BoLID: Body-Lightning ID Diffusion

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

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

Abstract

Большие генеративные модели продемонстрировали впечатляющие результаты в задаче генерации изображений по текстовому описанию. Одной из ключевых областей применения таких моделей является создание персонализированных изображений. Распространённые подходы, как правило, требуют дополнительных входных данных, таких как управляющие сигналы или несколько ракурсов объекта. В данной работе мы предлагаем метод, позволяющий выполнять высококачественную персонализированную генерацию изображений исключительно на основе текстового запроса, без использования вспомогательных данных. Ключевая идея заключается в обучении латентного вектора в рамках вариационного автокодировщика (VAE), способного кодировать информацию о форме и структуре тела пользователя. Наша модель, *Body Lightning ID Diffusion* (BoLID), расширяет архитектуру IP-Adapter, при этом сохраняет неизменной основную диффузионную модель. Вместо модификации базовой архитектуры, мы обогащаем латентное представление до этапа обучения адаптеров. Впоследствии адаптеры могут быть обусловлены как на основе изученного латентного вектора, так и напрямую от входного изображения. Предлагаемый подход демонстрирует конкурентоспособные результаты по стандартным метрикам оценки качества, таким как Fréchet Inception Distance (FID) и Inception Score (IS).

Ключевые слова: Machine Learning, Diffusion Models.

1 Введение

Значительный прогресс в области генерации изображений был достигнут благодаря таким архитектурам, как U-Net [1], DALL-E 2 [2], Imagen [3] и Stable Diffusion [4]. Первоначально пользователи могли генерировать изображения на основе текстового описания, что существенно ограничивало гибкость модели. Модель DALL-E 2 сделала шаг вперёд, интегрировав визуальные представления в процесс генерации при помощи диффузионных моделей. Дальнейшее развитие привело к появлению компактных адаптеров, таких как IP-Adapter [5], которые разделяют механизм *cross-attention* между текстовыми и визуальными признаками. Это дало возможность вводить элементы управляемой генерации, аналогично ControlNet [6], и использовать сторонние модели, дообученные на той же базовой архитектуре. Модели InstantID [7] и PuLID [8] предложили дополнительные механизмы контроля за идентичностью пользователя, включая признаки лица и ключевые точки, что значительно повысило идентификационную точность (ID fidelity). Тем не менее, данные подходы фокусируются исключительно на особенностях лица, игнорируя морфологические характеристики тела.

В настоящей работе мы предлагаем метод, учитывающий индивидуальные особенности телосложения пользователя путём внедрения в процесс генерации *Body-ID* представления. Кроме того, наше решение позволяет отказаться от всего, кроме текстового промпта, без ухудшения качества аватара. Ключевая задача заключается в сохранении в скрытом векторе информативных признаков с референсных изображений.

Для повышения ID-сходства мы дообучаем модель на специально собранных датасетах, сгруппированных по уникальным идентификаторам пользователей. Каждый идентификатор представлен серией изображений с различных ракурсов, что обеспечивает целостное восприятие фигуры при генерации на основе одного входного изображения.

Предлагаемое решение включает два основных модуля:

1. **Body-ID Encoder** — вариационный автоэнкодер (VAE), обучаемый на агрегацию признаков из нескольких ракурсов, взвешенных по степени уверенности в позах, выделенных моделью YOLOv11 [9]. Результатом является компактный латентный вектор, отражающий морфологию тела.
2. **Body-Condition Module** — адаптер, реализующий механизм *cross-attention* между признаками текста, лица и тела, по аналогии с IP-Adapter [5].

2 Related Works

Современные диффузионные модели значительно улучшили генерацию изображений по текстовому описанию, достигнув фотореалистичного качества. Тем не менее, персонализированная генерация изображений людей остаётся сложной задачей из-за высокой вариативности анатомических признаков и ограниченности индивидуальных данных. В данном разделе рассматриваются ключевые подходы, направленные на сохранение идентичности и адаптацию генеративных моделей под конкретного пользователя.

Авторы **IP-Adapter** [5] предложили раздельную обработку признаков текста и изображения путём модификации механизма *cross-attention*, что позволило интегрировать визуальную информацию в процесс генерации. 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.

