ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ "НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ "ВЫСШАЯ ШКОЛА ЭКОНОМИКИ" ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Янаков Дмитрий Спартакович

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Нейросетевой подход для оценки качества генерации видео

Neural Network Approach for Video Generation Quality Assessment

по направлению подготовки 01.04.02 Прикладная математика и информатика

образовательная программа «Финансовые технологии и анализ данных»

Научный руководитель

PhD, доцент базовой кафедры ПАО Сбербанк
"Финансовые технологии и анализ данных"

_____

А.А. Масютин

Соруководитель

ООО "Умное пространство",
руководитель группы
развития генеративных технологий

_____

Р.В. Лисов

Студент

_____

Д.С. Янаков

Москва 2025

Аннотация

В данной магистерской диссертации рассматривается задача оценки качества генерации видео. В последние годы заметно выросло количество моделей, способных генерировать видео на основе текстовых описаний и/или изображений. Однако оценка качества таких видео остается сложной задачей, поскольку существующие метрики плохо коррелируют с восприятием человека. Целью работы является разработка модели, которая сможет предсказывать качество сгенерированного видео, опираясь как на входное текстовое условие, так и на саму видеопоследовательность.

Для достижения поставленной цели были решены следующие задачи: проведен анализ методов, оценивающих изображения с текстовым описанием и сравнивающих изображения друг с другом. Собран и обработан оригинальный набор данных, состоящий из пар «текстовый запрос – видео» и соответствующих оценок качества, реализованы и обучены модели для регрессии и классификации оценок. Методологическая основа включает использование трансформерных архитектур нейросетей, многомодальных подходов к обработке информации и методов обучения с учителем. Эмпирической базой послужили данные от популярных open-source генераторов видео, а также человеко-ориентированные оценки, собранные с помощью краудсорсинга.

Предложенный подход может быть использован как компонент обратной связи при обучении генеративных моделей, что открывает возможности для их дальнейшего совершенствования без участия человека в цикле обучения. Работа выполнена на стыке компьютерного зрения, обработки естественного языка и машинного обучения, что подчеркивает ее научную и практическую значимость.

Abstract

This master's thesis addresses the problem of evaluating the quality of video generation. In recent years, there has been a significant increase in the number of models capable of generating videos based on textual descriptions and/or images. However, assessing the quality of such videos remains a challenging task, as existing metrics poorly correlate with human perception. The aim of this work is to develop a model that can predict the quality of generated videos based on both the input text condition and the video sequence itself.

To achieve the stated goal, the following tasks were addressed: an analysis of methods for evaluating images with textual descriptions and comparing images with each other was conducted. An original dataset consisting of "prompt – video" pairs and corresponding quality scores was collected and processed. Models for regression and classification of scores were implemented and trained. The methodological foundation includes the use of transformer-based neural network architectures, multimodal approaches to information processing, and supervised learning methods. The empirical basis consisted of data from popular open-source video generators, as well as human-oriented evaluations collected through crowdsourcing.

The proposed approach can be used as a feedback component during the training of generative models, opening up opportunities for their further improvement without human involvement in the training loop. The work lies at the intersection of computer vision, natural language processing and machine learning, highlighting its scientific and practical significance.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В последние годы наблюдается стремительное развитие методов искусственного интеллекта, особенно в области компьютерного зрения и обработки естественного языка. Одним из наиболее активно развивающихся направлений стало генеративное моделирование – создание новых данных, таких как изображения, тексты, аудио и видео, на основе заданных условий или случайных шумов. В компьютерном зрении исторически первыми на этом пути стали модели, способные генерировать изображения по текстовым описаниям (Text2Image) – такие как DALL-E [1] и Stable Diffusion [2]. Эти технологии быстро нашли применение в дизайне, рекламе и других сферах.

С развитием вычислительных мощностей и архитектур глубоких нейронных сетей началось активное продвижение от статических изображений к более сложным данным – видеопоследовательностям. Современные модели, такие как Runway, Pika, Sora, Qwen и другие, уже демонстрируют впечатляющие результаты в задачах Text2Video и Image2Video. Однако, несмотря на прогресс, качество генерируемых видео все еще остается сложным для оценки с точки зрения человеческого восприятия. Отсутствие устойчивых метрик и объективных критериев оценки качества генерации затрудняет сравнение моделей, их оптимизацию и внедрение в реальные приложения.

Актуальность данной работы обусловлена необходимостью разработки эффективных методов оценки качества сгенерированного видео. Такая оценка может служить важной обратной связью при обучении генеративных моделей, позволяя автоматически выбирать лучшие варианты, улучшать воспринимаемое качество вывода и повышать общую эффективность систем генерации.

Целью данной магистерской диссертации является разработка модели с использованием нейросетей, способной с хорошей точностью оценивать качество генерации видео на основе текстового запроса (промпта). Для достижения этой цели были поставлены следующие задачи:

1. Обзор существующих подходов к оценке качества генерации изображений на основе заданного промпта и методов, позволяющих сравнивать изображения друг с другом.
2. Сбор и подготовка репрезентативного набора данных, состоящего из пар «промпт – сгенерированное видео» и соответствующих оценок качества.
3. Разработка и обучение модели, способной давать согласованные и коррелирующие с человеческой оценкой прогнозы генерации видео, а также последующий анализ полученных результатов.

Структура дипломной работы следующая:

1. в первой главе рассказывается про современные подходы к генерации изображений и видео;
2. во второй главе представлен обзор используемых методов, которые легли в основу модели для оценки качества генерации видео;
3. третья глава посвящена проведенному экспериментальному исследованию с анализом полученных результатов.

На данный момент в научном сообществе отсутствуют устоявшиеся стандарты и общепринятые метрики, позволяющие объективно и точно оценивать качество видеогенерации. В связи с этим аналогов предложенному направлению исследования практически нет. Таким образом, работа посвящена решению актуальной и малоизученной проблемы, где представленный подход может стать основой для дальнейших исследований и создания стандартизированных метрик оценки качества генерации видео.

ГЛАВА 1. ВЕРОЯТНОСТНОЕ МОДЕЛИРОВАНИЕ

1.1. Дискриминативные и генеративные модели

Для начала определимся, что мы будем называть моделью. В машинном обучении мы обычно имеем дело с тремя видами переменных: x – наблюдаемые переменные, y – целевые переменные, θ – параметры алгоритма прогнозирования. Одна из распространенных постановок задач машинного обучения состоит в следующем. Дана выборка независимых одинаково распределенных объектов. Описание каждого объекта задается парой вида (x, y). Анализируя обучающую выборку, необходимо подобрать алгоритм (подстроить его параметры θ), который позволил бы по x спрогнозировать значение y. Для решения этой задачи часто вводят модель, описывающую способ порождения данных. На вероятностном языке такой моделью является совместное распределение на переменные x, t и θ. Традиционно выделяют 2 вида моделей:

1. Генеративная модель

$$p(x, y, \theta) = p(x, y \mid \theta)p(\theta) = p(y \mid x, \theta)p(x \mid \theta)p(\theta)$$

2. Дискриминативная модель

$$p(y, \theta \mid x) = p(y \mid x, \theta)p(\theta)$$

Генеративная модель более общая, поскольку если нам известно $p(x, y, \theta)$, то мы всегда можем получить $p(y, \theta \mid x)$. Обратное, вообще говоря, неверно. Кроме того, несомненным достоинством генеративной модели является возможность порождать новые x, или же пары (x, y). В рамках дискриминативной модели такое сделать не получится.

Однако, в традиционном машинном обучении чаще рассматривают дискриминативные модели. При этом на практике часто оказывается так, что пространство целевых переменных проще, чем пространство наблюдаемых переменных. Поэтому традиционные дискриминативные модели обычно на порядок проще генеративных, так как они решают гораздо более простую задачу. Например, пусть пространство наблюдаемых переменных – картины известных художников, а пространство целевых переменных – имена этих

художников. Тогда определить автора по картине (дискриминативная задача) проще, чем нарисовать картину в стиле автора (генеративная задача). Тем не менее, многие современные дискриминативные модели на практике такие же сложные, как и генеративные, потому что пространство целевых переменных не проще пространства наблюдаемых переменных. Например, в задаче машинного перевода с русского на английский: x – предложение на русском языке, y – предложение на английском.

