Cross Initialization for Face Personalization of Text-to-Image Models

Lianyu Pang[1], Jian Yin[1,2], Haoran Xie[3], Qiping Wang[4], Qing Li[5], Xudong Mao[1,2*]
[1]Sun Yat-sen University [2]Guangdong Key Laboratory of Big Data Analysis and Processing
[3]Lingnan University [4]East China Normal University [5]The Hong Kong Polytechnic University

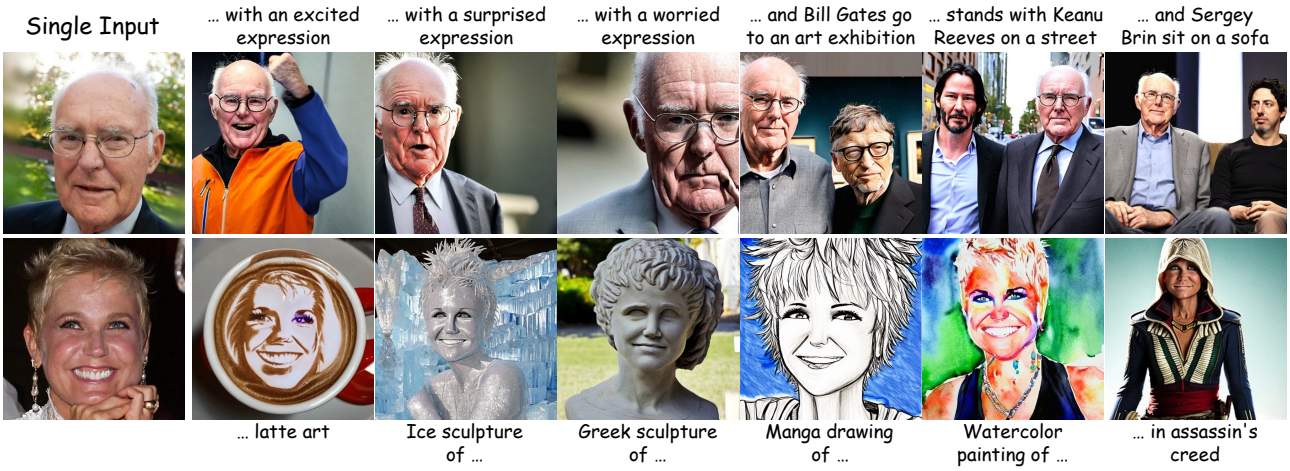

Figure 1. Personalization results of our method using a single input image. Our method enables a variety of novel personalized face generations with high visual fidelity, such as facial expression editing, interaction with other individuals, and stylization. Moreover, it significantly speeds up the personalization process by reducing the optimization steps from 5,000 to 320.

Abstract

Recently, there has been a surge in face personalization techniques, benefiting from the advanced capabilities of pretrained text-to-image diffusion models. Among these, a notable method is Textual Inversion, which generates personalized images by inverting given images into textual embeddings. However, methods based on Textual Inversion still struggle with balancing the trade-off between reconstruction quality and editability. In this study, we examine this issue through the lens of initialization. Upon closely examining traditional initialization methods, we identified a significant disparity between the initial and learned embeddings in terms of both scale and orientation. The scale of the learned embedding can be up to 100 times greater than that of the initial embedding. Such a significant change in the embedding could increase the risk of overfitting, thereby compromising the editability. Driven by this observation, we introduce a novel initialization method, termed Cross Initialization, that significantly narrows the gap between the initial and learned embeddings. This method not only improves both reconstruction and editability but also reduces the optimization steps from 5,000 to 320. Furthermore, we apply a regularization term to keep the learned embedding close to the initial embedding. We show that when combined with Cross Initialization, this regularization term can effectively improve editability. We provide comprehensive empirical evidence to demonstrate the superior performance of our method compared to the baseline methods. Notably, in our experiments, Cross Initialization is the only method that successfully edits an individual's facial expression. Additionally, a fast version of our method allows for capturing an input image in roughly 26 seconds, while surpassing the baseline methods in terms of both reconstruction and editability. Code is available at https://github.com/lyuPang/CrossInitialization.

1. Introduction

Recent advancements in large-scale diffusion models [43, 48, 51] have significantly advanced the field of text-to-image generation, paving the way for a variety of generative tasks [7, 17, 23]. Text-to-image personalization [17], when provided with several images of a target concept, enables users to produce personalized images in novel contexts or styles. This personalization is achieved either by inverting the target concept into the textual embedding

*Corresponding author (xudong.xdmao@gmail.com).

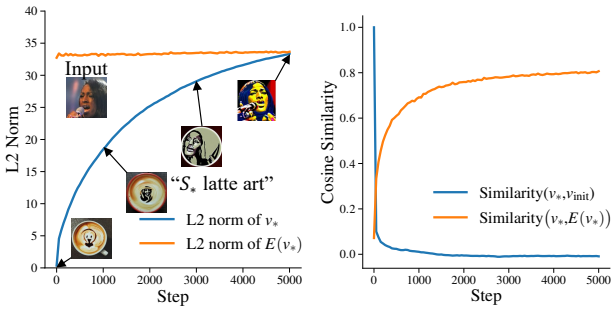

Figure 2. The average of scale (left) and orientation (right) of the textual embedding v_* of 10 examples, as initialized by the traditional method, along with some generated images. The term $E(v_*)$ represents the output vector of the text encoder, and v_{init} represents the initial state of the embedding. After optimization, both the scale and orientation of v_* undergo substantial alterations, aligning more closely with $E(v_*)$.

