DreamIdentity: Enhanced Editability for Efficient Face-Identity Preserved Image Generation

Zhuowei Chen[1,2*], **Shancheng Fang**[1], **Wei Liu**[2], **Qian He**[2], **Mengqi Huang**[1], **Zhendong Mao**[1†]

[1] University of Science and Technology of China
[2] ByteDance
{chenzw01, huangmq}@mail.ustc.edu.cn {zdmao, fangsc}@ustc.edu.cn {liuwei.jikun, heqian}@bytedance.com

Abstract

While large-scale pre-trained text-to-image models can synthesize diverse and high-quality human-centric images, an intractable problem is how to preserve the face identity and follow the text prompts simultaneously for conditioned input face images and texts. Despite existing encoder-based methods achieving high efficiency and decent face similarity, the generated image often fails to follow the textual prompts. To ease this **editability issue**, we present *DreamIdentity*, to learn edit-friendly and accurate face-identity representations in the word embedding space. Specifically, we propose self-augmented editability learning to enhance the editability for projected embedding, which is achieved by constructing paired generated celebrity's face and edited celebrity images for training, aiming at transferring mature editability of off-the-shelf text-to-image models in celebrity to unseen identities. Furthermore, we design a novel dedicated face-identity encoder to learn an accurate representation of human faces, which applies multi-scale ID-aware features followed by a multi-embedding projector to generate the pseudo words in the text embedding space directly. Extensive experiments show that our method can generate more text-coherent and ID-preserved images with negligible time overhead compared to the standard text-to-image generation process.

Introduction

Diffusion-based large-scale text-to-image (T2I) models (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022) have revolutionized the field of visual content creation recently. With the help of these T2I models, it is now possible to create vivid and expressive human-centric images easily. An exciting application of these models is that, given a specific person's face in our personal life (our family members, friends, etc.), they can create different scenes associated with this identity using textual prompts.

Deviated from the standard T2I task, as shown in Fig.1, this task requires the model to have the ability to preserve input face identity (*i.e.*, ID-preservation) while adhering to textual prompts (*i.e.*, editability). A plausible solution is to personalize a pre-trained T2I model (Gal et al. 2023a; Ruiz et al. 2023; Kumari et al. 2022), which involves learning to associate a unique word (denoted as S*) with the identity by optimizing its word embedding (Gal et al. 2023a) or tuning the model parameters (Ruiz et al. 2023; Kumari et al. 2022) for multiple images from the same face identity. Nevertheless, this process is highly inefficient, requiring at least minute-level optimization and substantial memory overhead. Therefore, to get rid of the optimization process for each face identity, several recent encoder-based methods (Shi et al. 2023; Wei et al. 2023) propose to directly map the global image features (typically, CLIP (Radford et al. 2021) features) into a word embedding ($S*$) and local grid features are incorporated by introducing extra injection branch in the T2I backbone. The model is then trained with the objective of reconstructing the input image itself. Despite their efficiency and decent performance in maintaining ID similarity, these methods may struggle to follow the text prompts.

We argue that the inferior editability problem of existing optimization-free works stems from reconstruction-biased and inaccurate identity feature representation in the word embedding space. On one side, given an input face image I, existing methods aim to learn a unique word $S*$ so that the $S*$ can reconstruct I. The learned $S*$ does incorporate key identity information. However, it also captures as many ID-irrelevant details as possible from the input image to minimize the reconstruction loss, such as its style and do-

Figure 1: Given only one facial image, our model *DreamIdentity* can efficiently generate countless identity-preserved and text-coherent images in different contexts without any test-time optimization. The unique word $S*$ denotes the input face-identity, which can be combined with text prompts to generate various identity-preserved images.

