Создание персонализированных генераций изображений

Казистова Кристина Михайловна

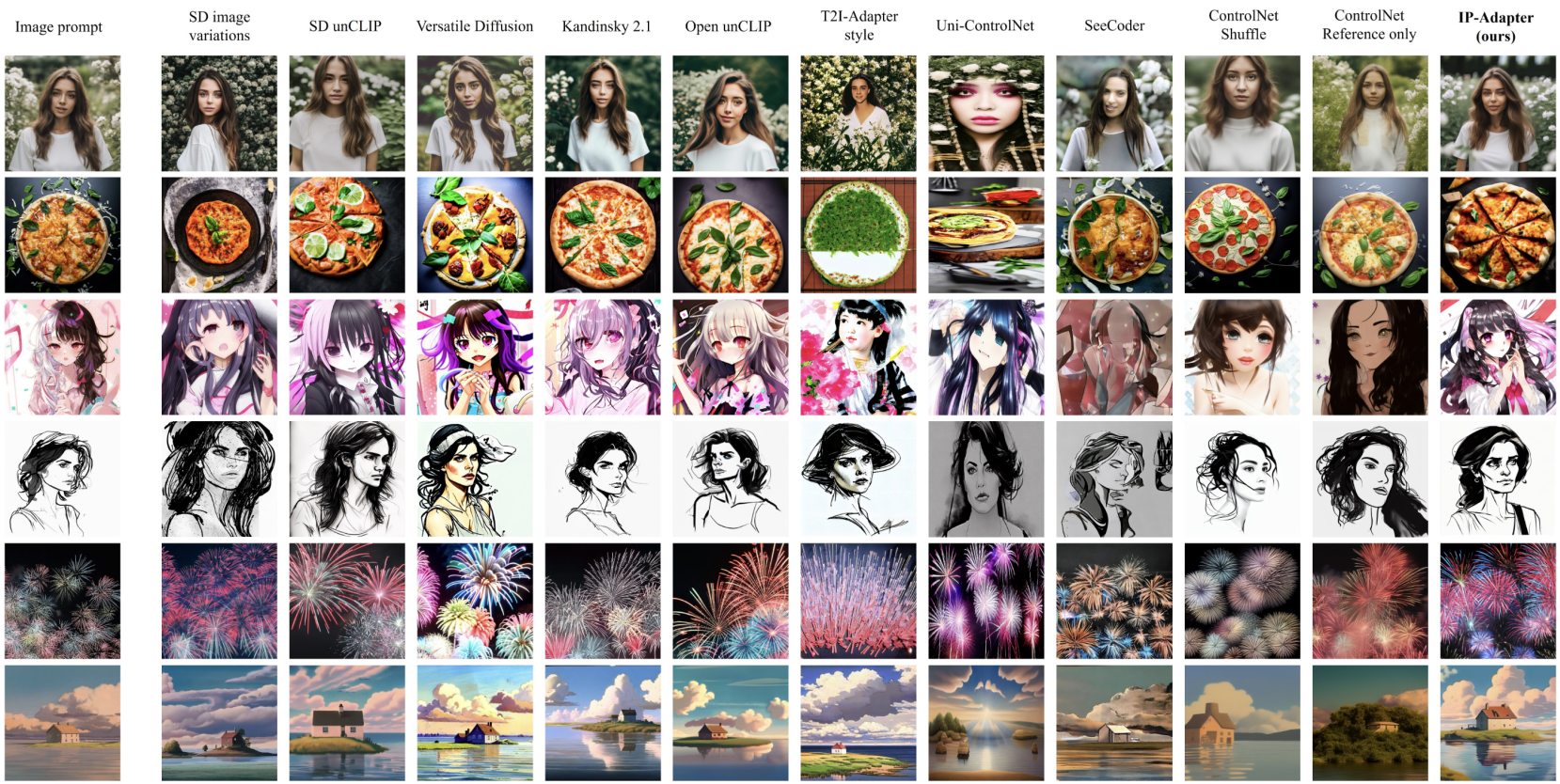