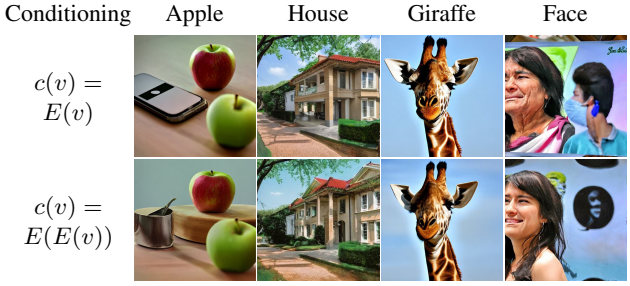

Figure 3. Top row: Images generated using standard textual embeddings as input for the text encoder, for instance, v_{apple}. Bottom row: Images generated using the output of the text encoder as its input, for instance, $E(v_{\text{apple}})$. Here, $c(v)$ denotes the conditioning vector in diffusion models. The images produced by v and $E(v)$ are remarkably similar.

space [2, 17, 64] or by fine-tuning the pretrained diffusion model [29, 49]. Among these, Textual Inversion [17] is one notable method that learns the target concept by inverting given images into textual embeddings.

Face personalization [18, 69, 70] focuses on the personalized generation of a particular individual. An effective face personalization model should be able to synthesize the individual in novel scenes or styles based on text prompts while preserving the individual's unique identity. However, many existing methods are prone to overfitting the whole input image [58], thus struggling to generate images that align with the prompts accurately.

In this work, we investigate the overfitting problem in Textual Inversion [17] through the lens of initialization. Traditional methods typically initialize the textual embedding with a super-category token (e.g., "face" or "person") [2, 17, 64]. However, after optimization, this approach often leads to significant deviations from the initial embedding in both scale and orientation, as depicted in Fig. 2. The significant gap between the initial and learned embeddings necessitates numerous optimization steps, which in turn increases the risk of overfitting.

To address this issue, our approach aims to minimize the disparity between the initial and learned embeddings. Our method is inspired by two main observations. Firstly, after optimization, the learned embedding tends to align with the output of the CLIP [40] text encoder in terms of both scale and orientation, as illustrated in Fig. 2. Secondly, using the text encoder's output as its input typically produces an image nearly identical to the original, as shown in Fig. 3. Drawing from these insights, we introduce *Cross Initialization*, a method where the textual embedding is initialized with the text encoder's output, as depicted in Fig. 4. This approach effectively narrows the gap between the initial and learned embeddings, facilitating more effective optimizations compared to traditional methods. Our results demonstrate that Cross Initialization not only enhances reconstruction quality and editability but also significantly speeds up the personalization process.

To further improve editability, we incorporate a regularization term designed to keep the learned embedding close to its initial state throughout the optimization process. In Textual Inversion, the effectiveness of this regularization is often limited due to the substantial disparity between the initial and learned embeddings. In contrast, when used in conjunction with Cross Initialization, this regularization strategy becomes significantly more effective. This improvement is primarily attributed to the reduced gap between the initial and learned embeddings facilitated by Cross Initialization.

We demonstrate the superior performance of Cross Initialization compared to the baseline methods through both qualitative and quantitative evaluations. Our method enables a variety of novel personalized face generations with high visual fidelity. Notably, in our experiments, Cross Initialization is the only method capable of editing an individual's facial expression. Furthermore, a fast version of our method allows for capturing an input image in roughly 26 seconds, while surpassing the baseline methods in terms of both reconstruction and editability.

2. Related Works

Text-to-Image Synthesis. Text-to-image synthesis is the task of generating realistic and diverse images from natural language descriptions. Various deep generative models have been widely explored for this task, such as GANs [44, 52], VAEs [15, 42], and Autoregressive Models [43, 68]. Recently, diffusion models [24, 48, 57] have demonstrated remarkable capabilities in generating high-fidelity images aligned with textual prompts [7, 36, 43, 48, 51].

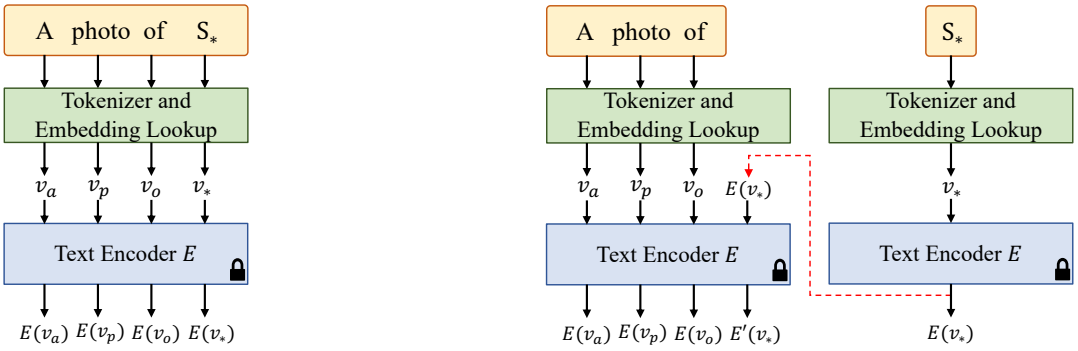

Figure 4. Comparison of Textual Inversion Initialization and Cross Initialization techniques. Textual Inversion [17] (left) initializes the textual embedding v_* with a super-category token (e.g., "face"). Cross Initialization (right) begins by obtaining the output vector from the text encoder $E(v_*)$, which is subsequently used to initialize the embedding. This approach reduces the disparity between the initial and learned embeddings.