Исторически развитие генеративных моделей прошло несколько этапов, начиная от простых вероятностных моделей, таких как наивный байесовский классификатор, до современных глубоких нейросетевых архитектур, способных генерировать данные высокого качества.

1.2. Основные подходы к генерации изображений

Наиболее известными семействами генеративных моделей в области компьютерного зрения являются генеративно-состязательные сети (Generative Adversarial Networks, GAN), вариационные автокодировщики (Variational Autoencoders, VAE) и диффузионные модели (Diffusion Models, DM).

1.2.1. Generative Adversarial Networks

Генеративно-состязательные сети были предложены в [3] и с тех пор стали одними из популярных подходов к задаче генерации данных. Основная идея GAN заключается в противостоянии двух нейросетевых моделей: генератора G и дискриминатора D. Целью генератора является создание реалистичных данных, способных обмануть дискриминатор, тогда как цель дискриминатора – различать реальные данные из обучающей выборки и сгенерированные образцы.

Формально процесс обучения GAN можно представить как игру с нулевой суммой, в которой минимизируется следующая функция потерь:

$$\min_G \max_D \mathbb{E}_{x\sim p(x)}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))]$$

Здесь x – реальные данные из обучающего набора, z – случайный шум, сгенерированный из стандартного нормального распределения, $G(z)$ –

сгенерированный образец, $D(x)$ – вероятность того, что x принадлежит реальным данным.

В ходе обучения дискриминатор учится повышать точность определения реальных и фальшивых примеров. Генератор учится "обманывать" дискриминатор, создавая все более реалистичные образцы. Этот процесс продолжается до тех пор, пока ни одна из сторон больше не может улучшить свои показатели без изменения стратегии другой стороны.

Современные вариации GAN (например, StyleGAN [4]) способны генерировать данные высокого разрешения и качества, часто неотличимые от реальных. К тому же существует множество модификаций GAN, адаптированных под конкретные задачи: cGAN [5] для условной генерации, CycleGAN [6] для перевода из одного домена в другой и т.д.

Существенным недостатком моделей с архитектурой GAN является так называемый mode collapse, при котором генератор может начать производить ограниченное количество типов выходов, игнорируя остальные возможные варианты. Такая проблема снижает разнообразие генерации.

1.2.2. Variational Autoencoders

VAE [7] представляют собой разновидность автоэнкодерных архитектур, адаптированных для задачи генерации данных через обучение вариационному приближению латентного пространства.

Основной идеей VAE является не просто кодирование входных данных в детерминированный вектор латентного представления, а построение распределения над этим вектором. Это позволяет модели генерировать новые данные, выбирая точки из латентного пространства, которые соответствуют обучающему распределению.

Цель обучения VAE заключается в максимизации нижней вариационной границы (ELBO – Evidence Lower Bound):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - D_{\text{KL}}\left(q_\phi(z \mid x) \parallel p(z)\right) =$$

$$\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - \int q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p(z)}\, dz$$

Здесь $p(z)$ – априорное распределение на латентные переменные (стандартное нормальное), $p_\theta(x|z)$ – параметризованный декодер, который восстанавливает данные по латентному представлению, а $q_\phi(z|x)$ – вариационный энкодер (в данном случае распределение q параметризовано некоторым семейством параметров $\Phi, \phi \in \Phi$), который аппроксимирует истинное апостериорное распределение $p(z|x)$. Первое слагаемое в ELBO отвечает за точность восстановления исходного объекта, а второе (дивергенция Кульбака-Лейблера) штрафует модель за отклонение от заданного априорного распределения на z, обеспечивая регуляризацию латентного пространства.

У VAE отсутствует mode collapse, однако качество генерации изображений хуже, чем у GAN. Например, возможны случаи генерации расплывчатых изображений. Тем не менее, вариационные автокодировщики могут применяться не только для генерации, но и для задач понижения размерности и обнаружения аномалий.

1.2.3. Diffusion Models

Диффузионные модели представляют собой мощный класс генеративных моделей, основанных на последовательном процессе зашумления и восстановления данных. Впервые идея диффузии была предложена в работе [8], а позже активно развивалась в работах и других исследователях. В последние годы эти модели стали доминировать в задачах генерации изображений и видео благодаря своей устойчивости к mode collapse, высокому качеству сгенерированных образцов и возможности условной генерации.

Диффузионные модели работают на основе двух ключевых этапов:

1. Форвардная диффузия (forward diffusion) – процесс постепенного добавления гауссовского шума к исходным данным до тех пор, пока они не превратятся в случайный шум.

Пусть $x_0 \sim q(x)$ – это исходный объект. Процесс форвардной диффузии определяется как марковская цепочка, в которой на каждом шаге $t = 1, \ldots, T$ к текущему состоянию x_{t-1} добавляется небольшое количество гауссовского шума:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t)\mathbf{I}\right),$$

где $\alpha_t \in (0,1)$ – коэффициент шума на шаге t. Как правило, $\alpha_1 > \alpha_2 > \ldots > \alpha_n$, то есть шум постепенно усиливается. Можно показать, что общий процесс зашумления имеет следующий вид:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\mathbf{I}\right),$$

где $\underline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$.

2. Обратная диффузия (reverse diffusion) – процесс последовательного удаления шума для восстановления данных из чистого шума.

Задача обратной диффузии – обучить модель, которая будет восстанавливать данные из шума. Это достигается с помощью параметризованной функции $p_\theta(x_{t-1}|x_t)$, которая аппроксимирует истинное распределение $q(x_{t-1}|x_t)$. Эта модель также представляет собой марковскую цепочку, где каждый шаг состоит в предсказании шума, добавленного на соответствующем этапе форвардной диффузии.

Более формально, модель учиться предсказывать $\widehat{\epsilon_\theta}(x_t, t) \approx \epsilon_t$, где $x_t = \sqrt{\underline{\alpha_t}} x_0 + \sqrt{1 - \underline{\alpha_t}}\epsilon_t$, а $\epsilon_t \sim N(0, I)$.

Обучение диффузионной модели сводится к минимизации разницы между истинным шумом ϵ_t и его оценкой $\widehat{\epsilon_\theta}$. Для этого используется следующая целевая функция:

$$\mathcal{L} = \mathbb{E}_{x_0, \epsilon_t, t}\left[\left\|\epsilon_t - \hat{\epsilon}_\theta(x_t, t)\right\|^2\right],$$

где математическое ожидание берется по начальным данным x_0, случайному шуму ϵ_t и равномерно выбранному временному шагу t.

Такая функция потерь позволяет обучать модель без явного вычисления полного логарифма правдоподобия, что делает ее эффективной и масштабируемой.

1.3. Генерация видео

В последние годы диффузионные модели стали основой ряда передовых методов генерации видео, которые делятся на два основных направления: генерация только по текстовому описанию (Text2Video) и генерация по тексту и входному изображению (Image2Video).

1.3.1 Подход Text2Video

Подход Text2Video направлен на генерацию последовательности видеокадров, используя только текстовый промпт в качестве входных данных. Формально задача заключается в аппроксимации условного распределения $p(v|t)$, где v – последовательность фреймов (например, тензор $v \in R^{T \times CxH \times W}$,где T – количество фреймов, C – количество каналов изображения, $H \times W$ – размер фрейма), а t – текстовое описание.

Современные архитектуры строятся на основе Latent Diffusion Models [2], работающих в сжатом пространстве, что позволяет значительно снизить вычислительные затраты. В них временная согласованность обеспечивается либо путем обучения в пространственно-временном латентном пространстве, либо через специализированные слои, моделирующие временные зависимости.

Одним из наиболее прогрессивных решений является модель Sora [9], архитектура которой включает в себя (см. **рисунок 1.1**):

- использование 3D VAE [9], расширенного для обработки видеоданных;
- Spatial-Temporal Diffusion Transformer – механизм, позволяющий разделять пространственные и временные механизмы внимания;
- нейросетевую модель T5 [10], через которую встраивается текстовое условие, после чего текстовые эмбеддинги интегрируются в процесс диффузии на каждом временном шаге.

