

**Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»**

**Факультет компьютерных наук
Основная образовательная программа
«Прикладная математика и информатика»**

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ
РАБОТА**

**Исследовательский проект на тему
Персонализированная генерация лиц с
помощью генеративных моделей**

**Выполнил студент группы 211, 4 курса,
Матосян Александр Акобович**

**Руководитель ВКР:
Кандидат Наук, Аланов Айбек**

Москва 2025

Содержание

Аннотация	4
1 Введение	6
2 Обзор литературы	8
2.1 Диффузионные модели	8
2.2 Fine-tuning based методы персонализированной генерации	8
2.2.1 Textual Inversion	9
2.2.2 DreamBooth	9
2.3 Encoder-based методы персонализированной генерации	9
2.3.1 IPAdapter	9
2.3.2 PhotoMaker	10
2.3.3 PuLID	11
3 Направление исследования	12
3.1 Мотивация	12
3.2 Гипотеза	12
3.3 Целевой метод	12
3.3.1 Разделенный промпт	13
3.3.2 Архитектура	13
3.3.3 Обучение	14
4 Протокол проведения экспериментов	14
4.1 Стартовый метод	14
4.2 Метрики	14
4.3 Данные	15
4.4 Детали имплементации	16
5 Эксперименты	17
5.1 Эксперименты по воспроизведению	17
5.1.1 Эксперимент 1	18
5.1.2 Эксперимент 2	18
5.1.3 Эксперимент 3	18
5.1.4 Эксперимент 4	18

5.1.5	Промежуточные выводы	19
5.1.6	Эксперимент 5	19
5.1.7	Эксперимент 6	19
5.1.8	Эксперимент 7	20
5.2	Построение целевого метода	20
5.2.1	Эксперимент 1	20
5.2.2	Эксперимент 2	21
5.2.3	Эксперимент 3	21
5.2.4	Эксперимент 4	21
6	Итоговый метод	22
6.1	Описание метода	22
6.2	Количественные результаты	22
6.3	Примеры работы	23
7	Заключение	25
	Список литературы	26
A	Аппендиц	28
A.1	Некорректные эксперименты	28
A.1.1	Эксперимент 1	28
A.1.2	Эксперимент 2	28
A.1.3	Эксперимент 3	28
A.1.4	Эксперимент 4	28
A.1.5	Эксперимент 5	29
A.1.6	Выводы	29

Аннотация

Персонализированная генерация лиц — одно из перспективных направлений в области компьютерного зрения и генеративного моделирования. Оно находит широкое применение в создании цифровых аватаров, виртуальной примерке и анимировании изображений. Ключевыми требованиями к методам персонализированной генерации лиц являются сохранение уникальных черт заданного лица и возможность управления компонентами изображения, не связанными с лицевыми характеристиками, такими как фон, выражение лица, поза, стиль.

Появление диффузионных моделей открыло новые возможности, позволив генерировать более разнообразные и реалистичные изображения, и способствовало развитию методов персонализации. Однако, существующие решения ещё не справились с удержанием баланса между контролируемостью генерации и сохранением определяющих черт заданного лица.

В данной работе рассматривается существующее состояние задачи и предлагается новый метод персонализированной генерации лиц, обеспечивающий более тонкий контроль над результатом. Наш подход основан на современных state-of-the-art методах, в частности на PhotoMaker [4] и IPAdapter [13]. Также, мы описываем процесс сбора, специализированного под нашу задачу, датасета.

Кодовая база нашей работы доступна по ссылке: <https://github.com/matosjan/persongen>

Abstract

Personalized face generation is one of the promising directions in the field of computer vision and generative modeling. It finds wide application in creating digital avatars, virtual try-on, and image animation. The key requirements for personalized face generation methods are preserving the unique features of the given face and the ability to control image parts unrelated to the identity, such as background, facial expression, pose, and style.

The rise of diffusion models has opened new possibilities, enabling the generation of more diverse and realistic images and advancing personalization methods. However, existing solutions have yet to achieve a balance between generation controllability and the preservation of identity.

This work examines the current state of the field and proposes a new method for personalized face generation that provides detailed control over results. Our approach builds on modern state-of-the-art methods, particularly PhotoMaker [4] and IPAdapter [13]. Additionally, we describe the process of collecting a specialized dataset for our task.

Ключевые слова

Глубинное обучение, компьютерное зрение, генеративные модели, диффузионные модели

1 Введение

Персонализированная генерация людей является одним из перспективных направлений в компьютерном зрении и генеративном моделировании. Данная задача имеет высокую прикладную ценность, так как широко применима в таких приложениях как: создание цифровых аватаров, виртуальная примерка, анимирование изображений.

Общепринятой формулировкой задачи персонализированной генерации является следующая - имея текстовый промпт и пример изображения, содержащего персонализируемый объект(далее концепт), требуется сгенерировать новое изображение, которое содержит нужный концепт и при этом соответствует текстовому запросу пользователя. В данной работе исследуется подзадача персонализированной генерации, которая концентрируется на синтезе изображений людей, в особенности человеческих лиц.

У методов персонализированной генерации лиц есть два основных качества. Первое, это сохранение *identity* - сгенерированное лицо должно точно передавать уникальные черты заданного человека. Второе, это контролируемость через текстовые описания - насколько гибко можно менять атрибуты не связанные с личностью человека, такие как выражение лица, фон, поза, стиль и т.д. Зачастую, возникает компромисс между этими двумя качествами. Чем больше согласованность генерации с текстовым промптом, тем менее точна передача индивидуальных черт человека. И наоборот, если акцент сделан на высокое качество сохранения *identity*, гибкость в управлении другими компонентами экспозиции заметно снижается.

Ранние методы персонализированной генерации лиц основывались на генеративно-состязательных сетях(GANs), и в силу этого страдали от малого разнообразия и слабого контроля над получаемыми изображениями. Появление диффузионных генеративных моделей, которые способны генерировать разнообразные и реалистичные изображения, способствовало созданию новых методов персонализированной генерации, отнаследовавших эти качества. Однако, так как обучение диффузионных моделей с нуля ресурсозатратно, существующие методы строятся над большими предобученными *text-to-image* моделями, такими как Stable Diffusion [10].

