When StyleGAN Meets Stable Diffusion: a $\mathcal{W}_+$ Adapter for Personalized Image Generation

Xiaoming Li Xinyu Hou Chen Change Loy
S-Lab, Nanyang Technological University
csxmli@gmail.com xinyu.hou@ntu.edu.sg ccloy@ntu.edu.sg

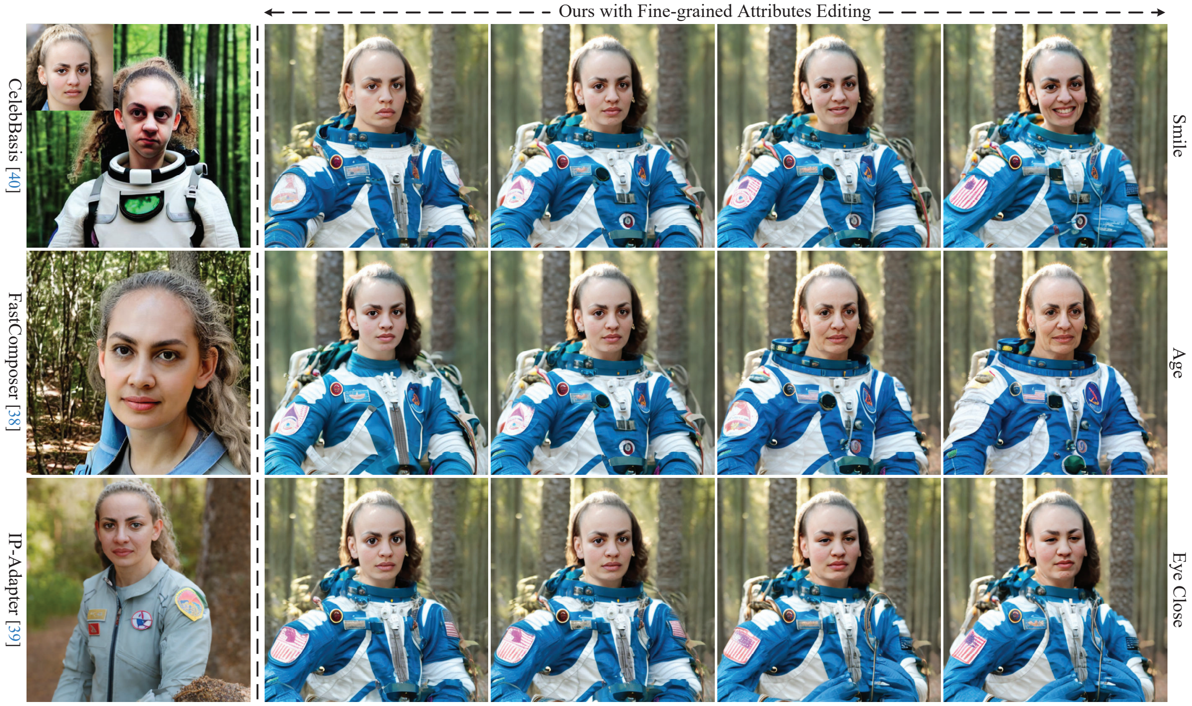


Figure 1. Given a single reference image (thumbnail in the top left), our $\mathcal{W}_+$ adapter not only integrates the identity into the text-to-image generation accurately but also enables modifications of facial attributes along the Δw trajectory derived from StyleGAN. The text prompt is "a woman wearing a spacesuit in a forest".

Abstract

Text-to-image diffusion models have remarkably excelled in producing diverse, high-quality, and photo-realistic images. This advancement has spurred a growing interest in incorporating specific identities into generated content. Most current methods employ an inversion approach to embed a target visual concept into the text embedding space using a single reference image. However, the newly synthesized faces either closely resemble the reference image in terms of facial attributes, such as expression, or exhibit a reduced capacity for identity preservation. Text descriptions intended to guide the facial attributes of the synthesized face may fall short, owing to the intricate entanglement of identity information with identity-irrelevant facial attributes derived from the reference image. To address these issues, we present the novel use of the extended StyleGAN embedding space $\mathcal{W}_+$, to achieve enhanced identity preservation and disentanglement for diffusion models. By aligning this semantically meaningful human face latent space with text-to-image diffusion models, we succeed in maintaining high fidelity in identity preservation, coupled with the capacity for semantic editing. Additionally, we propose new training objectives to balance the influences of both prompt and identity conditions, ensuring that the identity-irrelevant background remains negligibly affected during facial attribute modifications. Extensive experiments reveal that our method adeptly generates personalized text-to-image outputs that are not only compatible with prompt descriptions but also amenable to common StyleGAN editing directions in diverse settings. Our code and model are available at https://github.com/csxmli2016/w-plus-adapter.

