Body-Lightning ID Diffusion

Daniil Kazachkov, Andrew Filatov

20 марта 2025 г.

1 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.

keywords : Machine Learning, Diffusion Models.

2 Introduction

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 зна-

чимых деталей с референсных изображений. Как в InstantID restrict yourself to five key face points for a more generalized constraint, 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 экстрагировать особенности тела (body landmarks), аналогично подходу InstantID [1], но 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.

2.1 Loss Functions and Training Objective

- **Reconstruction Loss:** минимизирует расхождение между сгенерированным изображением и референсным образцом при наличии настраиваемого условия (bodyID).

- **Identity Consistency Loss:** аналогично InstantID [1], но расширено до корпуса и основных точек тела, для обеспечения правильной пропорциональности.

- **CLIP-based Similarity Loss:** использует EVA-CLIP [3] для обеспечения семантической близости к заданному тексту и общему стилю входного изображения.

3 Related Works

[Будет заполнено позже]

4 Method

[будет заполнено позже]

5 Computational experiment

5.1 Implementation Details

Мы строим нашу модель на основе SDXL [4] и (?) шагов SDXL-Lightning [5]. Для ID encoder мы используем (?) как модель распознавания тела, и EVA-CLIP [3] как CLIP Image encoder. Наш датасет состоит из приблизительно 50 тысяч картинок высокого разрешения, собранных из интернета, с подписями, автоматически сгенерированными с помощью BLIP-2 [6] суммарно. Наш процесс обучения состоит из (?) шагов.

[при получении финальной модели этот раздел будет дописан]

5.2 Basic experiment description

Бейзлайном выступает модель

5.3 Dataset

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.

5.4 Results

Ниже будет представлен макет таблицы с результатами эксперимента.

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] 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.

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

[6] 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.

[7] 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.