На данный момент есть два главных подхода к решению задачи персонализированной генерации: *fine-tuning based* подход и *encoder-based* подход. Первый подход предполагает, что для персонализации необходимо собрать десятки изображений одной и той же личности, на которых затем будет дообучаться основная генеративная модель, подстраивая свои параметры под конкретного человека. Во втором подходе вместо полноценного дообучения используются предобученные энкодеры изображений, которые преобразуют референсное изображе-

ние человека в эмбеддинг, на который затем обуславливается процесс генерации в основной модели.

Цели, которые стояли перед нами в данной работе - изучить текущие state-of-the-art решения задачи персонализированной генерации лиц и предложить новый метод, который позволит иметь более детализированный контроль над результатом, при этом не будет уступать существующим решениям в согласованности тексту и в качестве сохранения identity референсного лица.

В данной работе мы описываем целевой метод, построенный вокруг идеи смыслового разделения информации. Мы основываем его на таких решениях как PhotoMaker [4] и IPAdapter [13]. Далее мы проводим глубокую экспериментальную работу, целью которой является построение предложенного нами подхода. Одним из результатов работы является полученный в ходе экспериментов, промежуточный метод, который подтверждает разумность нашей гипотезы.

Работа организована следующим образом. В разделе 2 краткое введение в диффузионные модели, а также обзор релевантных статей по методам персонализированной генерации. Далее, в разделе 3 мы описываем подробнее нашу идею и строим вокруг неё целевой метод. В разделе 4 мы рассказываем про наш протокол проведения экспериментов. И далее в разделе 5 описаны сами эксперименты. В разделе 6 мы описываем метод, который получили в результате проведенной экспериментальной работы. В разделе 7 подведены итоги проделанной работы.

2 Обзор литературы

2.1 Диффузионные модели

Диффузионные модели - это класс генеративных моделей, которые позволяют аппроксимировать распределение данных, путем постепенного расшумления гауссовского шума. В основе этих моделей лежат два процесса: прямой, при котором исходные данные зашумляются и обратный, при котором шум постепенно удаляется для возвращения к исходным данным. Можно параметризовать обратный процесс и обучить расшумляющую сеть ϵ_θ , которая сможет предсказывать добавленный на шаге t прямого процесса шум ϵ . Для обучения такой модели используется следующая функция потерь:

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

Латентные диффузионные модели (LDM) сжимают входные данные в пространство меньшей размерности с помощью вариационного автокодировщика (VAE). Такой подход не ухудшает качество генерации и существенно снижает вычислительные затраты, ведь исходные данные зачастую имеют большую размерность (например, изображения в разрешении 1024 x 1024).

Text-to-image диффузионные модели (T2I) - это LDM, с возможностью обуславливания генерации на текстовый промпт. Одним из наиболее популярных представителей этого класса моделей является Stable Diffusion [10]. В этом методе механизм обуславливания на текст реализован с помощью добавления в архитектуру модели слоёв Cross-Attention [5], которые реализуют механизм "внимания". Стоит отметить, что методы персонализированной генерации основаны на предобученных T2I моделях.

2.2 Fine-tuning based методы персонализированной генерации

В основе fine-tuning based методов лежит простая идея: если научить диффузионную модель ассоциировать уникальный текстовый токен S^* с концептом, то можно будет генерировать его в желаемом контексте, просто добавив в промпт данный токен. Зачастую эта идея реализуется дообучением некоторых частей уже предобученной T2I модели на нескольких примерах нового концепта. Рассмотрим два наиболее популярных fine-tuning based метода.

2.2.1 Textual Inversion

В данной статье [1] авторы предлагают легковесный метод. Вводится специальный токен S^* , предназначенный для хранения информации о концепте, и во время дообучения оптимизируется только его латентное представление. Дообучение производится на нескольких изображениях концепта и их простых описаниях содержащих специальный токен. У такого подхода есть очевидный недостаток: обучаемого скрытого представления недостаточно для запоминания сложных концептов. Данная проблема особо заметна в генерации лиц, ведь человеческие лица очень разнообразны и сложно устроены.

2.2.2 DreamBooth

В данной работе [11] есть несколько ключевых изменений относительно Textual Inversion. Авторы предлагают дообучать всю диффузионную модель вместо скрытого представления специального токена. Такое решение увеличивает качество сохранения концепта, но приводит к тому, что модель начинает забывать свои старые генеративные способности. Чтобы избавиться от этой проблемы авторы вводят дополнительную функцию потерь, которая сравнивает обусловленные лишь на класс, к которому относится концепт, генерации оптимизируемой модели и замороженной базовой модели, что ослабляет эффект забывания. Из-за полного дообучения диффузионной модели, Dreambooth вычислительно дороже чем Textual Inversion, а также легче переобучается.

2.3 Encoder-based методы персонализированной генерации

Методы, придерживающиеся этого подхода, эксплуатируют латентные представления референсных изображений концепта, которые затем различными путями внедряются в процесс генерации. Сначала рассмотрим универсальный метод, который можно применять для персонализированной генерации любых классов концептов. Затем подробнее разберем два метода именно для генерации лиц.

