

# 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.

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

2025

# Количественная мера визуального реализма

## 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.

# Формализация задачи оценки реализма изображений

## Определение

*Странное изображение* — это изображение, визуальное содержание которого нарушает устойчивые представления об окружающем мире.

## Задача

Предложить функцию проверки реализма изображения

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

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

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

# Подход на основе логического вывода между фактами



## 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

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

Методы агрегации: min, abs max, clustering.

# Методы агрегации и формальные предпосылки

## Методы агрегации:

1. **Min:**  $\min(s_{ij})$
2. **AbsMax:**  $\text{sign}(s_{ij^*}) \cdot \max |s_{ij}|$
3. **Clust:**  $\min(\mu_1, \mu_2)$ , где  $\mu_i$  — центроиды 2x кластеров

## Допущения:

1. **Нормальные изображения:** все факты логически согласованы:  $s_{\text{nli}}(f_i, f_j) \geq 0$
2. **Странные изображения:** существует пара противоречивых фактов:  $s_{\text{nli}}(f_i, f_j) < 0$
3. NLI надёжно различает противоречия и согласования
4. Сгенерированных фактов хватает для определения реалистичности

$$f_{\text{reality}}(I_r) \geq 0 \rightarrow \tau = 0$$

## Теорема иерархии методов агрегации

**Теорема (Петрушина, 2025).** Для изображения  $I$  верна следующая цепочка неравенств:

$$f_{\text{reality}}^{\text{absmax}}(I) < 0 \Rightarrow f_{\text{reality}}^{\text{clust}}(I) < 0 \Rightarrow f_{\text{reality}}^{\text{min}}(I) < 0$$

**Доказательство:**

1. Пусть  $s_{\min} < 0$ ,  $|s_{\min}| = \max |s_{ij}| \rightarrow s_{\min} \in S_1$
2. Предположим  $\mu_1 > 0$ , тогда  $\rho(s_{\min}, \mu_1) > \rho(\mu_1, t)$ ,  $t \in S_2$
3. Противоречие, значит,

$$\mu_1 < 0 \Rightarrow f_{\text{reality}}^{\text{clust}}(I) < 0 \Rightarrow f_{\text{reality}}^{\text{min}}(I) < 0$$

# Линейный пробинг (LP)

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

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

2. Получение внутренних состояний модели

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

3. Обучение логистической регрессии

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

# Данные



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

Таблица: Характеристики рассматриваемых выборок

# Результаты

NLI модель	#	Метод агрегации		
		min	absmax	clust
nli-deberta-v3-large	304M	<u>63.73</u>	62.75	<b>72.55</b>
nli-deberta-v3-base	86M	<u>60.78</u>	55.39	<b>61.76</b>
nli-deberta-v3-small	47M	<b>61.27</b>	58.33	<u>60.78</u>

Таблица: Сравнение различных NLI моделей и методов агрегации для 5 атомарных фактов на выборке WHOOPS!.