Московский физико-технический институт

Курс: Научный трек Иннпрака ФПМИ/Группа Б05-124

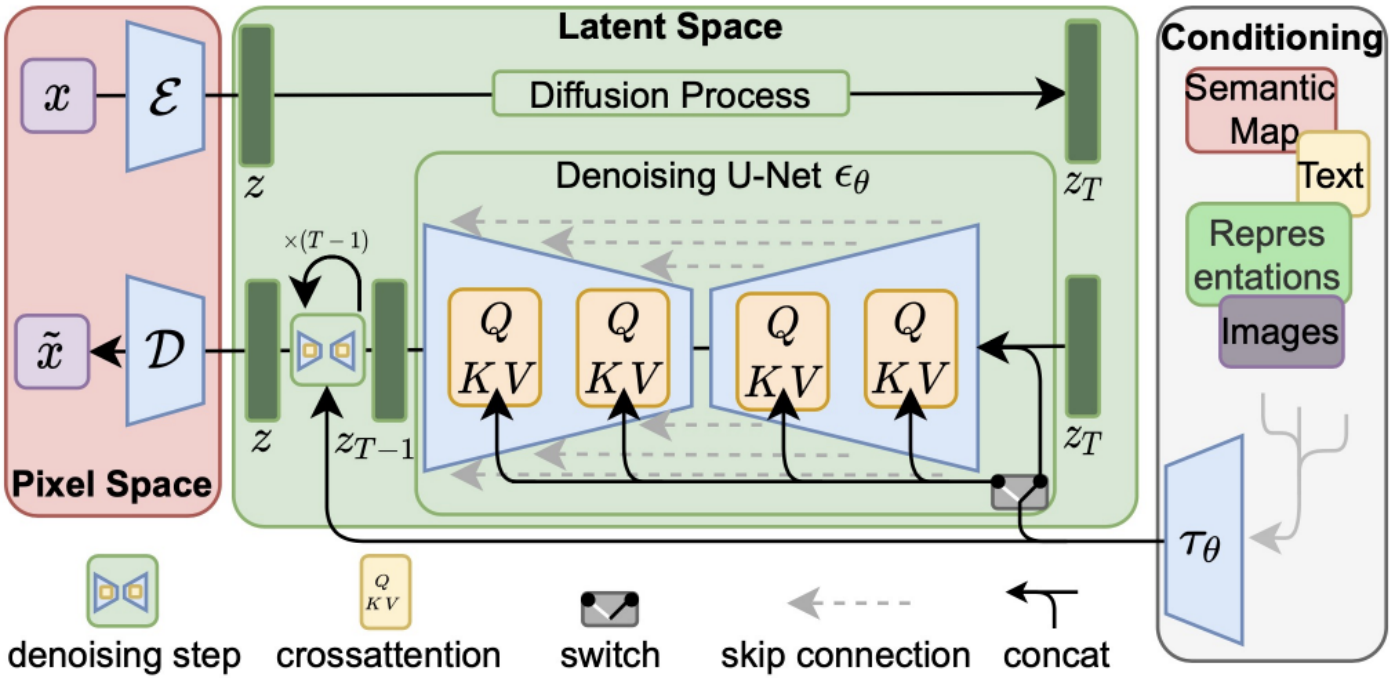
Эксперт: Филатов Андрей Викторович