Inversion. Image inversion involves reconstructing an image by mapping it into the latent space of a pretrained generator. This process can be accomplished either through direct optimization of the latent code [1, 19, 72] or by employing an encoder network to map the image into a latent space [6, 37, 39, 45, 59, 65, 71]. Image inversion has been applied to various image manipulation tasks [19, 38, 53]. In the context of diffusion models, image inversion aims to identify an initial noise latent code that can be denoised back to the input image [14, 35, 43]. This inverted noise latent code is then leveraged for text-guided image manipulation, as explored in recent studies [12, 23, 28, 30, 60].

Personalization. Personalization adapts pretrained generative models to capture new concepts depicted in several given images. In the realm of text-to-image diffusion models, this allows for the creation of personalized images guided by text prompts. Techniques for this task include optimizing textual embeddings to learn new concepts [2, 11, 16, 17, 62, 64], fine-tuning diffusion models for concept acquisition [4, 10, 11, 21, 22, 29, 49, 50, 56, 58], and training encoders for mapping new concepts to textual representations [3, 9, 18, 26, 33, 55, 70]. These methods facilitate applications like image editing [28, 61] and personalized 3D generation [31, 34, 41, 46]. Particularly, some studies [8, 18, 20, 25, 66, 69, 70] focus on the personalized generation of individual human images. However, existing methods often face the overfitting problem, hindering the creation of text-aligned personalized images. Our work addresses this challenge by examining the overfitting problem through the lens of initialization.

3. Preliminaries

Latent Diffusion Models. We implement our method on the publicly available Stable Diffusion (SD) model, a Latent Diffusion Model (LDM) [48] for text-to-image synthesis. This model is composed of an encoder, $\mathcal{E}$, which maps an image x to a latent code $z = \mathcal{E}(x)$, and a decoder, $\mathcal{D}$, which reconstructs the image from this code $\mathcal{D}(\mathcal{E}(x)) \approx x$. A Denoising Diffusion Probabilistic Model (DDPM) [24] is trained to generate latent codes within the latent space of a pretrained autoencoder. For text-to-image generation, the model is conditioned on a vector $c(y)$ derived from a text prompt y. The training objective of LDM is defined by:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z\sim\mathcal{E}(x),y,\varepsilon\sim\mathcal{N}(0,1),t}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, c(y)\right)\|_2^2\right]. \tag{1}$$

Given the timestep t, the noised latent z_t, and the conditioning vector $c(y)$, the denoising network ε_θ aims to remove the noise that was added to the original latent code z_0.

Text Embeddings. Given a text prompt y, the sentence is first tokenized into several tokens. Each token is then mapped to a textual embedding v_i using a predefined embedding lookup. Subsequently, these textual embeddings are passed through a pretrained CLIP text encoder E, which outputs a series of vectors that constitute the conditioning vector $c(y) = [E(v_1), \ldots, E(v_n)]$. For a textual embedding $v_i \in \mathbb{R}^{1024}$, its corresponding output of the text encoder is denoted by $E(v_i) \in \mathbb{R}^{1024}$. Note that in the SD v2.1 model, the dimensionality of both v_i and $E(v_i)$ is 1024.

Textual Inversion. Textual Inversion [17] is a technique that captures novel concepts from a few example images. It is achieved by injecting new concepts into the pretrained diffusion models. Specifically, Textual Inversion introduces a new token S_* and its corresponding textual embedding v_*, representing the new concept. To learn the new concept, Textual Inversion fixes the LDM and optimizes only v_*, minimizing the objective of LDM given in Eq. (1). The optimization objective is defined by:

$$v_* = \arg\min_v \mathbb{E}_{z,y,\varepsilon,t}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, c(y, v)\right)\|_2^2\right], \tag{2}$$

Figure 5. Images generated by Textual Inversion. This method fails to place the given individual in new styles, primarily due to its tendency to overfit the input image.

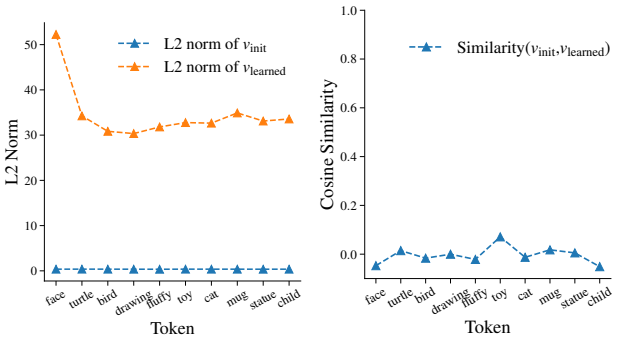

Figure 6. More examples illustrating that, after optimization, the textual embedding v_* experiences significant changes in both scale (left) and orientation (right). Here, v_{init} denotes the embedding's initial state, and v_{learned} denotes the embedding's final state.

where $c(y, v)$ is the conditioning vector obtained from the prompt y and the textual embedding v.

