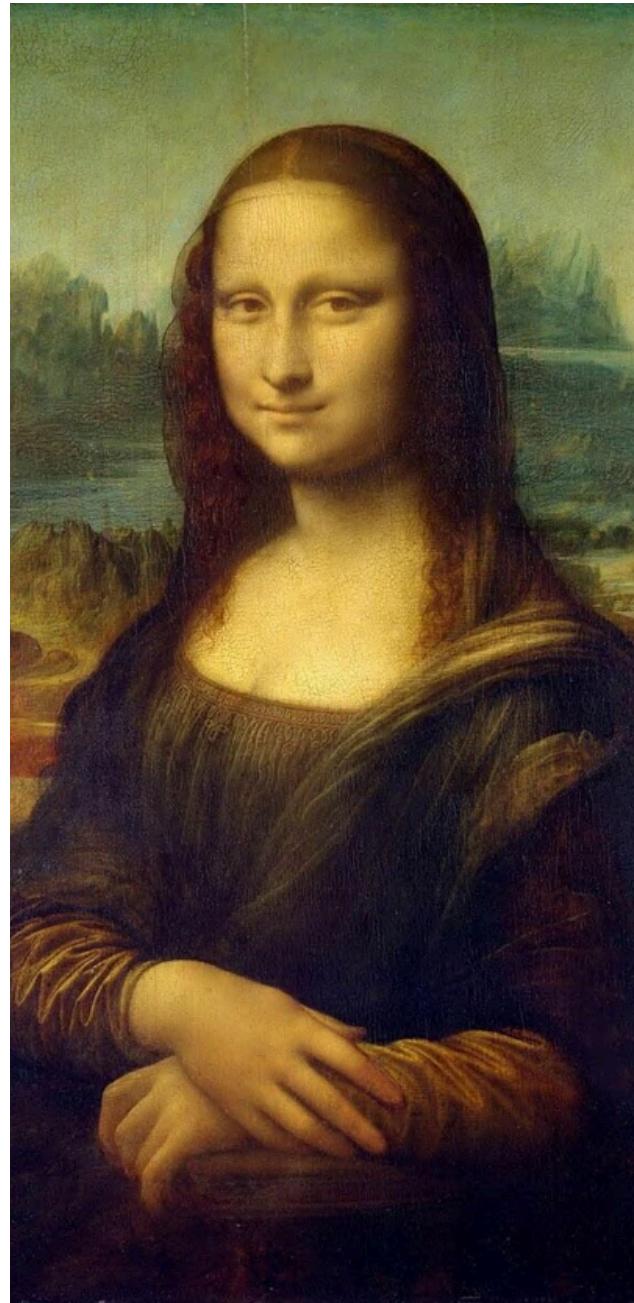


СЕРВИС РЕКОМЕНДАЦИЙ ПРОИЗВЕДЕНИЙ ИЗОБРАЗИТЕЛЬНОГО ИСКУССТВА

Левин Леонид

ПЛАН

- Вступление
- Цель проекта
- Основные этапы создания рекомендательной системы
- Обзор проделанной исследовательской работы
- Начальная версия прототипа. Анализ работы
- Улучшенная версия прототипа с LLama. Анализ работы
- Финальная версия прототипа с LLaVA. Анализ работы
- Планы по развитию проекта
- Личный вклад участников

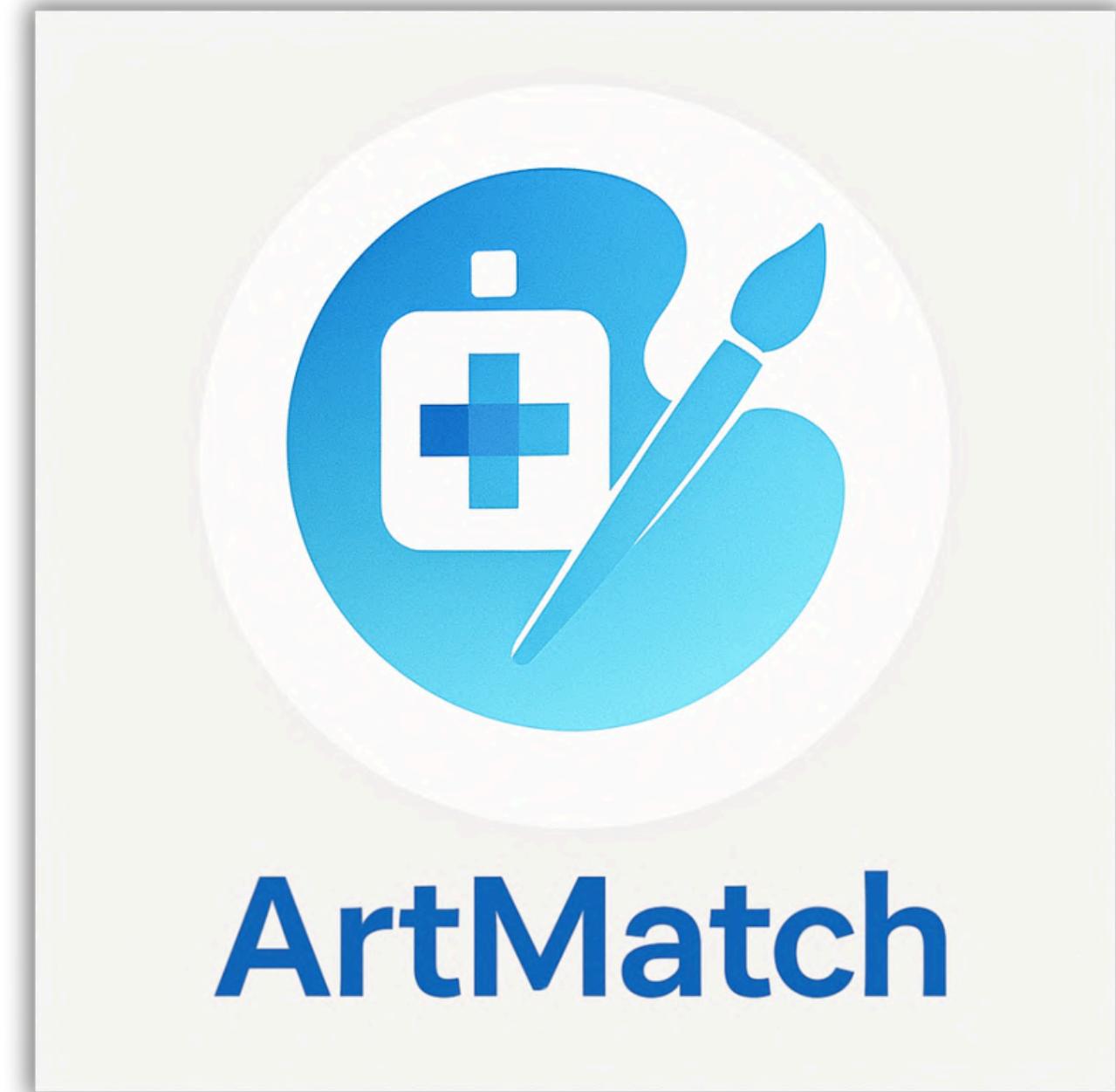


ВВЕДЕНИЕ

В современном цифровом пространстве ежедневно появляются тысячи новых произведений искусства — от классической живописи до цифровых иллюстраций и NFT. С ростом объёмов контента становится всё сложнее ориентироваться и находить именно те работы, которые откликаются на вкус и интересы конкретного пользователя. Это создаёт потребность в интеллектуальных системах, которые помогут каждому "найти своё искусство" среди огромного разнообразия.