## Основные результаты

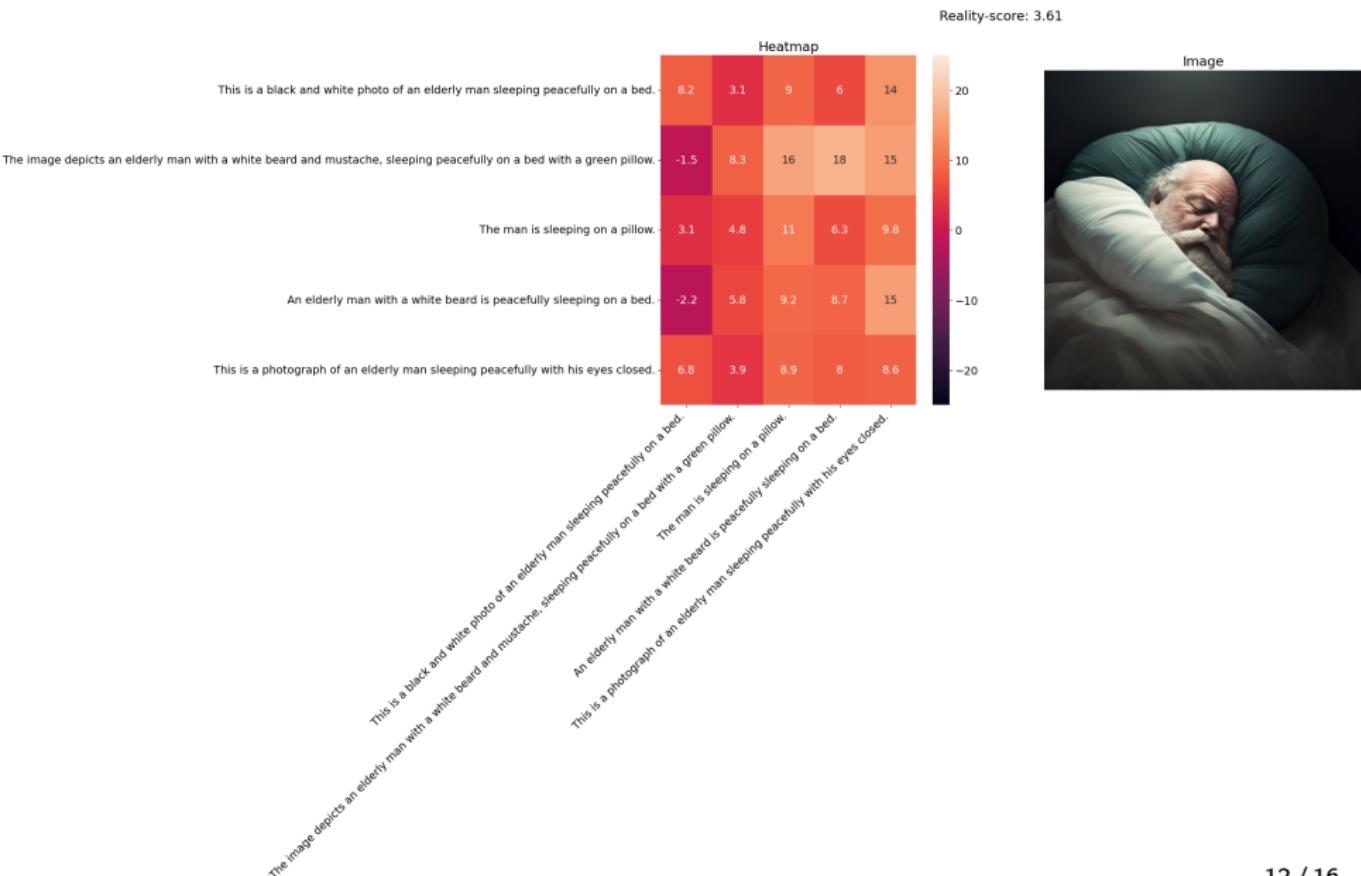
Метод	# Всего	Режим	WHOOPS!	WEIRD
Человек	—	—	92.00	82.22
BLIP2 FlanT5-XL	3.94B	fine-tuned	60.00	71.47
BLIP2 FlanT5-XXL	12.4B		73.00	72.31
BLIP2 FlanT5-XXL	12.4B	—	50.00	63.84
nanoLLaVA Qwen1.5 0.5B	1.05B	—	66.66	70.90
LLaVA 1.6 Mistral 7B	7.57B	—	56.86	61.18
LLaVA 1.6 Vicuna 7B	7.06B	zero-shot	65.68	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