4. Method

Our method is based on the Textual Inversion technique, in which the textual embedding is typically initialized with a super-category token (e.g., "face"). In this section, we analyze how Textual Inversion suffers from a severe overfitting problem through the lens of initialization, as detailed in Sec. 4.1. To address this issue, we propose a novel initialization method, named Cross Initialization, as described in Sec. 4.2. This method facilitates more efficient optimizations, enhancing both reconstruction and editability. To further improve editability, we introduce a regularization term in Sec. 4.3.

4.1. Analysis

In Fig. 5, we show several examples generated by Textual Inversion. This method fails to place the person in new styles and generates images similar to the input image, indicating a severe overfitting problem. In this section, we delve into this overfitting problem in Textual Inversion from the perspective of initialization. Existing methods based on Textual Inversion typically initialize the textual embedding with a super-category token [2, 17, 64]. However, our experiments consistently show that, after optimization, the learned embedding becomes significantly different from its initial state, both in scale and orientation. Figs. 2 and 6 show several examples where the scale of the learned embedding can be up to 100 times greater than that of the initial embedding. Such drastic changes in the embedding may increase the risk of overfitting and degrade the editability of the embedding.

Given that the learned embedding significantly differs from the initial embedding of a coarse descriptor, a question arises: How does the learned embedding manage to produce images that accurately represent the given concept? To investigate this, we examine the outputs of the intermediate layers in the text encoder. The text encoder comprises several self-attention blocks [54], with a LayerNorm layer [5] preceding the input of each sub-block. We observe that the LayerNorm layer normalizes the scale of the embedding, while the self-attention layer modifies its orientation. Fig. 7 illustrates this process: each sub-block progressively alters the scale and orientation of the embedding, and ultimately the output vectors of the initial and learned embeddings exhibit a similarity in both scale and orientation.

To mitigate the overfitting issue in Textual Inversion, this analysis motivates us to seek an initial embedding that can be close to the learned embedding.

4.2. Cross Initialization

Based on the analysis in Sec. 4.1, our goal is to design an initial embedding that meets two criteria: 1) it is close to the learned embedding, and 2) it roughly captures the target concept. Our method is inspired by two key observations. First, as shown in Fig. 2, the learned embedding becomes similar to the output of the text encoder after optimization. Second, when we use the text encoder's output as its input, the diffusion model produces an image nearly identical to the original, as shown in Fig. 3. The reason for these two phenomena is that the LayerNorm and self-attention layers in the text encoder gradually alter the scale and orientation of the embedding, making it converge to a specific vector, as discussed in Sec. 4.1. Based on these insights, we propose initializing the textual embedding with the output of the text encoder, a method we term Cross Initialization, as depicted in Fig. 4.

Formally, given a single face image, we first set the textual embedding to the mean of 691 well-known names' embeddings, denoted as $\bar{v}_{691}$. The computation of $\bar{v}_{691}$ is elaborated in the following subsection. Subsequently, we feed $\bar{v}_{691}$ into the text encoder E, obtaining the output vector $E(\bar{v}_{691})$. We then initialize the textual embedding v_{init} with

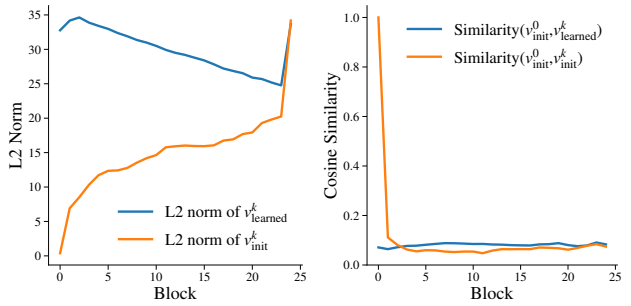

Figure 7. Scale (left) and orientation (right) of the textual embedding processed by the k-th self-attention block of the text encoder. The symbols v^k_{init} and v^k_{learned} denote the outputs of the k-th self-attention block using the initial and learned embeddings as inputs, respectively. Note that an additional LayerNorm layer is present after the final block. In each block, the LayerNorm layer and the self-attention layer gradually modify the scale and orientation of the embedding. After optimization, the output vectors derived from the initial and learned embeddings exhibit a similarity in both scale and orientation.

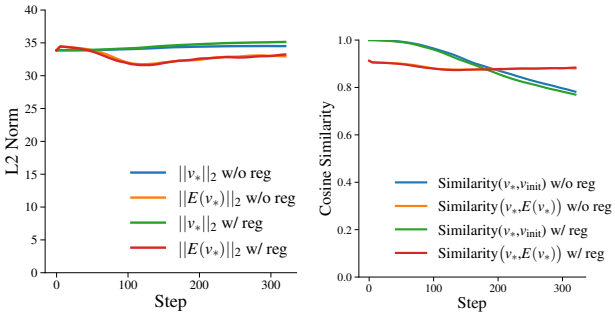

Figure 8. Scale (left) and orientation (right) of the textual embedding v_*, as initialized by Cross Initialization. Here, $E(v_*)$ represents the output vector of the text encoder, and v_{init} represents the initial state of the embedding. In contrast to the examples in Fig. 2, Cross Initialization maintains the learned embedding close to the initial state in terms of both scale and orientation.

this output vector:

$$v_{\text{init}} = E(\bar{v}_{691}). \tag{3}$$

Finally, we optimize the textual embedding by minimizing the LDM loss given in Eq. (2).

The aforementioned two observations ensure that the initial embedding $E(\bar{v}_{691})$ is close to the learned embedding, while also roughly representing the target concept. As shown in Fig. 8, using Cross Initialization, the learned embedding retains proximity to its initial state throughout the optimization process. This facilitates more efficient optimizations, leading to more identity-preserved, prompt-aligned, and faster face personalization.