ЦЕЛЬ ПРОЕКТА

Наша цель — разработать прототип рекомендательной системы, которая предлагает пользователю новые произведения искусства, опираясь на его визуальные предпочтения.



ОСНОВНЫЕ ЭТАПЫ

1

Получение эмбеддингов изображений

2

Генерация текстового описания изображения

3

Получение эмбеддингов текстового описания

4

**Формирование поисковой базы
(создание индексов)**

5

**Формирование рекомендаций
(поиск по индексам)**

ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА

ЭМБЕДДОРЫ

- Извлечение визуальных эмбеддингов осуществлялось с помощью модели CLIP.
- Для получения векторных представлений текстов были протестированы модели CLIP и SentenceTransformer (all-MiniLM-L6-v2).

ТЕКСТОВОЕ ОПИСАНИЕ

- Наибольшее внимание в исследовании было удалено генерации корректного текстового описания изображения, которое не только перечисляет объекты на картине, но также передаёт настроение, цветовую палитру и эмоциональный контекст.
- В рамках эксперимента были протестированы:
 - 1.BLIP
 - 2.BLIP-2
 - 3.GIT
 - 4.GIT + LLaMA-Tiny
 - 5.LLaVA

ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА

Подробнее про генерацию текстового описания

Вначале для генерации описания использовалась модель **BLIP**, предназначенная для создания коротких caption. В результате модель формировала простые подписи из 3–4 слов, указывающие лишь на основные объекты изображения.

Затем были протестированы модели с ≈3B параметров — **BLIP-2** и **GIT**.

- **BLIP-2** поддерживает генерацию по prompt, но на практике часто не давал ответа или выдавал некорректное описание (например, путала медведей с людьми, а лошадей с собаками).
- **GIT** продемонстрировал стабильную и логичную генерацию caption, поэтому был выбран для дальнейшей работы.

Поскольку **GIT** не поддерживает генерацию по prompt, была реализована собственная схема расширенного описания:

1. Один caption генерировался с **do_sample=False** — детерминированный, логичный результат.
2. Два caption — с **do_sample=True** для более описательных или второстепенных деталей.

Затем, с использованием Llama-Tiny, объединялись все три caption в одно связное описание. Специальный prompt подсказывал модели сохранить основную смысловую линию, добавляя детали без искажения смысла. В результате формировался полный и осмысленный текст.

На финальном этапе была выбрана мультимодальная модель **LLaVA-1.6** (в 4-битной квантизации для ускорения инференса):

<https://huggingface.co/unslloth/llava-v1.6-mistral-7b-hf-bnb-4bit>

Модель принимала изображение и текстовый prompt, генерируя подробное описание. Были протестированы различные формулировки prompt. Итоговый вариант: "**<image> Describe the artwork's genre, style, subject and colors in detail.**"

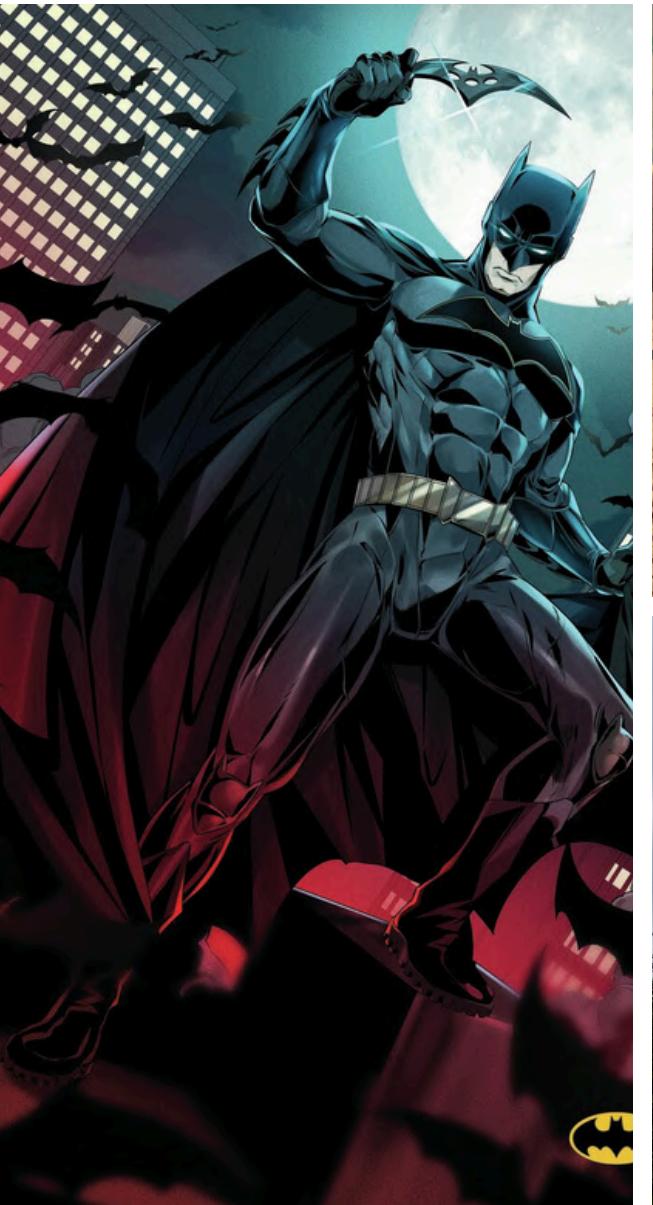
Он позволял получать не только перечень объектов, но и размышления о жанре, стиле и цветовой палитре произведения.