2024

Примеры генерации изображений



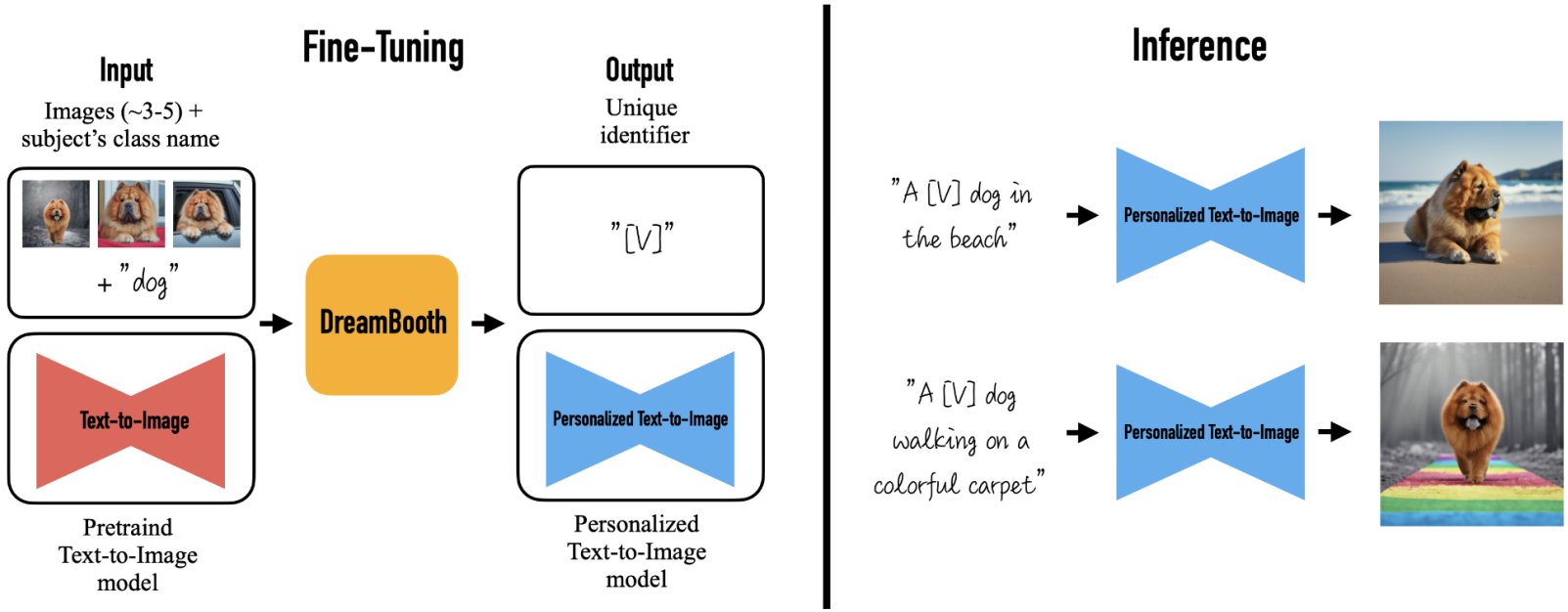
Latent Diffusion models



Основные этапы:

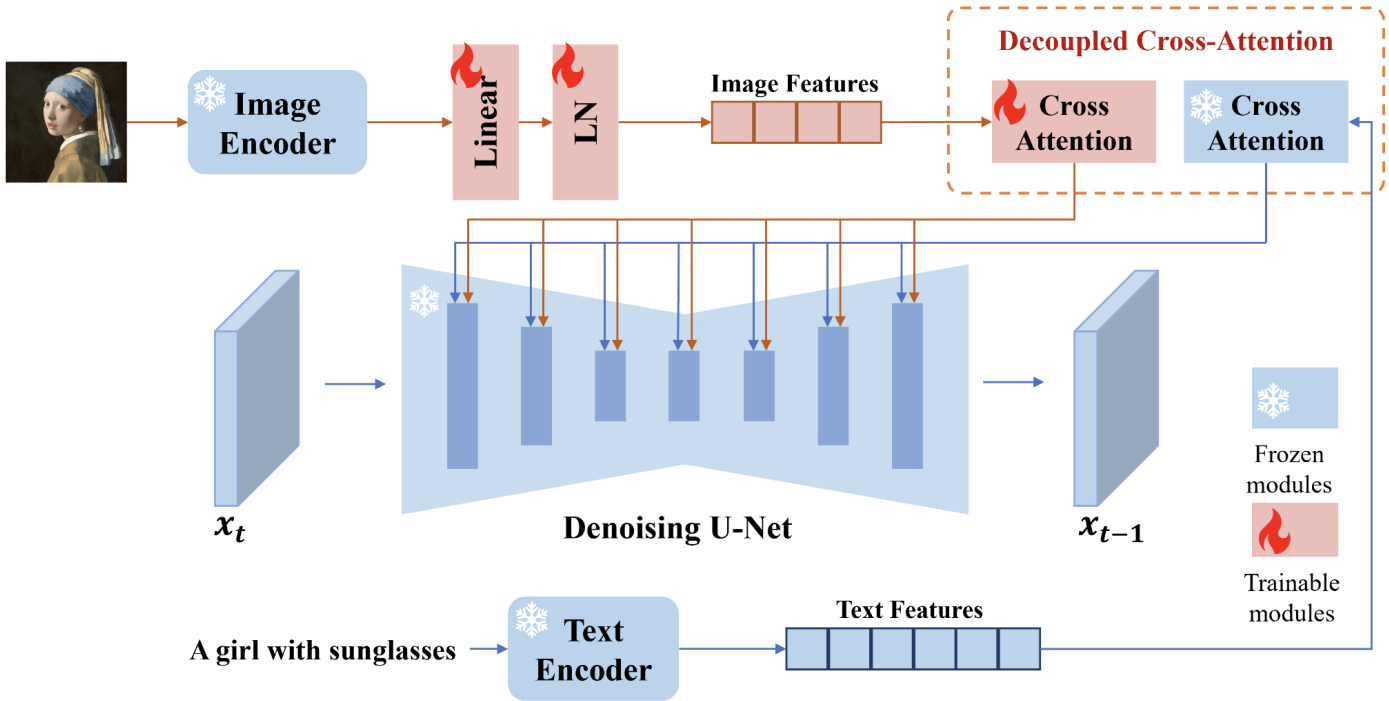
- Энкодер
- Скрытая диффузия
- Декодер

DreamBooth



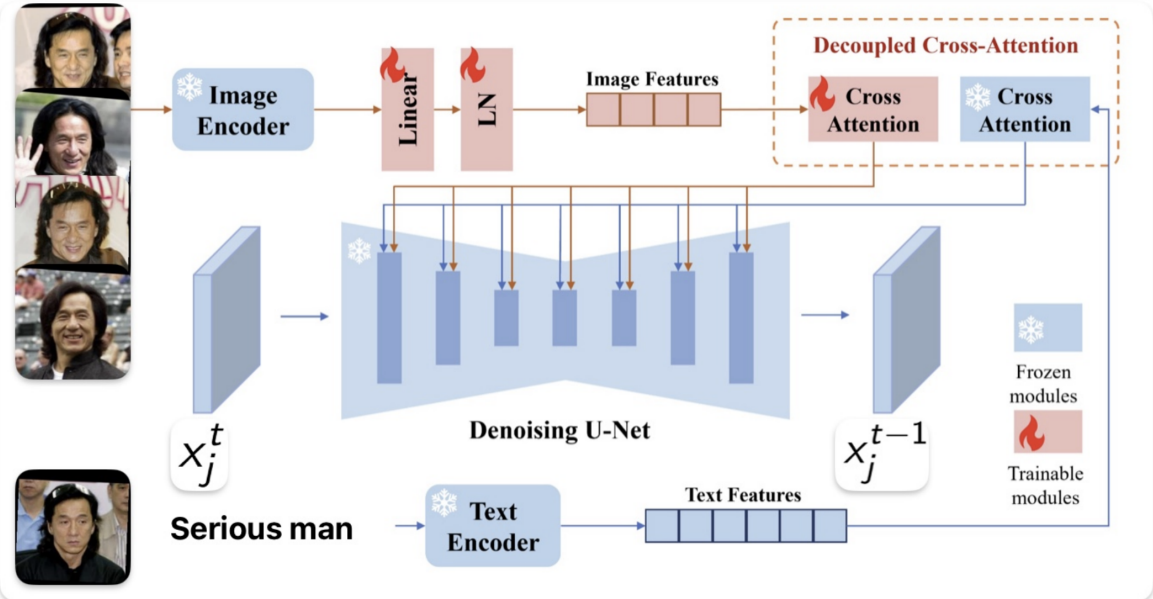
- Принимает на вход несколько изображений одного объекта вместе с соответствующим названием класса
- Возвращает специальный токен, идентифицирующий объект
- Токен встраивается в текстовую подсказку, по которой генерируется желаемое изображение