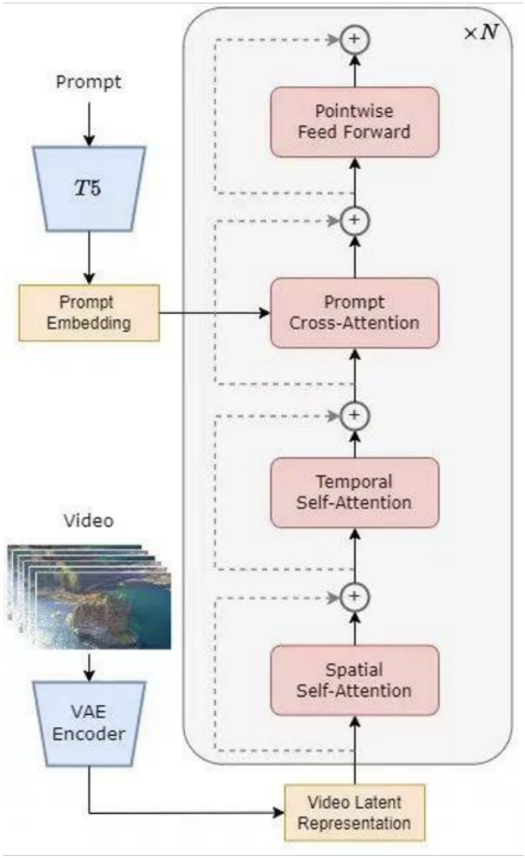

***Рисунок 1.1**. Архитектура Sora*

1.3.2 Подход Image2Video

В Image2Video генерация выполняется с учетом входного изображения x_0 и текстового описания t, таким образом задача формулируется как аппроксимация распределения $p(v|x_0, t)$. Этот подход позволяет управлять как стилем и содержанием первого кадра, так и общей тематикой видеоряда.

В отличие от Text2Video, где генерация видео осуществляется исключительно на основе текстового условия, подход Image2Video интегрирует входное изображение x_0 в процесс диффузии. Предварительно преобразованное энкодером изображение $E(x_0)$ задает начальное состояние латентного пространства, на основе которого осуществляется генерация.

ГЛАВА 2. ОБЗОР ИСПОЛЬЗУЕМЫХ МЕТОДОВ

2.1. Contrastive Language-Image Pre-Training (CLIP)

Модель Contrastive Language–Image Pretraining (CLIP), предложенная OpenAI в [11], представляет собой архитектуру, которая объединяет визуальные и текстовые данные в едином семантическом пространстве. Архитектура CLIP включает две нейронные сети: визуальный энкодер – обычно Vision Transformer (ViT) [12], и текстовый энкодер, реализованный на основе классических трансформеров [13].

Каждое изображение и текст преобразуются в числовые векторы (эмбеддинги) фиксированной размерности, которые затем нормализуются. Модель обучается на большом наборе пар (изображение, описание) с использованием симметричной кросс-энтропии по косинусному сходству соответствующих эмбеддингов. Данная функция потерь максимизирует сходство для положительных пар (где текстовое описание соответствует изображению) и минимизирует для отрицательных (где текстовое описание не соответствует изображению).

CLIP находит применение в таких задачах, как zero-shot классификация изображений по текстовому описанию и поиск изображений по текстовому описанию. В данной работе модель будет использована для вычисления косинусного сходства между фреймами видео и текстовым запросом, по которому генерировалось соответствующее видео, что позволит оценить, насколько видеопоследовательность соответствует заданному описанию.

На CLIP развитие направления, сравнивающего тексты и изображения не остановилось. Были предложены различные модификации и альтернативные подходы. Рассмотрим их поподробнее.

2.1.1 Sigmoid Loss for Language-Image Pre-Training (SigLIP)

Модель Sigmoid Loss for Language-Image Pre-Training (SigLIP), предложенная в [14], представляет собой модификацию архитектуры CLIP, нацеленную на устранение некоторых ограничений оригинальной функции

потерь. В отличие от классической CLIP, использующей softmax для нормализации вероятностей, SigLIP применяет сигмоидальную функцию активации и бинарную кросс-энтропию для обучения на положительных и отрицательных парах.

Основная идея заключается в том, чтобы интерпретировать задачу сопоставления изображений и текстов как задачу бинарной классификации для каждой пары. Модель должна предсказывать вероятность того, что пара (изображение, описание) является положительной. Это позволяет отказаться от необходимости вычисления нормализующих знаменателей (в формуле softmax) по всему батчу, что упрощает вычисления и делает обучение более стабильным.

2.1.2 Jina CLIP

Модель Jina CLIP [15], так же как и SigLIP, изменила оригинальную функцию потерь CLIP. Вместо симметричной кросс-энтропии использовалась InfoNCE (Information Noise-Contrastive Estimation) [16]. Кроме этого, изменились и архитектурные блоки:

- в качестве текстового энкодера использовалась модель Jina-XLM-RoBERTa [17];
- в качестве визуального энкодера использовалась модель EVA02 [18].

2.1.3 Human Preference Score (HPS)

Если предыдущие две модели меняли исходную архитектуру CLIP, то в случае Human Preference Score (HPS) [19] модель не изменилась. Ключевым отличием стал набор данных, на котором был дообучен оригинальный CLIP.

Авторы работы собрали уникальный датасет Human Preference Dataset (HPD) с предпочтениями людей для изображений, сгенерированных на основе промптов. В общей сложности набор данных содержит 798 тыс. пар бинарных предпочтений для 434 тыс. изображений. Каждая пара содержит два изображения, сгенерированных разными генеративными моделями с использованием одного и того же текстового запроса, и бинарную метку с предпочтением разметчика.

Дообучение на собранном наборе данных позволило скорректировать прогноз модели CLIP и сделать его более человеко-ориентированным.

2.2. Structural Similarity Index Measure (SSIM)

Structural Similarity Index Measure (SSIM) представляет собой одну из наиболее популярных немодельных метрик, предназначенных для оценки качества изображения на основе анализа структурного содержания. Впервые данный подход был предложен в [20], как альтернатива традиционным показателям, таким как MSE (Mean Squared Error), которые не всегда корректно отражают воспринимаемое человеком качество изображения.

Обычно SSIM рассчитывается локально – скользящим окном по изображению, после чего усредняется по всем фрагментам, чтобы получить итоговое значение показателя. Это позволяет учитывать не только глобальные, но и локальные особенности изображения. Формально, для двух окон изображения x и y, SSIM определяется как:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

где:

- μ_x, μ_y – средние значения в окнах x и y;
- ${\sigma_x}^2$, ${\sigma_y}^2$ – дисперсии в окнах x и y;
- σ_{xy} – ковариация окон x и y;
- C_1, C_2 – малые положительные константы, введенные для стабилизации деления при малых знаменателях.

Приведенная формула применима только для яркости изображения (то есть предполагается, что изображение черно-белое), по которой и происходит оценка качества. Индекс SSIM принимает значения от -1 до 1, где 1 соответствует идеальной идентичности окон x и y.

Пример на **рисунке 2.1** показывает две модификации входного изображения, каждая с одинаковым MSE, но с очень разными показателями SSIM.

