Body-Lightning ID Diffusion

Даниил Иванович Казачков
Научный руководитель: А. В. Филатов

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных

2025

Генеративные модели

Основные подходы в генерации изображений (от новых к старым):

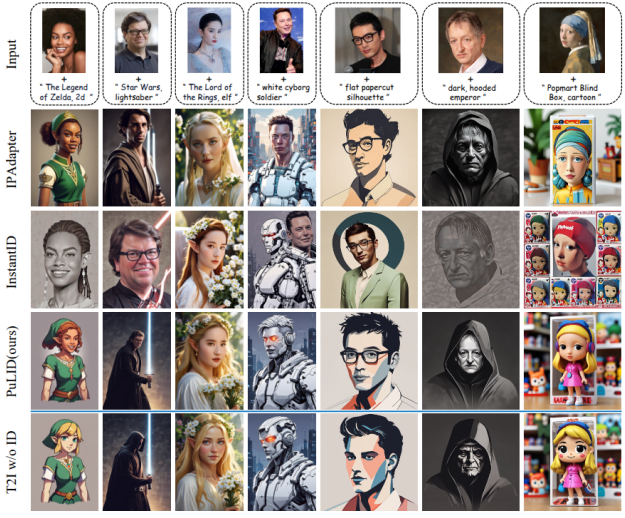
- Диффузионные модели
- Автокорреляционные модели и трансформеры
- Генеративные состязательные сети (GANs)
- Вариационные автокодировщики (VAEs)

  Диффузионные модели генерируют изображение путём последовательного удаления шума из случайного распределения. Современные модели: Stable Diffusion, DALL-E 2, Imagen.

Мотивация исследования

Существующие инструменты для генерации персонализированных изображений, используют лишь лицо пользователя.

Мотивация исследования

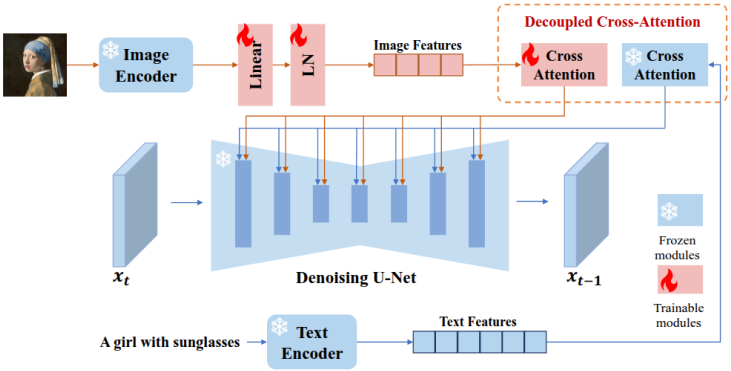


Источник: PuLID: Pure and Lightning ID Customization via Contrastive Alignment

Постановка задачи реконструкции стимулов

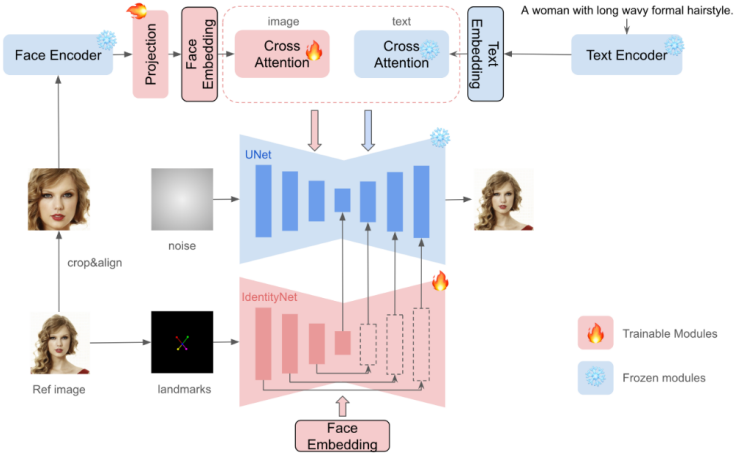
Наша главная задача - улучшить качество генерируемых аватаров, путем обуславливания картинки всем телом человека, причем человек может находиться в любой позе (сидеть, лежать и т.д., т.е. не обязательно стоять строго прямо). Мы проверяем гипотезу о том, что подходы в работах PuLID, InstantID, IP-Adapter распространяются не только на аватары лиц, но и на ростовые.