Дальнейшие исследования были направлены на расширение идеи легковесных адаптеров для улучшения идентичности и адаптацию моделей к индивидуальным признакам. **InstantID** [7] достигает высокой точности сохранения внешности за счёт использования предобученного энкодера лиц InsightFace. Модель работает в zero-shot режиме и требует всего одно референсное изображение, что делает её особенно удобной в применении, но при этом она ограничена только facial identity и не учитывает тело пользователя. В модели **PuLID** [8] предложен механизм контрастивного обучения для плавного объединения различных идентичностей. Это особенно эффективно при генерации изображений с комбинированной стилизацией, например, «лицо пользователя A в стиле B». Тем не менее, фокус также остаётся исключительно на лице. Подход **PhotoMaker** [10] предлагает стратегию *Stacked ID Embedding*, при которой используется несколько фотографий одного человека под различными ракурсами. Это позволяет добиться высокой степени сохранения идентичности без необходимости модифицировать параметры базовой диффузионной модели. Однако данный метод требует более трёх изображений и не адаптирован для моделирования позы тела.

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 идеи **IP-Adapter**, вводя дополнительный контроль за позой и формой тела пользователя. **Принципиальное отличие нашего подхода заключается в том, что для генерации достаточно лишь текстового промта — модель сохраняет высокое качество реконструкции тела даже при отсутствии изображения на входе.** В случае, если изображение всё же используется, модель не требует набора ракурсов или нескольких снимков: пространственные и кинематические характеристики тела восстанавливаются за счёт обучения скрытого вектора, извлекаемого VAE-энкодером.

Такой подход обеспечивает устойчивость генерации к ограничениям на количество или разнообразие референсных изображений и повышает универсальность системы в условиях реального применения.

3 Preliminary

3.1 Диффузионные модели. DDPM

Работа **диффузионной модели DDPM** [11] представляет из себя два процесса: прямой и обратный. Прямой процесс представляет собой последовательное зашумление входного изображения x_0 за T шагов, где x_t вычисляется по следующей формуле:

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

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

$$x_t|x_{t-1} \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{2}$$

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

$$x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon, \tag{3}$$

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

$$x_t|x_0 \sim \mathcal{N}(\sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I). \tag{4}$$

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

$$\hat{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\hat{\epsilon}_\theta(\hat{x}_t, t)\right) + \hat{\sigma}_t z, \tag{5}$$

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 $\hat{\epsilon}_\theta$ для $\hat{x}_t$; z — шум, который позволяет генерировать различные изображения, причем $z \sim \mathcal{N}(0, I)$, если $t > 1$, иначе $z = 0$.

3.2 Условная генерация: Classifier-free Guidance

Метод classifier-free guidance[12] позволяет увеличить степень, с которой модель ориентируется на идентификатор класса c. Во время семплирования предсказание получается путем линейной комбинации предсказаний обусловленной и необусловленной моделей:

$$\hat{\epsilon}_\theta(x_t, c, t) = (w+1)\epsilon_\theta(x_t, c, t) - w\epsilon_\theta(x_t, t), \tag{6}$$

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:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)} w_t \|\hat{x}_\theta(\alpha_t \mathbf{x} + \sigma_t\epsilon, \mathbf{c}) - \mathbf{x}\|^2, \tag{7}$$

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:

$$\mathbb{E}_{\epsilon,\epsilon'}[w_t\|\hat{x}_\theta(\alpha_t\mathbf{x} + \sigma_t\epsilon, \mathbf{c}) - \mathbf{x}\|^2 + \lambda w_{t'}\|\hat{x}_\theta(\alpha_{t'}\mathbf{x_{pr}} + \sigma_{t'}\epsilon', \mathbf{c_{pr}}) - \mathbf{x_{pr}}\|^2], \tag{8}$$

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

3.3 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.

3.4 Адаптация изображения через 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

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

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:

$$\mathcal{L}(\epsilon, \epsilon_\theta) = \mathbb{E}_{\mathbf{x_0},\epsilon,\mathbf{c_t},\mathbf{c_i},t}\|\epsilon - \epsilon_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t)\|^2, \tag{9}$$

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:

$$\hat{\epsilon}_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t) = w\epsilon_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t) + (1 - w)\epsilon_\theta(\mathbf{x_t}, t) \tag{10}$$

Если условие изображения отброшено, эмбеддинг соответствующего изображения зануляется.

