Создание персонализированных генераций изображений

Даниил И. Казачков
ФПМИ МФТИ
Долгопрудный
kazachkov.di@phystech.edu

Андрей В. Филатов
Сколковский Институт Технологий
filatovandreiv@gmail.com

Abstract

В данной работе мы представляем архитектуру BoLID - решение для интеграции body identity, позволяющее достигать более точной персонализации по сравнению с методами, учитывающими лишь лицо. Основные изменения заключены в энкодере пользовательского изображения.

Ключевые слова: Machine Learning, Diffusion Models.

1 Введение

Большой вклад в область генерации изображений внесли такие модели, как U-Net[link], DALL-E 2 [link], Imagen [link], Stable Diffusion [link]. Сначала пользователи смогли получать картинки по текстовому промту, что является сильным ограничением. В DALLE-2 [link] пошли дальше и сделали попытку встроить изображение в диффузионную модель. Дальнейшее развитие архитектур привело к созданию легковесных адаптеров, например IP-Adapter [link], разделяющих механизм cross-attention для текстовых признаков и признаков изображения. Это позволило ввести элементы контроля генерации, подобно ControlNet [link], и использовать other custom models fine-tuned from the same base model. В работе InstantID [link] и PuLID [link] авторы включают дополнительный контроль за facial и landmark identity (ID), что дает сильный прирост в ID fidelity.

Однако каждая из этих работ концентрируется лишь на учете особенностей лица, когда как тело пользователя остается без внимания. В данной работе я устраняю этот недостаток, встраивая в процесс генерации bodyID пользователя. Важной задачей является выделение значимых деталей с референсных изображений. Как в InstantID ограничиваются 5 ключевыми точками для лица, так и мы выделяем 12 points для туловища, что позволяет учитывать пропорции и форму тела. Для улучшения ID similarity мы дообучаем модель на датасетах, сгруппированных по уникальным идентификаторам пользователя, где каждый ID представлен серией снимков из разных ракурсов. Это обеспечивает более целостную генерацию, основанную на единичной входной фотографии, сохраняя физические пропорции пользователя вместе с его чертами лица.

Ключевые изменения касаются двух модулей:

1. Body-ID Encoder, обучаемый выделять особенности тела (body landmarks), аналогично подходу InstantID [1], но уже для 12 основных ключевых точек туловища.
2. Body-Condition Module, действующий как адаптер, схожий с IP-Adapter [2], где cross-attention разделяется между признаками текста, лица и тела пользователя.

1.1 Loss Functions and Training Objective

- Reconstruction Loss: лосс модели диффузии.
- Identity Consistency Loss: для сохранения пропорций тела.
- CLIP-based Similarity Loss: использует EVA-CLIP [3] для обеспечения семантической близости к заданному тексту и общему стилю входного изображения.

2 Related Works

Последние достижения в области генеративных моделей позволили создавать изображения высокого качества по текстовым описаниям. Однако персонализированная генерация, особенно для изображений людей, остается сложной задачей. Ниже мы рассматриваем ключевые подходы к ее решению.

Авторы IP-Adapter [2] разделили блок cross-attention на два подблока, один из которых отвечает за image-prompt, а второй - за text-prompt. Такой подход позволил обуславливать диффузионный процесс генерации не только текстом, но и картинкой для достижения лучшей консистентности запросу пользователя. Однако учет пространственных характеристик референсного изображения не предусмотрен, что не позволяет достичь качественной генерации тела человека.

В последующих работах идея легковесного адаптера была модернизирована для достижения лучшего сходства с запросом.

Авторы InstantID [1] достигли лучшего сохранения идентичности за счет дополнительного введения энкодера лиц InsightFace. Работает в "zero-shot"режиме, требует одно референсное изображения.

В работе PuLID [4] используется контрастивное обучение для плавного смешивания нескольких идентичностей. Особенно полезен для запросов типа "лицо человека A в стиле человека B".

Попытки точного вычленения черт пользователя привели к методу Stacked ID Embedding в работе PhotoMaker [5], где использовали фотографии одного человека под разными ракурсами. Это обеспечило высокую сохранность идентичности без тонкой настройки базовой диффузионной модели. Метод эффективен при использовании более трех референсов, но не заточен под учет тела пользователя.

Таблица 1: Сравнительная характеристика методов персонализированной генерации.

Метод	Референсы	Сохранение ID	Модальность
DreamBooth	3–5	Высокое	Только текст
IP-Adapter	1	Среднее	Изображение + Текст
PhotoMaker	3+	Очень высокое	Изображение + Текст
InstantID	1	Высокое	Изображение + Текст
PuLID	1	Очень высокое	Изображение + Текст

Наш подход развивает IP-Adapter, добавляя контроль позы всего тела, что объединяет преимущества PhotoMaker (работа с несколькими референсами) и гибкость адаптеров с изображением в качестве условия. В отличие от предыдущих работ, мы явно учитываем кинематику тела для повышения точности позы.