[*] Author did this work during his internship at ByteDance.
[†] Zhendong Mao is the corresponding author.

main. This results in S* being biased towards reconstruction and less favorable for editing. On another side, the common object encoder (*i.e.*, CLIP) used by concurrent works (Wei et al. 2023; Shi et al. 2023; Ma et al. 2023; Gal et al. 2023b; Xiao et al. 2023) is unsuitable for ID-preservation as evident by the fact that the current best CLIP model is still much worse than the face recognition model on top-1 face identification accuracy (80.95% vs. 87.61% (Bhat and Jain 2023)). Additionally, the last layer feature from a classification/image-text matching network struggles to preserve the identity information since it primarily contains high-level semantics for recognition, lacking detailed facial descriptions. Therefore, in order to keep the identity, these methods attempt to change the pre-trained T2I model architecture to inject grid features, sacrificing the model's original editability. As a result, the generated images by ELITE (Wei et al. 2023) in Fig 4 struggles to follow the text prompt.

To mitigate this issue, we introduce a novel encoder-based approach (dubbed as *DreamIdentity*) with *edit-friendly and accurate representations* in the projected word embedding space to enhance the editability for efficient face-identity preserved image generation. Specifically, to ensure that the projected word embedding is edit-friendly rather than reconstruction-biased, Self-Augmented Editability Learning is devised to take the editing task into the training phase. It exploits the T2I model itself to construct a self-augmented dataset by generating celebrity faces along with a wide range of target-edited celebrity images. Subsequently, we design a dedicated Multi-word Multi-scale ID encoder named as M^2 ID encoder for identity encoding, whose backbone is a ViT-based network (Dosovitskiy et al. 2020) and pre-trained on a large-scale face dataset. Then multi-word embeddings are projected from its multi-scale coarse-to-fine features. Benefiting from accurate identity representation, we can keep the original T2I model untouched, thereby preserving its inherent editing capabilities without compromise. M^2 ID encoder is then trained with a combination of the self-augmented dataset and the typical face-only dataset to learn the edit-friendly identity and accurate word embedding S*.

The main contributions of our work can be summarized as follows:

Conceptually, we point out that current encoder-based methods fail for high editability due to their reconstruction-biased and inaccurate word embedding representation.

Technically, (1) For edit-friendly representation, we introduce self-augmented editability learning to generate a high-quality editing dataset by the foundation T2I model itself. (2) for accurate representation, we propose a delicate M^2 ID Encoder with multi-scale feature and multi-embedding projection.

Experimentally, extensive experiments demonstrate the superiority of our method, which can efficiently achieve flexible text-guided generation while preserving high ID-similarity.

Related Work

Text-to-image Generation

Text-to-Image generation aims to generate realistic and semantically consistent images with natural language descriptions. Early works mainly adopted GAN (Goodfellow et al. 2014) as the foundational generative model for this task. Various works have been proposed (Zhang et al. 2021; Zhu et al. 2019; Xu et al. 2018; Zhang et al. 2017; Zhang, Xie, and Yang 2018; Liang, Pei, and Lu 2020; Cheng et al. 2020; Ruan et al. 2021; Tao et al. 2020; Li et al. 2019; Huang et al. 2022) with well-designed textual representations, elegant text-image interactions, and effective loss functions. However, GAN-based models often suffer from training instability and model collapse, making it hard to be trained on large-scale datasets (Brock, Donahue, and Simonyan 2018; Kang et al. 2023; Schuhmann et al. 2021). Witnessed by the scalability of large language models (Radford et al. 2019), autoregressive methods like DALL-E and Parti (Yu et al. 2022; Ramesh et al. 2021) where the images are quantized into discrete tokens are scaled to learn more general text-to-image generation. More recently, diffusion models, such as GLIDE (Nichol et al. 2021), Imagen (Saharia et al. 2022), DALL-E 2 (Ramesh et al. 2022), LDM (Rombach et al. 2022) have demonstrated the ability on generating unprecedentedly high-quality and diverse images. However, it is infeasible to generate a specified face/person identity within the context described by the text with the text-to-image model alone.