ПРОТОТИП CLIP + BLIP

- Эмбеддор изображения - **CLIP**
- Получение текстового описания - **BLIP**
- Эмбеддор текстового описание - **CLIP**
- Посторонение индексов - **FAISS**
(по косинусному сходству)

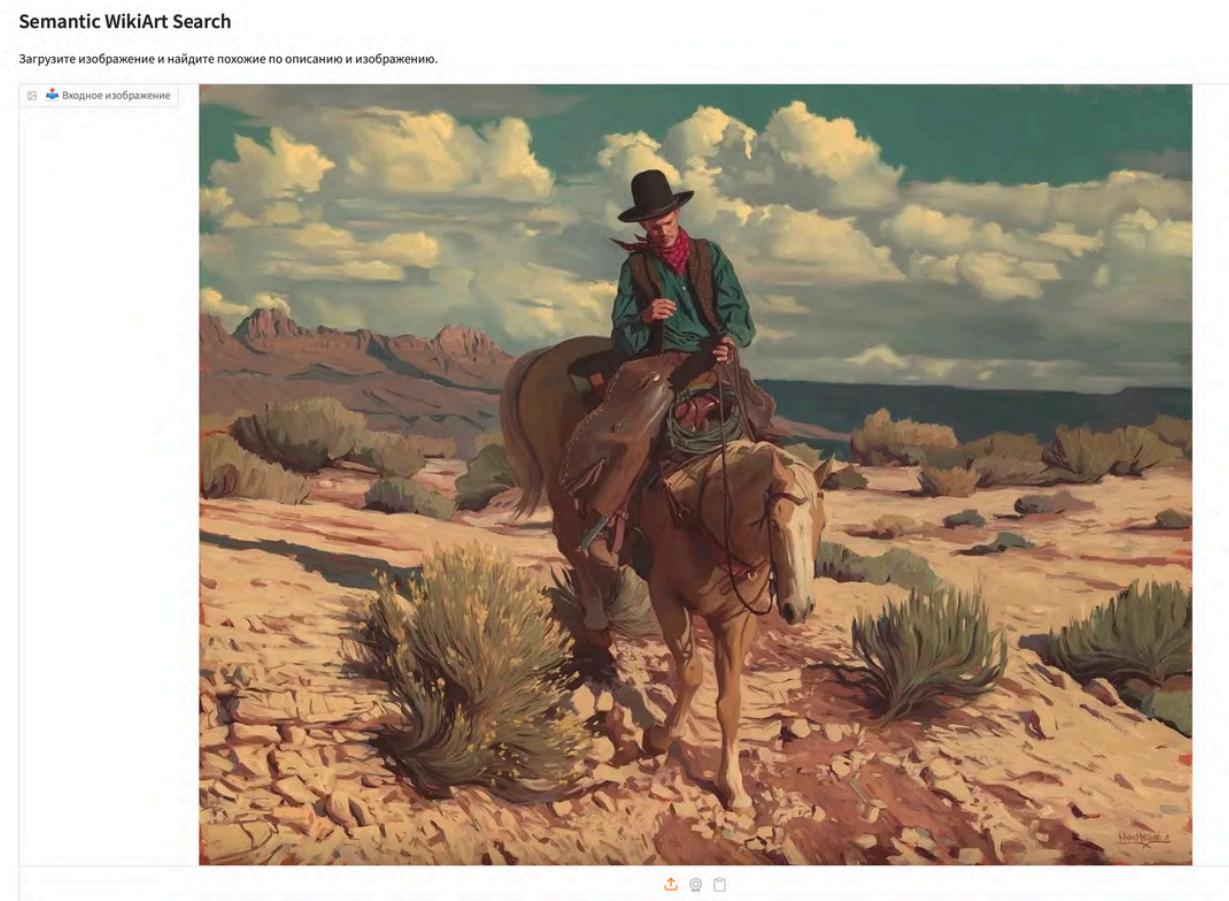
Для демонстрации работы рекомендательной системы были выбраны изображения разных стилей и жанров.

[Ссылка: видео работы модели BLIP](#)



ПРИМЕР 1

Входное изображение



Caption

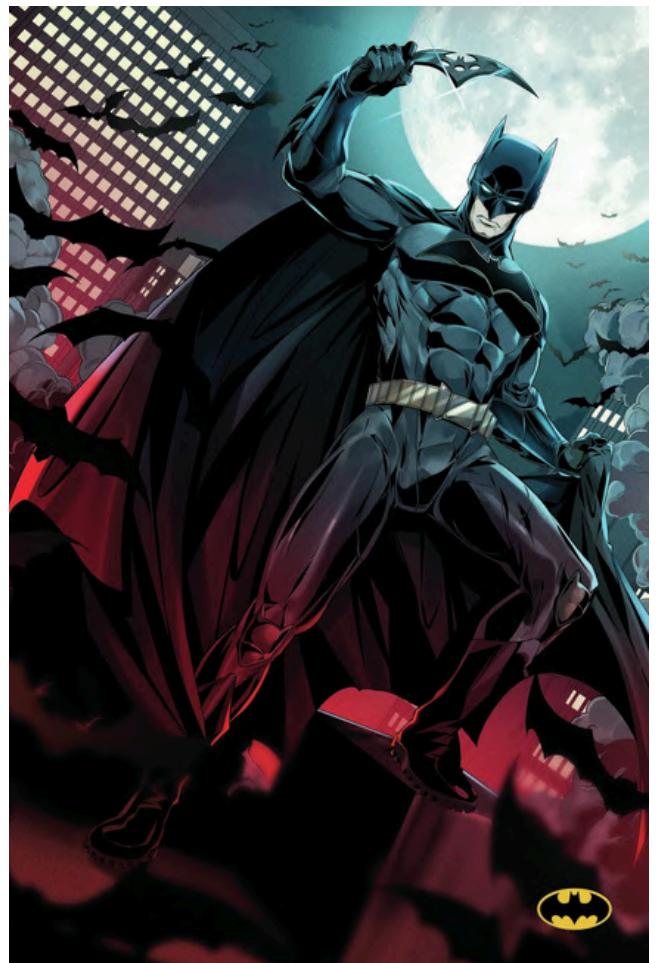
a painting of a man riding a horse

Рекомендация

The image shows two sections of recommended paintings from the Semantic WikiArt Search results. The top section, titled 'Похожие по описанию (текстовое сходство)', displays five images with IDs 3651, 2991, 2247, 905, and 629. The bottom section, titled 'Похожие по изображению (визуальное сходство)', displays five more images with IDs 42072, 63042, 34170, 46061, and 45011. Both sections include checkboxes for filtering by description or image similarity.