1. Introduction

What if we could be the protagonists of our own fantasies? This paper primarily addresses personalized text-to-image (T2I) generation, a field that is attracting growing attention due to users' desire to craft their unique content. Overlapped with customized T2I [8, 17, 28, 36], which integrates various visual concepts like human faces and objects into the generated imagery, our focus is on tailoring image generation to specific identities. This focus is motivated by applications such as storyboarding, where a consistent identity needs to be maintained across all images, despite variations in expressions or ages. Additionally, we believe that human faces possess unique, fine-grained intrinsic features that merit special attention and present a worthwhile area for exploration.

Recent personalized image generation approaches utilize a small reference set of a target identity and embed it into a specific space. The common choice of embedding space has been the textual embedding space used by large language models (LLMs). While existing methods based on textual embedding [8, 28, 36] have proven capable of maintaining target identity, they are often limited by inherent trade-offs. Specifically, these methods face challenges in simultaneously preserving identity, generating varied facial attributes, and creating identity-irrelevant content that aligns with the text description. We observe that these issues predominantly stem from the entangled nature of the textual embedding space, where a single pseudo word $\mathcal{S}^*$ struggles to distinctly isolate identity-related features from the reference image. Efforts to separate such information include approaches like encoding a visual concept into multiple word embeddings [36] or employing a separate branch for extracting identity-irrelevant details [3]. However, these methods tend to overlook the nuanced facial features critical to an individual's identity, resulting in incomplete disentanglement and sub-optimal identity preservation.

In this study, we aim to more effectively separate identity-relevant and -agnostic features for better identity preservation, while also ensuring editability. To achieve this, we propose inverting the visual concept of a target identity into StyleGAN's [15] $\mathcal{W}_+$ latent space, as opposed to using the textual embedding space. In particular, we introduce a mapping network to integrate the $\mathcal{W}_+$ space with the diffusion model and a residual cross-attention module to add the w_+ vector as an additional identity condition. This mapping network, once trained with *(image, w_+)* pairs, can generalize to unseen individuals during inference without the need for a separate identity-specific model for each person. We further present novel regularized training to ensure that edits have a negligible effect on identity-irrelevant regions and that the overall generation remains aligned with the prompt conditions. It is noteworthy that although CelebBasis [40] explores a similar concept of creating a face-specific latent space through PCA of selected celebrity name embeddings, we argue that relying on such a limited dataset of celebrity names is insufficient for a comprehensive face latent space due to potential under-representation. Our experimental results confirm that our approach exhibits a superior capacity for identity preservation compared to CelebBasis.

The contributions are summarised as follows: 1) We introduce extended StyleGAN $\mathcal{W}_+$ space as a target inversion latent space to better encode the identity-preserving facial concepts for personalized text-to-image synthesis. This is the first study that considers the fusion between StyleGAN $\mathcal{W}_+$ space and diffusion-based image generation. 2) Embedding target identity in $\mathcal{W}_+$ space enables smooth and controllable semantic editing on facial attributes in our text-to-image model. 3) The effective disentanglement of identity-relevant and -irrelevant information in our model facilitates an identity-preserving generation that is not only diverse but also adaptable to a wide range of prompt instructions.

2. Related Work

StyleGAN Latent Space. GANs have drawn enormous attention due to the well-disentangled latent space [20]. StyleGAN [15] proposes to map input latent code to an intermediate latent space $\mathcal{W}$ to prevent warping of the training data distribution to fit in a particular probability distribution. The resulting $\mathcal{W}$ space is a semantically disentangled space that allows fine-grained controls over image synthesis. InterFaceGAN [29] and GANSpace [10] further interpret the latent space and identify several control directions such as age, gender, face angle, smile, etc. The powerful semantic editing ability has motivated research on inverting and editing real images in the StyleGAN space [1, 19, 24, 32]. Abdal *et al.* [1] utilize a direct optimization framework to embed a given real image to the extended StyleGAN space. Tov *et al.* [32] deploy an encoder to perform the inversion and keep the concatenated latent codes close to the original StyleGAN space to maintain high perceptual quality and editability. Motivated by the disentanglement and editability of StyleGAN latent space, we aim to introduce it to the T2I diffusion models to achieve more controllable synthesis. We follow Image2StyleGAN to denote the extended $\mathcal{W}$ space represented by multiple nonidentical w vectors as $\mathcal{W}_+$.