Personalized Image Synthesis

Recently personalization methods (Gal et al. 2023a; Ruiz et al. 2023; Kumari et al. 2022) have shown promising results in customized concept generation. We can apply these methods to our tasks when the concept is a specified face identity. Textual Inversion (Gal et al. 2023a) optimizes a new word embedding to represent the given specific concept. (Ruiz et al. 2023) (Kumari et al. 2022) associate the concept with a rare word embedding by fine-tuning all or part of the parameters in the generator. However, the necessity for multiple images to define a concept, in conjunction with the time-intensive optimization process for per identity (which requires at least several minutes), limits its broader application. In this work, we present an optimization-free method that directly encodes a face identity as the word embedding given only one image.

Similar to our goal, some recent works utilize an embedding encoder for efficient personalized image synthesis. Specifically, ELITE (Wei et al. 2023), UMM-Diffusion (Ma et al. 2023) and InstantBooth (Shi et al. 2023) encode a common object as a word embedding with the last layer feature from the CLIP encoder. Additionally, ELITE and InstantBooth augment finer details with a local mapping network. Our work differs in several aspects: 1) At the encoder level: We design a dedicated ID encoder for accurate face encoding with multi-scale features along with multiple word embeddings mapping, whereas the concurrent works use a last layer feature to predict a single word embedding with a common object encoder (CLIP). 2)Instead of training the en-