3 Постановка задачи

Определим датасет как $\mathfrak{D} = \{(x_i, \tau_i) : i = 1, \ldots, n\}$, x_i — изображение, τ_i – соответствующий текстовый промпт. Рассматривается модель ϵ_θ из класса диффузионных моделей. На этапе обучения на каждом шаге из $\mathfrak{D}$ удаляется изображение $x_j, j \sim \mathcal{U}\{1, \ldots, n\}$, и модель учится восстанавливать его по оставшимся изображениям.

Определим функцию потерь:

$$\mathcal{L}(\epsilon, \epsilon_\theta) = \mathbb{E}_{\epsilon \sim N(0,I), \mathbf{c}_\tau, \mathbf{c}_i, t, \mathbf{c}_t^j} \|\epsilon - \epsilon_\theta(\mathbf{c}_\tau, \mathbf{c}_i, t, \mathbf{c}_t^j)\|^2, \tag{1}$$

где $\mathbf{c}_\tau = \Gamma_\tau(\tau_j)$ — текстовые признаки удаленного изображения, полученные путем применения текстового энкодера Γ_τ к текстовому промпту τ_j; $\mathbf{c}_i = G(\Gamma_i(x_1), \ldots, \Gamma_i(x_{j-1}), \Gamma_i(x_{j+1}), \ldots, \Gamma_i(x_n))$ — признаки оставшихся изображений, являющиеся результатом применения агрегирующей функции G к эмбеддингам изображений, полученным с помощью image-энкодера Γ_i; $\mathbf{c}^j = \Gamma_i(x_j)$ — признаки удаленного изображения; $t \in [0, T]$ — временной шаг диффузионного процесса, $\mathbf{c}_t^j = \alpha_t \mathbf{c}^j + \sigma_t \epsilon$ — зашумленные данные удаленного изображения на шаге t; α_t, σ_t — предопределенные функции от t, определяющие диффузионный процесс.

Решается следующая оптимизационная задача:

$$\epsilon_\theta^* = \arg\min_{\epsilon_\theta} \mathcal{L}(\epsilon, \epsilon_\theta), \tag{2}$$

4 Предпосылки

4.1 Диффузионные модели. DDPM

Процесс диффузии состоит из двух процессов: прямого и обратного.

Прямой процесс представляет собой последовательное зашумление входного изображения x_0 за T шагов, где x_t получается по следующей формуле:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon, \tag{3}$$

где $\varepsilon \sim \mathcal{N}(0, I)$, а β_t гиперпараметр подбираемый так, чтобы каждое следующее изображение x_t было сильнее зашумлено,

$$x_t|x_{t-1} \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{4}$$

При $T \to \infty, x_T \to \mathcal{N}(0, I)$, то есть на последнем шаге итераций получается гауссовский шум.

Положим $\alpha_t = 1 - \beta_t, \overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$. Тогда

$$x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon, \tag{5}$$

где $\varepsilon \sim \mathcal{N}(0, I)$,

$$x_t|x_0 \sim \mathcal{N}(\sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I). \tag{6}$$

Во время обратного процесса исходное изображение восстанавливается из шума. Знаем $x_T \sim \mathcal{N}(0, I)$. Семплирование происходит итеративно:

$$\hat{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\hat{\epsilon}_\theta(\hat{x}_t, t)\right) + \hat{\sigma}_t z, \tag{7}$$

где $\hat{x}_t$ — восстановленное изображение на итерации t, причем $\hat{x}_T = x_T$; $\hat{\epsilon}_\theta(\hat{x}_t, t)$ — реконструкция шума, полученная моделью для $\hat{x}_t$; z — шум, который позволяет генерировать различные изображения, причем $z \sim \mathcal{N}(0, I)$, если $t > 1$, иначе $z = 0$.

4.2 Classifier-free Guidance

Можно регулировать силу влияния условия c без специального классификатора. Метод classifier-free guidance[6] позволяет увеличить степень, с которой модель ориентируется на промпт. Во время семплирования предсказание получается путем линейной комбинации предсказаний обусловленной и необусловленной моделей:

$$\hat{\epsilon}_\theta(x_t, c, t) = (w+1)\epsilon_\theta(x_t, c, t) - w\epsilon_\theta(x_t, t), \tag{8}$$

где w — весовой коэффицент, $t \in [0, T]$ — временной шаг диффузионного процесса, x_t — зашумленное изображение на шаге t.