Personalized Image Synthesis. Customization has been extensively studied in the context of T2I to generate images of specified objects or individuals [2, 8, 28]. Gal *et al.* [8] directly apply learnable text embedding optimization. Ruiz *et al.* [28] fine-tune the diffusion model for the target concept while preserving its class prior of the concept. Further improvements have been made to invert multiple concepts simultaneously with cross-attention fine-tuning [17], and disentangle background irrelevant information [3]. Encoder-based approaches [4, 9, 13, 30, 36–38] are employed for their efficiency. The target visual concept is encoded to embedding space as additional conditions. SingleInsert [37] adopts a two-stage scheme to insert concepts from a single image into the foreground region exclusively and enables the editing of the concept by text prompts. However, it is worth noticing that our method achieves more fine-grained editing with a single reference image at inference and does not necessitate identity-specific fine-tuned models.

With special attention on identity inversion utilizing the distinct features of human faces, Yuan *et al.* [40] define a celebrity space by applying PCA on selected celebrity name embeddings from CLIP and embedding new identities into the space via learnable coefficients. Valevski *et al.* [33] utilize a pre-trained face recognizer as the encoder and further project the face embedding to CLIP space. FaceChain [21] trains separate face and style LoRAs [12] to synthesize specific faces in specific styles. They follow a data preprocessing pipeline and apply face fusion of the best reference face on the synthesized results. However, we observe that previous personalization approaches inherit a common limitation that the synthesized faces share similar attributes as the reference, indicating that the model cannot disentangle irrelevant information well and overfits the particular reference.

Diffusion Model Adapters. Light-weight adapters have been adopted to avoid the laborious work of fine-tuning large models and add additional controls [23, 39, 42] to the diffusion model. ControlNet [42] proposes to train task-specific adapters while freezing the original diffusion model to adapt to various input conditions. Concurrent work T2I-Adapter [23] constructs a lighter-weight adapter for controls on structure and color. Ye *et al.* [39] employ a decoupled cross-attention module to consider both text and image prompts in the denoising process. In our work, we consider identity information as an additional condition and align it with the T2I model for identity-preserving synthesis.

3. Methodology

Our approach is capable of generating images that preserve identity while allowing for semantic edits, requiring just a single reference image for inference. This capability is realized by innovatively aligning StyleGAN's $\mathcal{W}_+$ latent space with the diffusion model. The training of our $\mathcal{W}_+$ adapter is divided into two stages. In Stage I (Sec. 3.2), we establish a mapping from $\mathcal{W}_+$ to SD latent space, using the resulting projection as an additional identity condition to synthesize center-aligned facial images of a specified identity. In Stage II (Sec. 3.3), this personalized generation process is expanded to accommodate more dynamic, “in-the-wild” settings, ensuring adaptability to a variety of textual prompts.

3.1. Preliminary

In our implementation, we use Stable Diffusion [26] as the foundational T2I model, which is one of the most commonly used latent space-based text-to-image generation model [6, 34]. During the training phase, the model architecture includes: 1) a pre-trained encoder $\mathcal{E}(\cdot)$, which transforms an image I into a latent representation with reduced dimensionality, 2) a conditional diffusion model tasked with predicting the latent code of a previous timestep, based on the text condition c_{txt} and the current timestep's latent code, and 3) a pre-trained decoder $\mathcal{D}(\cdot)$ that reconstructs the final latent code into the synthesized image. The learning objective of this model is formulated as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,1),t,c_{txt}}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_{txt}(c_{txt})\right)\|_2^2\right] \quad (1)$$

where z_0 is $\mathcal{E}(I)$, ϵ is the ground truth noise added for current timestep t, z_t is the latent code of timestep t. The τ_{txt} denotes the pre-trained CLIP text encoder [25] that converts the text prompt c_{txt} to textual embeddings. The ϵ_θ represents the UNet [27] denosing network. In the inference stage, the denoising network is applied iteratively to denoise a random sampled noise z_T to z_0 with condition c_{txt}. The final result is then generated through the pre-trained decoder, *i.e.*, $\mathcal{D}(z_0)$.