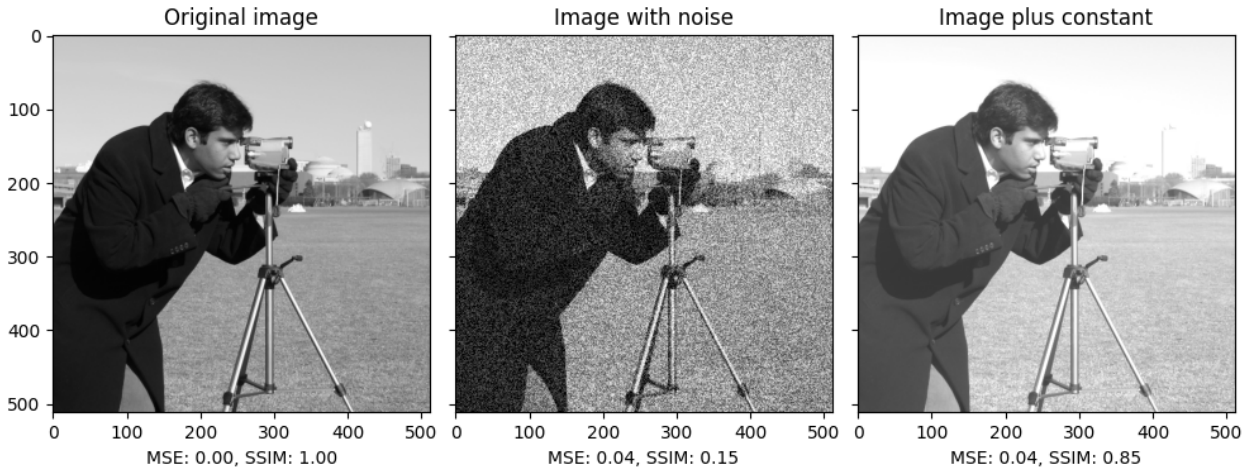

***Рисунок 2.1**. Изменение MSE и SSIM при модификации исходного изображения (слева). В изображение посередине добавлен шум, в крайнее правое – константа.*

В контексте анализа видеороликов SSIM используется для оценки схожести пары кадров, что позволяет судить о временной согласованности видео. Например, высокое значение SSIM между соседними кадрами может указывать на плавное движение или небольшие изменения в сцене, тогда как резкие скачки могут свидетельствовать о разрыве временного потока или ошибке генерации.

2.3. Learned Perceptual Image Patch Similarity (LPIPS)

Learned Perceptual Image Patch Similarity (LPIPS) – это показатель качества, предложенный для измерения различия между изображениями на основе их высокоуровневых признаков. В отличие от метрик, таких как MSE или SSIM, которые оперируют пиксельными значениями или статистиками локальных окон, LPIPS оценивает схожесть изображений с использованием признаков, извлеченных из сверточных нейронных сетей, предобученных на больших датасетах, например, на ImageNet [21]. Метод был предложен в [22] и быстро стал стандартом в задачах оценки качества генерации изображений.

Формально, для пары изображений x и y, метрика LPIPS определяется как:

$$\mathrm{LPIPS}(x, y) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (\phi_l(x)_{h,w} - \phi_l(y)_{h,w}) \|_2^2$$

где:

- $\varphi_l(\cdot)$ – активации сверточной нейронной сети на слое l;
- w_l – вектор весов для слоя l, определяющий значимость признаков;
- H_l, W_l – размерность карты признаков (feature map) на слое l;
- $\odot$ – произведение Адамара.

Часто используются сети VGG [23] или AlexNet [24] в качестве базовой архитектуры для извлечения признаков. При этом веса w_l могут быть обучены на основе человеко-ориентированных оценок схожести изображений, что делает метрику ещё более согласованной с субъективным восприятием.

В контексте анализа видеороликов LPIPS применяется аналогично SSIM.

2.4. Recurrent All-Pairs Field Transforms (RAFT)

Recurrent All-Pairs Field Transforms (RAFT) представляет собой одну из наиболее популярных моделей оптического потока, предложенную в [25]. Оптический поток – это векторное поле, которое описывает движение пикселей между двумя последовательными кадрами видео. RAFT является инновационным подходом к решению этой задачи, сочетающим рекуррентные механизмы и эффективные вычисления корреляций между пикселями. Архитектура модели включает два основных модуля: экстрактор признаков и рекуррентный обновляющий блок.

Экстрактор признаков представляет собой сверточную нейронную сеть, которая преобразует каждое входное изображение в карту признаков высокой размерности. После этого формируется 4-х мерный тензор (all-pairs correlation volume), где каждая точка соответствует схожести признаков из первой и

второй карты. Формально, пусть $f(I_1) \in R^{H \times W \times C}$ и $f(I_2) \in R^{H \times W \times C}$ – карты признаков двух изображений, полученные после прохождения через экстрактор признаков. Тогда all-pairs correlation volume $C \in R^{H \times W \times H \times W}$ определяется как перемножение матриц $f(I_1)$ и $f(I_2)$:

$$C_{ijkl} = \sum_h f(I_1)_{igh} \cdot f(I_2)_{klh}$$

Затем RAFT использует этот all-pairs correlation volume как вход для рекуррентного модуля (в оригинальной статье используется GRU [26]), который итеративно уточняет поле оптического потока, на каждом шаге выбирая наиболее вероятные смещения на основе корреляции признаков. Это обеспечивает более точное определение движения, особенно в сложных или неоднозначных регионах изображения.

RAFT помогает анализировать и визуализировать движения объектов в видео, что важно для задач трекинга и анализа действий. Более того, модель оптического потока применяется в системах автоматического вождения для распознавания дорожных ситуаций и прогнозирования движения объектов.

В рамках данной работы RAFT может быть использована для анализа временной согласованности видеопоследовательностей. В частности, можно использовать следующие характеристики оптического потока, вычисленные с помощью модели:

- Магнитуда. Отражает общее количество движения между кадрами. Высокая магнитуда может указывать на резкие изменения в сцене, которые могут быть связаны с неестественностью генерации.
- Угол. Характеризует направление движения объектов. Непоследовательные углы между соседними кадрами также могут свидетельствовать о нарушении согласованности всего видео.

ГЛАВА 3. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

3.1. Сбор данных

Для проведения экспериментального исследования, направленного на разработку модели оценки качества генерации видео, необходима была организация сбора репрезентативного набора данных (датасета), состоящего из сгенерированных видео, связанных с соответствующими текстовыми промптами, а также числовых оценок их качества.

В рамках работы были собраны видеоролики, сгенерированные с помощью современных Text2Video и Image2Video моделей. Длительность видео в среднем составляла 5 секунд. Использовались следующие источники:

- VidProM [27] – датасет, содержащий в себе промпты для Text2Video генерации;
- Pika [28] и Sora [29] – готовые видео, взятые из публично доступных демонстрационных наборов;
- Qwen [30] и Kandinsky [31] – видеоролики, сгенерированные на основе текстовых промптов;
- DeepAction [32] – готовые видео, предоставленные в открытом доступе на платформе Hugging Face;
- Runway Gen-4 [33] – видеоролики, сгенерированные с использованием комбинации текстового промпта и входного изображения.

Таким образом, общий набор данных охватывает широкий спектр современных подходов к генерации видео, что позволяет учесть разнообразие визуальных стилей, структур сцены и условий генерации.

Для получения оценок качества была организована процедура краудсорсинга, в которой участвовало 10 человек. Каждому участнику предлагалось просмотреть видеоролик и оценить его качество по шкале от 1 до 10, где 1 соответствовало крайне низкому качеству (например, логическая несогласованность, размытость, артефакты), а 10 – высокому уровню

реалистичности, временной согласованности и соответствию текстовому описанию.

Полученные индивидуальные оценки усреднялись для формирования предварительной оценки качества. Однако анализ показал, что в ряде случаев среднее значение не всегда адекватно отражает объективное качество видео, особенно в случаях наличия выбросов или систематических ошибок оценки со стороны части участников. Для повышения надежности оценок была применена экспертная коррекция: группа специалистов в области компьютерного зрения и генеративных моделей перепроверила наиболее сложные и неоднозначные случаи, после чего финальная оценка была скорректирована.

В результате был собран датасет из 200 уникальных пар «промпты – видео» и числовых оценок качества в диапазоне [1, 10]. Все данные были выложены в открытый доступ на платформе Hugging Face [34], чтобы способствовать дальнейшим исследованиям в области автоматической оценки качества генерации видео. Примеры текстовых запросов представлены в **таблице 3.1**. Распределение оценок представлено на **рисунке 3.1**.

***Таблица 3.1**. Примеры текстовых запросов для генерации видео*

A person walking through a park
A group of young people racing motorcycles
forest, rain, early morning
a blue whale swimming in desert at night under starry night with fireworks in sky
a modern flying car above the city skyscrapers

Этот датасет представляет собой ценную эмпирическую базу, поскольку сочетает в себе разнообразие генеративных моделей, вариативность тестовых условий и объективно оцененные примеры, что делает его пригодным для использования как в академической, так и в прикладной сфере.