В основе данной модели лежит предобученная диффузионная text-to-image модель $\hat{x}_\theta$, функция потерь которой определяется как:

$$\mathbb{E}_{\mathbf{x},\epsilon\sim\mathcal{N}(0,I),\mathbf{c},t} w_t \|\hat{x}_\theta(\alpha_t\mathbf{x} + \sigma_t\epsilon, \mathbf{c}) - \mathbf{x}\|^2, \tag{9}$$

где $\mathbf{x}$ — латентное представление исходного изображения, $\mathbf{c} = \Gamma(P)$ — вектор условия, полученный с помощью текстового энкодера Γ и текстового промпта P, $t \in [0, T]$ обозначает временной шаг диффузионного процесса; α_t, σ_t, w_t — предопределенные функции от t, определяющие процесс диффузии. Исходная диффузионная модель дообучается на нескольких входных изображениях одного объекта в паре с текстовым промптом, содержащим название класса, к которому принадлежит данный объект. Генерируются данные $\mathbf{x_{pr}} = \hat{x}(z, \mathbf{c_{pr}})$ с использованием сэмплера на основе предобученной диффузионной модели со случайным начальным шумом $z \sim \mathcal{N}(0, I)$ и вектором условия $\mathbf{c_{pr}} := \Gamma(f(\text{"a [class noun]"}))$, где f — токенизатор. Функция потерь принимает следующий вид:

$$\mathbb{E}_{\mathbf{x},\epsilon,\epsilon',\mathbf{c},t}[w_t\|\hat{x}_\theta(\alpha_t\mathbf{x} + \sigma_t\epsilon, \mathbf{c}) - \mathbf{x}\|^2 + \lambda w_{t'}\|\hat{x}_\theta(\alpha_{t'}\mathbf{x_{pr}} + \sigma_{t'}\epsilon', \mathbf{c_{pr}}) - \mathbf{x_{pr}}\|^2], \tag{10}$$

где λ — весовой коэффициент. Генерация изображений происходит путем встраивания уникального идентификатора в текстовый промпт в виде: "a [identifier] [class noun]".

4.3 Stable Diffusion

Stable Diffusion — это модель text-to-image, которая на вход получает текстовый промпт. CLIP преобразует его в эмбеддинг, который направляет генерацию. Далее генерируется случайный шум в латентном пространстве. Модель U-Net в ходе обратного диффузионного процесса убирает шум, на каждом шаге учитывая текстовый эмбеддинг, чтобы изображение соответствовало описанию. После завершения этого этапа VAE-декодер переводит латентное представление в изображение исходного размера.

4.4 IP-Adapter

IP-Adapter состоит из двух частей: image-энкодера для извлечения признаков изображения из промпта и адаптированных модулей с механизмом изолированного перекрестного внимания для встраивания признаков изображения в предобученную text-to-image модель.

Пусть даны признаки изображения $\mathbf{c_i}$ и текста $\mathbf{c_t}$. Поскольку в основе метода лежит идея отдельно использовать текст и изображение, то получим следующую формулу для перекрестного внимания

$$\mathbf{Z}^{new} = Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \lambda \cdot Attention(\mathbf{Q}, \mathbf{K}', \mathbf{V}'), \tag{11}$$

$$\mathbf{Z}^{new} = softmax(\frac{\mathbf{Q}\mathbf{K}^\mathbf{T}}{\sqrt{d}})\mathbf{V} + \lambda \cdot softmax(\frac{\mathbf{Q}(\mathbf{K}')^\mathbf{T}}{\sqrt{d}})\mathbf{V}',$$

где λ — весовой коэффициент на изображение; $\mathbf{Q} = \mathbf{Z}\mathbf{W_q}, \mathbf{K} = \mathbf{c_t}\mathbf{W_k}, \mathbf{V} = \mathbf{c_t}\mathbf{W_v}$ — матрицы запросов, ключей и значений механизма внимания для текстовых признаков соответственно, а $\mathbf{W_q}, \mathbf{W_k}, \mathbf{W_v}$ — соответствующие матрицы весов; $\mathbf{K}' = \mathbf{c_i}\mathbf{W}'_\mathbf{k}, \mathbf{V}' = \mathbf{c_i}\mathbf{W}'_\mathbf{v}$ — матрицы запросов, ключей и значений механизма внимания для признаков изображения соответственно, а $\mathbf{W}'_\mathbf{k}$, $\mathbf{W}'_\mathbf{v}$ — соответствующие матрицы весов.

Поскольку модель UNet заморожена, только $\mathbf{W}'_\mathbf{k}$ и $\mathbf{W}'_\mathbf{v}$ являются обучаемыми параметрами.

В процессе обучения минимизируется следующая функция потерь:

$$\mathcal{L}(\epsilon, \epsilon_\theta) = \mathbb{E}_{\mathbf{x_0}, \epsilon, \mathbf{c_t}, \mathbf{c_i}, t}\|\epsilon - \epsilon_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t)\|^2, \tag{12}$$

где $\mathbf{x_t}$ — зашумленное изображение на шаге t.