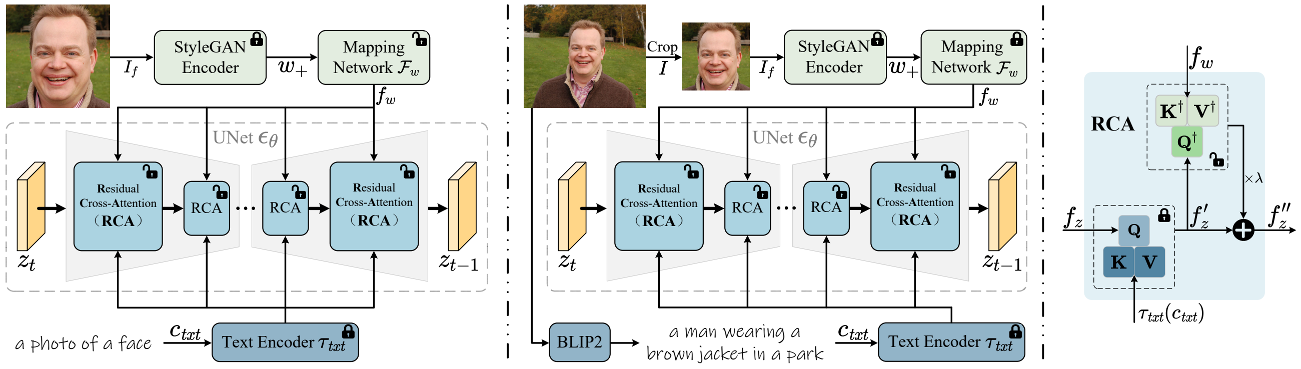
3.2. Stage I: Aligning $\mathcal{W}_+$ with Stable Diffusion

The goal of Stage I is to align $\mathcal{W}_+$ space with SD to carry identity information in T2I generation while retaining its editing ability. To do so, we train a mapping network to project a w_+ embedding to SD latent space and inject this extra condition into SD by modifying the cross-attention module. The framework of Stage I is illustrated in Fig. 2. Details are introduced below.

Training Pair Construction. We adopt 100K discrete training samples to fit the continuous distribution of $\mathcal{W}_+$ and align it with SD. For each face image I_f, we use the pre-trained e4e [32] encoder from StyleGAN inversion task to get its corresponding $w_+ \in \mathbb{R}^{18\times512}$ vector. The pairs of $\{I_f, w_+\}$ constitute our training data in this stage. Specifically, to generalize on real-world face images and improve w_+ diversity, two types of training pairs are built, *i.e.*, synthetic face images from StyleGAN2 [16] and real-world face images from FFHQ [15]. Note that in Stage I, we only consider the aligned face images in order to exclude any extraneous influences.

Mapping Network. Since the original $\mathcal{W}_+$ space is designated for StyleGAN generation, we project the vector $w_+ \in \mathbb{R}^{18\times512}$ to four tokens of dimension 768 ($\mathbb{R}^{4\times768}$) to align with the input dimension of SD condition. The mapping network consists of four groups of linear layers and is denoted as $\mathcal{F}_w(\cdot)$. The latent space after mapping inherits the editability and disentanglement properties of $\mathcal{W}_+$ space and is compatible with the SD generation process.

Residual Cross-Attention. To incorporate the identity condition from the projected w_+ embedding into the pre-trained SD model, we introduce a residual cross-attention module. In the standard SD framework, the output of cross-attention is determined using query features f_z from the hidden state



(a) Stage I: Aligning w_+ with Stable Diffusion (b) Stage II: Generation in the Wild (c) Residual Cross-Attention

Figure 2. Overview of $\mathcal{W}_+$ adapter training stages. *Left*: Stage I for aligning $\mathcal{W}_+$ space with Stable Diffusion. *Middle*: Stage II for generating in-the-wild images with w_+ embeddings. *Right*: details of residual cross-attention module. Open lock indicates trainable parts.

and the text condition $\tau_{txt}(c_{txt})$. The process is defined as

$$f'_z = \text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \tag{2}$$

where $\mathbf{Q} = f_z\mathbf{W}_q$, $\mathbf{K} = \tau_{txt}(c_{txt})\mathbf{W}_k$, $\mathbf{V} = \tau_{txt}(c_{txt})\mathbf{W}_v$ are the query, key, value matrices of the attention module, respectively. $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ denote the corresponding projection matrices. The cross-attention module is where the latent noise interacts with text condition embeddings inside the denoising process. Following previous methods [36, 39], we add an additional condition by a separate cross-attention module. Our residual cross-attention is defined as:

$$f''_z = f'_z + \lambda \cdot \text{Attention}(\mathbf{Q}^\dagger,\mathbf{K}^\dagger,\mathbf{V}^\dagger) \tag{3}$$