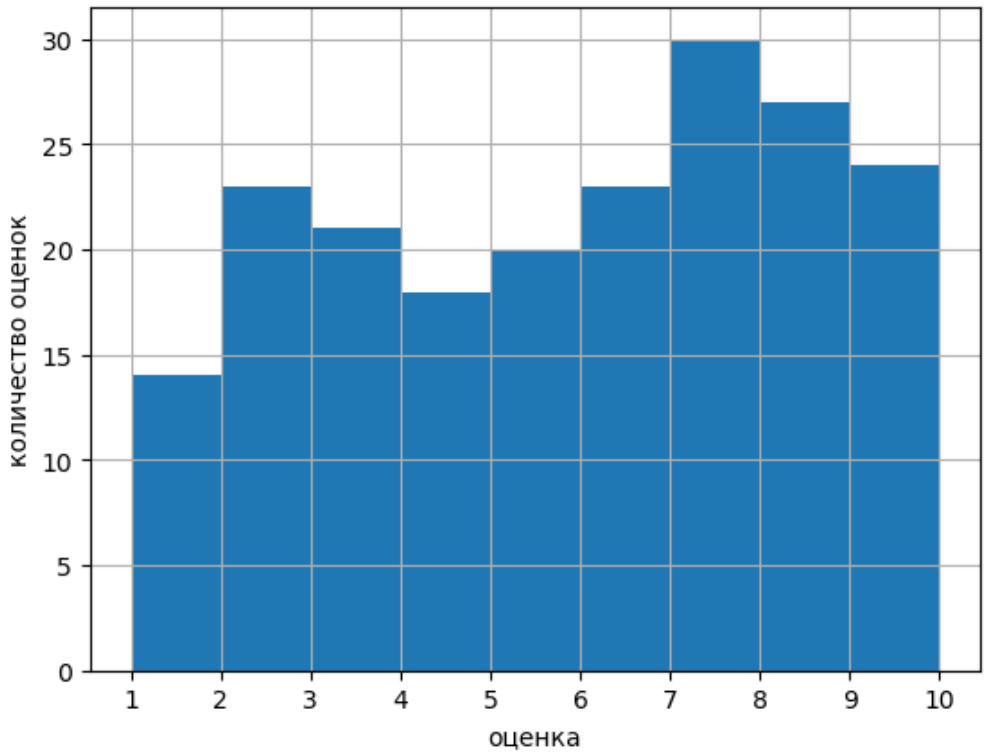

Рисунок 3.1. *Распределение оценок собранного датасета*

3.2. Генерация признаков

Для последующего обучения модели оценки качества генерации видео необходимо преобразовать собранные данные в числовое представление, пригодное для машинного обучения. Для этого был реализован процесс извлечения признаков, в основном основанный на применении современных предобученных моделей компьютерного зрения.

Для начала каждый видеоролик разбивается на кадры (фреймы), что позволяет получить выборку изображений, характеризующую всю видеопоследовательность. Дальнейшая обработка заключается в извлечении двух групп признаков: тех, которые характеризуют семантическое соответствие между промптом и отдельными кадрами, и тех, которые описывают взаимосвязь между соседними кадрами внутри видео.

Для извлечения признаков из первой группы использовались модели (см. раздел 2.1), ориентированные на мультимодальное сравнение изображения и текста: CLIP (в вариантах clip-vit-base-patch32 и clip-vit-large-patch14), SigLIP (siglip-so400m-patch14-384, siglip-large-patch16-384), Jina CLIP

(jina-clip-v2) и HPS. Эти модели позволяют вычислять эмбеддинги изображений и текста в общем латентном пространстве, что дает возможность количественно оценить степень их схожести. Для каждого кадра видео вычислялось значение близости к текстовому промпту, после чего эти значения объединялись в массив, характеризующий изменение соответствия на протяжении всего ролика.

Вторая группа признаков была направлена на анализ временной структуры видео и включала в себя метрики, оценивающие схожесть и движение между соседними кадрами (см. раздел 2.2-2.4). Для этой цели применялись такие подходы, как SSIM, LPIPS с использованием AlexNet в качестве базовой архитектуры, а также модель оптического потока RAFT (raft_large).

Для LPIPS и SSIM схожесть двух кадров оценивалась одним числом. В случае модели RAFT использовался выход предпоследнего слоя, который представляет из себя массив размера $2 \times H \times W$ ($H \times W$ – размер изображения), где первый элемент соответствует горизонтальному смещению каждого пикселя от первого изображения ко второму, а второй – вертикальному. По полученному потоку к тому же рассчитывались скалярные характеристики: его магнитуда и угол. Три полученные величины агрегировались (поскольку являлись двумерными массивами) и использовались в качестве признаков, характеризующих динамику между двумя кадрами. Использовалось 6 функций агрегации (см. **приложение 1**):

- сумма;
- среднее значение;
- медиана;
- минимальное значение;
- максимальное значение;
- среднеквадратическое отклонение.

Полученные массивы признаков (как для первой, так и для второй группы) отличались по размерности, поскольку количество кадров в

различных видеороликах варьировалось. Для устранения этого несоответствия и подготовки унифицированного входного представления были сформированы три типа наборов признаков: агрегированные, неагрегированные и объединенные.

Агрегированные признаки формировались путем применения функций агрегации к каждому массиву, полученному с помощью вышеуказанных моделей. Такой подход позволил создать компактное числовое описание, инвариантное к длине видеоролика. Использовались следующие 15 функций агрегации:

- экспоненциальное взвешенное скользящее среднее со значениями сглаживающей константы 0.1, 0.3, 0.5, 0.7 и 0.9;
- сумма;
- среднее значение;
- медиана;
- минимальное значение;
- максимальное значение;
- среднеквадратическое отклонение;
- 25-% процентиль;
- 75-% процентиль;
- межквартильный размах;
- отношение минимального значения к максимальному.

Неагрегированные признаки сохраняли информацию о динамике изменения характеристик по кадрам. Для обеспечения одинаковой длины векторов в этом случае использовалось ограничение: из всех кадров каждого видео выбиралось фиксированное число – 100 первых кадров. Если оригинальный массив был длиннее – он обрезался, если короче – дополнялся нулями. Неагрегированные признаки формировались только для первой группы.

Объединенный набор признаков содержал в себе как агрегированный набор, так и неагрегированный.

Таким образом, было сформировано 21 набора признаков:

1. Для каждой из 6 моделей, оценивающих близость промпта и фрейма – 3 набора признаков: агрегированные, неагрегированные и объединенные. Итого, $6 \times 3 = 18$ наборов.
2. Для каждой из 3 моделей, анализирующих временную согласованность видео – 1 набор признаков: агрегированный. Итого, $3 \times 1 = 3$ набора.

Полученные в результате описанной процедуры наборы признаков легли в основу последующего обучения модели оценки качества генерации видео.

3.3. Модель, обучение и целевые метрики

Для решения поставленной задачи на основе извлеченных признаков была выбрана модель случайного леса (Random Forest), реализованная в библиотеке scikit-learn [35]. Данный выбор обусловлен устойчивостью алгоритма к переобучению, а также интерпретируемостью получаемых результатов. Рассматривались две задачи: регрессия, где целевая переменная представляла собой числовую оценку качества в диапазоне [1, 10], и бинарная классификация, где целевая переменная равнялась 0, если оценка ниже 5.5 и 1 в противном случае.

Обучение проводилось на различных комбинациях наборов признаков, полученных в разделе 3.2, что позволило оценить информативность каждого из них. Итоговое признаковое пространство включало в себя:

1) обязательно один из трех набор признаков (агрегированный, неагрегированный, объединенный) моделей CLIP, SigLIP, Jina CLIP и HPS;
2) необязательно комбинацию от одного до трех агрегированных наборов признаков моделей LPIPS, SSIM и RAFT.

Перед началом обучения вся совокупность данных из 200 объектов разделялась на обучающую и тестовую выборки в соотношении 80/20 для настройки гиперпараметров модели. Для этого применялся инструмент Optuna

[36], осуществляющий поиск оптимальных значений по заданному пространству параметров. Оптимизация проводилась с учетом специфики решаемых задач: для регрессии минимизировалось значение средней абсолютной ошибки (Mean Absolute Error, MAE), а для классификации – максимизировалось значение точности (accuracy).