Таким образом, диффузионная модель Stable Diffusion с IP-Adapter позволяет нам обучать компактный модуль на изображениях и промптах, сохраняя остальную модель фиксированной.

4 Method

В данном разделе подробно описывается предложенная архитектура *BoLID* (Body Language Identity Diffusion) и процедура её обучения. Основная цель метода — интеграция признаков тела пользователя в процесс текстово-ориентированной генерации изображений без потери качества моделируемой фигуры.

Рассмотрим датасет $\mathfrak{D} = \{(I_i, \tau_i) \mid i = 1, \ldots, n\}$, где I_i — изображение пользователя, а τ_i — соответствующий текстовый промпт. Предполагается, что каждое I_i содержит информацию как о лице, так и о теле пользователя. Пусть ϵ_θ — модель из класса диффузионных моделей.

4.1 Gaussian weighting

Мы будем использовать гауссово взвешивание (Gaussian weighting) для каждой ключевой точки тела. Это позволит присвоить высокий вес фрагментам изображения, близким к ключевым точкам, и эффективно игнорировать фон.

Для ключевой точки с координатами (x_i, y_i), $i = \overline{1, 17}$ вес пикселя с координатами (x, y) определяется гауссовой функцией:

$$w_i(x, y) = \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right), \tag{11}$$

где σ параметр, контролирующий радиус влияния точки. Чем больше σ, тем шире область влияния.

Пусть у нас K ключевых точек, итоговый вес для пикселя (x, y) - это максимум весов от всех точек:

$$W(x, y) = \max_{i=1,\dots,K} w_i(x, y) \tag{12}$$

Чтобы веса находились в диапазоне $[0, 1]$ и не искажали интенсивность изображения, нормализуем:

$$W_{\text{norm}(x,y)} = \frac{W(x, y)}{\max(W(x, y))} \tag{13}$$

Тогда взвешивание изображения зададим формулой $I_w(x, y) = I(x, y) \cdot \big(\alpha + (1 - \alpha)W_{\text{norm}(x,y)}\big)$, где $\alpha \in [0, 1]$ позволяет полностью не занулять фон изображения.

4.2 Latent knowledge vector

На первом этапе необходимо выделить вектор $c_{\text{body}} \in \mathbb{R}^{756}$. Каждое изображение I_j из поданного на вход множества ракурсов $\{I_j\}_{j=1}^{100}$ приводится к размеру 1024×1024, а потом обрабатывается моделью Γ_{body} для выделения структурных элементов контроля $\{s_j\}_{j=1}^{100}$, $s_j \in \mathbb{R}^{17}$. В нашем случае Γ_{body} это модель YOLO11-pose, если какая-то точка тела не видна, соотвествующее значение компоненты вектора зануляется. Для выделения фрагментов изображения мы применяем гауссово взвешивание, получая $\{I_w^j\}_{j=1}^{100}$. Все изображения I_w^j объединяются в один тензор, который подается на вход VAE-энкодеру, выходом которого является вектор $c_{\text{body}} \in \mathbb{R}^{756}$.

4.3 Body-Condition Module

Модуль условной генерации принимает на вход вектор c_{body}, текстовое описание τ_i и признаки референсного изображения I_{ref} при его наличии. Далее необходимо обучить адаптеры, интегрирующие их в генеративный процесс. Для этого используется модуль, аналогичный IP-Adapter [5], расширенный для одновременного учета нескольких источников информации:

$$\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}_t, \mathbf{V}_t) + \lambda_{\text{ref}} \,\text{Attention}(\mathbf{Q}, \mathbf{K}_{\text{ref}}, \mathbf{V}_{\text{ref}}) + \lambda_b \,\text{Attention}(\mathbf{Q}, \mathbf{K}_b, \mathbf{V}_b),$$

где

- $(\mathbf{K}_t, \mathbf{V}_t)$ — пары ключей и значений из текстового энкодера $\Gamma_\tau(\tau)$;
- $(\mathbf{K}_{\text{ref}}, \mathbf{V}_{\text{ref}})$ — пары ключей и значений из image-энкодера для референсного изображения;
- $(\mathbf{K}_b, \mathbf{V}_b)$ — пары ключей и значений из Body-ID Encoder.