## Обобщающая способность методов

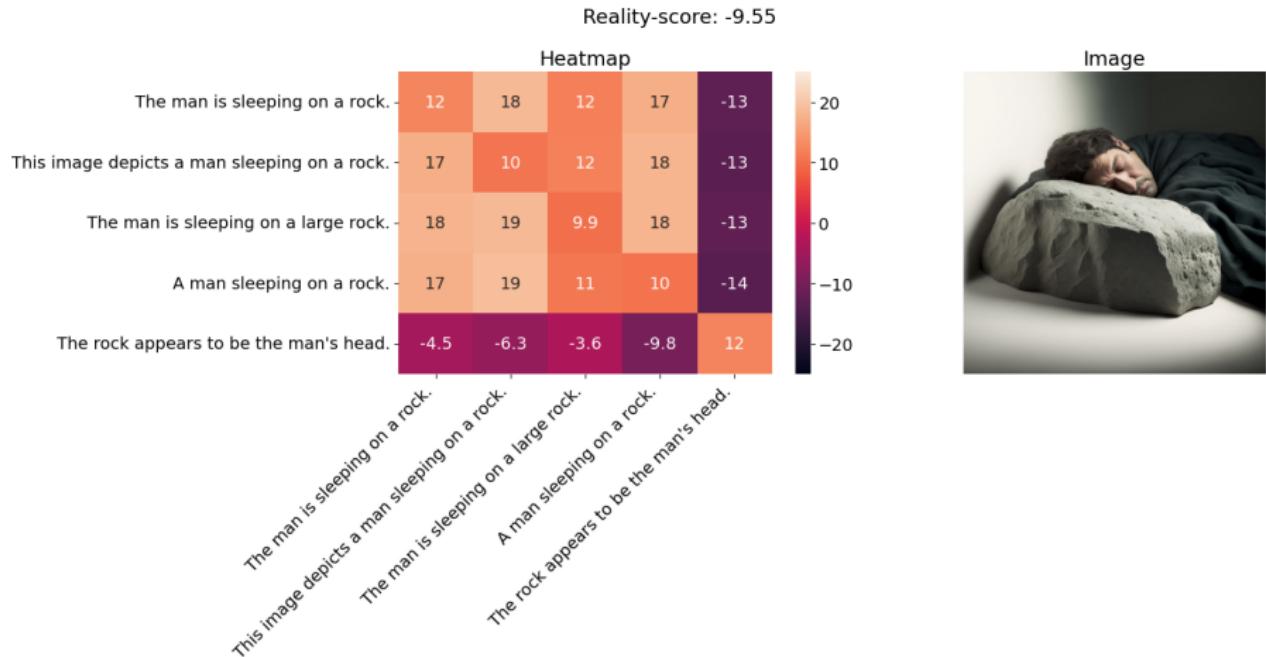
Method	#	Accuracy
<b>WEIRD → WHOOPS!</b>		
LP (+Prompt)	13B	72.06
LP (Image only)	13B	<b>75.00</b>
TLG	8B	<u>74.02</u>
<b>WHOOPS! → WEIRD</b>		
LP (+Prompt)	13B	74.69
LP (Image only)	13B	<u>79.61</u>
TLG	8B	<b>83.05</b>

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

# Применение NLI метода к нормальному изображению



# Применение NLI метода к странному изображению



## Феномены, лежащие в основе предсказаний модели

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>

## Результаты, которые выносятся на защиту

1. Предложен метод на основе NLI для детектирования странных изображений благодаря противоречиям и галлюцинациям в тексте, сгенерированном визуально-языковой моделью.
2. NLI показывает конкурентоспособный результат на WHOOPS! - 72.55. Линейный пробинг достигает наилучшего результата на WHOOPS! при обучении на WEIRD - 75.00, а также является второй по качеству методом на WEIRD.
3. Проведен анализ противоречий в моделях, вызванных галлюцинациями и маркерными словами.