ПРИМЕР 2

Входное изображение



Caption

batman and bat in the city

Рекомендация

Похожие по описанию (текстовое сходство)

По описанию

ID: 47023 ID: 48534 ID: 41184 ID: 77183 ID: 55699

Похожие по изображению (визуальное сходство)

По изображению

ID: 67105 ID: 68105 ID: 65753 ID: 62757 ID: 66417

ПРИМЕР 3

Входное изображение



Caption

a castle with a tower and a walkway

Рекомендация

Похожие по описанию (текстовое сходство)

По описанию

ID: 33983 ID: 7807 ID: 4671 ID: 4362 ID: 2234

Похожие по изображению (визуальное сходство)

По изображению

ID: 33677 ID: 5459 ID: 58532 ID: 4261 ID: 35141

АНАЛИЗ РЕЗУЛЬТАТА BLIP

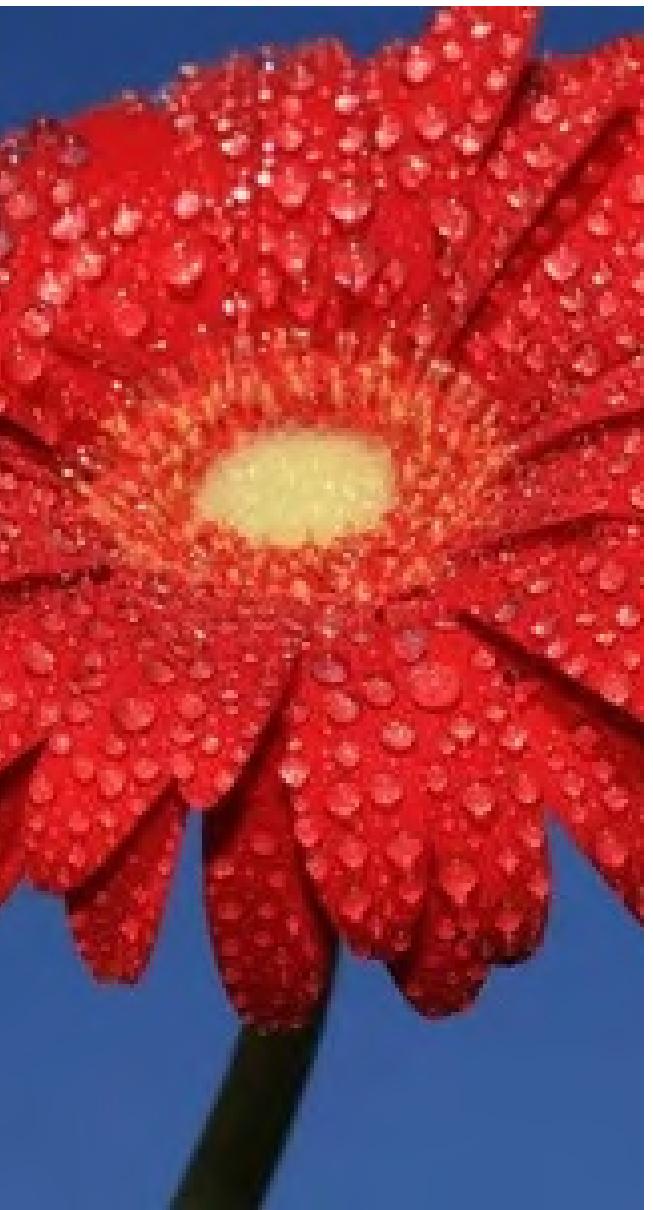
- BLIP хорошо справился с определением главных объектов на изображении. Однако может некорректно работать с случае известных картин: он выдает настоящее название картины, которое может не иллюстрировать, что изображено на изображении.
- Изображения для рекомендаций берутся из датасета WikiArt и там преимущественно хранятся изобразительные искусства прошлых веков, поэтому для рекомендаций современного творчества нужно добавить соответствующие датасеты.
- BLIP работает быстро и не занимает много ресурсов.

ПРОТОТИП CLIP + GIT + LLAMA

- Эмбеддор изображения - **CLIP**
- Получение текстового описания - **GIT + LLAMA**
- Эмбеддор текстового описание - **SentenceTransformer**
- Построение индексов - **FAISS**
(по косинусному сходству)

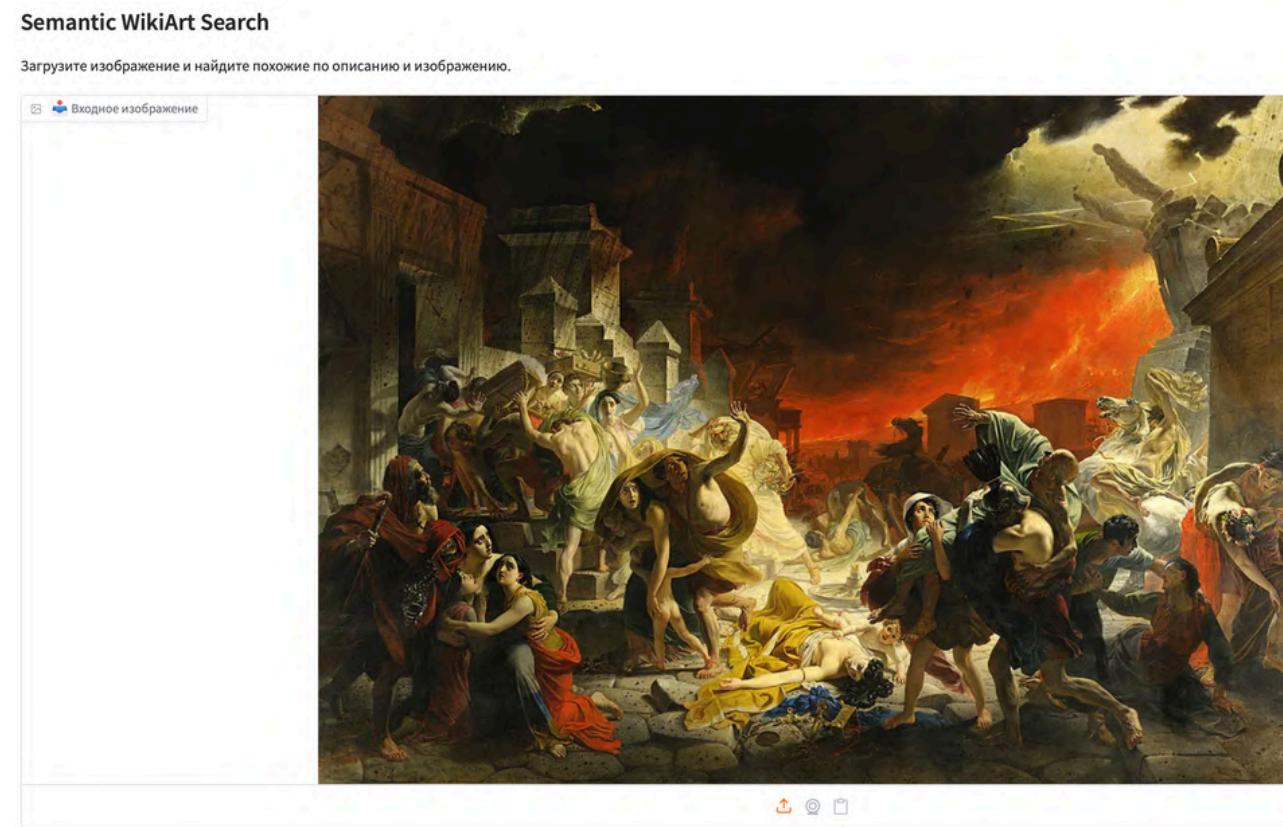
Для демонстрации работы рекомендательной системы были выбраны изображения разных стилей и жанров.