IP-Adapter



- Энкодер для извлечения признаков изображения
- Адаптированные модули с механизмом перекрестного внимания

Наш метод

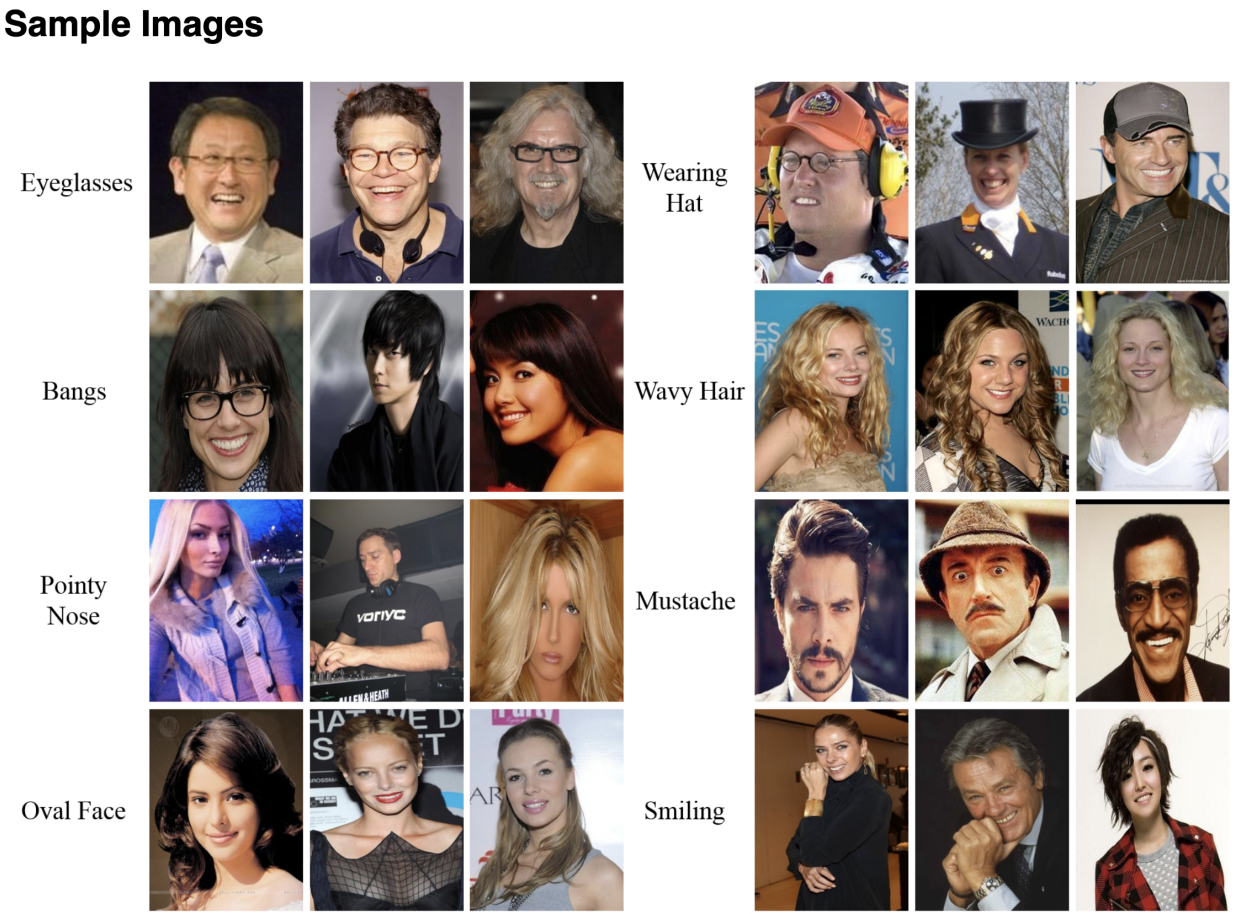


На вход подается несколько изображений вместо одного, каждому изображению соответствует своя текстовая подсказка. Ко входным изображениям применяется агрегирующая функция.

Процедура обучения: на каждом шаге из множества входных изображений случайно выбирается одно, удаляется из рассмотрения, и модель учится восстанавливать выброшенное изображение по его текстовому описанию и оставшимся картинкам.

Датасет

CelebA — набор данных по атрибутам лиц, содержащий более 200 тыс. изображений знаменитостей, каждое из которых снабжено 40 аннотациями к атрибутам.



Постановка задачи

Определим датасет как $X_0 = \{(x_i, t_i) : i = 1, \ldots, n\}$, где x_i — входное изображение, t_i — соответствующая ему текстовая подсказка. Рассматривается модель из класса диффузионных моделей. На этапе обучения на каждом шаге из датасета X_0 удаляется изображение $x_j, j \sim \mathcal{U}\{1, \ldots, n\}$.
Определим функцию потерь:

$$\mathcal{L} = \mathbb{E}_{X_0 \setminus \{x_j\}, \epsilon, c_t, c_i \setminus \{c_j\}, t} \|\epsilon - \epsilon_\theta(X_t, c_t, c_i \setminus \{c_j\}, t)\|^2,$$

где G — агрегирующая функция, применяемая ко входным данным; c_t — текстовые признаки; c_i — признаки изображений; c_j — признаки удаленного изображения; $t \in [0, T]$ — временной шаг диффузионного процесса; $X_t = \alpha_t G(X_0 \setminus \{x_j\}) + \sigma_t \epsilon$ — зашумленные данные на шаге t; α_t, σ_t — предопределенные функции от t, определяющие диффузионный процесс; ϵ_θ — цель обучения модели диффузии.
