ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК 004.8

Отчет об исследовательском проекте на тему
Исследование методов персонализированной генерации с помощью диффузионных моделей

Выполнен студентом:

группы #БПМИ222
Гришина Елена Романовна  (подпись) 29.04.25 (дата)

Проверен руководителем проекта:

Аланов Айбек _____ (подпись) _____ (дата)

Кандидат компьютерных наук
Заведующий Центром глубинного обучения и байесовских методов

Москва 2025

Оглавление

Аннотация

Персонализированная генерация изображений — активно развивающееся направление в области компьютерного зрения, позволяющее создавать консистентные визуальные представления объектов в разнообразных сценах на основе текстовых описаний или референсных фотографий. Однако существующие методы сталкиваются с рядом ограничений: высокой вычислительной сложностью, снижением качества при работе со специфичными доменами и недостаточной гибкостью в адаптации к пользовательским запросам.

В данной работе было предложено избавиться от маскирования и использовать сингулярное разложение матриц механизма внимания для контроля над консистентностью объекта при генерации в различных сценах. Для нивелирования проблемы протекания фона было предложено несколько стратегий зануления младших сингулярных чисел. Также были реализованы и исследованы различные техники перевзвешивания сингулярных чисел для повышения визуального сходства объекта внутри батча.

Ключевые слова

Машинное обучение, диффузионные модели, персонализированная генерация изображений

1 Введение

1.1 Актуальность работы

Задача персонализированной генерации подразумевает создание нескольких картинок, на которых изображен один и тот же объект, но в разных сценах и позах. Диффузионные модели - это один из наиболее перспективных классов генеративных моделей, и они демонстрируют выдающиеся результаты в задачах синтеза изображений. Однако, несмотря на значительные успехи, существующие подходы к персонализированной генерации имеют ряд проблем.

Персонализированная генерация изображений имеет широкий спектр применений: от иллюстрирования книг до создания индивидуального контента в медиа и рекламе. В связи с этим улучшение существующих методов или разработка новых подходов, способных генерировать изображения с одним и тем же объектом в различных сценах, является крайне актуальной задачей.

1.2 Цель работы

Современные методы персонализированной генерации изображений обладают рядом ограничений, включая высокие требования к вычислительным ресурсам, риск снижения качества генерации при работе со специфичными наборами данных и адаптации под них, а также недостаточную гибкость в учете пожеланий пользователей.

Целью данного исследования является разработка усовершенствованного подхода персонализированной генерации, направленного на устранение указанных недостатков. Предполагается модификация существующих методов или создание нового, способного повысить консистентность внешнего вида объекта между сгенерированными изображениями с сохранением детализации.

1.3 Постановка задачи

Основные этапы работы:

1. Ознакомиться с существующими подходами к задаче персонализированной генерации и изучить релевантные статьи.

2. Провести анализ SoTA-методов, выявить их преимущества и недостатки. Исследовать влияние различных частей методов на их качество.

3. Выдвинуть гипотезы по улучшению существующих подходов, провести необходимые эксперименты для их подствержения или опровержения.

4. Разработка нового метода персонализированной генерации с использованием диффузионных моделей.

5. Сравнение предложенного подхода с существующими SoTA-методами и анализ полученных результатов.

2 Обзор литературы

Ряд более ранних методов при генерации изображений с новыми объектами опирались на их текстовое описание. Такой подход плох тем, что зависит от качества и полноты текстового описания предметов.

В ряде ранних статей ([17], [4], [5]) авторы клонировали референсный объект на желаемый фон и старались гармонично встраивать его в изображение. Методы [7] и [3] используют 3D-реконструкцию для генерации изображения с объектом с нового ракурса или в измененной позе, однако для этого им необходимо большое количество референсных изображений объекта. Эти подходы не позволяют изменять объекты из обучающей выборки и могут испытывать сложности с освещением и контрастом на изображениях.

Методы на основе GANs и CLIP не универсальны и хорошо себя показывают только на датасетах, специально подобранных под специфичную задачу (например, для редактирования лиц). Также они позволяют локально изменять изображение, но не могут сгенерировать тот же объект в новом контексте. Такие модели, как Imagen [2], DALL-E2 [1] и Stable Diffusion [13] при генерации ориентируются только на текстовый промпт, поэтому бывает очень сложно или невозможно сгенерировать на новом изображении конкретный желаемый объект.

В некоторых работах используются инверсия или пользовательские маски для ограничения редактируемой области и сохранения объекта, но эти методы не подходят для генерации нового идентичного объекта или модификации его внешнего вида. Метод статьи An Image is Worth One Word [11] ограничен полнотой и репрезентативностью замороженной диффузионной модели.

Заново обучать модель на расширенном датасете, в который добавлены изображения с новым концептом, - долго и неэффективно, а дообучение модели на нескольких новых примерах зачастую приводит к забыванию. При замораживании модели и дообучении дополнительных модулей для новых концептов тоже велик шанс забывания ранее выученных знаний.

В статье Textual Inversion [12] предложен новый метод, благодаря которому модель может научиться воспроизводить новый концепт (например, объект или стиль) на основе нескольких изображений-примеров, причем их достаточно всего лишь 3-5. Идея подхода заключается в том, чтобы в пространстве эмбеддингов предобученной и замороженной text-to-image модели найти эмбеддинг, соответствующий новому концепту S_*. На самом деле обучается не единственный новый эмбеддинг, а целиком эмбединговый слой энкодера, но после обучения все вектора кроме целевого возвращаются к исходным значениям. Впоследствии псевдослово S_* можно использовать в промтах как обычное слово языка (например, так: “A photo of S_* on the beach”). Получается, что для представления нового концепта достаточно всего одного ‘слова’ - S_*.