2.3.1 IPAdapter

Авторы данной статьи [13] предлагают добавлять к предобученной T2I модели легковесный адаптер, который позволяет обуславливать генерацию не только на текстовый промпт, но и на изображение. Архитектуру метода можно увидеть на рисунке 2.1:

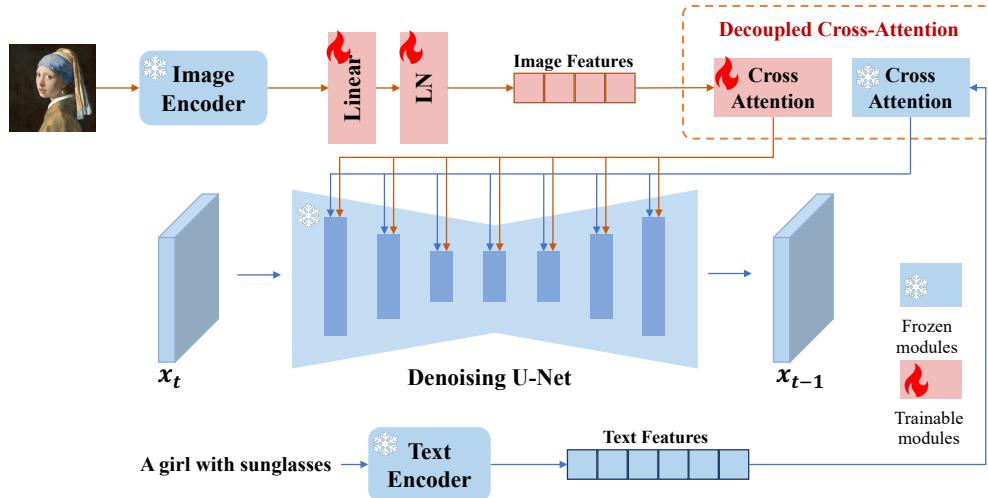


Рис. 2.1: Архитектура IPAdapter

IPAdapter состоит из двух частей:

- Энкодер изображений: для извлечения из референсного изображения богатых латентных представлений
- Модули адаптера: по сути, в дополнение к каждому имеющемуся в базовой модели Cross-Attention слою добавляется еще один такой же слой, который работает с эмбеддингом референсного изображения

Такая архитектура позволяет встраивать визуальную информацию в процесс генерации, не смешивая её с текстовой информацией, что сохраняет гибкость исходной модели. При обучении IPAdapter оптимизируются только голова энкодера и добавленные Cross-Attention слои, а предобученная T2I диффузионная модель остается замороженной.

2.3.2 PhotoMaker

Метод предложенный авторами статьи [4] является одним из state-of-the-art решений задачи персонализированной генерации лиц. Архитектура показана на рисунке 2.2.

Основой метода является Stacked ID Embedding - единое представление, которое формируется следующим образом:

- Вычисляются эмбеддинги референсных изображений e_1, \dots, e_n
- Вычисляется эмбеддинг токена класса (например, "man" или "person") из текстового промпта
- С помощью обучаемого MLP-слоя e_1, \dots, e_n "смешиваются" с эмбеддингом токена класса, и полученные векторы конкатенируются

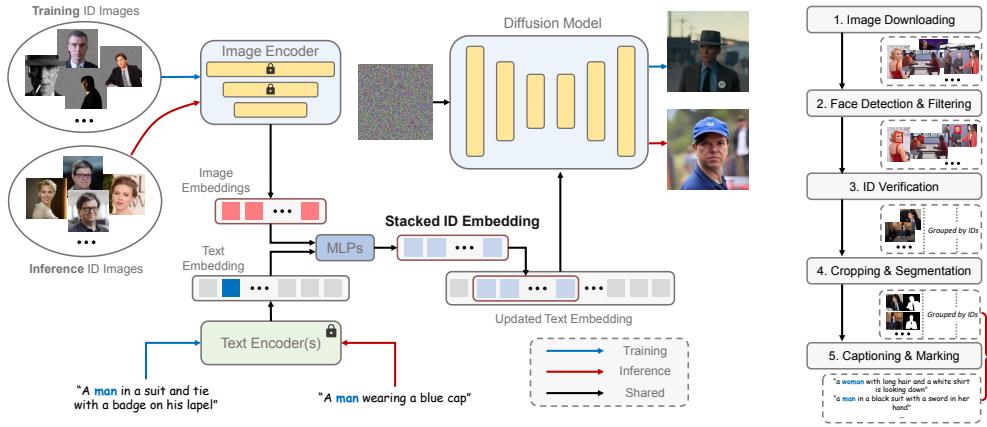


Рис. 2.2: Архитектура PhotoMaker

Stacked ID Embedding вставляется в эмбеддинг текстового промпта вместо токена класса, и дальнейшая генерация обуславливается на такое обновленное представление текста. Формирование такого представления позволяет модели учитывать разнообразие поз, выражений и атрибутов, избегая запоминания случайных деталей(например, фона), а также "привязывает" identity к конкретному токену, что позволяет учитывать ID-информацию в Cross-Attention слоях T2I модели(данный метод использует большую модель Stable Diffusion XL [8]). Для лучшего восприятия информации о лицах, дополнительно обучаются LoRA (Low-Rank Adapter) [2] добавки в Attention слоях, а также с некоторой вероятностью используется маскированная версия диффузионной функции потерь, которая вычисляется для области лица. Для обучения метода авторы собрали ID-ориентированный датасет, который содержит несколько разнообразных изображений для каждого человека. Обучаемыми являются некоторые слои CLIP Image Encoder [9], LoRA веса, а также MLP-слой для смешивания эмбеддингов.

2.3.3 PuLID

Ещё один современный метод для генерации лиц, главным новшеством которого является использование двух веток генерации. Одна ветка обуславливается на identity, а вторая нет. Такое разделение позволяет следить за тем, чтобы модель изменяла лишь лицевую часть и не портила базовые генеративные способности предобученной T2I модели. Авторы предлагают две функции потерь: Contrastive Alignment Loss и ID Loss. Первая штрафует за сильное влияние identity на базовую модель, а вторая следит за качеством сохранения лицевых черт. Также, авторы используют в качестве T2I модели SDXL-Lightning [6], которая позволяет генерировать качественные изображения за 4 шага.

3 Направление исследования

3.1 Мотивация

Как отмечалось ранее, ключевая задача методов персонализированной генерации лиц — сохранить identity лица, не нарушая при этом соответствия текстовому промпту. К сожалению, зачастую возникает компромисс: повышение точности воспроизведения identity приводит к снижению качества соответствия описанию, и наоборот. Также, стоит отметить, что проблема соответствия текстовому запросу возникает и в базовых T2I диффузионных моделях, особенно это заметно на подробных и детализированных промптах. А так как методы персонализированной генерации используют предобученные T2I модели, то данная проблема перетекает и к ним. Внедрение признаков identity в процесс генерации может сделать эту проблему еще более ярко выраженной.