[Ссылка: видео работы модели GIT+LLama](#)



ПРИМЕР 1

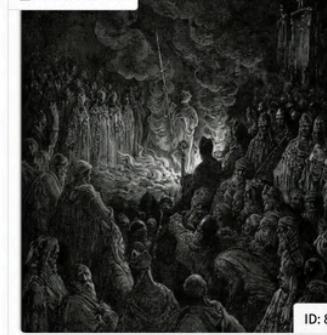
Входное изображение



Рекомендация

Похожие по описанию (текстовое сходство)

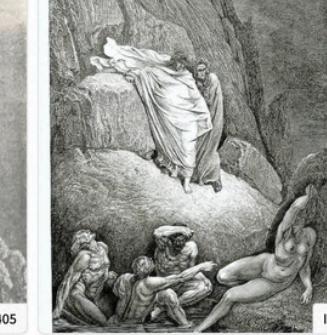
По описанию



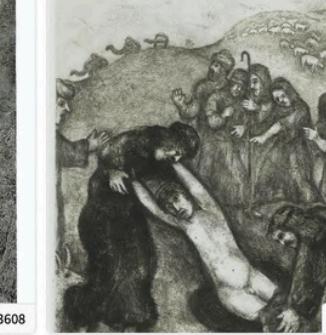
ID: 8940



ID: 6405



ID: 3608



ID: 3816



ID: 3036

Похожие по изображению (визуальное сходство)

По изображению



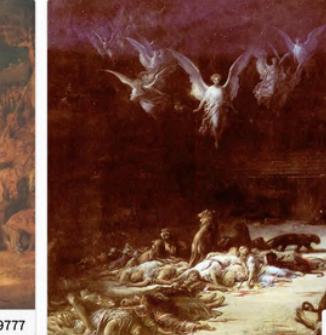
ID: 2647



ID: 5518



ID: 9777



ID: 7058



ID: 120

Caption

The burning of the city:

The city of Rome was engulfed in flames as the Roman army marched towards the city. The streets were littered with debris and the air was thick with smoke. The people of Rome were terrified, as they knew that the city was about to be destroyed. The soldiers were determined to bring the city down, and they were willing to risk their lives to achieve their goal. The fire was so intense

ПРИМЕР 2

Входное изображение



Рекомендация

Похожие по описанию (текстовое сходство)

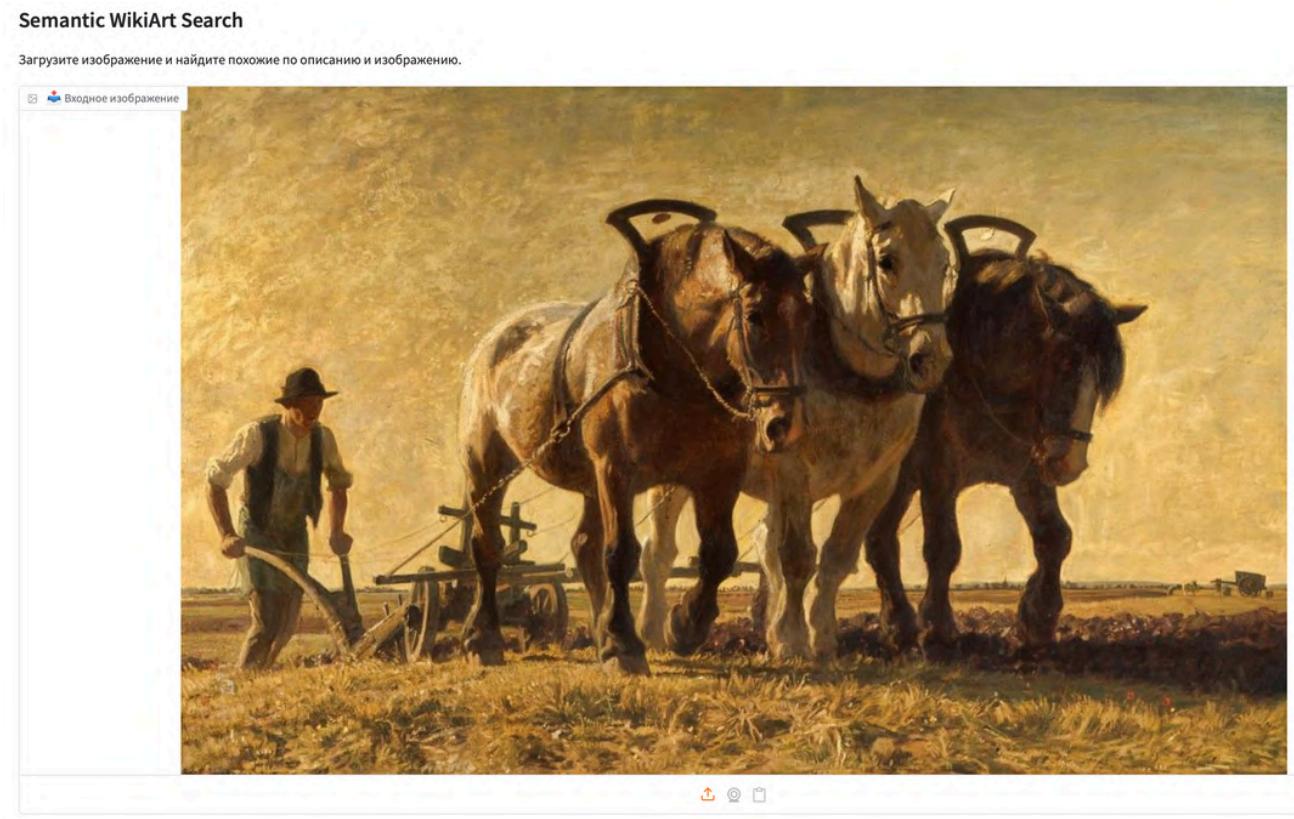
Похожие по изображению (визуальное сходство)

Caption

The morning sun falls on the flower, casting a soft, golden glow over its petals. The dew drops glisten in the light, creating a delicate, iridescent effect. The petals are soft and velvety, with a subtle, delicate texture that contrasts with the rough, sharp edges of the leaves. The flower's delicate beauty is a testament to the beauty of nature, a reminder that even the smallest