Метод DreamBooth [8] сохраняет детали объекта при переносе его в новые сцены. Авторы добиваются такого результата, привязывая объект к уникальному идентификатору из текстового промпта и используя при fine-tuning предобученной text-to-image модели такой лосс, который штрафует модель как за значительное изменение объекта из обучающей выборки, так и за забывание ранее выученных знаний. Для обучения модели новому объекту достаточно $3 - 5$ референсных изображений. Этот метод также позволяет по запросу изменять внешний вид нового объекта (например, цвет, позу, ракурс).

CustomDiffusion [9] - метод персонализированной генерации для создания новых высококачественных изображений с уникальным объектом, который модель видела только на небольшом числе референсных изображений. Суть предложенного подхода заключается в оптимизации лишь малой части параметров предобученной модели при ее дообучении на новый объект, что вычислительно эффективнее и быстрее раннее предложенных методов и требует меньших затрат по памяти. Во избежание забывания моделью ранее выученных знаний, авторы используют реальные изображения для регуляризации. Предложенный подход также позволяет дообучать модель на несколько новых объектов одновременно и впоследствии гармонично генерировать их на одном изображении.

3 Воспроизведение основных бейзлайнов

3.1 Датасеты

Исследуемые методы не требуют ни дообучения, ни использования референсных изображений для генерации объектов в новых сценах, что исключает необходимость использования датасетов. В ходе работы были подобраны по 10 примеров, иллюстрирующих как успешные, так и плохие результаты работы бейзлайнов. Данные кейсы будут использоваться для тестирования разрабатываемого метода и анализа его качества. Более подробно про выделенные примеры - в разделе 3.5.

3.2 Метрики

Результаты экспериментов будут оцениваться преимущественно на основе визуального сравнения.

Также будут использованы критерии соответствия сгенерированных изображений текстовым промптам (prompt-alignment) и консистентности объектов (subject consistency), а при анализе результатов предлагается смотреть на их соотношение. Для оценки соответствия промптам используется модель CLIP, чтобы измерить сходство между каждым сгенерированным изображением и соответствующим промптом. Результатом служит усредненный CLIPScore [6]. Для оценки консистентности сгенерированных объектов может быть использована модель DreamSim [14], которая, как показали исследования, лучше коррелирует с человеческой оценкой сходства между изображениями.

3.3 Список бейзлайнов

ConsiStory

ConsiStory[1] [16] - это *training-free* метод персонализированной генерации, который показывает хорошее качество генерации одного и того же объекта в разных сценах благодаря нескольким идеям:

1. SDSA (Subject-driven self-attention). Изменение стандартного self-attention так, чтобы матрицы запросов W^Q влияли на матрицы ключей W^K и значений W^V. Для реализации этой идеи авторы используют в механизме внимания матрицы K^+, V^+, полученные конкатенацией матриц K_i, V_i батча.

[1] https://research.nvidia.com/labs/par/consistory/

Также авторы добавляют в этой составляющей метода маскирование, чтобы обмен информацией происходил только между областями, содержащими объект.

2. Взвешенное смешивание матриц Q стандартного блока self-attention и из SDSA. Применяется для повышения разнообразия изображений.
3. Self-Attention Dropout. Случайное зануление патчей в маске self-attention блока. Контролирует похожесть объектов.
4. Feature injection. Использование DIFT-матрицы для нахождения похожих патчей на сгенерированных изображениях для их последующего взвешенного смешивания с целью дополнительного обмена информацией между деталями внешнего вида объекта.

- Для оптимизации вычислений авторы метода ConsiStory используют в компонентах SDSA и Feature Injection ‘якорные’ изображения - подмноженство сгенерированных изображений, которое можно считать референсным. Только эти изображения могут влиять на другие в данных механизмах: они оказывают влияние на матрицы ключей и значений остальных изображений, а их патчи взвешенно суммироваться для повышения консистентности деталей объектов. Это также позволяет повысить качество генерации изображений при больших размерах батча.

3.4 Конфигурация и параметры

Для воспроизведения результатов бейзлайнов используется авторский код с дефолтными гиперпараметрами, если не указано иное. Эксперименты запускаются с теми же версиями библиотек и установленными зависимостями, которые указаны в авторских репозиториях.

Для проведения экспериментов использоввался суперкомпьютерный комплекс НИУ ВШЭ [10] с видеокартами V100.

3.5 Результаты и выводы

Примеры работы ConsiStory

В ходе проведенного исследования были проанализированы результаты бейзлайна. В частности, было отобрано 10 примеров успешной работы метода ConsiStory, демонстрирующих высокую степень согласованности внешнего вида объекта в различных сценах. Эти примеры подтверждают, что он способен сохранять облик объекта и его характерные черты при генерации его в разных контекстах, заданных текстовыми промптами. Например, на Рис. 3.1 в первой строке представлены три изображения, на которых изображена одна и та же девушка в разных контекстах, что свидетельствует о хорошей работе метода.

Также были выявлены 10 случаев, в которых ConsiStory показал недостаточное качество. В этих примерах наблюдаются значительные расхождения во внешнем виде объекта на сгенерированных изображениях, что указывает на ограничения метода в обеспечении консистентности. Так, во второй строке Рис. 3.1 продемонстрирован пример неудачной работы бейзлайна: на сгенерированных изображениях внешний вид подарочной коробки существенно различается. Выделенные случаи подчеркивают необходимость дальнейшего совершенствования метода.