После получения оптимальных гиперпараметров проводилась процедура кросс-валидации с пятью фолдами. Для обеспечения сбалансированности фолдов была применена стратификация. Поскольку исходная целевая величина являлась непрерывной, она была предварительно дискретизирована путем разбиения ее значений на интервалы, соответствующие квантилям распределения. Каждый объект получил дискретную метку, определяемую положением его значения в этом распределении. Это позволило сохранить репрезентативность выборок и повысить устойчивость модели к возможным перекосам в данных.

Итого, исследовалось 144 признаковых пространства, как для задачи классификации, так и для задачи регрессии.

3.4. Результаты

Рассмотрим результаты для случая только агрегированных признаков в задаче классификации (см. **таблицу 3.2**). Для наборов признаков используются следующие обозначения:

- ONLY – только признаки, полученные с помощью соответствующей модели, оценивающей близость промпта и изображения;
- L – ONLY и признаки, полученные с помощью LPIPS;
- S – ONLY и признаки, полученные с помощью SSIM;
- OF – ONLY и признаки, полученные с помощью модели оптического потока RAFT;
- L_S – ONLY и признаки, полученные с помощью LPIPS и SSIM;
- L_OF – ONLY и признаки, полученные с помощью LPIPS и модели оптического потока RAFT;

***Таблица 3.2**. Качество классификации (accuracy) при использовании только агрегированных признаков на кросс-валидации с 5 фолдами, %*

Модель	**Набор признаков**							
	ONLY	**L**	**S**	**OF**	**L_S**	**L_OF**	**S_OF**	**L_S_OF**
HPS	61.50 ± 5.39	66.00 ± 6.82	64.00 ± 6.63	59.00 ± 5.39	67.50 ± 7.07	62.50 ± 6.71	58.00 ± 7.31	59.50 ± 4.30
jina-clip-v2	67.00 ± 3.32	66.50 ± 5.83	70.00 ± 6.32	57.00 ± 9.27	**74.50 ± 4.85**	64.50 ± 6.78	58.50 ± 8.46	60.50 ± 6.00
clip-vit-base-patch32	59.50 ± 1.87	67.50 ± 3.54	65.00 ± 6.12	59.00 ± 8.00	61.00 ± 3.00	61.50 ± 10.32	59.00 ± 6.04	61.50 ± 6.63
clip-vit-large-patch14	67.00 ± 7.48	72.50 ± 3.87	70.00 ± 3.16	60.00 ± 7.07	73.50 ± 4.36	64.00 ± 9.82	63.00 ± 11.34	63.00 ± 6.20
siglip-so400m-patch14-384	62.50 ± 3.54	64.00 ± 3.39	62.00 ± 5.79	58.50 ± 7.68	63.00 ± 7.31	58.50 ± 3.00	57.50 ± 9.87	61.50 ± 6.82
siglip-large-patch16-384	59.50 ± 2.92	65.00 ± 4.74	63.00 ± 2.92	56.00 ± 5.15	68.00 ± 5.10	57.50 ± 8.22	61.00 ± 3.00	62.50 ± 5.24

- S_OF – признаки ONLY и признаки, полученные с помощью SSIM и модели оптического потока RAFT;
- L_S_OF – признаки ONLY и признаки, полученные с помощью LPIPS, SSIM и модели оптического потока RAFT;

Наилучшее качество в данном случае показала модель, использующая признаки, полученные Jina CLIP, LPIPS и SSIM – в среднем 74.5% accuracy. Причем, если убрать признаки моделей LPIPS и SSIM, то качество станет хуже – 67%. Наименее информативными оказались признаки от RAFT – качество модели для любого признакового пространства ухудшилось при их добавлении.

Для понимания того, какие из используемых признаков внесли наибольший вклад в предсказательную способность модели, был применён метод интерпретации моделей машинного обучения SHAP (SHapley Additive exPlanations) [37]. Данный подход позволяет количественно оценить вклад каждого отдельного признака в финальное предсказание обученной модели путем расчета его среднего абсолютного значения SHAP, которое характеризует степень его влияния на прогноз. Чем выше это значение, тем более существенным считается признак для принятия решения моделью.

Полученные результаты (см. **рисунок 3.2**) позволили выделить наиболее информативные признаки. Например:

- **ssim_sum** – сумма значений SSIM между соседними кадрами демонстрирует самый высокий вклад в предсказание. Среднее абсолютное значение SHAP составляет около 0.12, что делает его ключевым фактором для оценки временной согласованности видеопоследовательности.
- **jina-clip-v2_sum** – сумма значений семантической близости между текстовым промптом и фреймами видео, используя модель Jina CLIP, показывает второй по величине вклад среди всех признаков. Среднее абсолютное значение SHAP составляет около 0.08, что указывает на

важность соответствия содержания видео заданному текстовому запросу.

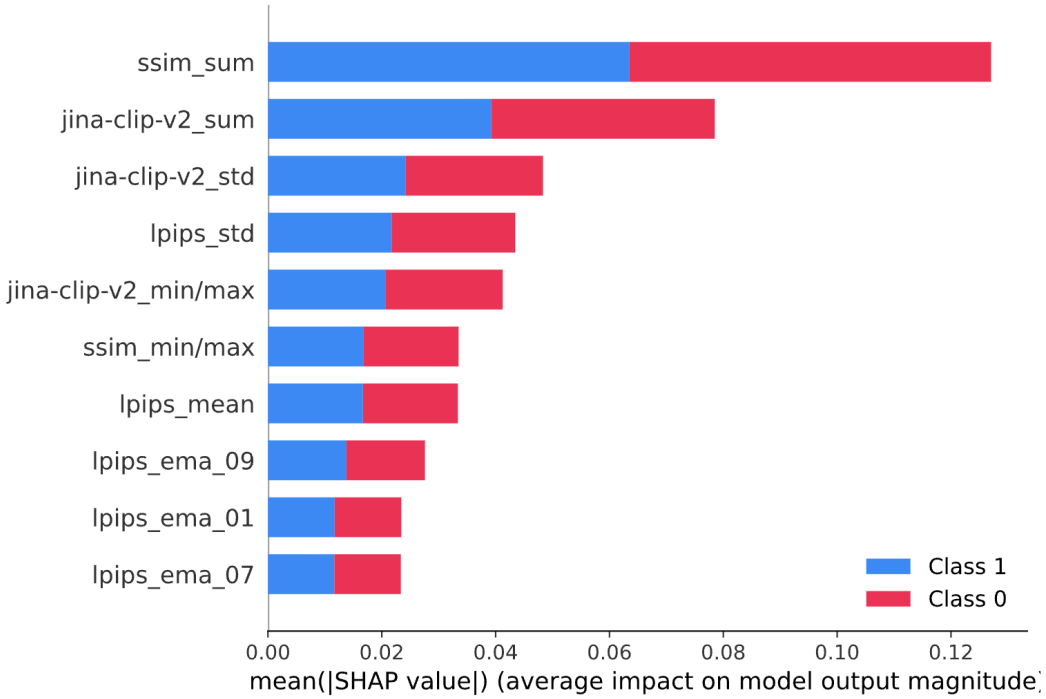

***Рисунок 3.2**. Среднее абсолютное значение SHAP топ-10 признаков для задачи классификации в случае использования агрегированных признаков Jina CLIP, LPIPS и SSIM*

Качество классификации и регрессии для оставшихся случаев приведено в **приложении 2**. Наилучшее качество в них достигается с использованием признаков модели HPS, что свидетельствует о важности использования мультимодальных моделей, дообученных на человеко-ориентированных оценках.

ЗАКЛЮЧЕНИЕ

В рамках данной работы была предпринята попытка к созданию подхода к оценке качества генерации видео. Был проведен анализ методов, включая модели сравнения изображения и текста, а также метрики, характеризующие схожесть между видеокадрами. На основе полученных данных была разработана методика извлечения признаков, позволяющая формализовать качественные характеристики видеороликов в количественное представление, пригодное для машинного обучения.

С целью эмпирической проверки предложенного подхода был собран и подготовлен оригинальный датасет, состоящий из пар «промпты – видео» и соответствующих числовых оценок качества в диапазоне от 1 до 10, полученных с помощью процедуры краудсорсинга. Данный датасет был опубликован в открытом доступе на платформе Hugging Face с целью способствовать дальнейшим исследованиям в области оценки качества генерации видео.