ПРИМЕР 3

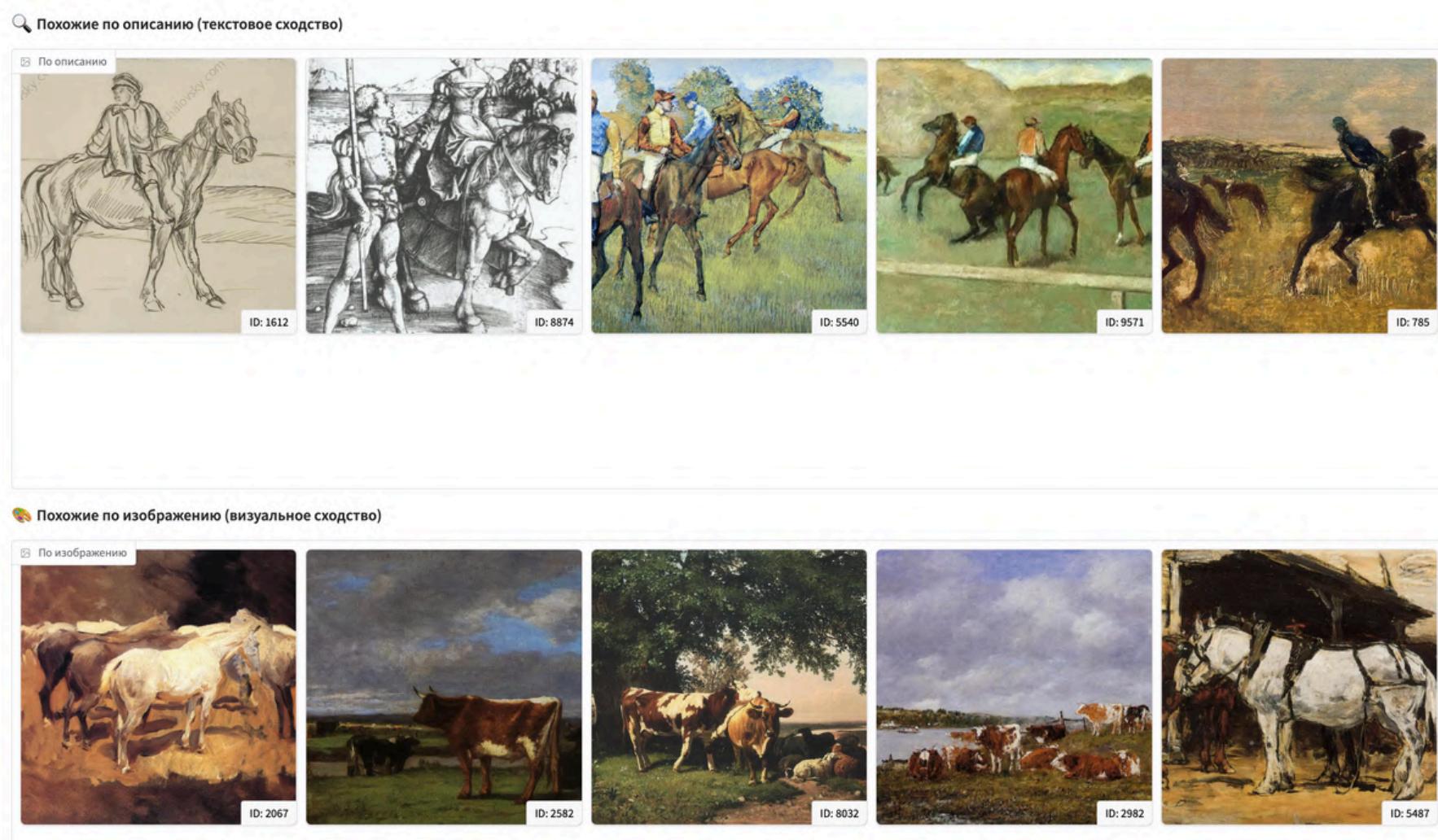
Входное изображение



Caption

The sun is shining brightly, casting a warm glow over the fields. The horses are pulling a plow, their hooves pounding the earth with each step. The air is crisp and fresh, and the scent of freshly turned soil fills the air. The plow is in the field, ready to begin the harvest.

Рекомендация



АНАЛИЗ РЕЗУЛЬТАТА GIT+LAMMA

- Git+LLama хорошо справились с задачей развернутого описания, передав не только основные объекты картины, но и описав их, также в большинстве случаев есть описание второстепенных деталей и указана цветовая палитра.
- Иногда в caption может встречаться мусор в виде специальных символов или слов, но такие случаи можно обработать.
- Данная модификация GIT+LAMMA отлично справляется с задачей создания текстового описания, при этом не требуем много ресурсов

ПРОТОТИП CLIP + LLaVA-1.6

- Эмбеддор изображения - **CLIP**
- Получение текстового описания - **GIT + LLaVA**
- Эмбеддор текстового описание - **SentenceTransformer**
- Построение индексов - **FAISS**
(по косинусному сходству)

Для демонстрации работы рекомендательной системы были выбраны изображения разных стилей и жанров.

Ссылка: [видео работы модели LLaVA](#)