Рис. 3.1: Примеры хорошей и плохой работы метода ConsiStory

Эксперименты показали, что метод как правило демонстрирует высокое качество при генерации изображений с животными (например, если объектом являются леопард или дельфин), сохраняя консистентность их визуальных особенностей между изображениями в различных сценах. Однако для неодушевленных объектов (автобус, коробка) зачастую наблюдается снижение стабильности: их внешний вид варьируется в зависимости от контекста.

Влияние гиперпараметров на работу метода

Было проведено исследование влияния ключевых параметров метода и его частей на соответствие сгенерированных изображений текстовым промптам, качество генерации объектов и их консистентность между картинками. Результаты приведены на Рис 3.3 и 3.2.

same_latent. `default: False`. Этот параметр определяет, из одного ли шумового вектора генерируется серия изображений одного объекта в различных сценах. По умолчанию каждое изображение в батче получается в путем расшумления уникального латента, что способствует повышению визуального разнообразия результатов. При значении параметра `same_latent=True` все изображения в серии генерируются из одного и того же гауссова шума. Это приводит к увеличению консистентности внешних особенностей объекта, но снижает вариативность его поз.

perform_sdsa. `default: True`. Данный параметр управляет активацией компоненты метода, реализующей идею SDSA. Обмен информацией между матрицами механизма внимания повышает консистентность изображений, а если параметр отключен (`perform_sdsa=False`), то при генерации изображений реализуется стандартный self-attention, что приводит к снижению похожести объектов. Авторы метода отмечают, что данная составляющая метода является ключевой, что подтверждается проведенными экспериментами.

mask_dropout. `default: 0.5`. Для повышения стабильности генерации и управления балансом между сохранением идентичности объекта и вариативностью авторы метода предлагают применять зануление mask_dropout случайно выбранных патчей

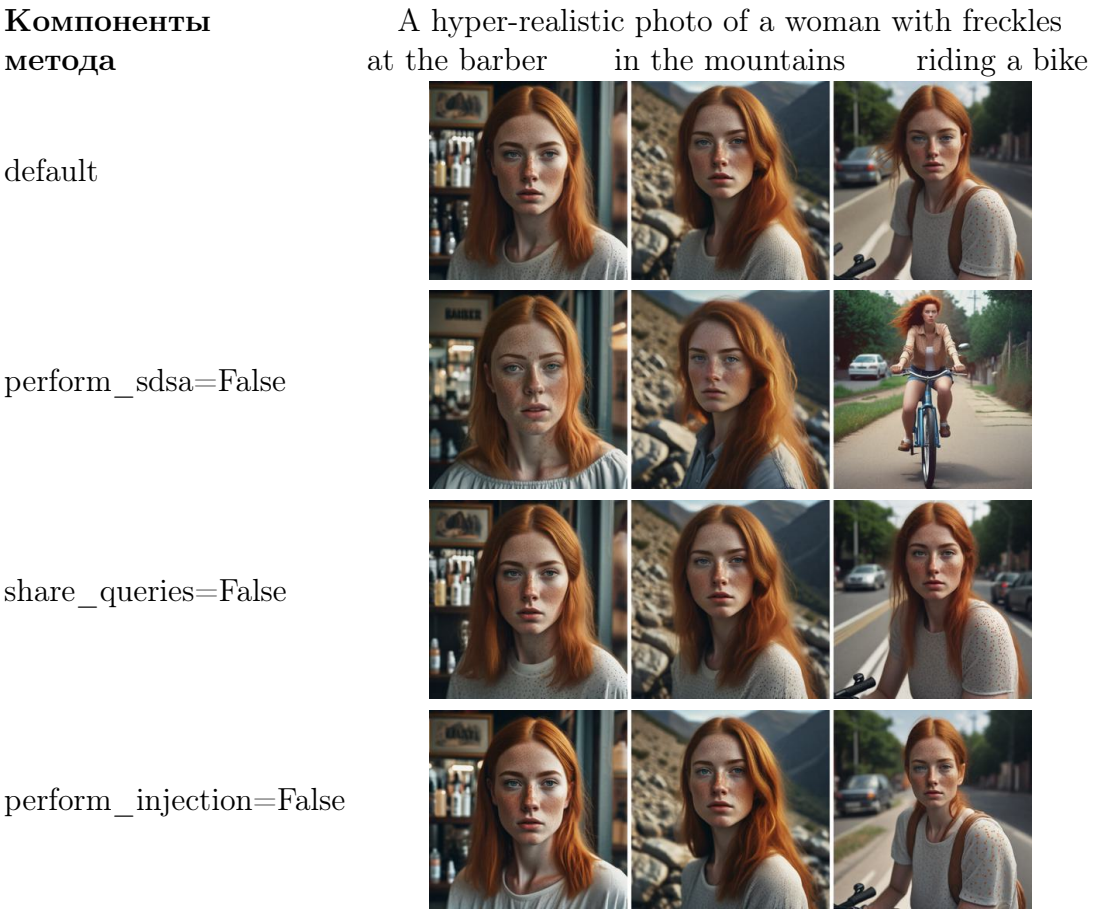

Рис. 3.2: Влияние компонент метода ConsiStory на его качество

в маске, используемой в SDSA. Снижение значения данного параметра усиливает консистентность особенностей внешнего вида объекта, но значительно снижает разнообразие поз и ракурсов.