## Список работ автора по теме диплома

1. Kseniia Petrushina<sup>1</sup>, Elisei Rykov<sup>1</sup>, 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.

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

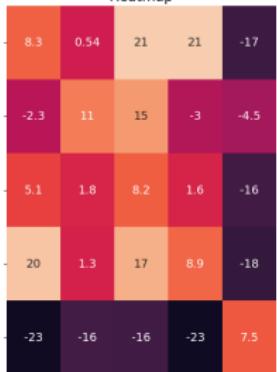
---

<sup>1</sup>Равный вклад

# FINE

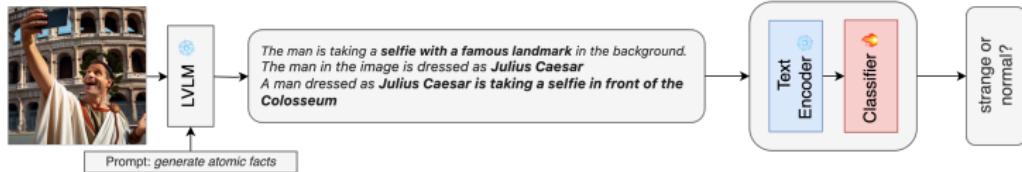
Reality-score: -9.6

Heatmap



This is a digitally created image of a cat with a lobster's body.  
The cat in the image has a lobster head.  
A stylized depiction of a cat with a lobster's body.  
The image is a digital artwork depicting a cat with a lobster's body.  
This is an artistic representation of a cat with the body of a shrimp, set in an underwater environment.

# Through the Looking Glass



## 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

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

# Подробные результаты по линейному пробингу

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.

## Анализ линейного пробинга

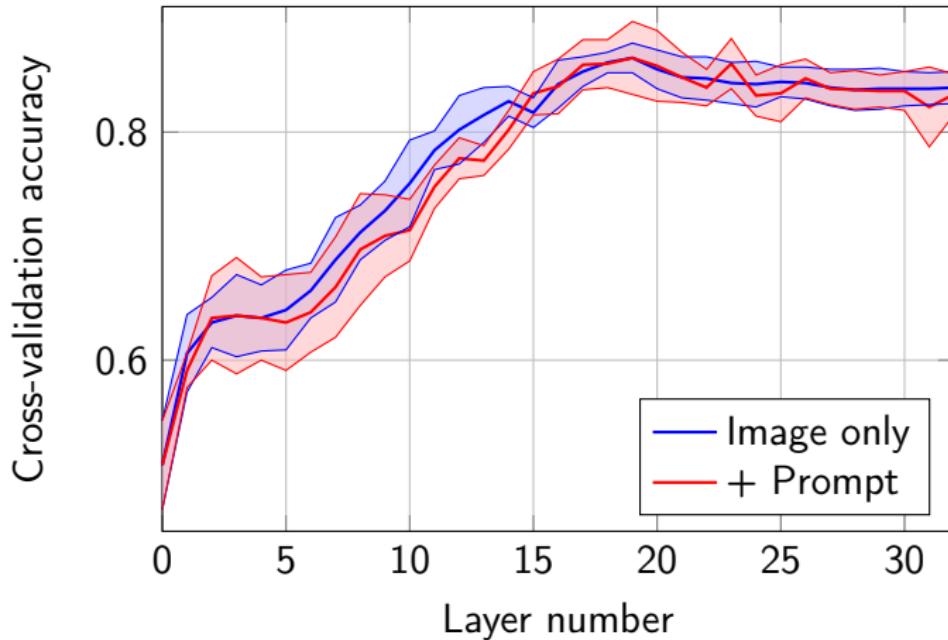


Рис.: Кросс-валидационная точность в зависимости от индекса слоя модели LLaVA 1.6 Vicuna 13B для линейного пробинга на WEIRD. Слои с наиболее релевантной информацией находятся посередине декодировщика.