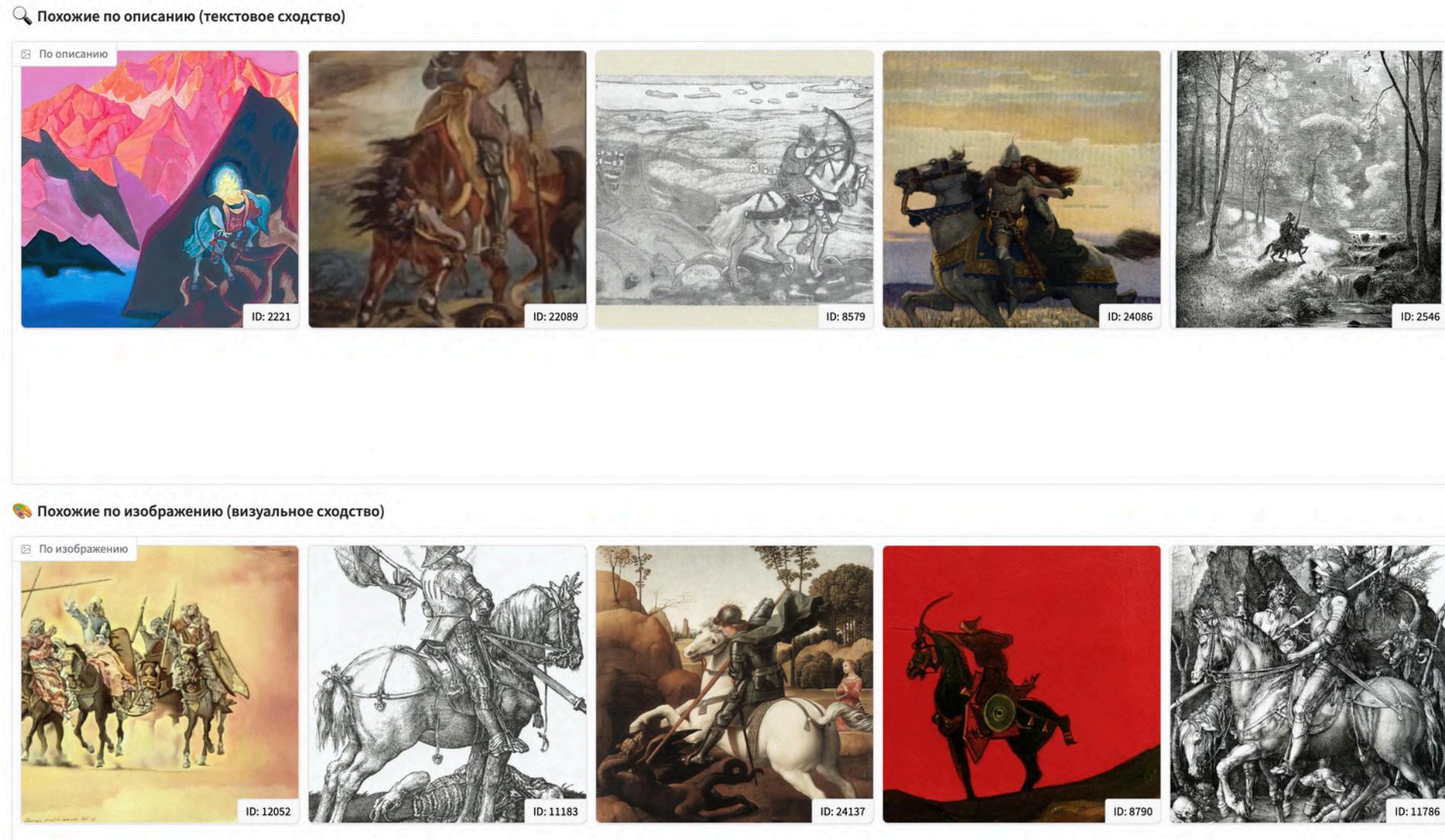


ПРИМЕР 1

Входное изображение



Рекомендация



Caption

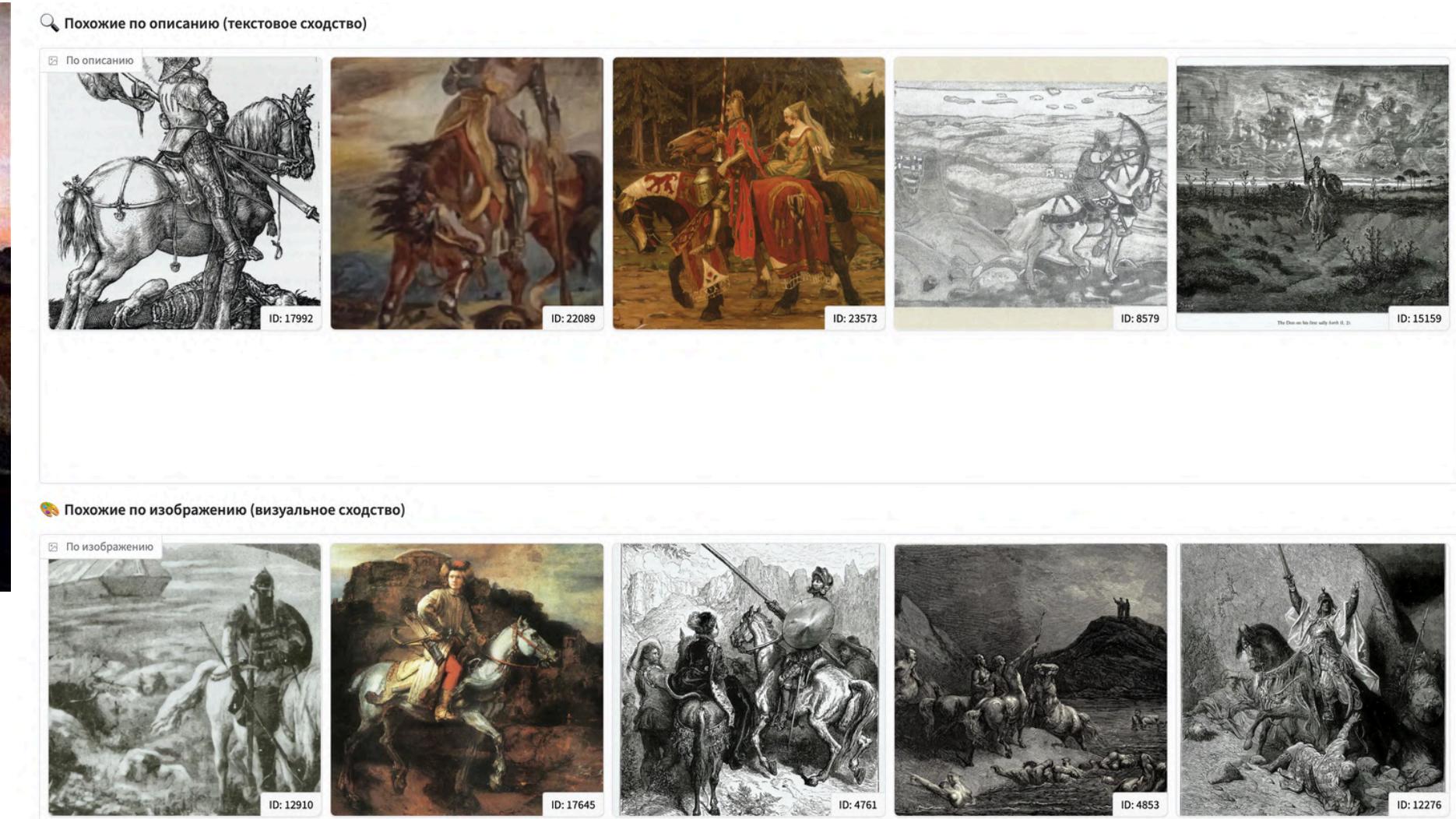
The artwork is a digital painting that falls under the fantasy genre. The style of the painting is realistic with a focus on dramatic lighting and shadows, which gives the image a sense of depth and atmosphere. The subject of the painting is a knight riding a white horse. The knight is wearing a suit of armor and is holding a flag. The colors used in the painting are predominantly dark and muted, with the white horse and armor standing out against the darker background. The overall mood of the painting is one of intensity and drama.