share_queries. `default: True`. Данный параметр показывает, будет ли применяться взвешенное комбинирование матриц Q из механизма SDSA и стандартного self-attention путем их взвешенного суммирования. При его деактивации матрица Q формируется только на основе модифицированного механизма внимания, что повышает похожесть внешнего вида объекта в различных сценах, но снижает вариативность изображений в батче.

perform_injection. `default: True`. При активации данного параметра реализуется идея Feature Injection, направленного на дополнительное повышение идентичности внешнего вида объекта на сгенерированных изображениях. При выключении этой составляющей метода (`perform_injection=False`) наблюдается снижение консистентности и визуального сходства объектов. В большинстве случаев это не оказывает критического влияния на качество генерации, но может привести к незначительным различиям в детализации.

n_anchors. `default: 2`. Параметр n_anchors определяет количество ‘якорных’ изображений, используемых в компонентах SDSA и feature injection для повышения консистентности внешнего вида объекта в различных сценах. При использовании 3 референсных изображений сходство снижается, и особенности объекта различаются между изображениями. При `n_anchors=1` подход недостаточно переносит его ключевые детали. Дефолтное значение `n_anchors=2` является оптимальным и обеспечивает

Параметры	A hyper-realistic photo of a woman with freckles at the barber	in the mountains	riding a bike
default			
same_latent=True			
mask_dropout=0.2			
mask_dropout=0.8			
perform_injection=False			
n_anchors=1			
n_anchors=3			

Рис. 3.3: Влияние параметров метода ConsiStory на его качество

баланс, способствуя консистентности деталей объекта и сохраняя разнообразие сцен.

3.6 Анализ недостатков бейзлайнов и возможных улучшений методов

На Рис. 3.1 приведен пример неудачной работы метода. Также экспериментально было выявлено, что ConsiStory хуже справляется с генерацией таких неодушевленных объектов, как коробка или автобус. Например, объект не гармонично встраивается в сцену, или его внешний вид существенно отличается на различных изображениях.

Одним из ключевых недостатков ConsiStory является зависимость его ключевого компонента (SDSA) от точности локализации объекта в изображении. В случае нетипичных объектов это может приводить к усилению расхождений во внешнем виде объектов. В данной работе предполагается преодолеть эту проблему и усовершенствовать метод, предложив другой способ модифицировать стандартный self-attention для повышения консистентности объекта на различных изображениях.

4 Эксперименты и гипотезы

В данной работе мы проводим эксперименты на основе метода ConsiStory и меняем его ключевую составляющую SDSA, стараясь избавиться от необходимости маскирования объектов в методе, но сохранить их консистентность между изображениями и качество генерации.

4.1 Гипотезы

Основная гипотеза состоит в том, что старшие сингулярные числа матриц K и V в self-attention несут информацию о самых значимых частях соответствующих изображений и, в частности, об объектах на них. Тогда зануление младших сингулярных чисел для ‘якорных’ (референсных) изображений в процессе применения механизма внимания может снизить протекание фона, но не уменьшит консистентность ключевых объектов.

Следующая гипотеза заключается в том, что увеличение старших сингулярных чисел поможет сохранить концепт на изображениях ярче. В результате этого может увеличиться похожесть объектов между изображениями.

4.2 Эксперименты

Измененный механизм внимания в методе ConsiStory выглядит следующим образом:
$K^+ = [K_1 \oplus ... \oplus K_N]$
$V^+ = [V_1 \oplus ... \oplus V_N]$
$A_i^+ = softmax(Q_i {K^+}^T / \sqrt{d_k} + \log M_i^+)$
$h_i = A_i^+ \cdot V^+$

Здесь $\oplus$ - оператор конкатенации, Q_i, K_i, V_i - матрицы запросов, ключей и значений для $i^{\text{ого}}$ изображения батча соответственно, M_i^+ реализует маскирование объектов. Для проверки гипотез мы убираем из метода маскирование, чтобы оценивать зависимость между сингулярными числами и деталями генерируемых изображений.

Имплементация всех проведенных экспериментов доступна в репозитории на GitHub[2].

[2] https://github.com/GrishHelen/personalized-generation

Зануление младших сингулярных чисел

В процессе тестирования первой гипотезы при генерации $i^{\text{ого}}$ изображения в модифицированном self-attention были использованы матрицы K^+, V^+, в которых занулены младшие сингулярные числа. Также были протестированы различные тактики контроля количества оставленных главных компонент в матрицах.

В первом блоке экспериментов оставляются N старших сингулярных чисел, а остальные приравниваются к 0. Основными достоинствами этого подхода являются простота реализации и четкий контроль над степенью сжатия изображения. Однако в различных сеттингах количество значимых сингулярных чисел может существенно различаться. Таким образом, стратегия сохранения фиксированного числа главных компонент накладывает жесткие ограничения на количество оставляемой информации с исходного изображения, и ставит перед необходимостью подбора оптимального значения параметра N для каждого отдельного случая. Чтобы усовершенствовать этот подход, предлагается выделять количество значимых сингулярных чисел по квантилю.

Следующий эксперимент тестирует предположение о том, что определять количество главных компонент, которые содержат информацию об объекте на изображении, можно на основе значения квантиля для выборки из сингулярных чисел. Для этого предполагается вычислять q-квантиль значений сингулярных чисел матрицы и занулять те, что меньше полученной величины. Данный подход дает больше гибкости в контроле сохраняемых главных компонент, а также легко интерпретируемый, хотя не учитывает распределение значений сингулярных чисел и структуру матриц. Для преодоления недостатков этой тактики можно рассмотреть критерий, опирающийся на значения сингулярных чисел при подборе количества зануляемых компонент.