# Пример

Reality-score: -18.21

Heatmap

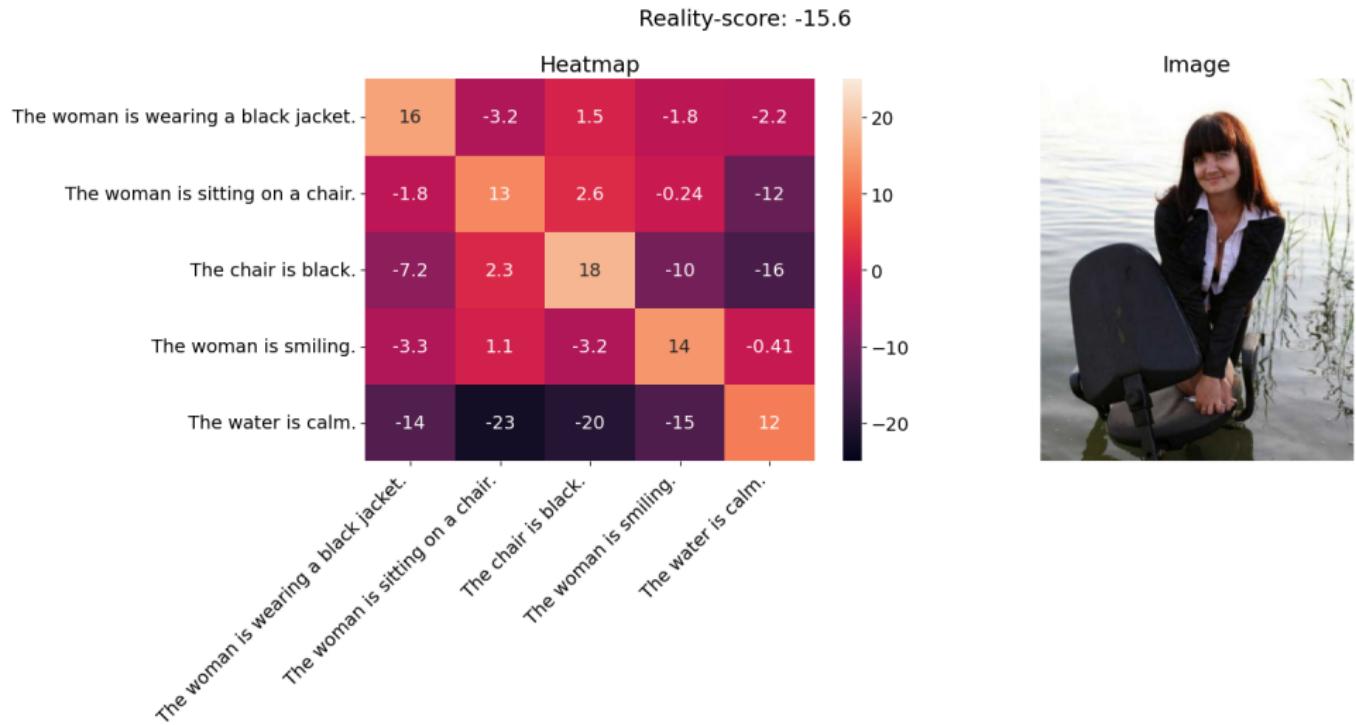


The man is ironing clothes on the back of a moving car.  
The car is yellow and is driving on a city street.  
The man is wearing a yellow shirt.  
There are pink banners hanging from the street lights.  
The car is passing a taxi cab.

Image



# Пример

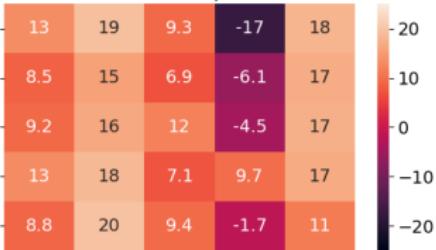


# Пример

Reality-score: 2.68

Heatmap

The image shows a series of traffic signs.

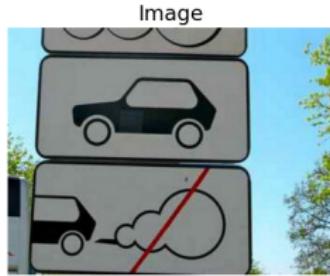


The signs are designed to convey information to drivers.

The signs are mounted on a pole.

The signs depict a car, a truck, and a cloud.