Mean Textual Embedding. We follow [69] to construct the mean textual embedding $\bar{v}_{691}$. A total of 691 well-known names are used to form an embedding set $C = \{v_1, \ldots, v_m\}$, where $m = 691$ and each textual embedding v_i is obtained from the pre-defined embedding lookup. The mean textual embedding is calculated as $\bar{v}_{691} = \frac{1}{m}\sum_{i=1}^{m} v_i$. Moreover, we represent each name with two tokens (i.e., the first and last names), resulting in the final mean textual embedding as $\bar{v}_{691} = [\bar{v}^f_{691}, \bar{v}^l_{691}]$, where $\bar{v}^f_{691}$ and $\bar{v}^l_{691}$ are calculated using the embedding sets of the first and last names, respectively.

Comparison with Directly Optimizing $E(v)$. In Cross Initialization, we set the text encoder's output as its input, i.e. $v_{\text{init}} = E(\bar{v})$, and optimize the input vector v_{init}. An alternative method is to directly optimize the output vector $E(\bar{v})$. However, this approach eliminates the interaction between the new concept and other prompt tokens, as the new concept is not passed through the text encoder along with the other prompt tokens, leading to poor editability. This issue is also indicated in [2]. In contrast, Cross Initialization optimizes the input vector, thereby preserving the ability to create new compositions for the new concept.

4.3. Regularization

As illustrated in Sec. 4.2, the initial embedding is constructed using the mean center of embeddings from 691 well-known names. We assume that the region around this central embedding represents the subspace corresponding to the concept of the individual. High editability is expected when the learned embedding lies close to this subspace. Therefore, we introduce a regularization term to keep the learned embedding close to the central embedding throughout the optimization process. Specifically, we minimize the L2 distance between them, defined as:

$$\mathcal{L}_{\text{reg}} = ||v - v_{\text{init}}||_2^2. \tag{4}$$

Overall, our final optimization objective is defined as:

$$v_* = \arg\min_{v} \mathcal{L}_{\text{diffusion}} + \lambda \mathcal{L}_{\text{reg}}. \tag{5}$$

Note that this regularization approach, also investigated in [17], faces challenges when applied in Textual Inversion. This is primarily due to the significant disparity between the initial and learned embeddings, as well as the coarseness of the super-category token. These factors limit the effectiveness of this regularization approach.

5. Experiments

In this section, we first present the implementation details of our method. Subsequently, we demonstrate its effectiveness by conducting a comparative analysis with four state-of-the-art personalization methods, focusing on aspects such as identity preservation, editability, and optimization time.