Для того чтобы задействовать classifier-free guidance на этапе вывода, во время обучения случайным образом отбрасываются условия изображения:

$$\hat{\epsilon}_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t) = w\epsilon_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t) + (1 - w)\epsilon_\theta(\mathbf{x_t}, t) \tag{13}$$

Если условие изображения отброшено, эмбеддинг соответствующего изображения зануляется.

5 Метод

В данной работе представлена архитектура BoLID (Body Language Identity Diffusion), разработанная для интеграции идентичности тела пользователя в процесс генерации изображений. Предлагаемый метод улучшает персонализацию по сравнению с существующими подходами, которые учитывают только черты лица, за счет включения характеристик тела пользователя, таких как пропорции и поза.

5.1 Архитектура метода

Архитектура BoLID состоит из двух основных модулей:

- Body-ID Encoder: Данный модуль обучается извлекать особенности тела пользователя, включая 12 ключевых точек туловища, таких как плечи, локти, бедра и колени. Подход аналогичен используемому в InstantID [1] для лица, но адаптирован для учета пропорций тела.
- Body-Condition Module: Выступает в роли адаптера, схожего с IP-Adapter [2], и разделяет механизм перекрестного внимания (cross-attention) между текстовыми признаками, признаками лица и признаками тела, обеспечивая их совместную обработку в процессе генерации.

5.2 Функции потерь

Для обучения модели используются следующие функции потерь:

- Reconstruction Loss: Минимизирует расхождение между сгенерированным изображением и референсным образцом с учетом условия bodyID:
$$\mathcal{L}_{\text{rec}} = \|x_{\text{gen}} - x_{\text{ref}}\|^2,$$
где x_{gen} — сгенерированное изображение, x_{ref} — референсное изображение.
- Identity Consistency Loss: Расширена для тела на основе подхода InstantID [1], обеспечивая согласованность пропорций и ключевых точек:
$$\mathcal{L}_{\text{id}} = \sum_{k=1}^{12} \|p_{\text{gen},k} - p_{\text{ref},k}\|^2,$$
где $p_{\text{gen},k}$ и $p_{\text{ref},k}$ — координаты k-й ключевой точки на сгенерированном и референсном изображениях соответственно.
- CLIP-based Similarity Loss: Использует EVA-CLIP [3] для обеспечения семантической близости к текстовому промпту и стилю референсного изображения:
$$\mathcal{L}_{\text{CLIP}} = 1 - \cos(\text{CLIP}(x_{\text{gen}}), \text{CLIP}(\tau)),$$
где cos — косинусное сходство, $\text{CLIP}(x_{\text{gen}})$ — признаки сгенерированного изображения, $\text{CLIP}(\tau)$ — признаки текстового промпта.

6 Метрики

Для оценки качества модели BoLID предлагаются следующие метрики:

- FID (Fréchet Inception Distance): Измеряет расстояние между распределениями реальных и сгенерированных изображений в пространстве признаков InceptionV3:
$$FID = ||\mu_p - \mu_q||^2 + Tr(\mathbf{\Sigma_p} + \mathbf{\Sigma_q} - 2(\mathbf{\Sigma_p}\mathbf{\Sigma_q})^{1/2}) \tag{14}$$
где μ_p и μ_q — средние значения признаков в реальных и сгенерированных изображениях соответственно, $\mathbf{\Sigma_p}$ и $\mathbf{\Sigma_q}$ — ковариационные матрицы для распределений признаков на реальных и сгенерированных изображениях соответственно.
- Pose Accuracy: Среднеквадратичное отклонение координат ключевых точек тела:
$$\text{Pose}_{\text{acc}} = \frac{1}{12} \sum_{k=1}^{12} \|p_{\text{gen},k} - p_{\text{ref},k}\|^2.$$
- CLIP Score: Косинусное сходство между CLIP-признаками сгенерированного изображения и текстового промпта:
$$\text{CLIP}_{\text{score}} = \cos(\text{CLIP}(x_{\text{gen}}), \text{CLIP}(\tau)).$$

7 Датасет

Поскольку наша цель - встроить тело пользователя, важно дообучить модель на множестве ракурсов одного человека. Поэтому в качестве датасета были выбраны высококачественные (с разрешением не менее 1000×1000 px) фотографии знаменитостей. Для каждого человека собирался пакет из не менее 100 фотографий с сайта theplace с помощью написанного парсера. Датасет доступен по ссылке.

8 Вычислительный эксперимент

9 Заключение

Список литературы

[1] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519, 2024.

[2] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[3] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. 2023.

[4] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. In Advances in Neural Information Processing Systems, 2024.

[5] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023.

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[7] Xiao Yang Shanchuan Lin, Anran Wang. Sdxl-lightning: Progressive adversarial diffusion distillation. 2024.

[8] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. 2023.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 19730–19742. PMLR, 23–29 Jul 2023.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.