Во многих работах по персонализированной генерации предлагаются различные способы борьбы с этими проблемами. В нашем исследовании мы бы хотели использовать эти идеи в комбинации с нашими собственными для получения нового метода, который не только будет превосходить существующие методы по двум ключевым критериям, но и потенциально откроет новые возможности для пользователей.

3.2 Гипотеза

Мы сформулировали одну основную гипотезу, вокруг которой построено наше исследование. Предлагается использовать вместо одного общего текстового промпта, несколько отдельных промптов, каждый из которых отвечает за конкретную часть генерируемого изображения(например, фон). Предполагается, что такое разделение позволит более гранулярно контролировать результаты генерации. А это потенциально может улучшить соответствие промпту. Более того, если удастся подобрать правильный метод внедрения identity, то можно улучшить и качество его сохранения.

Для большей ясности, опишем целевой метод, который учитывает эту гипотезу.

3.3 Целевой метод

Целевой метод во многом основан на идеях из статей про PhotoMaker [4] и IPAdapter [13], а также на нашей гипотезе. Сначала, раскроем идею разделенного промпта, затем опишем архитектуру и процесс обучения.

3.3.1 Разделенный промпт

Мы предлагаем разбить содержимое сгенерированного изображения на три смысловые группы: фон, тело и лицо. Лицо в большей степени контролируется референсным изображением, а фон и тело полностью зависят от текстового промпта. Соответственно, теперь в дополнение к общему описанию $P_{general}$ метод будет получать еще два текстовых условия:

- P_{bg} - описание всего, что происходит на заднем плане изображения
- P_{body} - описание всего, что относится к телу генерируемого человека, т.е. физические параметры, одежда, поза и т.д.

А в $P_{general}$ будет содержаться вся остальная информация.

3.3.2 Архитектура

Руководствуясь тем, что диффузионные модели в процессе генерации сначала формируют глобальные концепции, а только затем локальные, мы предлагаем ввести некоторый порядок внедрения наших дополнительных условий. Было бы логично, подавать модели информацию от большего к малому: сначала о фоне, затем о человеке и только в конце о лице. Соединяя эту идею с идеей разделенного внедрения информации, которую предложили авторы IPAdapter, мы предлагаем свою собственную архитектуру адаптера, общую схему которого можно видеть на рисунке 3.1.

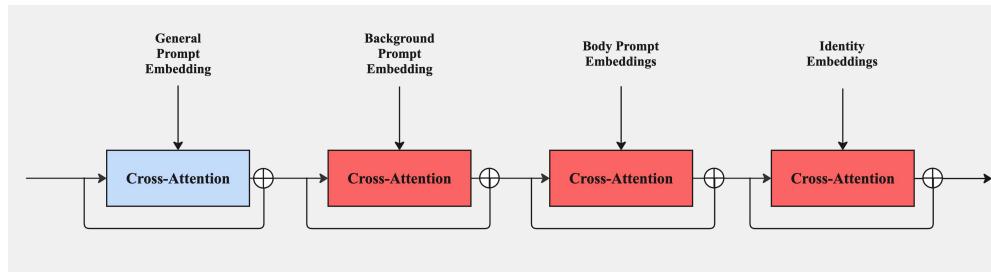


Рис. 3.1: **Архитектура предложенного адаптера.** Слева направо расположены Cross-Attention слои для обуславливания: на $P_{general}$, P_{bg} , P_{body} , identity. Между каждым блоком есть сквозные связи. Обучаются только модули отмеченные красным

Наш адаптер представляет из себя три Cross-Attention [5] слоя, которые соединены последовательно, то есть выход одного переходит на вход другому. Такой способ соединения позволит постепенно внедрять условия, при этом учитывать их логический порядок. Между слоями также присутствуют сквозные связи(skip connection), благодаря которым модель сможет работать даже при отсутствии одного из условий.

Так как мы уже ввели желаемый порядок внедрения информации, стоит отметить, что на вход первым двум добавленным Cross-Attention слоям будут поступать эмбеддинги промптов P_{bg} и P_{body} , соответственно, а третьему будут подаваться эмбеддинги референсных изображений.

3.3.3 Обучение

Предполагается, что обучаемыми будут добавленные слои, а также некоторые слои текстовых и визуальных энкодеров. Для обучения целевого метода понадобится ID-ориентированный датасет, наподобие тому, который использовался для обучения PhotoMaker. Следуя примеру этой же статьи, мы будем использовать стандартную диффузионную функцию потерь, с некоторой вероятностью применения её маскированной версии.

4 Протокол проведения экспериментов

4.1 Стартовый метод

Стартовым методом для проведения экспериментов мы выбрали PhotoMaker [4], так как он лежит в основе предлагаемого нами подхода. В рамках проведенной экспериментальной работы мы постепенно двигались к построению целевого метода, проверяя промежуточные гипотезы.

4.2 Метрики

В наших экспериментах мы хотим следить за тем, насколько хорошо генерация согласована с промптом и насколько хорошо в ней сохранено identity. Для этого мы вычисляем две стандартные для задачи персонализированной генерации лиц метрики: CLIP-T и Face-Sim.

CLIP-T оценивает соответствие изображения I целевому текстовому промпту T и вычисляется, как косинусная близость двух векторов:

$$\text{CLIP-T}(I, T) = \frac{\mathbf{v}_I \cdot \mathbf{v}_T}{\|\mathbf{v}_I\| \cdot \|\mathbf{v}_T\|},$$

где \mathbf{v}_I и \mathbf{v}_T — эмбеддинги изображения и текста, полученные из CLIP энкодеров [9], соответствующих модальностей.

Метрика Face-Sim оценивает насколько сгенерированное лицо соответствует референсному. Для ее расчета выделяются квадраты лиц из обоих изображений, затем для их

эмбеддингов вычисляется косинусная близость аналогично CLIP-Т.