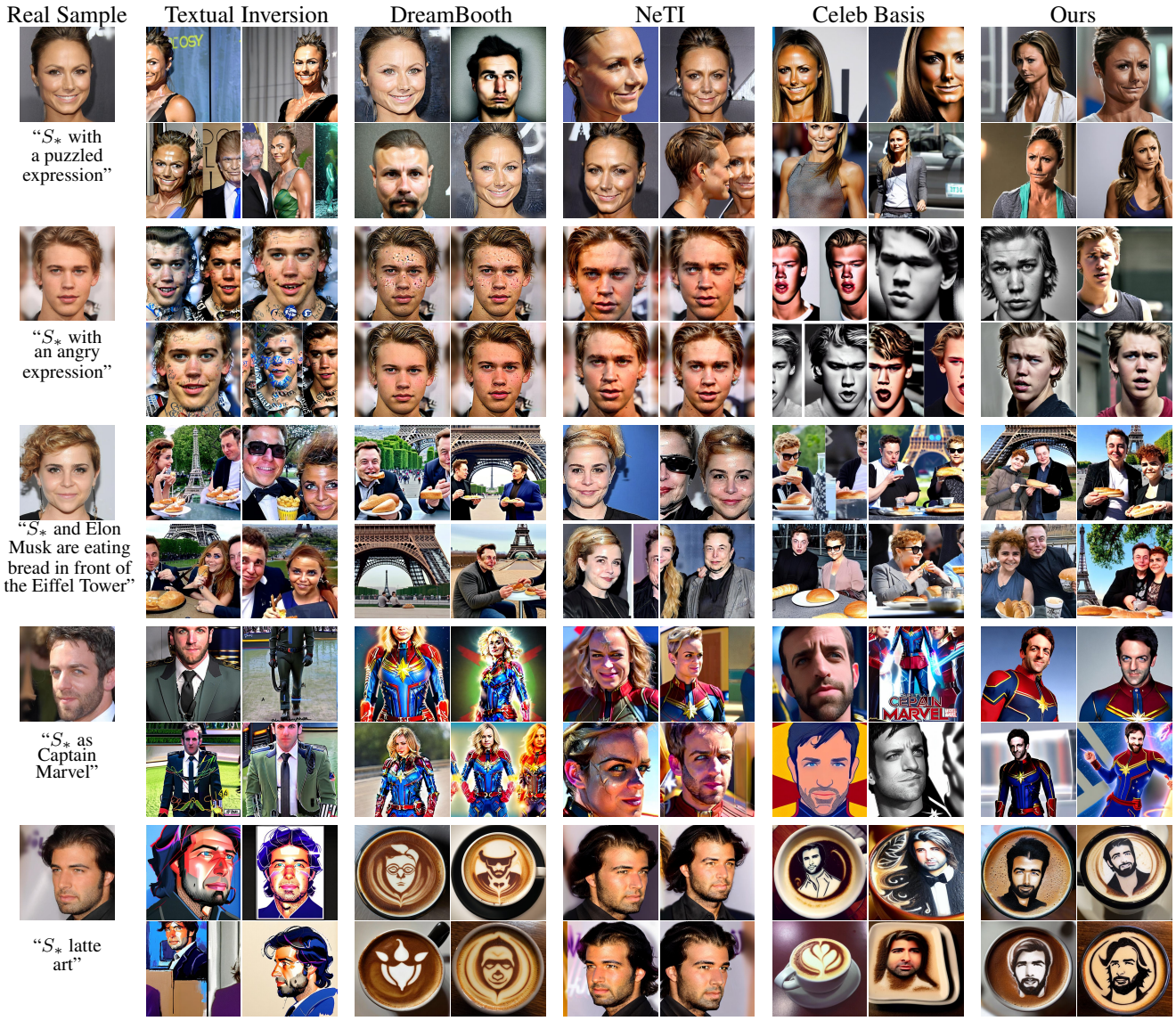

Figure 9. Qualitative comparisons. Given a single input image, we present four images generated by each method using identical random seeds. Our approach demonstrates superior performance in identity preservation and editability. Notably, Cross Initialization is the only method that successfully edits an individual's facial expression.

5.1. Implementation and Evaluation Setup

Implementation. We utilize the publicly available Stable Diffusion v2.1 [48] as our base model. Images are generated at a resolution of 512×512. The hyper-parameter λ is set to 10^{-5} for all experiments. Given a single image as input, our experiments are conducted on a single A800 GPU, using a batch size of 8 and a learning rate of 0.005. All results are obtained using 320 optimization steps.

Evaluation Setup. We evaluate each method using the images from CelebA-HQ test set [27, 32]. The prompts used are primarily sourced from [69] and [18]. We compare our method with four state-of-the-art personalization methods: Textual Inversion [17], DreamBooth [49], NeTI [2], and Celeb Basis [69]. The implementation details of baselines are presented in Appendix A. All methods are implemented for one-shot personalization. For quantitative evaluation, each method is evaluated on the first 200 images from CelebA-HQ test set using two metrics, including identity similarity and prompt similarity. For identity similarity, ArcFace [13], a pretrained face recognition model, is used to measure the identity preservation in generated images. Prompt similarity is measured by computing the CLIP score between generated images and text prompts. We exclude the prompts for stylization in the identity similarity assessment, as ArcFace is trained on real images.