Энергетический критерий позволяет автоматически определять количество компонент, содержащих основную информацию об изображении. Его использование дает возможность оставлять такое минимальное количество сингулярных чисел N, что их кумулятивная энергия $E(k)$ превышает заданный порог q: $N = \min\{k|E(k) > q\}$, где $E(k) = \frac{\sum\limits_{i=1}^{k} \sigma_i^2}{\sum\limits_{i=1}^{r} \sigma_i^2}$. Главное достоинство подхода на основе энергии заключается в том, что это математически строгий и эффективный способ выбора значимых компонент. Также он учитывает распределение сингулярных чисел матрицы и позволяет достигать заданного уровня сохранения информации, поэтому является оптимальным для тестирования выдвинутой гипотезы.

Влияние матриц K и V на результат

Ранее предложенная идея заключается в том, чтобы оставлять только небольшую часть главных компонент матриц K и V из self-attention, зануляя остальные. Эти матрицы называются матрицами ключей и значений соответственно и имеют разную интерпретацию: можно считать, что K (keys) формирует “карту значимости”, показывая, каким векторам латентного представления релевантен некоторый вектор (из матрицы Q), а V (values) отвечает за содержание и визуальное наполнение выбранных релевантных областей. В механизме внимания эти две матрицы по-разному влияют на результат преобразования и отличаются по смыслу, поэтому изменение соотношения количества оставляемых в них сингулярных чисел может играть суще-

ственную роль. Для исследования этого предположения предлагается использовать два независимых параметра n_svd_K и n_svd_V, каждый из которых будет регулировать долю сингулярных чисел, сохраняемых в соответствующих матрицах. Такая параметризация предоставляет возможность экспериментально исследовать влияние различных степеней сжатия информации в K и V на консистентность объекта между изображениями, протекание фона и качество генерации в целом.

Перевзвешивание сингулярных чисел

Согласно второй выдвинутой гипотезе, усиление старших сингулярных чисел позволит улучшить выраженность ключевого концепта и повысить его сходство между изображениями. Предполагается, что управление значениями сингулярных чисел может помочь улучшить консистентность объекта.

В целях повышения выраженности ключевого концепта и снижения эффекта протекания фона между изображениями предлагается модификация механизма внимания, основанная на перевзвешивании сингулярных чисел матриц K и V для 'якорных' изображений. Ожидается, что ускорение их затухания путем перевзвешивания позволит увеличить фокус на объект. В статье [15] был предложен метод, использующий экспоненциальное преобразование сингулярных чисел для матрицы латентрых представлений промптов. Мы используем эту идею и предлагаем для нашей задачи тоже применять перевзвешивание сингулярных чисел по экспоненте: $\hat{\Sigma} = \Sigma \cdot e^{\alpha\Sigma}$. Данный подход направлен на усиление акцента на ключевом концепте при одновременном подавлении нежелательных фоновых артефактов, что может привести к повышению консистентности объекта при генерации в различных сценах.

С другой стороны, ускорение затухания сингулярных чисел фокусирует внимание метода на небольшом количестве главных компонент матриц K и V в self-attention, что может привести к утрате информации об уникальных чертах объекта, сосредоточенной дальше в спектре. Для сохранения информации как о концепте в целом, так и о более мелких визуальных особенностях, требуется коррекция динамики затухания сингулярных чисел, направленная на их выравнивание. Такой подход способствует сглаживанию перепада и замедлению их затухания. В этой модификации механизма внимания предлагается использовать для перевзвешивания сингулярных чисел логарифмическую функцию для сглаживания спектра матриц K и V по Алгоритму 1. Однако этот прием повышает риск протекания фона и возникновения нежелательных артефактов на изображениях. Чтобы нивелировать данный эффект, можно дополнить метод ранее рассмотренными техниками обнуления младших сингулярных чисел, ответственных за фоновые и мелкие элементы. Таким образом, предлагается рассмотреть и комбинированный подход: замедлять затухание сингулярных чисел матриц K и V в механизме внимания, выравнивая их значения, и одновременно занулять младшие сингулярные числа, которые несут информацию о второстепенных деталях изображения. Ожидается, что этот метод обеспечит усиление акцента на объекте с сохранением его визуальных особенностей между изображениями и поможет избавиться от проблемы протекания фона.

4.3 Результаты и выводы

При исключении механизма маскирования из метода ConsiStory, наблюдаются две существенные проблемы на генерируемых изображениях. Во-первых, начинает

Algorithm 1 Перевзвешивание сингулярных чисел по логарифму

$\hat{\Sigma} = \log(\Sigma)$ # перевзвешивание σ_i
$\hat{\Sigma}[\hat{\Sigma} < 0] = 0$
$\hat{\Sigma} = \hat{\Sigma} \cdot \frac{\|\Sigma\|}{\|\hat{\Sigma}\|}$

протекать фон между изображениями батча, что приводит к визуальным артефактам. Второй проблемой является снижение консистентности объекта, изображенного в различных сценах.

Зануление младших сингулярных чисел

Проведенные эксперименты подтвердили справедливость первой гипотезы: старшие сингулярные числа действительно отвечают за основную информацию о внешнем виде объекта на изображении. Вклад последующих компонент носит уточняющий характер, влияя на детализацию.

Стратегия оставлять фиксированное число N главных компонент помогла быстро протестировать гипотезу и получить результаты, которые ее подтвердили. Тем не менее, этот вариант не подходит для использования на практике, потому что тяжело интерпретируем.