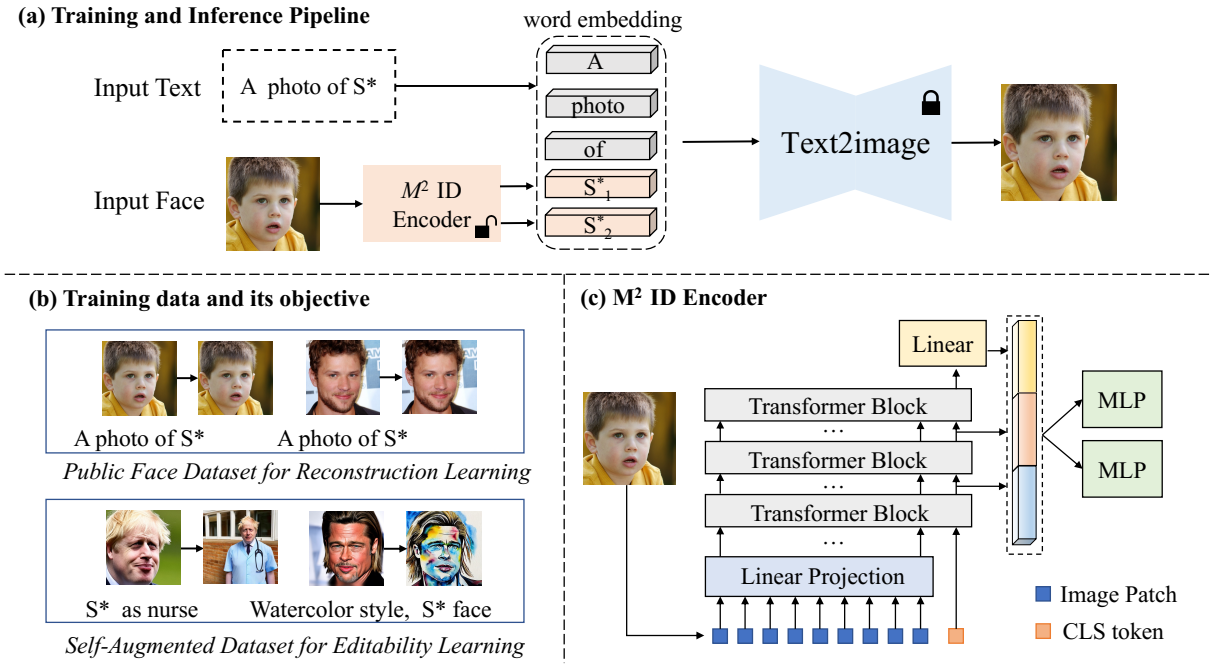

Figure 2: Overview of the proposed *DreamIdentity*: (a) The training and inference pipeline. The input face image is first encoded into multi-word embeddings (denoted by S^*) by our proposed M^2 ID encoder. Then S^* are associated with the text input to generate an image that is consistent with the scene depicted by the text and preserves input face identity. (b) The composition of the training data and its objectives. The training data consists of a public face dataset for reconstruction and a self-augmented dataset for editability learning. (c) The architecture of M^2 ID encoder, where a ViT-based face identity encoder is adopted as the backbone and the extracted multi-scale features are projected to multi-word embedding.

coder solely under the reconstruction objective, We propose a self-augmented editability learning method to improve the editability for face identity.

Methods

Given a single facial image of an individual, our objective is to endow the pre-trained T2I model with the ability to efficiently re-contextualize this unique identity under various textual prompts. These prompts may include clothing, accessories, styles, or background variations.

The overall framework is shown in Fig.2, given a pre-trained T2I model, to achieve fast and identity-preserved image generation, we first precisely encode the target identity into the word embedding space (represented as the pseudo word $S*$) with the proposed M^2 ID encoder. Afterward, $S*$ is integrated with the input prompt for generating the text-guided image. To avoid S* being reconstruction-biased, a novel self-augmented editability learning is further introduced to train the M^2 ID encoder with the editability objective.

In the following parts, we first briefly introduce the pre-trained diffusion-based text-to-image model used in our work, then describe our proposed self-augmented editability learning and M^2 ID encoder in detail.

Preliminary

In this work, we adopt the open-sourced Stable Diffusion 2.1-base (SD) as our text-to-image model, which has been trained on billions of images and shows excellent image generation quality and prompt understanding.

SD is a kind of Latent Diffusion Model (LDM) (Rombach et al. 2022). LDM firstly represents the input image x in a lower resolution latent space z via a Variational Auto-Encoder (VAE) (Kingma and Welling 2013). Then a text-conditioned diffusion model is trained to generate the latent code of the target image from text input c. The loss function of this diffusion model can be formulated as:

$$\mathcal{L}_{diffusion} = \mathbb{E}_{\epsilon,z,c,t}[\|\epsilon - \epsilon_\theta(z_t, c, t)\|_2^2], \quad (1)$$

where ϵ_θ is the noise predicted by the model with learnable parameters θ, ϵ is noise sampled from standard normal distribution, t is the time step, and z_t is noisy latent at the time step t.

During inference, the latent code is generated through the diffusion model. Subsequently, the decoder maps the latent code into the image space.

Self-Augmented Editability Learning

The vanilla reconstruction objective could result in the learned $S*$ tends to be reconstruction-biased, which causes inferior editability for the input face. An intuitive way is to

Figure 3: The samples from self-augmented dataset.

take the editing task into the training phase to improve editability. However, collecting such pair data for the editing task is challenging. To tackle this issue, we propose a simple yet effective self-augmented editability learning to generate a dataset by the model itself for editability learning.

Experimentally, We observe that current state-of-the-art text-to-image models can generate images of celebrities (e.g., Boris Johnson, Emma Watson) in various contexts, maintaining high levels of identity preservation and text-coherence. With this insight, The self-augmented editability learning utilizes the pre-trained model itself to construct a self-augmented dataset by generating various celebrity faces along with the target edited celebrity images, which will be used to train the M^2 ID encoder with the editability objective. Formally, the construction of the dataset includes the following four steps:

Step 1: Celebrity List Generation. Firstly, we collect a candidate celebrity list. The large language model (*i.e.*, ChatGPT) is used to generate the most famous 400 names in four fields (*i.e.*, sports players, singers, actors, and politicians). After filtering duplicate ones, we finally got 1015 celebrity names.