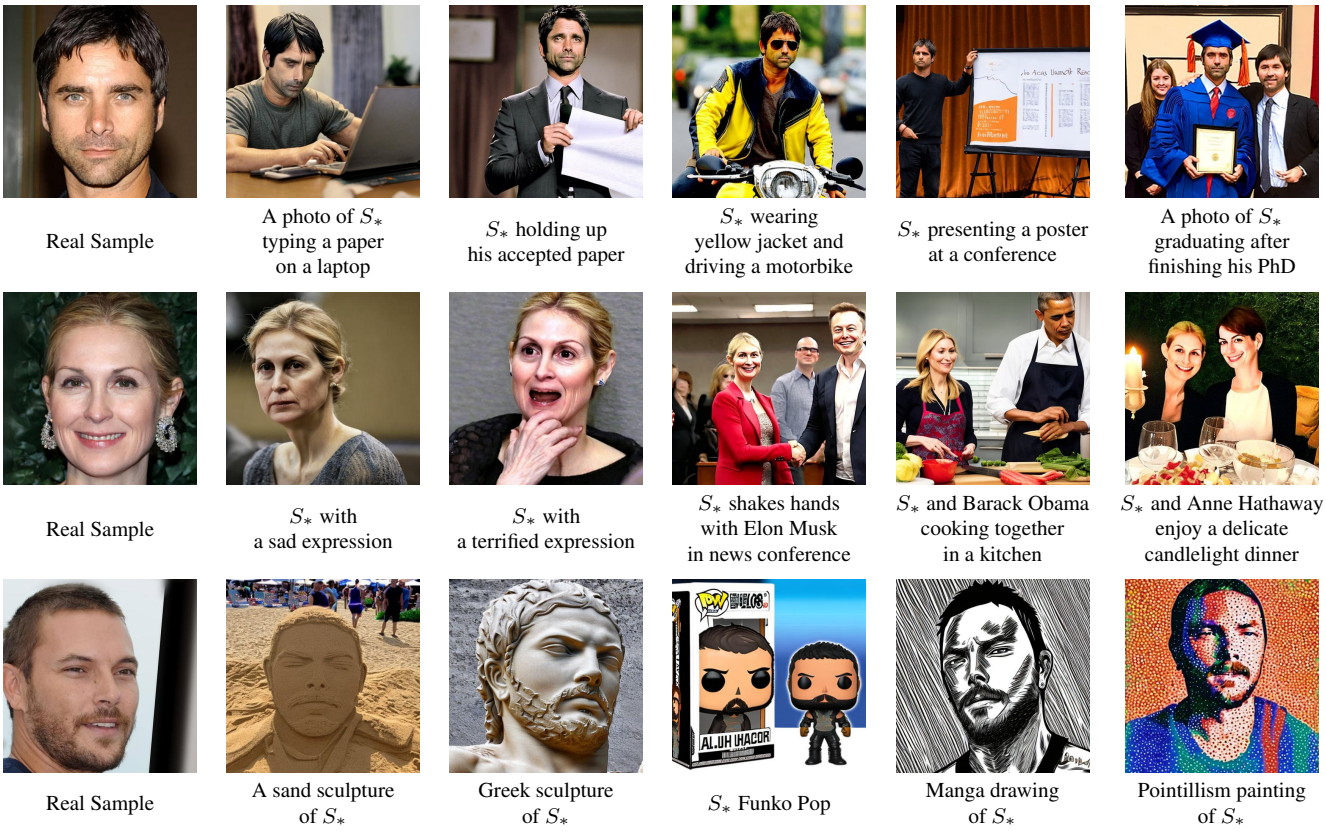

Figure 10. Examples of personalized text-to-image generation obtained with Cross Initialization.

5.2. Results

Qualitative Evaluation. In Fig. 9, we present a visual comparison of personalized generation using four types of prompts: expression editing, background modification, individual interaction, and artistic style. Textual Inversion exhibits an overfitting problem, failing to compose the given individual in novel scenes. DreamBooth struggles to reconstruct the individual for complex editing prompts such as background modification and artistic style. It tends to disregard the new concept and generate images based solely on the remaining prompt tokens. In contrast, NeTI generates images based solely on the new concept without incorporating the other prompt tokens, indicating a severe overfitting problem. Both Celeb Basis and our method are capable of generating novel compositions of personalized concepts. Compared to Celeb Basis, our method shows superior identity preservation and excels in editing the individual's expression. For all prompts, Cross Initialization achieves high-fidelity reconstruction of the individual's identity while providing superior editability. Notably, it is the only method that successfully edits an individual's facial expression. Fig. 10 shows more results with different prompts from our method. Additional qualitative results can be found in Appendices D and F. We also provide results on synthetic facial images in Appendix G.

Quantitative Evaluation. We quantitatively evaluate our approach in two aspects: 1) identity similarity between the generated and input images, and 2) prompt similarity between the generated image and the given text prompt. All methods are evaluated over 20 text prompts, see Appendix B for a full list. These prompts cover expression editing (e.g., "S_* with a sad expression"), background modification (e.g., "S_* on the beach"), individual interaction (e.g., "S_* shakes hands with Anne Hathaway in news conference"), and artistic style (e.g., "S_* latte art"). For each prompt, we generate 32 images using the same random seed for all methods.

The results are shown in Tab. 1. DreamBooth excels in prompt similarity but ranks lowest in identity similarity. This is consistent with the qualitative observations, where DreamBooth often overlooks the new concept, focusing solely on the other prompt tokens. In contrast, NeTI achieves the highest identity similarity scores but ranks lowest in prompt similarity, as NeTI tends to overfit the input image. Besides these two extreme cases, our method demonstrates superior performance in both identity and prompt similarity metrics.

Table 1. Quantitative comparisons. "Identity" denotes the identity similarity between the generated and input images. "Prompt" denotes the prompt similarity between the generated image and the given text prompt. "Time" denotes the average personalization time in seconds.

Methods	Identity↑	Prompt↑	Time↓
Textual Inversion [17]	0.2115	0.2498	6331
DreamBooth [41]	0.2053	**0.3015**	623
NeTI [2]	**0.3789**	0.2325	1527
Celeb Basis [69]	0.2070	0.2683	<u>140</u>
Ours-fast	0.2225	0.2800	**26**
Ours	<u>0.2517</u>	<u>0.2859</u>	346

Table 2. User study results. We asked the participants to select the image that better preserves the identity and matches the prompt.

Baselines	Prefer Baseline	Prefer Ours
Textual Inversion [17]	22.0%	**78.0%**
DreamBooth [41]	9.3%	**90.7%**
NeTI [2]	24.7%	**75.3%**
Celeb Basis [69]	26.7%	**73.3%**

Personalization Time. The average time for personalization using each method is reported in Tab. 1. Compared to Textual Inversion, our method significantly reduces the optimization time from 106 minutes to 6 minutes. Additionally, We develop a fast version of our method, denoted as "Ours-fast", with a learning rate of 0.08. This fast version allows for learning the new concept in merely 25 optimization steps, taking only 26 seconds. As demonstrated in Tab. 1, this fast version achieves the quickest personalization while surpassing Celeb Basis and Textual Inversion in both identity similarity and prompt similarity. The visual results of this fast version are presented in Appendix E.

User Study. We also evaluate our method from a human perspective by conducting a user study. We randomly selected one prompt from the prompt set and one image from the CelebA-HQ test set. These were used to generate personalized images for each method. In each question of the study, participants were presented with the input image and text prompt, as well as two generated images: one from our method and another from the baseline method. Participants were asked to select the image that better preserves the identity and matches the prompt. In total, we collected 600 responses from 30 participants, as shown in Tab. 2. The results show a clear preference for our method.

5.3. Ablation Study

We conduct an ablation study by separately removing each sub-module from our method. Specifically, we sequentially remove the following sub-modules: 1) Cross Initialization, 2) mean textual embedding, and 3) the regularization term. In Fig. 11, we present a visual comparison of the personalized images generated by each variant. The results indicate that all sub-modules are crucial for achieving identity-preserved and prompt-aligned personalized face generation. Specifically, the model without Cross Initialization produces results similar to those by Textual Inversion. This variant tends to generate images focusing either solely on the given concept or exclusively on the other prompt tokens. The models without mean textual embedding or the regularization term lead to degradation in editability, struggling to create consistent scenes as described in the prompt. More ablation study results are provided in Appendix H.