В отличие от предыдущего подхода, квантильный метод автоматически определяет пороговое значение на количество сингулярных чисел, зануляя компоненты, величина которых ниже рассчитанного квантиля. Как и ожидалось, эксперименты показали, что в зависимости от текстовых промптов, количества генерируемых изображений и сложности сцен метод требует различных порогов на число сохраняемых главных компонент, потому что не учитывает их распределение. Эта вариативность приводит к необходимости ручного подбора параметра q-квантиля для каждого отдельного случая, что существенно снижает универсальность подхода. Однако для большинства примеров сохранение 5% ($q = 0.05$) старших сингулярных чисел оказалось оптимальным, позволяя соблюсти баланс между нивелированием проблемы протекания фона и сохранением внешнего вида объекта.

Энергетический критерий демонстрирует устойчивость к вариативности данных и обеспечивает контроль над уровнем сохраняемой информации, что делает его предпочтительным для задач генерации изображений. Эксперименты показали, что для того, чтобы шэринг матриц ключей и значений в механизме внимания повышал похожесть внешнего вида объектов между изображениями, необходимо оставлять сингулярные числа, в которых содержится хотя бы 90% их общего веса. На Рис. 4.1 показано распределение сингулярных чисел матриц K и V для некоторых итераций диффузионного процесса. Графики позволяют понять, что основная энергия содержится в небольшом количестве главных компонент, а младшие сингулярные числа имеют незначительный вес, поэтому дальнейшее увеличение параметра q добавляется сразу большое число менее значимых компонент. Это немного улучшает консистентность объекта, но одновременно провоцирует появление фоновых артефактов. Таким образом, энергетический критерий крайне чувствителен к значению параметра, который контролирует долю сохраняемых сингулярных чисел на основе их веса, и требует очень тонкого подбора его величины. В ходе экспериментов было установлено, что оптимальное значение данного параметра равно $q \approx 0.95$.

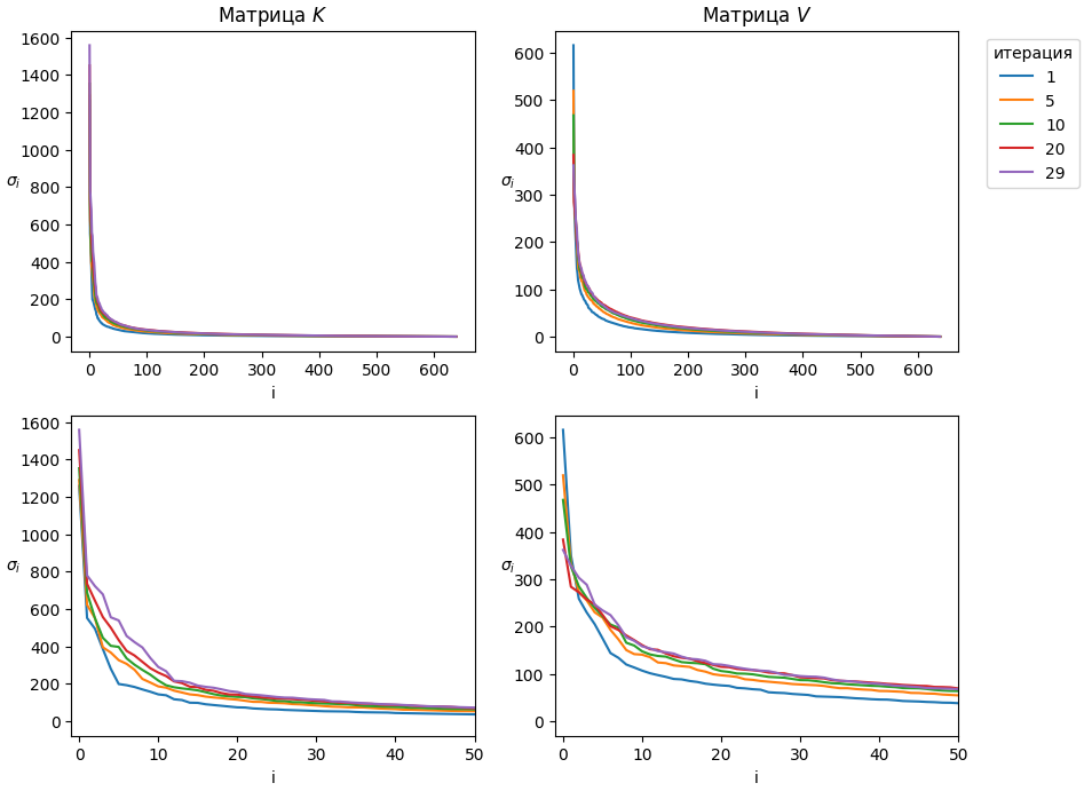

Рис. 4.1: Сингулярные числа матриц K и V

Влияние матриц K и V на результат

Проведенные эксперименты подтвердили, что влияние матриц ключей и значений в self-attention на результат генерации различается.

Матрица K, ответственная за формирование 'карты значимости', оказывает существенное влияние на фоновые элементы изображения. Так, увеличение доли сохраняемых главных компонент в матрице ключей (n_svd_K) приводит к усилению эффекта протекания фона между изображениями в батче, поэтому подбор оптимального значения данного параметра является критически важным при генерации объекта в различных сценах. Количественный анализ позволил выявить, что при значениях $\text{n\_svd}_K \geq 0.1$ появляются нежелательные фоновые артефакты, а сохранение $1 - 5\%$ сингулярных чисел обеспечивает баланс между их предотвращением и сохранением сцены и объекта на изображении.