Step 2: Celebrity Face Generation. We use generated face images rather than real images because the model has its own understanding of the celebrities. Specifically, the celebrities who appeared less frequently in the Stable Diffusion training dataset are not very similar to the real person, while these generated faces maintain a high level of identity resemblance. We use the prompt template "[celebrity-name] face, looking at the camera" to produce the source images, followed by face crop and alignment operation to get face-only images. Furthermore, a face-only image is kept if its short size is larger than 128 pixels.

Step 3: Edit Prompts and Edited Images Generation. We manually design a variety of prompts that contain images of celebrities in different jobs, styles, and accessories (*e.g.*, "[celebrity-name] as a chef", "oil painting style, [celebrity-name] face"). Then these prompts are transformed to images by the T2I model as edited images, and the [celebrity-name] in prompts is replaced by the pseudo word S^* as Editing Prompts.

Step 4: Data Cleaning. After the above procedures, we can get the initial self-augmented dataset consisting of a set of triplets, [identity face, editing prompt, edited image]. Due to the instability of the current diffusion model, the edited images do not always follow the editing instructions. Therefore, we need to filter out the noise data in the self-augmented dataset. We employ ID Loss and CLIP score, which reflect identity similarity and text-image consistency as the metrics, to rank the edited images for every prompt, then the top 25% triplets at kept as the final training set.

Finally, as shown in Fig.3, we construct a high-quality self-augmented dataset consists of around 36k training examples from the pre-trained T2I model itself. This dataset is then used for edit-oriented training.

M^2 ID Encoder

To accurately represent the input face identity in the word embedding space without modifying the T2I model architecture, we propose a novel Multi-word Multi-scale embedding ID encoder (M^2 ID encoder), which is achieved by the multi-scale ID features extracted from a dedicated backbone, followed by multiple-word embedding projection.

Backbone. An accurate representation of the facial identity is crucial. However, the common image encoder CLIP (adopted by *all* existing works) fails for that purpose since it can not capture the identity feature as accurately as the face ID encoder, which has been trained for face identification tasks on large-scale face datasets. As (Bhat and Jain 2023) shows, the current best CLIP VIT-L/14 model is still much worse than the face recognition model on top-1 face identification accuracy (80.95% vs 87.61%). Therefore, we employ a ViT backbone (Dosovitskiy et al. 2020) pre-trained on a large-scale face recognition dataset to extract ID-aware features for input face faithfully.

Multi-scale Feature. However, naively mapping the final layer's output identity vector v_{final} could only bring sub-optimal identity preservation. The reason lies in that v_{final} mainly contains the high-level semantics, which is suitable for discriminative tasks (*e.g.*, face identification) rather than generative tasks. For example, the same identity with different expressions should share similar representation under the face recognition training loss, while the generation requests more detailed information like facial expressions, *etc*. Hence, only mapping the last layer representation could become an information bottleneck for the image generation task. To address the above problem, we utilize multi-scale features from the face encoder to represent an identity more faithfully. Specifically, the identity vector is augmented by four CLS embeddings (v_3, v_6, v_{12}, v_{12}) from the 3rd, 6th, 9th, and 12th layer, respectively. Formally, the multi-scale feature from the ID encoder is depicted as follows:

$$V = [v_3, v_6, v_9, v_{12}, v_{final}]. \tag{2}$$

Multi-word Embeddings. The multi-scale feature is further projected into the word embedding domain. To maintain the original large-scale T2I model's generalization and editability, we leave all its parameters and structure unchanged. As a result, it raises the problem that a single word embedding is hard to represent the face's identity faithfully. Therefore, we further propose a multi-word projection mechanism to represent a face with multi-word embedding:

$$s_i = MLP_i(V), \text{for } i = 1, ..., k, \tag{3}$$

where k is the number of embeddings. Experimentally, we set $k = 2$ as depicted in Fig.2. Following (Gal et al. 2023b), l_2 regularization is further adapted to constrain the output embedding:

$$\mathcal{L}_{reg} = \sum_{i=1}^{k} \|s_i\|. \tag{4}$$

Benefiting from the above-dedicated ID feature, we can facilitate highly identity-preservation control in the embedding space without sacrificing the pre-trained T2I model's editability caused by feature injection.