Стоит отметить, что мы также вручную отсматриваем сгенерированные картинки, так как обе метрики не учитывают мелкие детали, и поэтому могут пропустить нежелательные артефакты.

4.3 Данные

Для наших экспериментов нами был собран ID-ориентированный датасет, по примеру того, на котором обучался PhotoMaker. В нашем датасете порядка 125000 изображений из открытых источников. Он разбит на примерно 3800 персон, каждая из которых представлена в среднем 30 изображениями. К каждому изображению было сгенерировано общее описание $P_{general}$, из которого затем были выделены описания фона и тела, P_{bg} и P_{body} соответственно.

Процесс сбора данных во многом похож на предложенный в статье про PhotoMaker. Можно выделить следующие этапы:

- **Скачивание:** с различных открытых источников были собраны имена знаменитостей. Затем по каждому имени производился сбор изображений с помощью поисковых движков
- **Фильтрация:** для начала отсеивались изображения, у которых одна из сторон имеет разрешение меньше 512. Затем, с помощью RetinaNet [7], выделялись прямоугольники лиц, и если изображение не содержало прямоугольника с размером больше 256 x 256, то оно отсеивалось
- **Верификация:** на данном этапе проверялась принадлежность изображение к своей identity группе аналогично алгоритму из статьи
- **Форматирование:** оставшиеся изображения центрировались и обрезались до квадрата
- **Генерация подписей:** к каждому изображению была сгенерирована подпись с помощью Qween2.5-VL [12]. Та же модель, затем использовалась, для выделения дополнительных описаний

Чтобы отслеживать качество обучений в наших экспериментах мы придумали 4 текстовых промпта средне-высокой детализированности, а также собрали три валидационных набора референсных изображений:

- **In-train:** 10 изображений разных персон из нашего датасета, которые встречаются во время обучения

- **Out-of-train:** 10 изображений разных персон из нашего датасета, которые не встречаются во время обучения
- **Out-of-domain(OOD):** 7 изображений, наших коллег и знакомых, которых точно не видела предобученная T2I модель, лежащая в основе обучаемого метода

В Таблице 4.1 представлены валидационные промпты

Таблица 4.1: Валидационные промпты

Photo of a person looking into the camera, wearing a black cloak and red hat. Daytime, 3 mountains in the background, a medieval castle can be seen on the far right mountain
A photo of an angry businessman in a yellow suit, talking on the phone and looking into the camera. He is in the street of a big city: to the left behind him is a bank building with a big sign above it saying "Neon Bank"
A photo of a man in a space suit, his face is seen very surprised, he is in the desert near Oasis with lake, palm trees and a couple of camels behind him
A photo of a middle-aged man in a dark green sweater looking at the viewer, he is in a room with white walls, there is a portrait behind him and a books on a shelf

На валидации для каждой пары референсного изображения и промпта проводилось пять генераций. Метрики усреднялись внутри каждого валидационного набора независимо. Стоит отметить, что во время валидации используется одно референсное изображение на каждую персону, так как это наиболее распространенный сценарий использования подобных методов.

4.4 Детали имплементации

Для экономии вычислительных ресурсов, наши обучения проводились в таргетном разрешении 512 x 512 (в PhotoMaker использовалось разрешение 1024 x 1024). Это решение также повлияло на выбор предобученных весов для SDXL - в отличие от PhotoMaker, мы используем веса, дообученные под генерацию в разрешении 512x512. Еще одно отличие от стартового метода - это то, что в качестве референсных изображений мы подаем в модель обрезанные квадраты лиц.

Мы имплементировали распределенное обучение, следуя парадигме Distributed Data Parallel, и все наши эксперименты обучались на 2 видеокартах(либо NVIDIA A100, либо NVIDIA H100) с эффективным размером батча 24. Для корректного сравнения, валидация во всех экспериментах проводилась каждые 1000 итераций обучения.

Основная часть гиперпараметров обучения и инференса такие же как в статье. Зададим их:

- Обучение:
 - Ранг LoRA адаптеров: 64 (если не указано иначе)
 - Оптимизатор Adam [3]
 - Learning rate: 1e-4 для LoRA слоев, 1e-5 для остальных обучаемых весов (если не указано иначе)
 - Вероятность использования маскированной функции потерь: 0.5
 - Вероятность сброса обуславливаний: 0.1
- Инференс:
 - 50 шагов расшумления
 - identity внедряется начиная с 10 шага расшумления
 - classifier-free-guidance scale: 5

5 Эксперименты

В связи с отсутствием официального кода для обучения PhotoMaker-a, чтобы убедиться, что метод воспроизводим с нуля, нам пришлось имплементировать его следуя информации из статьи. Поэтому эксперименты разделены на две части: первая посвящена воспроизведению бейзлайна, а во второй мы пытаемся его улучшить, следуя нашей гипотезе.

5.1 Эксперименты по воспроизведению

Первые несколько экспериментов по воспроизведению стартового метода, как оказалось, были проведены не совсем корректно. Во-первых, во время обучения таргетные изображения были в разрешении 512 x 512, а на валидации мы генерировали картинки размера 1024 x 1024. Во-вторых, мы использовали ту же версию SDXL, что и PhotoMaker, а как оказалось, она была обучена для работы с большим разрешением изображений. Всё это привело к занижению результатов, обученных в этих экспериментах, моделей. Тем не менее, некоторые полезные выводы были получены. Подробнее про эти эксперименты можно прочитать в A.1.

Отметим на какие значения метрик мы ориентировались в процессе воспроизведения исходного метода. Для их получения, мы взяли веса предоставленные авторами PhotoMaker

Таблица 5.1: Метрики оригинальных весов PhotoMaker

Dataset	CLIP-T	Face-Sim
In-train	28.62	0.19
Out-of-train	28.87	0.18
OOD	29.35	0.28

и использовали их поверх используемой нами версии SDXL. Метрики для каждого валидационного набора данных указаны в таблице 5.1.