ПРИМЕР 2

Входное изображение



Рекомендация



Caption

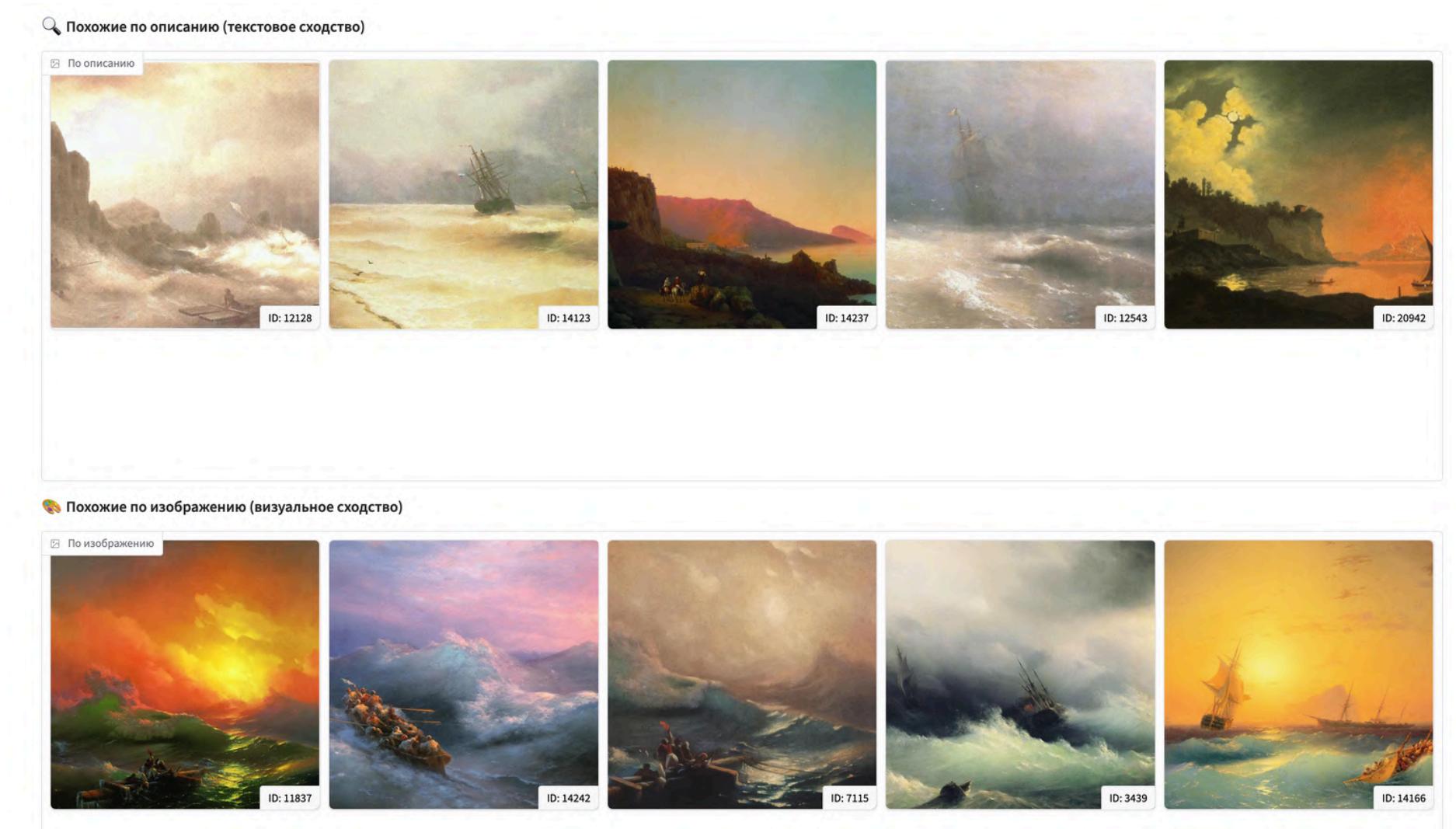
The artwork is a painting that falls under the genre of fantasy or historical art. The style of the painting is realistic, with a focus on detail and accuracy in the depiction of the subject matter. The subject of the painting is a knight, who is riding a white horse. The knight is dressed in armor and is holding a lance. The background of the painting features a rocky landscape with a stone monument. The colors used in the painting are predominantly earth tones, with the knight and horse standing out against the muted colors of the background. The overall composition of the painting suggests a narrative, possibly depicting a moment of triumph or victory for the knight.

ПРИМЕР 3

Входное изображение



Рекомендация



Caption

The artwork is a painting that falls under the genre of landscape art. The style of the painting is realistic, with a focus on capturing the natural beauty of the scene. The subject of the painting is a large body of water, possibly an ocean or a sea, with a boat floating on its surface. The colors used in the painting are predominantly dark and green, with the sky and water creating a dramatic contrast. The sun is depicted as a bright, fiery orange, adding a sense of warmth and light to the otherwise cool and dark palette. The overall composition of the painting, with its emphasis on the interplay of light and shadow, suggests a deep appreciation for the natural world and the power of nature.

АНАЛИЗ РЕЗУЛЬТАТА LLaVA

- В рамках проекта была использована квантизованная версия модели LLaVA-1.6, что позволило значительно снизить требования к вычислительным ресурсам без существенной потери качества генерации. Для достижения стабильных и высокоточных результатов также был тщательно подобран промпт, обеспечивающий комплексное описание изображений. Сгенерированные тексты охватывают не только основные объекты сцены, фон, цветовую палитру и мелкие детали, но также идентифицируют художественный стиль, жанр произведения и эмоциональное воздействие, которое оно оказывает на зрителя.
- Поскольку локальный запуск модели сопровождается высокой задержкой обработки, для ускорения инференса была использована видеокарта NVIDIA A100 с объёмом памяти 80 ГБ. Это обеспечило необходимую производительность при работе с высокоточной генерацией описаний.

ПЛАНЫ ПО РАЗВИТИЮ ПРОЕКТА

ВЗАИМОДЕЙСТВИЕ С ПОЛЬЗОВАТЕЛЕМ

Для объективной оценки качества эмбеддингов планируется реализовать механизм сравнения различных эмбеддеров: пользователю будет предоставлена возможность выбора наиболее релевантного варианта. Аналогичная стратегия будет применяться при интеграции более сложных моделей.

УЛУЧШЕНИЕ СИСТЕМЫ

В рамках улучшения системы планируется расширение метаинформации. Каждому изображению будут присваиваться дополнительные теги, такие как жанр, художественный стиль, эмоциональная окраска и другие характеристики.

ЗАКЛЮЧЕНИЕ

- В рамках проекта была протестирована и оптимизирована цепочка моделей для генерации текстового описания изображений.
- Переход от BLIP к связке GIT + LLama позволил существенно повысить информативность описаний при умеренных ресурсных затратах.
- Финально была интегрирована мультимодальная модель LLaVA-1.6, способная по изображению и промпту генерировать подробное описание, включая жанр, стиль, цветовую палитру и эмоциональную составляющую.
- Использование 4-битной квантизации и GPU A100 позволило ускорить инференс без потери качества.
- Результаты легли в основу системы, пригодной для построения рекомендательной системы произведений изобразительных искусств.

личный вклад

Левин Леонид

- Предобработка датасета **WikiArt**
- Исследование и сравнение моделей генерации описаний: **BLIP, BLIP-2, GIT, LLama-Tiny, LLaVA-1.6**
- Разработка схемы объединения caption'ов в связный текст с помощью **LLama-Tiny**
- Настройка prompt'ов для улучшения качества описаний
- Построение эмбеддингов и индексов с использованием **FAISS** для дальнейших рекомендаций
- Разработка графического интерфейса с использованием фреймворка **Gradio**
- Тестирование моделей
- Подготовка презентации, текстовых материалов

ССЫЛКИ НА МАТЕРИАЛЫ



GitHub



Видео

**СПАСИБО
ЗА
ВНИМАНИЕ**