Training

We directly combine the FFHQ (Karras, Laine, and Aila 2019) and the self-augmented dataset to form 106k training examples (70k from FFHQ, 36k from the model itself) for training our proposed M^2 ID encoder. The total loss consists of noise prediction loss of the diffusion model and the embedding regularization loss:

$$\mathcal{L}_{total} = \mathcal{L}_{diffusion} + \lambda \mathcal{L}_{reg}, \tag{5}$$

where λ is the embedding regularization weight.

Experiments

Experiment Settings

Dataset. Our experiments are conducted on the widely used FFHQ dataset (Karras, Laine, and Aila 2019), which contains $70k$ high-resolution human face images. We resize the images to 512x512 for training. The test set consists of 100 faces from (Liu et al. 2015). We carefully check every test image to ensure no overlap between the test set and the self-augmented celebrity dataset.

Metrics. We evaluate our method on Text-alignment and Face-similarity. Text-alignment refers to the extent to which the generated image reflects the semantics specified in the editing prompts, calculated by the cosine distance in the CLIP text-image embedding space. Face-similarity is used to measure whether the face ID is preserved. We use the ID feature from Arcface (Deng et al. 2018), a model pre-trained on face recognition tasks, to represent the face identity. Then the cosine distance of ID features between the input face and the face cropped from the edited image is calculated to measure ID-similarity. The encoding time is adapted to measure the time to obtain the S*.

For each editing prompt and face identity, four images are generated to mitigate the randomness. We provided the editing prompts in the supplementary materials.

Implementation Details. We choose Stable Diffusion 2.1-base as our base text-to-image model. The learning rate and batch size are set to $5e-5$ and 64. The encoder is trained for 60,000 iterations. The embedding regularization weight λ is set to $1e-4$. Our experiments are trained on a server with eight A100-80G GPUs, which takes about one day to complete our experiment. We use the DDIM (Song, Meng, and Ermon 2020) sampler with 30 steps during inference. The guidance scale is set to 7.5.

Methods	Text-alignment ↑	Face-similarity ↑	E-Time ↓
TI	0.213	0.326	20 min
DB	0.217	0.425	4 min
E4T	0.220	0.420	20 s
Elite	0.196	0.450	0.05 s
Ours	**0.228**	**0.467**	**0.04 s**

Table 1: Quantitative comparisons with optimization-based and encoder-based methods. E-Time means the time cost to obtain the unique $S*$. Our method achieves better results in terms of text-alignment, face-similarity, and encoding time. TI, DB denotes Textual Inversion (Gal et al. 2023a) and DreamBooth (Ruiz et al. 2023), respectively.

Recon	self-aug	Text-alignment ↑	Face-similarity ↑
✓		0.213	0.380
	✓	0.216	0.348
✓	✓	**0.228**	**0.467**

Table 2: Ablation study on self-augmented editability learning. Recon denotes reconstruction training. self-aug denotes self-augmented editability learning, the editability gets improved after applying self-aug.

Comparison to SOTA Methods

In this section, we compare our method with optimization-based methods: Textual Inversion (Gal et al. 2023a), DreamBooth (Ruiz et al. 2023) and concurrent works on efficient personalized model: E4T (Gal et al. 2023b), which requires optimization for around 15 iterations for each face, and ELITE (Wei et al. 2023), an encoder-based method. We adopt the widely-used open-sourced Diffusers codebase for Textual Inversion, DreamBooth, and re-implemented E4T and ELITE. All experiments are conducted with a single-face image input to ensure a fair comparison.