5.1.1 Эксперимент 1

В первом эксперименте мы попробовали обучить исходный метод в соответствии со статьей. То есть, обучались CLIP Image Encoder, MLP слой смешивания(learning rate 1e-5) и LoRA адаптеры ранга 64 (learning rate 1e-4). К сожалению, в такой постановке модель не обучалась, и генерировала для разных референсных лиц одинаковые изображения. Это означает, что сигнал от identity не проникал в процесс генерации. Мы решили проверить выходы из энкодера изображений после нескольких тысяч итераций обучения. Как оказалось, для разных референсных лиц энкодер выдавал очень близкие друг к другу эмбеддинги, это и приводило к тому, что модель не различала identity.

5.1.2 Эксперимент 2

Тут мы решили убрать LoRA слои и обучать только весь энкодер и MLP сеть. Сигнал от референсных изображений увеличился, так как на генерациях начали различаться лицевые области. Несмотря на это, модель почти не прогрессировала и не выучивала identity.

5.1.3 Эксперимент 3

В данном эксперименте мы решили ограничиться лишь обучением MLP слоя и проекционных голов CLIP Image Encoder-a. Такая постановка оказалась удачнее предыдущей. Модель не только реагировала на сигнал от identity, но и постепенно училась его сохранять.

5.1.4 Эксперимент 4

Далее мы решили добавить LoRA слои с рангом 1 к предыдущей конфигурации и обучать их с уменьшенной до 1e-5 длиной шага. Таким образом, мы хотели проверить , как

наличие слоев адаптера влияет на качество сохранения identity. И действительно, обучение LoRA добавок заметно увеличило качество передачи identity относительно предыдущей постановки. Однако модель сошлась на значениях Face-Sim хуже желаемых.

5.1.5 Промежуточные выводы

Из первых экспериментов можно сделать вывод, что нет смысла обучать весь CLIP Image Encoder, можно ограничиться лишь его проекционными головами. Также, мы поняли, что обучаемые LoRA слои важны для сохранения identity. Однако, эксперименты в некорректной постановке показали, что начиная с какого-то значения ранга адаптеров, метрика Face-Sim не получает прироста. Мы считаем, что это связано с количеством, представленных в нашем обучающем наборе данных персон, которых примерно в 4 раза меньше, чем в датасете, на котором был обучен PhotoMaker. Возможно именно поэтому адаптера с рангом в несколько раз меньше 64 достаточно, чтобы выделить всю полезную информацию из нашего ID-ориентированного набора данных.

Учитывая всё высказанное, в следующих экспериментах, мы начинаем наши обучения с предобученных весов PhotoMaker-a. Это поможет нам достичь целевых метрик и построить бейзлайн для дальнейшего улучшения.

5.1.6 Эксперимент 5

В данном эксперименте, мы опять-таки пробуем обучить метод в соответствии со всеми деталями из статья, но начинаем мы с предобученных весов. Несмотря на такую инициализацию, метрики резко деградировали, а затем модель и вовсе разучилась переносить identity. Возможно, это связано с дообучением всего энкодера на новом датасете, из-за чего предобученные веса смещаются из своего оптимума, и поэтому модель теряет обобщающую способность.

5.1.7 Эксперимент 6

Далее мы решили оставить обучаемыми головы энкодера, MLP сеть и LoRA слои ранга 64. Также, мы уменьшили learning rate для адаптеров до 1e-5, так как теперь они проинициализированы обученными под нашу задачу весами. В начале обучения опять произошло резкое падение Face-Sim, но затем метрика начала расти и вышла на уровень целевого значения.

5.1.8 Эксперимент 7

Так как в начале обучения в оптимизаторе Adam еще не сформирована история градиентов, то делая большие шаги мы можем удалиться от оптимального положения обученных весов PhotoMaker-а. Мы считаем, что именно с этим связано резкое падение качества сохранения identity в предыдущей конфигурации. Поэтому, в этом эксперименте мы на протяжении первых 500 итераций линейно увеличиваем learning rate до 1e-5 для всех обучаемых частей модели. Такое решение оказалось успешным и модель с первых же эпох показала рост относительно целевых метрик.

Поэтому, полученный метод будем считать нашим **бейзлайном** и в дальнейшем будем сравниваться с его траекторией обучения.

5.2 Построение целевого метода

В этой части экспериментов мы пытаемся грамотно и постепенно подойти к построению целевого метода.

5.2.1 Эксперимент 1

Для начала мы решили проверить, насколько хорошо будет работать внедрение identity не через текстовое пространство(как это делается в PhotoMaker), а через дополнительный Cross-Attention слой, как в IPAdapter. Изменения относительно бейзлайна:

- Убрали LoRA слои
- Наподобие IPAdapter к каждому Cross-Attention слою добавили ещё один, который на вход получал эмбеддинги референсных изображений. Новые слои инициализировали старыми
- Отметим, что в отличие от IPAdapter, наша имплементация позволяет обуславливаться на несколько референсов
- Теперь нет нужды в MLP сети, так как мы не смешиваем визуальные признаки с текстовыми
- Обучаемыми являются: проекции энкодера и добавленные Cross-Attention слои

Энкодер был инициализирован весами из PhotoMaker-а. Сначала обучали с длиной шага для новых слоев равной 1e-4, однако изображения содержали очень много артефактов.

Поэтому попробовали уменьшить длину шага до $1e-5$ и это помогло: модель стала генерировать адекватные изображения и учиться сохранять identity. Однако по Face-Sim метод не достиг значений бейзлайна.

5.2.2 Эксперимент 2

Далее мы попробовали смешать обе техники внедрения identity: через текстовое пространство и через дополнительный Cross-Attention. Изменения относительно предыдущего эксперимента:

- Вернули MLP сеть, а также добавили ещё одну голову проекции в энкодер, чтобы в разных путях внедрения эмбеддинги изображений отличались. Новую проекцию проинициализировали имеющейся
- Не стали возвращать LoRA слои, чтобы посмотреть как метрики отреагируют на двойное внедрение в базовом случае

Если смотреть по траекториям обучения, то этот метод чуть улучшил Face-Sim, но при этом чуть ухудшил согласованность с текстом.