The signs are located outdoors.

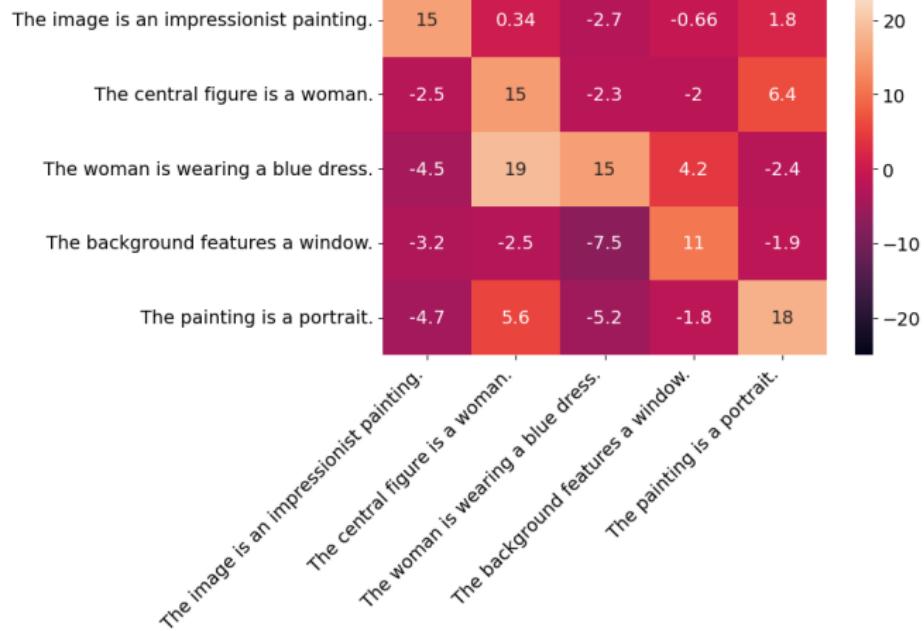


The image shows a series of traffic signs.  
The signs are designed to convey information to drivers.  
The signs are mounted on a pole.  
The signs depict a car, a truck, and a cloud.  
The signs are located outdoors.

# Пример

Reality-score: -1.36

Heatmap



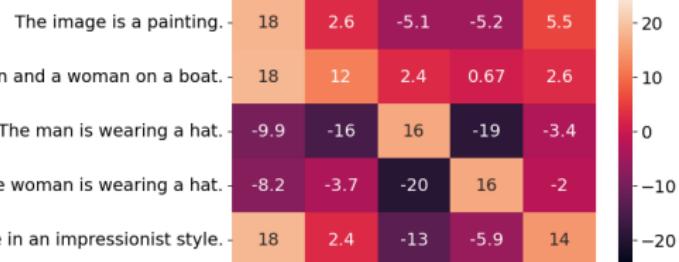
Image



# Пример

Reality-score: -5.28

Heatmap



Image



The image is a painting.  
The painting depicts a man and a woman on a boat.  
The man is wearing a hat.  
The woman is wearing a hat.  
The painting is done in an impressionist style.

# Пример

Reality-score: -2.46

Heatmap

The image is a painting.	18	2.6	-1.5	-1.8	5.5
The painting depicts a man and a woman on a boat.	18	12	1.4	0.86	2.6
The man in the center is wearing a hat.	-4.7	-17	11	-16	0.46
The woman on the left is wearing a hat.	-3.1	-0.9	-17	11	0.56
The painting is done in an impressionist style.	18	2.4	0.61	0.96	14



Image



The image is a painting.  
The painting depicts a man and a woman on a boat.  
The man in the center is wearing a hat.  
The woman on the left is wearing a hat.  
The painting is done in an impressionist style.

# Пример

Reality-score: -17.13

Heatmap

The image is a surrealist painting.

It features a horse with a human body.

The horse is holding a chalice.

The background is a cloudy sky.

The painting is by Salvador Dali.