Quantitative and Qualitative Results. As demonstrated in Tab.1, our work *DreamIdentity* outperforms recent methods across all the metrics, demonstrating superior performance in terms of *editability*, *ID-preservation*, and *encoding speed*. We show that *DreamIdentity* improves the text-alignment by 7% compared to the second-best E4T (Gal et al. 2023b) thanks to the edit-friendly word embedding S*. Meanwhile, *DreamIdentity* surpasses the second-best model (Wei et al. 2023) on *ID-preservation* by 3.7% due to the accurate identity representation while enjoying much better editability. Benefiting from the direct encoding rather than optimization for unique embeddings $S*$, the additional computation cost is only 0.04 s for every face identity, which can be negligible compared to the seconds-level time cost for a standard diffusion-based text-to-image process. The conclusion is further validated by the qualitative results in Fig.4. In particular, DreamIdentity outperforms the encoder-based method, ELITE, by generating more text-coherent images. For instance, it successfully captures the chef's outfit in row 1, funko pop style in row 2 and the snowy context in row 4. Conversely, ELITE exhibits bias towards the input image

Figure 4: Qualitative comparisons with state-of-the-art methods. *DreamIdentity* can generate comparable or better text-aligned and ID-preserved images.

ID	MS	ME	Text-alignment ↑	Face-similarity ↑
			0.229	0.266
✓			0.228	0.302
✓	✓		**0.229**	0.412
✓	✓	✓	0.228	**0.467**

Table 3: Ablation study on M^2 ID Encoder. ID encoder (ID) with multi-scale feature (MS) and multiple word embeddings (ME) achieves the best Face-similarity while maintaining a comparable result on the text-alignment metric.

Emb Num	Text-alignment ↑	Face-similarity ↑
1	**0.229**	0.412
2	0.228	0.467
3	0.188	**0.472**

Table 4: Ablation study on the number of word embeddings (Emb Num). Single-word embedding could limit the face-similarity, while excessive ones may hinder text-alignment.

and falls short in following the text prompts effectively.

Ablation Studies

In this section, we conduct ablation studies to verify the effectiveness of our proposed self-augmented editability learning and M^2 ID feature.

Self-Augmented Editability Learning. We can observe from Fig.5 that if the model is only trained under the reconstruction objective, the editability of embeddings will be limited. To be specific, the model trained without the editability learning objective fails to edit the input identity to a police. Besides, if we only use the limited generated editing dataset, face similarity will be degraded because there are only around 1000 face IDs in the self-augmented dataset. Combining the reconstruction data (i.e., FFHQ) and the generated self-augmented dataset is a better choice to preserve face similarity while following the textual instruction. The quantitative results in Tab.2 further confirm our conclusion.

M^2 ID Encoder. As shown in Tab.3, upon switching from the CLIP encoder to the face-specific ID encoder, the *ID-preservation* is improved from 0.266 to 0.302. Integrating the multi-scale features further boosts the *ID-preservation* to 0.412. Fig.6 further demonstrates the effectiveness of M^2

Figure 5: Qualitative comparisons on the self-augmented dataset for editability learning. The editing prompt is "S* as a police, looking at the camera". "w/o edit" and "w/o recon" denote for the encoder is trained without editability learning objective and without reconstruction learning, respectively. We show that the generated images can not follow the prompt properly without the editability learning. Meanwhile, the face similarity will be lower without the reconstruction learning on FFHQ.

Figure 6: Qualitative comparisons between ID Encoder and the multi-scale features. The editing prompt is "S* as a chef, looking at the camera". We could conclude that both ID Encoder and the multi-scale features greatly improve the ID preservation (*i.e.*, face-similarity).

ID Encoder.

Fig.7 indicates that multi-word embeddings could enhance ID-preservation. As shown in Tab.4, when we increase the number of embedding to 2, the Face-similarity is improved by 12% with marginal change of 0.4% on text-alignment. However, when we further increase the number of word embeddings, text-alignment is dropped by 17%. We

Figure 7: Qualitative comparisons of multiple word embeddings. The editing prompt is "S* as a police, looking at the camera", and "NUM" denotes the number of embeddings.