5.2.3 Эксперимент 3

В этом эксперименте мы вернулись к конфигурации с внедрением визуальных признаков, только через новые слои внимания. Так как LoRA веса PhotoMaker-а содержат в себе много информации об identity, мы решили вернуть их обратно, а также скопировать их в новые Cross-Attention слои. Обучали мы только головы в энкодере и все LoRA добавки.

По Face-Sim данный метод обогнал первые два, но всё ещё оказался хуже бейзлайна. Также, если посмотреть на генерируемые изображения, то можно заметить нежелательные артефакты, которые не отражаются в метрике CLIP-T, так как она у этого метода находится на уровне бейзлайна.

5.2.4 Эксперимент 4

Так как попытки внедрить identity методом похожим на IPAdapter не увенчались успехом, мы вернулись к способу из PhotoMaker-а. В этом эксперименте мы решили проверить нашу гипотезу про разделение промпта, но пока только с дополнительным описанием фона P_{bg} . Некоторые детали:

- Обуславливание на P_{bg} реализовано через дополнительный Cross-Attention(как в IPAdapter)

- Валидационные промпты были разбиты на две части: $P_{general}$ и P_{bg} без потери информации
- Обучались энкодер, MLP слой, LoRA, а также новые слои

Разделение информации дало интересные результаты. Face-Sim значительно улучшился относительно бейзлайна (график 6.1). Это говорит о том, что благодаря разделению информации, identity учитывается лучше в процессе генерации. Однако метрика согласованности с промптом просела (график 6.2). Но сравнивая генерации(рисунки 6.4 и 6.3) можно заметить, что у бейзлайна часто встречаются лишние объекты и есть признаки непонимания промпта. А полученный в этом эксперименте метод, наоборот, иногда не генерирует нужные объекты, но зато в большинстве случаев правильно понимает текстовые условия.

6 Итоговый метод

6.1 Описание метода

Получившийся в последнем эксперименте метод будем считать итоговым. В нем используется предложенная нами идея разделения информации: в модель в качестве дополнительного условия подается текстовое описание фона P_{bg} . Данный метод является некоторым слиянием двух существующих методов персонализированной генерации лиц - PhotoMaker и IPAdapter:

- От первого метода у нас осталась идея внедрения identity в процесс генерации через текстовое пространство
- У второго мы позаимствовали идею разделенного Cross-Attention слоя, чтобы учесть в генерации дополнительное условие из P_{bg}

6.2 Количественные результаты

Судя по траекториям метрик во время обучения: итоговый метод превосходит бейзлайн в качестве сохранения identity, и проигрывает в согласованности с текстовым описанием. Ситуация идентична на всех валидационных наборах данных. Траектории обучения метрик Face-Sim и CLIP-T можно увидеть на графиках 6.1 и 6.2 соответственно.

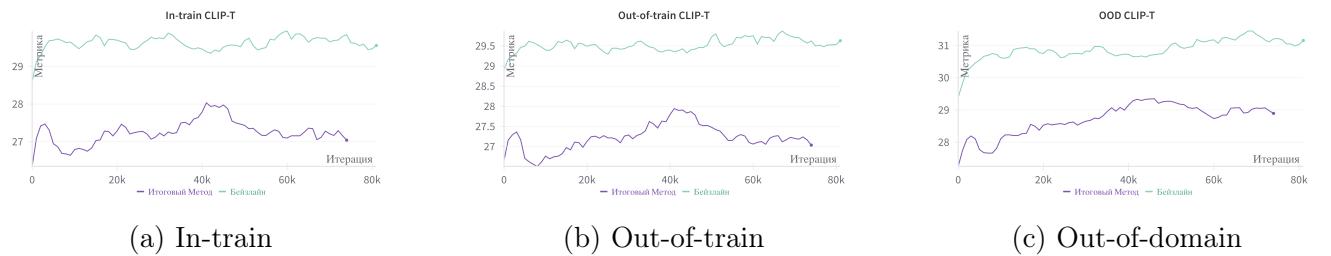


(a) In-train

(b) Out-of-train

(c) Out-of-domain

Рис. 6.1: Траектории Face-Sim итогового и бейзлайн методов на разных валидационных данных



(a) In-train

(b) Out-of-train

(c) Out-of-domain

Рис. 6.2: Траектории CLIP-T итогового и бейзлайн методов на разных валидационных данных

6.3 Примеры работы

На рисунке 6.3 представлены примеры работы итогового метода. А на рисунке 6.4 для тех же промптов и референсных лиц представлены примеры работы бейзлайн метода. Для этих примеров использовались валидационные промпты, которые указаны в таблице 4.1.



Рис. 6.3: Примеры работы итогового метода. Слева - референсное лицо, справа - генерация



Рис. 6.4: Примеры работы бейзлайн метода. Слева - референсное лицо, справа - генерация

Глазами видно, что итоговый метод лучше сохраняет identity. И несмотря на значения CLIP-T, мы считаем, что в большинстве случаев текстовым описаниям соответствует лучше именно финальный подход. Так например:

- В левом верхнем примере на фоне, судя по промпту, должен находиться средневековый замок ("...a medieval castle can be seen on the far right..."). Как можно заметить, итоговый метод сгенерировал замок, а бейзлайн нет
- В левом нижнем примере, судя по промпту, должна висеть картина на белой стене ("...in a room with white walls, there is a portrait behind him..."). В генерации итогового метода видна часть картины на белой стене. Тогда как, бейзлайн неправильно понял промпт и сгенерировал что-то наподобие картины в руках человека
- В правом верхнем примере, промпт указывает на то, что у человека должно быть удивленное лицо ("...his face is seen very surprised..."). Однако, видно, что бейзлайн просто скопировал выражение лица с референса, а итоговый метод показывает правильные эмоции на лице
- В правом нижнем примере, человек сгенерированный бейзлайном разговаривает по двум телефонам одновременно

7 Заключение

В данной исследовательской работе, мы изучили как базовые, так и передовые методы персонализированной генерации. Затем сформулировали ключевую гипотезу нашего исследования, суть которой заключается в разделении информации. Также, мы предложили целевой метод, основанный на существующих работах и учитывающий нашу идею. Для проверки нашей гипотезы мы собрали ID-ориентированный датасет и провели ряд экспериментов для обучения качественного бейзлайна и дальнейшего его улучшения. В результате нами был получен промежуточный метод, который увеличивает одну из целевых метрик и в некоторой степени подтверждает разумность нашей гипотезы.