Также исследование показало, что количество сохраняемых сингулярных чисел матрицы значений в механизме внимания напрямую влияет на степень сходства генерируемого объекта между изображениями. В результате экспериментов было установлено, что для достаточной консистентности объектов необходимо оставлять $5 - 10\%$ главных компонент матрицы V. При дальнейшем увеличении параметра n_svd_V, который контролирует количество сохраняемых сингулярных чисел, существенного увеличения сходства деталей объектов не наблюдается, но может снижаться вариативность его поз и возникать проблема протекания фона.

Полученные результаты указывают на различие между матрицами K и V в модифицироанном self-attention и их ролью в формировании выходных изображений. Таким образом, изменение параметров n_svd_K и n_svd_V, отвечающих за количество сохраняемых сингулярных значений для этих матриц, позволяет одновременно достичь минимизации фоновых артефактов между изображениями и повышения

консистентности объекта при генерации в различных сценах.

Перевзвешивание сингулярных чисел

Результаты экспериментального исследования продемонстрировали, что использование экспоненциального преобразования сингулярных чисел не только не приводит к ожидаемому повышению консистентности объектов, но и оказывает противоположный эффект. Домножение сингулярных чисел на их экспоненту ускоряет их затухание, что влечет за собой сокращение количества главных компонент, которые оказывают существенное влияние на результат обмена информацией между матрицами в механизме внимания. Вопреки первоначальным предположениям, данный подход усиливает различия внешнего вида объекта в различных сценах и приводит к снижению их визуального сходства, что подтверждает необходимость исследования альтернативных техник перевзвешивания сингулярных чисел.

Также было предложено использовать логарифмическую функцию для перевзвешивания сингулярных чисел. На Рис. 4.2 показано распределение спектра матрицы K при различных техниках их перевзвешивания. Графики показывают, что использование логарифма сглаживает сингулярные числа и замедляет их затухание. Также результаты экспериментов подтвердили эффективность предложенного подхода логарифмического перевзвешивания сингулярных чисел матриц K и V в механизме внимания. В то время как стандартная реализация метода ConsiStory показывает существенные расхождения объекта между изображениями между кадрами, предложенное решение обеспечивает более хорошие результаты. Как видно из примеров, показанных на Рис. 4.3, применение данной методики привело к заметному повышению визуального сходства генерируемых объектов, а протекание фона незначительно либо не критично.

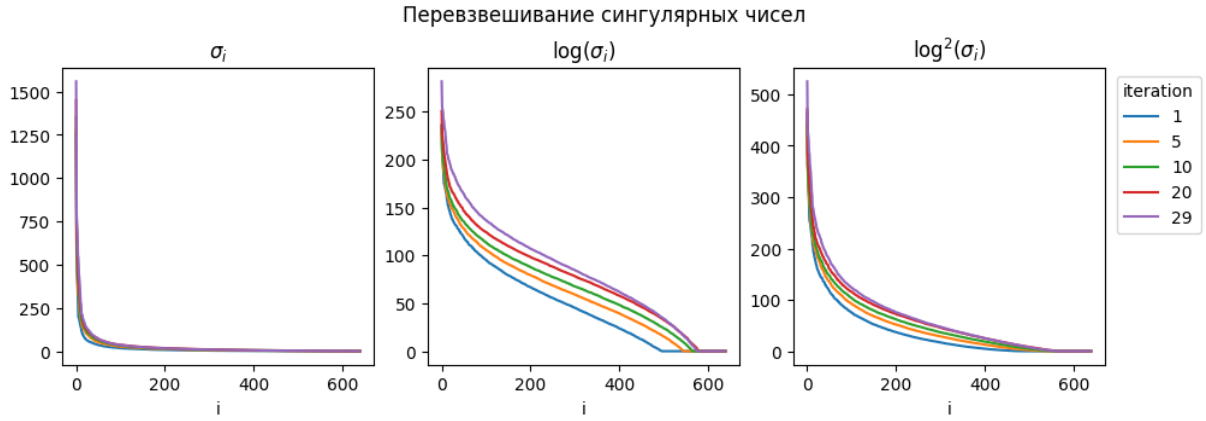

Рис. 4.2: Спектр матрицы K при перевзвешивании

Рис. 4.3: Сравнение методов

Из примера с автобусом видно, что на первых двух изображениях он стал выглядеть одинаково, а при логарифмическом перевзвешивании сингулярных чисел он имеет такую же форму на третьем изображении, но отличается цветом и другими визуальными особенностями. Хотя в данном случае результат не идеальный, он демонстрирует значительный прогресс по сравнению с исходными результатами и потенциал предложенной методики. Приведенный на этом же рисунке пример с врачом доказывает перспективность функции $log^2(\sigma_i)$ для перевзвешивания сингулярных чисел: благодаря ее использованию удалось добиться консистентности внешнего вида объекта на всех трех изображениях. В данном случае наблюдается умеренное протекание фона, но в дальнейших экспериментах применение подходов по зануleнию младших сингулярных чисел может помочь преодолеть эту проблему. Таким образом, мы получили удачный результат работы предложенного метода, который указывает на улучшение бейзлайна.