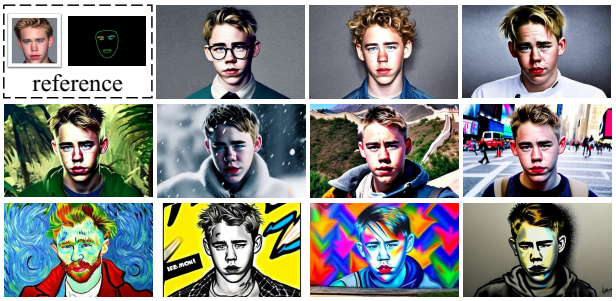

Figure 8: Given a face identity and its gaze location on the canvas, our method can generate a series of images that maintain the same identity while following the editing prompts in the same location.

argue that excessive word embeddings may include more information beyond the ID feature, which hinders editability. Therefore, we choose the embedding number as 2 to avoid degraded editability.

ID-preserved Scene Switch. As shown in Fig.8, given the input face ID and its location in the canvas indicated by the gaze location, we can generate a series of different scene images which share the same identity information and head location with the help of ControlNet (Zhang and Agrawala 2023). The generated scene is specified by the text description and can encompass different accessories, hairstyles, backgrounds, and styles. With this method, we may achieve the effect of "everything and everywhere all at once".

Conclusion

In this study, we present an edit-friendly encoder-based approach for generating a specified person in new scenes with only a single facial image. The self-augmented editability learning mechanism is proposed to endow the T2I model with the ability to achieve high editability. Moreover, The novel M^2 ID encoder is proposed to project the identity into multiple word embeddings with multi-scale ID-aware features for the accurate representation of the human with negligible time costs. Extensive experiments demonstrate the effectiveness of the proposed methods.

Acknowledgments

We thank Tianhao Qi for helping to revise our first draft. This work is supported by the National Science Fund for Excellent Young Scholars under Grant 62222212 and the National Science Fund under Grant 62102384.

References

Bhat, A.; and Jain, S. 2023. Face Recognition in the age of CLIP & Billion image datasets. *arXiv preprint arXiv:2301.07315*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Cheng, J.; Wu, F.; Tian, Y.; Wang, L.; and Tao, D. 2020. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In *CVPR*, 10911–10920.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2018. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In 2019 IEEE. In *CVPR*, 4685–4694.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023a. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR*.

Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023b. Designing an encoder for fast personalization of text-to-image models. *ACM Transactions on Graphics*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. *arXiv preprint arXiv:1406.2661*.

Huang, M.; Mao, Z.; Wang, P.; Wang, Q.; and Zhang, Y. 2022. DSE-GAN: Dynamic Semantic Evolution Generative Adversarial Network for Text-to-Image Generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4345–4354.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. *arXiv preprint arXiv:2303.05511*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*, 4401–4410.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2022. Multi-Concept Customization of Text-to-Image Diffusion. *arXiv preprint arXiv:2212.04488*.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. 2019. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32.

Liang, J.; Pei, W.; and Lu, F. 2020. CPGAN: Content-parsing generative adversarial networks for text-to-image synthesis. In *ECCV*, 491–508.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Ma, Y.; Yang, H.; Wang, W.; Fu, J.; and Liu, J. 2023. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ruan, S.; Zhang, Y.; Zhang, K.; Fan, Y.; Tang, F.; Liu, Q.; and Chen, E. 2021. DAE-GAN: Dynamic aspect-aware GAN for text-to-image synthesis. In *ICCV*, 13960–13969.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CVPR*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. *arXiv preprint arXiv:2304.03411*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tao, M.; Tang, H.; Wu, S.; Sebe, N.; Jing, X.-Y.; Wu, F.; and Bao, B. 2020. DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *ICCV*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. *arXiv preprint arXiv:2305.10431*.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 1316–1324.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, 833–842.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 5907–5915.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, Z.; Xie, Y.; and Yang, L. 2018. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 6199–6208.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 5802–5810.