На основе извлеченных признаков были обучены модели случайного леса как для задачи регрессии, так и для задачи классификации (определение принадлежности оценки к категории выше или ниже порогового значения). Проведённая серия экспериментов позволила выявить наиболее информативные наборы признаков и оценить их вклад в общую предсказательную способность модели.

Полученные результаты демонстрируют работоспособность предложенного подхода и открывают возможности для его дальнейшего улучшения. В качестве перспективных направлений развития можно выделить необходимость сбора более обширного датасета, что позволит повысить обобщающую способность модели. Также представляет интерес использование исходного изображения в задачах Image2Video в качестве референса для последующего анализа временной согласованности кадров, как дополнительного сигнала для оценки качества генерации.

СПИСОК ЛИТЕРАТУРЫ

[1] Ramesh A. et al. Zero-shot text-to-image generation //International conference on machine learning. – Pmlr, 2021. – С. 8821-8831.

[2] Rombach R. et al. High-resolution image synthesis with latent diffusion models //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2022. – С. 10684-10695.

[3] Goodfellow I. J. et al. Generative adversarial nets //Advances in neural information processing systems. – 2014. – Т. 27.

[4] Karras T., Laine S., Aila T. A style-based generator architecture for generative adversarial networks //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2019. – С. 4401-4410.

[5] Mirza M., Osindero S. Conditional generative adversarial nets //arXiv preprint arXiv:1411.1784. – 2014.

[6] Zhu J. Y. et al. Unpaired image-to-image translation using cycle-consistent adversarial networks //Proceedings of the IEEE international conference on computer vision. – 2017. – С. 2223-2232.

[7] *Chen, Yankun & Liu, Jingxuan & Peng, Lingyun & Wu, Yiqi & Xu, Yige & Zhang, Zhanhao. (2024). Auto-Encoding Variational Bayes. Cambridge Explorations in Arts and Sciences. 2. 10.61603/ceas.v2i1.33.*

[8] Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models //Advances in neural information processing systems. – 2020. – Т. 33. – С. 6840-6851.

[9] Zheng Z. et al. Open-sora: Democratizing efficient video production for all //arXiv preprint arXiv:2412.20404. – 2024.

[10] Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //Journal of machine learning research. – 2020. – Т. 21. – №. 140. – С. 1-67.

[11] Radford A. et al. Learning transferable visual models from natural language supervision //International conference on machine learning. – PmLR, 2021. – С. 8748-8763.

[12] Dosovitskiy A. et al. An image is worth 16x16 words: Transformers for image recognition at scale //arXiv preprint arXiv:2010.11929. – 2020.

[13] Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – T. 30.

[14] Zhai X. et al. Sigmoid loss for language image pre-training //Proceedings of the IEEE/CVF international conference on computer vision. – 2023. – C. 11975-11986.

[15] Koukounas A. et al. jina-clip-v2: Multilingual multimodal embeddings for text and images //arXiv preprint arXiv:2412.08802. – 2024.

[16] Oord A., Li Y., Vinyals O. Representation learning with contrastive predictive coding //arXiv preprint arXiv:1807.03748. – 2018.

[17] Sturua S. et al. jina-embeddings-v3: Multilingual embeddings with task lora //arXiv preprint arXiv:2409.10173. – 2024.

[18] Fang Y. et al. Eva-02: A visual representation for neon genesis //Image and Vision Computing. – 2024. – T. 149. – C. 105171.

[19] Wu X. et al. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis //arXiv preprint arXiv:2306.09341. – 2023.

[20] Wang Z. et al. Image quality assessment: from error visibility to structural similarity //IEEE transactions on image processing. – 2004. – T. 13. – №. 4. – C. 600-612.

[21] Deng J. et al. Imagenet: A large-scale hierarchical image database //2009 IEEE conference on computer vision and pattern recognition. – Ieee, 2009. – C. 248-255.

[22] Zhang R. et al. The unreasonable effectiveness of deep features as a perceptual metric //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – C. 586-595

[23] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition //arXiv preprint arXiv:1409.1556. – 2014.

[24] Krizhevsky A., Sutskever I., Hinton G. E. Imagenet classification with deep convolutional neural networks //Advances in neural information processing systems. – 2012. – T. 25.

[25] Teed Z., Deng J. Raft: Recurrent all-pairs field transforms for optical flow //Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. – Springer International Publishing, 2020. – С. 402-419.

[26] Chung J. et al. Empirical evaluation of gated recurrent neural networks on sequence modeling //arXiv preprint arXiv:1412.3555. – 2014.

[27] Wang W., Yang Y. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models //arXiv preprint arXiv:2403.06098. – 2024

[28] Pika: [Электронный ресурс]. URL: https://pika.art/ (Дата обращения: 25.05.2025)

[29] Sora: [Электронный ресурс]. URL: https://openai.com/sora/ (Дата обращения: 25.05.2025)

[30] Qwen: [Электронный ресурс]. URL: https://chat.qwen.ai/ (Дата обращения: 25.05.2025)

[31] Kandinskiy: [Электронный ресурс]. URL: https://t.me/kandinsky21_bot (Дата обращения: 25.05.2025)

[32] Bohacek M., Farid H. Human Action CLIPS: Detecting AI-generated Human Motion //arXiv preprint arXiv:2412.00526. – 2024.

[33] Runway Gen-4: [Электронный ресурс]. URL: https://runwayml.com/research/introducing-runway-gen-4 (Дата обращения: 25.05.2025)

[34] Собранный датасет, содержащий промпты, видео и оценки: [Электронный ресурс]. URL: https://huggingface.co/datasets/yanakidis/Text2Video (Дата обращения: 25.05.2025)

[35] Библиотека scikit-learn: [Электронный ресурс]. URL: https://scikit-learn.org/stable/api/sklearn.ensemble.html (Дата обращения: 25.05.2025)

[36] Библиотека optuna: [Электронный ресурс]. URL: https://optuna.readthedocs.io/en/stable/index.html (Дата обращения: 25.05.2025)

[37] Библиотека shap: [Электронный ресурс]. URL: https://shap.readthedocs.io/en/latest/ (Дата обращения: 25.05.2025)

ПРИЛОЖЕНИЕ

1. Генерация признаков для модели оптического потока RAFT

```python
from torchvision.models.optical_flow import Raft_Large_Weights, raft_large
import torch

def get_features(x):
    return torch.tensor((
        x.sum(),
        x.median(),
        x.mean(),
        x.std(),
        x.min(),
        x.max()
    ))

video_frames = # список изображений в RGB формате
weights = Raft_Large_Weights.DEFAULT
model = raft_large(weights=weights, progress=False)
transform = weights.transforms()

for j in range(len(video_frames) - 1):
    frame1_tensor = to_tensor(video_frames[j]).unsqueeze(0)
    frame2_tensor = to_tensor(video_frames[j + 1]).unsqueeze(0)

    frame_1, frame_2 = transform(frame1_tensor, frame2_tensor)

    with torch.no_grad():
        flow = model(frame_1, frame_2)[-1]

    flow = flow.squeeze(0)
    flow_features = get_features(flow) # признаки оптического потока

    magnitude = torch.sqrt(flow[0]**2 + flow[1]**2)
    magnitude_features = get_features(magnitude) # признаки магнитуды

    angle = torch.arctan2(flow[0], flow[1])
    angle_features = get_features(angle) # признаки угла
```

2. Качество классификации и регрессии для различных признаковых пространств

Для каждой из задач (классификации и регрессии) представлены по 3 таблицы с качеством, соответствующие агрегированным, неагрегированным и объединенным наборам признаков моделей, оценивающих близость промпта и изображения (см. раздел 3.2). В случае классификации – это accuracy, в

случае регрессии – это MAE. Замеры производились с помощью процедуры кросс-валидации на 5 фолдах. Для наборов признаков используется аналогичная разделу 3.4 нотация. Качество классификации при использовании агрегированных признаков представлено в **таблице 3.2**.