Мы планируем продолжить данное исследование и постепенными шагами прийти к построению целевого метода.

Список литературы

- [1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik и Daniel Cohen-Or. *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion*. 2022. arXiv: [2208.01618 \[cs.CV\]](https://arxiv.org/abs/2208.01618). URL: <https://arxiv.org/abs/2208.01618>.
- [2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang и Weizhu Chen. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: [2106.09685 \[cs.CL\]](https://arxiv.org/abs/2106.09685). URL: <https://arxiv.org/abs/2106.09685>.
- [3] Diederik P. Kingma и Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: [1412.6980 \[cs.LG\]](https://arxiv.org/abs/1412.6980). URL: <https://arxiv.org/abs/1412.6980>.
- [4] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng и Ying Shan. *PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding*. 2023. arXiv: [2312.04461 \[cs.CV\]](https://arxiv.org/abs/2312.04461). URL: <https://arxiv.org/abs/2312.04461>.
- [5] Hezheng Lin, Xing Cheng, Xiangyu Wu, Fan Yang, Dong Shen, Zhongyuan Wang, Qing Song и Wei Yuan. *CAT: Cross Attention in Vision Transformer*. 2021. arXiv: [2106.05786 \[cs.CV\]](https://arxiv.org/abs/2106.05786). URL: <https://arxiv.org/abs/2106.05786>.
- [6] Shanchuan Lin, Anran Wang и Xiao Yang. *SDXL-Lightning: Progressive Adversarial Diffusion Distillation*. 2024. arXiv: [2402.13929 \[cs.CV\]](https://arxiv.org/abs/2402.13929). URL: <https://arxiv.org/abs/2402.13929>.
- [7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He и Piotr Dollár. *Focal Loss for Dense Object Detection*. 2018. arXiv: [1708.02002 \[cs.CV\]](https://arxiv.org/abs/1708.02002). URL: <https://arxiv.org/abs/1708.02002>.
- [8] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna и Robin Rombach. *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*. 2023. arXiv: [2307.01952 \[cs.CV\]](https://arxiv.org/abs/2307.01952). URL: <https://arxiv.org/abs/2307.01952>.
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger и Ilya Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: [2103.00020 \[cs.CV\]](https://arxiv.org/abs/2103.00020). URL: <https://arxiv.org/abs/2103.00020>.

- [10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser и Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: [2112.10752 \[cs.CV\]](https://arxiv.org/abs/2112.10752). URL: <https://arxiv.org/abs/2112.10752>.
- [11] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein и Kfir Aberman. *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation*. 2023. arXiv: [2208.12242 \[cs.CV\]](https://arxiv.org/abs/2208.12242). URL: <https://arxiv.org/abs/2208.12242>.
- [12] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou и Junyang Lin. *Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution*. 2024. arXiv: [2409.12191 \[cs.CV\]](https://arxiv.org/abs/2409.12191). URL: <https://arxiv.org/abs/2409.12191>.
- [13] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han и Wei Yang. *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models*. 2023. arXiv: [2308.06721 \[cs.CV\]](https://arxiv.org/abs/2308.06721). URL: <https://arxiv.org/abs/2308.06721>.

A Аппендикс

A.1 Некорректные эксперименты

В этом разделе приведены эксперименты, которые были проведены в некорректной постановке.

A.1.1 Эксперимент 1

Первым экспериментом было обучение исходного метода идентично статье. В такой постановке модель не обучалась, и генерировала для разных референсных лиц одинаковые изображения.

A.1.2 Эксперимент 2

Дальше мы решили попробовать обучать только головы энкодера изображений и MLP слой. Модель начала обучаться, и начали появляться различия в генерациях от разных референсов. Однако identity обучался очень медленно и в итоге модель сошлась на значениях Face-Sim сильно хуже желаемых.

A.1.3 Эксперимент 3

В этом эксперименте мы решили взять конфигурацию из предыдущего и добавить обратно LoRA слои, но с рангом 1. Также, мы снизили learning rate для адаптеров до 1e-5. Мотивация была в том, чтобы проверить насколько обучение дополнительных LoRA адаптеров влияет на качество сохранения identity.

Действительно, обучение LoRA добавок оказалось важным для сохранения identity, так как Face-Sim заметно увеличился относительно Эксперимента 2, но желаемые значения всё ещё были сильно впереди.

A.1.4 Эксперимент 4

В этом эксперименте мы исследовали идею поэтапного обучения частей модели. Для этого мы взяли обученные в Эксперименте 2 головы энкодера, MLP слои и продолжили их обучение, при этом добавив в модель LoRA слои ранга 1. Данная модель сошлась примерно к тем же метрикам, что и модель из Эксперимента 3. Таким образом, нет смысла в поочередном обучении разных компонент модели.

A.1.5 Эксперимент 5

На данном этапе мы решили продолжить обучение модели из Эксперимента 3, заменив LoRA слои на новые с рангом 16. Мы ожидали повышения качества передачи identity, однако этого не произошло. Возможно, это связано с малым количеством различных персон в нашем ID-ориентированном датасете. В нашем наборе данных представлены порядка 3800 различных персон, тогда как в датасете использованном для обучения PhotoMaker-а их примерно в 4 раза больше. Возможно, поэтому адаптера с малым рангом достаточно, чтобы выделить всю полезную информацию из наших обучающих данных.

A.1.6 Выводы

Несмотря на наличие некорректностей в постановке данных экспериментов, были сделаны полезные выводы. Во-первых, не стоит обучать весь CLIP Image Encoder. Во-вторых, стоит задуматься об увеличении количества персон представленных в нашем ID-ориентированном датасете.