5 Заключение

5.1 Выводы

В данной работе было проведено исследование о влиянии изменения сингулярных чисел матриц ключей K и значений V в механизме внимания на качество генерации и консистентность объекта в различных сценах. Проведенные эксперименты подтвердили справедливость первой гипотезы, доказав, что главные компоненты содержат основную информации о концепте, а частичное зануление младших сингулярных чисел в процессе применения модифицированного self-attention позволяет снизить протекание фона при отсутствии маскирования объектов. Также было установлено, что матрица K преимущественно определяет фоновые детали, а матрица V больше влияет на сходство объектов. Тестирование различных стратегий сохранения сингулярных чисел показало превосходство квантильного подхода и энергетического критерия.

Первоначальное предположение о том, что увеличение скорости затухания сингулярных чисел будет способствовать усилению концепта и увеличению визуального сходства объекта между изображениями батча, не подтвердилось. Напротив, сглаживание спектра матриц путем логарифмического преобразования позволило повысить конситентность. Полученные результаты показывают, что предложенный метод логарифмического перевзвешивания сингулярных чисел, несмотря на существующие ограничения, демонстрирует значительное улучшение в сравнении с бейзлайном и может рассматриваться как перспективное направление для дальнейших исследований.

5.2 Направления дальнейшей работы

В данной работе были исследованы различные техники, направленные на усиление консистентности внешнего вида объекта между изображениями батча. Для этого были рассмотрены несколько идей, основанные на изменение распределения сингулярных чисел матриц ключей и значений из self-attention. Однако во всех экспериментах неименным оставалось то, что в начале генерации фиксировались несколько ‘якорных’ изображений, на которые делался шэринг в механизме внимания. Таким

образом, мы старались приблизить внешний вид объекта на остальных (таргетных) изображениях к представленному на этих референсных изображениях.

В работе [18] было предложено при каждом вызове self-attention использовать случайную картинку батча как референсную, и мы хотим попробовать применить эту идею в нашей работе. Благодаря такому подходу мы будем приближать объект со всех изображений к некоторому усредненному виду, а не тому, который представлен на заранее зафиксированных референсных изображениях. В совокупности с этим методом можно использовать и все подходы, рассмотренные в нашей работе, для повышения внешнего сходства объектов между изображениями и нивелирования проблемы протекания фона.

Литература

[1] Alex Nichol Casey Chu Mark Chen Aditya Ramesh, Prafulla Dhariwal. Hierarchical text-conditional image generation with clip latents. 2022.

[2] Saurabh Saxena Lala Li Jay Whang Emily Denton Seyed Kamyar Seyed Ghasemipour Burcu Karagol Ayan S. Sara Mahdavi Rapha Gontijo Lopes Tim Salimans Jonathan Ho David J Fleet Mohammad Norouzi Chitwan Saharia, William Chan. Photorealistic text-to-image diffusion models with deep language understanding. 2022.

[3] Ben Mildenhall Todd Zickler Jonathan T. Barron Pratul P. Srinivasan Dor Verbin, Peter Hedman. Ref-nerf: Structured view-dependent appearance for neural radiance fields. 2021.

[4] ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing. Chen-hsuan lin, ersin yumer, oliver wang, eli shechtman, simon lucey. 2018.

[5] Junge Zhang Kaiqi Huang Huikai Wu, Shuai Zheng. Gp-gan: Towards realistic high-resolution image blending. 2017.

[6] Maxwell Forbes Ronan Le Bras Yejin Choi Jack Hessel, Ari Holtzman. Clipscore: A reference-free evaluation metric for image captioning. 2021.

[7] Abhishek Kar Yuanzhen Li Deqing Sun Jonathan T. Barron Hendrik P. A. Lensch Varun Jampani Mark Boss, Andreas Engelhardt. Samurai: Shape and material from unconstrained real-world arbitrary image collections. 2022.

[8] Varun Jampani Yael Pritch Michael Rubinstein Kfir Aberman Nataniel Ruiz, Yuanzhen Li. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

[9] Richard Zhang Eli Shechtman Jun-Yan Zhu Nupur Kumari, Bingliang Zhang. Multi-concept customization of text-to-image diffusion. 2022.

[10] Viachesvlav Kozyrev Pavel S. Kostenetskiy, Roman Chulkevich. Hpc resources of the higher school of economics. *Journal of Physics Conference Series*, 2021.

[11] Yuval Atzmon Or Patashnik Amit H. Bermano Gal Chechik Daniel Cohen-Or Rinon Gal, Yuval Alaluf. An image is worth one word: Personalizing text-to-image generation using textual inversion. 2022.

[12] Yuval Atzmon Or Patashnik Amit H. Bermano Gal Chechik Daniel Cohen-Or Rinon Gal, Yuval Alaluf. An image is worth one word: Personalizing text-to-image generation using textual inversion. 2022.

[13] Dominik Lorenz Patrick Esser Björn Ommer Robin Rombach, Andreas Blattmann. High-resolution image synthesis with latent diffusion models. 2021.

[14] Shobhita Sundaram Lucy Chai Richard Zhang Tali Dekel Phillip Isola Stephanie Fu, Netanel Tamir. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. 2023.

[15] Senmao Li Joost van de Weijer Fahad Shahbaz Khan Shiqi Yang-Yaxing Wang Jian Yang Ming-Ming Cheng Tao Liu, Kai Wang. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. 2025.

[16] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024.

[17] Li Niu Liu Liu Zhixin Ling Weiyuan Li Liqing Zhang Wenyan Cong, Jianfu Zhang. Dovenet: Deep image harmonization via domain verification. 2019.

[18] Ming-Ming Cheng Jiashi Feng Qibin Hou Yupeng Zhou, Daquan Zhou. Storydiffusion: Consistent self-attention for long-range image and video generation. 2024.