***Таблица 1**. Accuracy на неагрегированных признаках, %*

Модель	**Набор признаков**							
	ONLY	**L**	**S**	**OF**	**L_S**	**L_OF**	**S_OF**	**L_S_OF**
HPS	59.00 ± 7.35	59.50 ± 7.97	69.50 ± 5.79	60.50 ± 12.39	**73.00 ± 4.85**	60.00 ± 10.61	61.00 ± 8.00	63.50 ± 10.32
jina-clip-v2	63.50 ± 7.35	66.50 ± 6.63	67.00 ± 6.20	62.50 ± 11.29	68.00 ± 4.85	59.50 ± 10.30	62.00 ± 9.27	65.50 ± 10.30
clip-vit-base-patch32	57.00 ± 2.92	61.50 ± 4.36	62.50 ± 4.47	54.50 ± 6.96	64.50 ± 4.85	60.00 ± 3.16	60.00 ± 10.84	58.00 ± 10.30
clip-vit-large-patch14	64.50 ± 5.10	66.00 ± 4.64	64.50 ± 4.30	60.50 ± 3.32	60.00 ± 7.58	58.50 ± 7.00	61.50 ± 7.68	58.50 ± 6.63
siglip-so400m-patch14-384	58.00 ± 5.79	56.50 ± 4.64	54.00 ± 9.30	61.50 ± 7.00	60.50 ± 6.60	63.00 ± 6.60	62.50 ± 11.83	60.00 ± 6.52
siglip-large-	59.50 ±	63.50 ±	63.50 ±	60.00 ±	65.50 ±	65.00 ±	65.00 ±	64.50 ±

patch16-384	6.78	3.74	5.15	6.52	1.87	7.42	6.32	5.57

***Таблица 2**. Accuracy на объединенных признаках, %*

Модель	**Набор признаков**							
	ONLY	**L**	**S**	**OF**	**L_S**	**L_OF**	**S_OF**	**L_S_OF**
HPS	66.50 ± 6.63	66.50 ± 6.44	69.00 ± 4.64	57.50 ± 9.87	**73.00 ± 6.60**	63.50 ± 8.00	58.00 ± 6.00	64.00 ± 3.74
jina-clip-v2	67.00 ± 8.57	67.50 ± 6.71	70.00 ± 4.47	64.50 ± 6.78	71.00 ± 3.39	66.50 ± 9.03	63.50 ± 8.15	69.50 ± 8.12
clip-vit-base-patch32	60.50 ± 2.45	64.00 ± 3.39	62.00 ± 4.00	55.00 ± 6.52	66.00 ± 4.06	61.00 ± 5.15	62.50 ± 10.84	64.00 ± 6.04
clip-vit-large-patch14	65.50 ± 6.20	66.50 ± 4.64	61.50 ± 6.04	63.50 ± 4.36	67.50 ± 4.18	63.00 ± 5.79	66.00 ± 5.15	63.50 ± 5.61
siglip-so400m-patch14-384	56.00 ± 6.63	62.50 ± 6.12	60.00 ± 4.18	58.50 ± 8.75	63.50 ± 6.44	56.00 ± 6.82	61.00 ± 7.18	60.00 ± 7.42
siglip-large-patch16-	58.00 ±	62.00 ±	64.00 ±	61.50 ±	63.50 ±	64.50 ±	61.50 ±	65.00 ±

384	4.85	3.67	6.63	10.56	6.04	4.85	8.60	2.24

***Таблица 3**. MAE на агрегированных признаках*

Модель	**Набор признаков**							
	ONLY	**L**	**S**	**OF**	**L_S**	**L_OF**	**S_OF**	**L_S_OF**
HPS	1.84 ± 0.19	1.88 ± 0.22	1.78 ± 0.21	1.98 ± 0.19	**1.77 ± 0.19**	1.99 ± 0.18	1.88 ± 0.18	1.90 ± 0.16
jina-clip-v2	1.89 ± 0.07	1.88 ± 0.14	1.79 ± 0.11	1.95 ± 0.15	1.79 ± 0.10	1.95 ± 0.13	1.88 ± 0.15	1.87 ± 0.15
clip-vit-base-patch32	1.84 ± 0.14	1.93 ± 0.14	1.79 ± 0.19	1.99 ± 0.14	1.85 ± 0.16	1.97 ± 0.15	1.91 ± 0.16	1.92 ± 0.15
clip-vit-large-patch14	1.78 ± 0.19	1.88 ± 0.16	1.81 ± 0.22	1.96 ± 0.16	1.80 ± 0.15	1.96 ± 0.16	1.88 ± 0.17	1.92 ± 0.13
siglip-so400m-patch14-384	2.13 ± 0.09	2.05 ± 0.12	1.85 ± 0.15	2.02 ± 0.17	1.91 ± 0.17	2.03 ± 0.14	1.92 ± 0.14	1.95 ± 0.14
siglip-large-patch16-384	2.13 ± 0.09	2.15 ± 0.13	1.84 ± 0.11	2.08 ± 0.14	1.87 ± 0.12	2.03 ± 0.15	1.93 ± 0.15	1.94 ± 0.13

***Таблица 4**. MAE на неагрегированных признаках*

Модель	**Набор признаков**							
	ONLY	**L**	**S**	**OF**	**L_S**	**L_OF**	**S_OF**	**L_S_OF**
HPS	2.17 ± 0.06	2.11 ± 0.13	**1.73 ± 0.14**	1.98 ± 0.14	1.78 ± 0.11	1.97 ± 0.14	1.86 ± 0.15	1.87 ± 0.15
jina-clip-v2	2.03 ± 0.16	2.04 ± 0.14	1.78 ± 0.11	2.07 ± 0.13	1.82 ± 0.12	2.00 ± 0.16	1.88 ± 0.18	1.89 ± 0.16
clip-vit-base-patch32	2.18 ± 0.10	2.13 ± 0.10	1.99 ± 0.10	2.06 ± 0.12	1.90 ± 0.11	2.07 ± 0.11	1.96 ± 0.14	1.98 ± 0.14
clip-vit-large-patch14	2.10 ± 0.15	2.11 ± 0.14	1.91 ± 0.18	2.06 ± 0.13	1.88 ± 0.13	2.06 ± 0.12	1.95 ± 0.16	1.96 ± 0.14
siglip-so400m-patch14-384	2.11 ± 0.10	2.04 ± 0.17	1.89 ± 0.13	2.04 ± 0.14	1.91 ± 0.08	2.02 ± 0.14	1.93 ± 0.13	1.93 ± 0.12
siglip-large-patch16-384	2.20 ± 0.09	2.15 ± 0.13	1.84 ± 0.16	2.10 ± 0.20	1.87 ± 0.11	2.03 ± 0.15	1.94 ± 0.16	1.94 ± 0.16

***Таблица 5**. MAE на объединенных признаках*

Модель	**Набор признаков**							
	ONLY	**L**	**S**	**OF**	**L_S**	**L_OF**	**S_OF**	**L_S_OF**
HPS	1.79 ± 0.18	1.81 ± 0.19	1.79 ± 0.16	1.96 ± 0.20	**1.76 ± 0.15**	1.96 ± 0.20	1.85 ± 0.16	1.87 ± 0.16
jina-clip-v2	1.89 ± 0.11	1.90 ± 0.10	1.77 ± 0.09	1.93 ± 0.17	1.79 ± 0.09	1.93 ± 0.17	1.84 ± 0.19	1.89 ± 0.18
clip-vit-base-patch32	1.83 ± 0.12	1.87 ± 0.15	1.84 ± 0.14	1.99 ± 0.14	1.87 ± 0.14	2.00 ± 0.14	1.91 ± 0.15	1.93 ± 0.16
clip-vit-large-patch14	1.82 ± 0.23	1.90 ± 0.16	1.81 ± 0.20	1.99 ± 0.18	1.83 ± 0.15	1.98 ± 0.14	1.89 ± 0.17	1.91 ± 0.15
siglip-so400m-patch14-384	2.02 ± 0.09	2.00 ± 0.11	1.85 ± 0.10	2.12 ± 0.13	1.85 ± 0.14	2.02 ± 0.15	1.92 ± 0.14	1.93 ± 0.14
siglip-large-patch16-384	2.16 ± 0.07	2.06 ± 0.07	1.85 ± 0.11	2.05 ± 0.19	1.87 ± 0.12	2.04 ± 0.15	1.94 ± 0.16	1.94 ± 0.15