Варьирование гиперпараметров $\lambda_{\text{ref}}, \lambda_b$ дает дополнительные степени свободы. Это позволит модели делать больший упор либо на лицо, либо на тело в процессе генерации. Заметим, что в случае отсутствия референсного изображения, слагаемое с λ_{ref} зануляется. Такой подход позволяет декомпозировать визуальное представление тела через агрегированные признаки других изображений, что делает возможной генерацию по одному текстовому описанию. При наличии изображения на входе модель использует его как прямой источник эмбеддинга тела, не требуя множества ракурсов, так как структурная информация уже закодирована в скрытом пространстве VAE.

4.4 Full objective

Обучение всей системы производится посредством минимизации взвешенной суммы потерь:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \alpha_{\text{id}} \,\mathcal{L}_{\text{id}} + \beta_{\text{CLIP}} \,\mathcal{L}_{\text{CLIP}}, \tag{14}$$

где

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\epsilon} \big\| \epsilon - \epsilon_\theta(x_t, \mathbf{c}_\tau, \mathbf{c}_f, \mathbf{c}_{\text{body}}, t) \big\|^2,$$
$$\mathcal{L}_{\text{id}} = \sum_{k=1}^{17} \big\| p_k^{\text{gen}} - p_k^{\text{ref}} \big\|^2,$$
$$\mathcal{L}_{\text{CLIP}} = 1 - \cos\big(\text{CLIP}(x_{\text{gen}}), \,\text{CLIP}(\tau)\big).$$

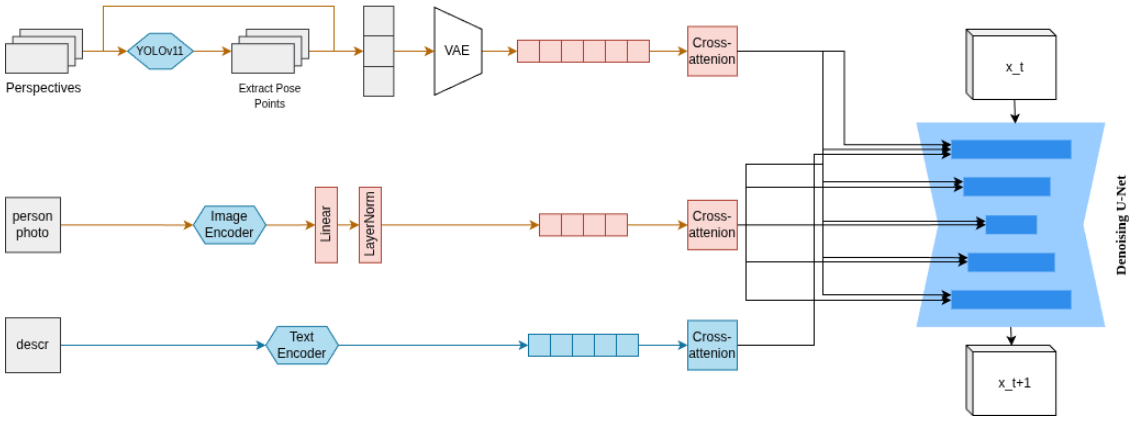

Рис. 1: Схема архитектуры *BoLID*.

5 Метрики

Для оценки качества модели *BoLID* предлагаются следующие метрики:

- **FID (Fréchet Inception Distance)**[13]: Измеряет расстояние между распределениями реальных и сгенерированных изображений в пространстве признаков InceptionV3:

$$FID = ||\mu_p - \mu_q||^2 + Tr(\mathbf{\Sigma_p} + \mathbf{\Sigma_q} - 2(\mathbf{\Sigma_p}\mathbf{\Sigma_q})^{1/2}) \tag{15}$$

  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.

- **IS (Inseption Score)** [14]

6 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.

Обработка датасета состояла в удалении изображений с водяными знаками и более чем одним человеком на изображении. Для обнаружения водяных знаков использовалась модель watermark-detection, а модель YOLOv11n-face-detection для проверки количества людей.

7 Вычислительный эксперимент

8 Заключение

Список литературы

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

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

[3] Chitwan Saharia, William Chan, 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, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[5] 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.

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[7] 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.

[8] 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.

[9] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.

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

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

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

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

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

[16] 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.

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

[18] 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.

[19] 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.