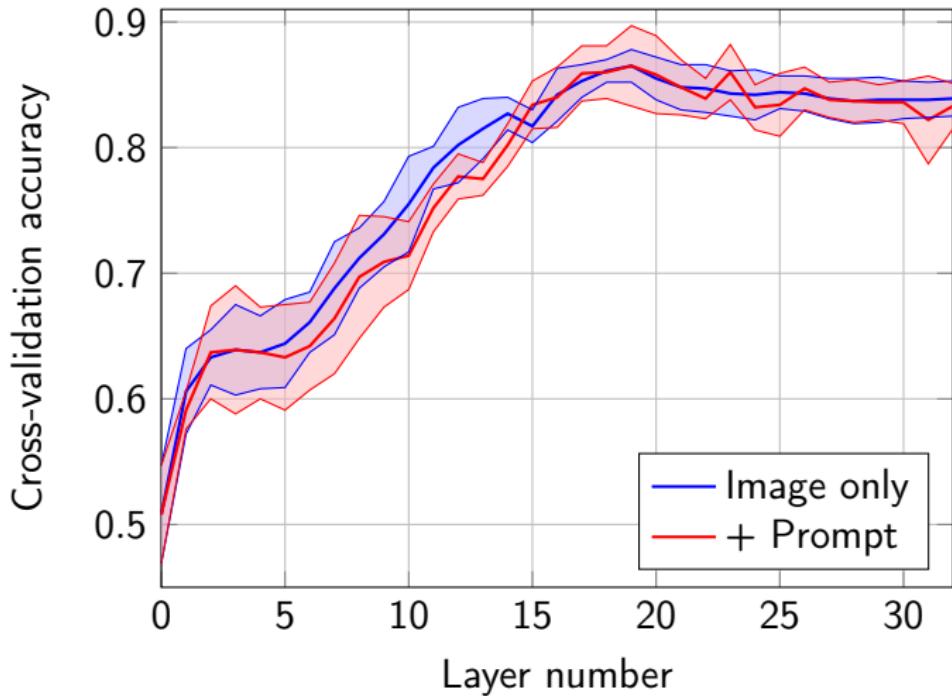


Рис.: Кросс-валидационная точность в зависимости от индекса слоя модели LLaVA 1.6 Vicuna 13B для линейного пробинга на WEIRD. Слои с наиболее релевантной информацией находятся посередине

## Результаты, которые выносятся на защиту

- ▶ Предложен метод на основе NLI для детектирования странных изображений благодаря противоречиям и галлюцинациям в тексте, сгенерированном визуально-языковой моделью.
- ▶ Продемонстрировано, что скрытые состояния визуально-языковых моделей несут в себе сигналы реализма. Линейный пробинг достигает наилучшего результата на WHOOPS! при обучении на WEIRD.
- ▶ Проведен анализ противоречий в моделях, вызванных галлюцинациями и маркерными словами, а также выяснена информативность признаков по слоям модели.

## Список работ автора по теме диплома

- ▶ Petrushina, K.<sup>1</sup>, Rykov, E.<sup>1</sup>, Titova, K., Panchenko, A., and Konovalov, V. Don't Fight Hallucinations, Use Them: Estimating Image Realism using NLI over Atomic Facts //arXiv preprint arXiv:2503.15948. - 2025.
- ▶ 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.

Вклад: разработка идеи статьи, базовые подходы, NLI подход, линейный пробинг, анализ влияния галлюцинаций на предсказания.

---

<sup>1</sup>Равный вклад

# Благодарность

Елисей Рыков  
Василий Коновалов