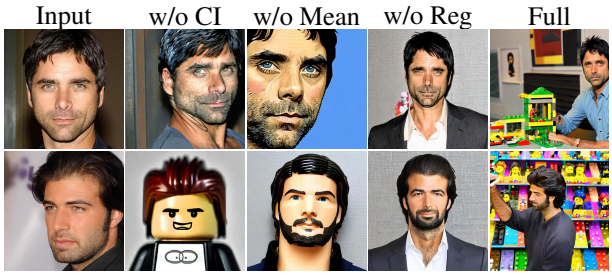

Figure 11. Ablation study. The prompt is "S_* plays the LEGO toys". We compare the models trained without Cross Initialization (w/o CI), without mean textual embedding (w/o Mean), and without regularization (w/o Reg). As can be seen, all sub-modules are essential for achieving identity-preserved and prompt-aligned personalized face generation.

6. Conclusions and Future Work

We introduced a new initialization method for personalized text-to-image generation. We identified a significant disparity between the initial and learned embeddings in Textual Inversion, which often leads to an overfitting problem. Our approach, "Cross Initilization", addresses this issue by initializing the textual embedding with the output of the text encoder. Cross Initialization enables more identity-preserved, prompt-aligned, and faster face personalization. In this work, we mainly examined the performance of Cross Initialization on the human being concept. For general concepts, we found that Cross Initialization is not as effective as it is for the human being concept. In future work, we plan to further investigate the applicability of Cross Initialization to a broader range of concepts.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62176223, No. 62302535, and No. 72201100) and Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012897).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, pages 4432–4441, 2019. 3

[2] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *arXiv preprint arXiv:2305.15391*, 2023. 2, 3, 4, 5, 6, 8, 12

[3] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06925*, 2023. 3

[4] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 3

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[6] Qingyan Bai, Yinghao Xu, Jiapeng Zhu, Weihao Xia, Yujiu Yang, and Yujun Shen. High-fidelity gan inversion with padding space. In *ECCV*, pages 36–53. Springer, 2022. 3

[7] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 2

[8] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, and Min Zheng. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023. 3, 12, 15

[9] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 3

[10] Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*, 2023. 3

[11] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *ECCV*, pages 558–577. Springer, 2022. 3

[12] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 6

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3

[15] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, pages 19822–19835, 2021. 2

[16] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 3

[17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2, 3, 4, 5, 6, 8, 12

[18] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *TOG*, 42(4):1–13, 2023. 2, 3, 6, 12, 15

[19] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, pages 3012–3021, 2020. 3

[20] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023. 3

[21] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K. Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 3

[22] Xingzhe He, Zhiwen Cao, Nicholas Kolkin, Lantao Yu, Helge Rhodin, and Ratheesh Kalarot. A data perspective on enhanced identity preservation for diffusion personalization. *arXiv preprint arXiv:2311.04315*, 2023. 3

[23] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 3

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 3

[25] Junha Hyung, Jaeyo Shin, and Jaegul Choo. Magicapture: High-resolution multi-concept portrait customization. *arXiv preprint arXiv:2309.06895*, 2023. 3

[26] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 3

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2018. 6

[28] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 3

[29] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customiza-

tion of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 2, 3

[30] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022. 3

[31] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, pages 300–309, 2023. 3

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 6

[33] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 3

[34] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *CVPR*, pages 12663–12673, 2023. 3

[35] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 3

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[37] Gaurav Parmar, Yijun Li, Jingwan Lu, Richard Zhang, Jun-Yan Zhu, and Krishna Kumar Singh. Spatially-adaptive multilayer selection for gan inversion and editing. In *CVPR*, pages 11399–11409, 2022. 3

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094, 2021. 3

[39] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *CVPR*, pages 14104–14113, 2020. 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[41] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 3, 8

[42] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021. 2

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2, 3

[44] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069, 2016. 2

[45] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 3

[46] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 3

[47] Oren Rippel, Michael Gelbart, and Ryan Adams. Learning ordered representations with nested dropout. In *ICML*, pages 1746–1754, 2014. 12

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 6

[49] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2, 3, 6, 12

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 3

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 1, 2

[52] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 2

[53] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *ICML*, pages 9489–9502, 2021. 3

[54] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 4

[55] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 3

[56] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023. 3

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[58] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *SIGGRAPH*, 2023. 2, 3

[59] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *TOG*, 40(4):1–14, 2021. 3

[60] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 3

[61] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 3

[62] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *arXiv preprint arXiv:2305.18203*, 2023. 3

[63] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 12

[64] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2, 3, 4

[65] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *CVPR*, pages 11379–11388, 2022. 3

[66] Zijie Wu, Chaohui Yu, Zhen Zhu, Fan Wang, and Xiang Bai. Singleinsert: Inserting new concepts from a single image into text-to-image models for flexible editing. *arXiv preprint arXiv:2310.08094*, 2023. 3

[67] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 12, 15

[68] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[69] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. In *NeurIPS*, 2023. 2, 3, 5, 6, 8, 12, 20

[70] Yufan Zhou, Ruiyi Zhang, Tong Sun, and Jinhui Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*, 2023. 2, 3, 12, 15

[71] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, pages 592–608, 2020. 3

[72] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 3