Решается следующая оптимизационная задача:

$$\theta^* = \arg\min_\theta \mathcal{L}(G(X_0 \setminus \{x_j\}), \epsilon, c_t, c_i \setminus \{c_j\}, t, x_t, \epsilon_\theta),$$

Постановка задачи

Frechet Inception Distance (FID), Inception Score (IS) — это метрики, которые используются для оценки качества генерации изображений
Формула для FID:

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

где μ_p и μ_q — средние значения признаков в реальных и сгенерированных изображениях соответственно, Σ_p и Σ_q — ковариационные матрицы для распределений признаков в реальных и сгенерированных изображениях соответственно.
Формула для IS:

$$IS(x) = \exp(\mathbb{E}_x\left[D_{KL}(p(y|x)||p(y))\right]),$$

где D_{KL} — дивергенция Кульбака-Лейблера для двух распределений $p(y|x)$ — вероятность класса y для изображения x и $p(y)$ — равномерное распределение на множестве классов, $\mathbb{E}_x$ — математическое ожидание по всем изображениям x.

Вычислительный эксперимент

- Написаны скрипты для запуска DreamBooth и IP-Adapter

План работы

- Подобрать агрегирующие функции
- Подобрать метрики идентичности и метрики разнообразия
- Реализовать наш метод с разными агрегирующими функциями
- Сравнить IP-Adapter и DreamBooth с нашей моделью, используя значения вышеупомянутых метрик