Обзор существующих решений: IP-adapter



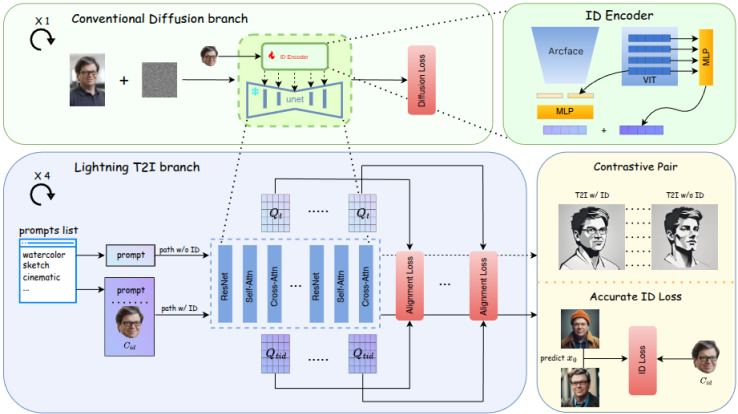
Источник: IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models

Обзор существующих решений: InstantID



Источник: InstantID : Zero-shot Identity-Preserving Generation in Seconds

Обзор существующих решений: PuLID



Источник: PuLID: Pure and Lightning ID Customization via Contrastive Alignment

О модели. Вход и выход. Метрика

вход: Ref Image + TextPrompt + Image Prompt (optional)

выход: Output Image

В качестве метрики буду использовать аналогичную той, что использовалась в Stable Diffusion модели. Это естественная функция для проверки того, насколько хорошо диффузионная модель "расшумляет"данные.

$$L = \mathbb{E}_{x_0,\epsilon,c_t,c_i,t}\left[\|\epsilon - \epsilon_\theta(x_t, c_t, c_i, t)\|^2\right], \tag{1}$$

где

- x_0 незашумленные данные
- ϵ случайный шум из распределения $\mathcal{N}(\mathbf{0}, \mathbf{I})$
- c_t, c_i дополнительное условие на текст (text) и картинку (image)
- t шаг (time step) диффузионного процесса

О модели. Валидация

Первой идеей является взять способ проверки как в IP-Adapter, пункт 4.2.1, а именно для использовать валидационную выборку COCO2017 (5 000 изображений с подписями). Для каждой подписи генерируется изображение (одинаковый случайный seed), всего получается 20 000 сгенерированных изображений на метод. Сравниваем качество генерации разными метриками, которые измеряют, насколько сгенерированное изображение соответствует

`CLIP-I` насколько векторное представление (эмбеддинг) сгенерированного изображения похоже на векторное представление реального изображения (из датасета COCO)

`CLIP-T` оценивают сходство между векторным представлением сгенерированного изображения и текстовой подписью (prompt).

Для вычисления CLIP-I и CLIP-T используют модель *CLIP ViT-L/14*.

О модели. Данные

Данные я собираю с сайта theplace с помощью написанного парсера. Он лежит в репозитории проекта.

О модели. Бейзлайн

IP-Adapter 6 с измененным датасетом.

О модели. Улучшения

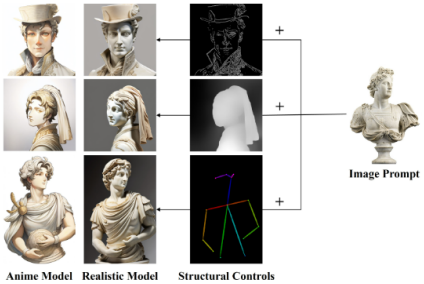


Рис. 1: Пример контролирующего элемента

Архитектура модели представлена на картинке 6, я попытаюсь поднять ее качество до уровня PuLID. Еще можно модифицировать энкодер для извлечения не только FaceID, но и BodyID. А еще докинуть элемент контроля "геометрии"(structural controls на картинке 1).

О модели. Внедрение

Способы примения моей модели - приложения типа Snapchat, чат-боты с API к моей модели и подобные сервисы, позволяющие поместить себя в разный контекст, улучшить внешность.