where $\mathbf{Q}^\dagger = f'_z\mathbf{W}^\dagger_q$, $\mathbf{K}^\dagger = f_w\mathbf{W}^\dagger_k$, and $\mathbf{V}^\dagger = f_w\mathbf{W}^\dagger_v$. The $\mathbf{W}^\dagger_q$, $\mathbf{W}^\dagger_k$, and $\mathbf{W}^\dagger_v$ are the projection matrices for query, key, and value, respectively. The $f_w = \mathcal{F}_w(w_+)$ is mapped from w_+ by the mapping network. We use λ as a scale parameter to balance the influence of text and identity conditions on the generation. We incorporate the decoupled cross-attention module in a residual fashion instead of the parallel approach as in previous works [36, 39] to avoid performance degradation on the original text condition. When λ is set to 0, our w_+ vector has no impact on the pre-trained SD model. We set λ to 1 during training. The $\mathbf{Q}^\dagger$, $\mathbf{K}^\dagger$ and $\mathbf{V}^\dagger$ are initialized from their corresponding $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, respectively. The residual cross-attention module is performed on all the cross-attention layers of SD.

Learning Objectives. The trainable parts in this stage include the mapping network $\mathcal{F}_w$ and $\{\mathbf{Q}^\dagger,\mathbf{K}^\dagger,\mathbf{V}^\dagger\}$ matrices in all residual cross-attention modules. The optimization target is formulated as:

$$\mathcal{L}_{w_+} = \mathbb{E}_{z_0,\epsilon,t,c_{txt},w_+}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_{txt}(c_{txt}), \mathcal{F}_w(w_+)\right)\|_2^2\right] \tag{4}$$

Eqn. (4) is similar to Eqn. (1) except for the additional $\mathcal{F}_w(w_+)$ serving as identity condition and cross-attention structure modified inside ϵ_θ. The text prompt c_{txt} is randomly selected from several neutral templates describing a human face, *e.g.*, “a photo of a face”, or “a face” (see suppl.).

Through mapping the $\mathcal{W}_+$ distribution into an embedding space compatible with SD, our approach effectively uses the w_+ vector to condition the personalized generation of aligned facial images. Furthermore, attribute directions Δw derived from the original $\mathcal{W}_+$ space, encompassing features like smile, age, eye-opening, and others, remain applicable to our projected w_+ vector. This allows for personalized, fine-grained face editing, as demonstrated in Fig. 1.

3.3. Stage II: Generation in the Wild

In order to refine the performance of our $\mathcal{W}_+$-adapted SD model from Stage I for scenarios beyond controlled environments, we continue on a second stage of training. This phase is dedicated to further fine-tuning the weights of $\mathbf{Q}^\dagger,\mathbf{K}^\dagger,\mathbf{V}^\dagger$ across all residual cross-attention modules. To ensure that the projected $\mathcal{W}_+$ space exclusively encapsulates identity-related facial attributes, while remaining uninfluenced by irrelevant distractions from the background, we keep the mapping network fixed during this stage. The architecture and workflow of Stage II are depicted in the middle of Fig. 2.

Training Data Construction. For Stage II, in-the-wild images are used for training. For each image I, its corresponding text caption, c_{txt}, is extracted using an off-the-shelf captioning tool [18]. The aligned face image I_f is cropped from I, and w_+ is obtained from I_f by the e4e encoder (same as Stage I). A face region mask, M, is also obtained from the image I, with 0 denoting the face region and 1 representing the non-face region.

Learning Objectives. To keep high identity fidelity while encouraging high diversity of identity-irrelevant context, three losses are employed jointly in this stage.

First, the reconstruction loss $\mathcal{L}_{rec}$ is used to guide the denoising process:

$$\mathcal{L}_{rec} = \mathbb{E}_{z_0,\epsilon,t,c_{txt},w_+}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_{txt}(c_{txt}), \mathcal{F}_w(w_+)\right)\|_2^2\right] \tag{5}$$

Note that Eqn. (5) is different from Eqn. (4) for that: 1) z_0 here is encoded from the in-the-wild image I rather than the face image I_f, so the reconstruction goal differs, and 2) c_{txt} is the caption describing the in-the-wild image rather than a simple neutral prompt, therefore containing more information about the image context to facilitate reconstruction of I with w_+ from I_f.

Second, when semantic edits Δw are applied in the $\mathcal{W}_+$ space, the objective is to exclusively modify the facial attributes, leaving the remaining regions unchanged. In order to accomplish this, we propose a w_+ disentanglement loss to limit the editability of w_+ outside the face region:

$$\begin{aligned} \mathcal{L}_{disen} = \|&M \cdot \epsilon_\theta(z_t, t, \tau_{txt}(c_{txt}), \mathcal{F}_w(w_+)) - \\ &M \cdot \epsilon_\theta(z_t, t, \tau_{txt}(c_{txt}), \Psi(\mathcal{F}_w(w_+)))\| \end{aligned} \tag{6}$$

where Ψ is the augmentation operation. During training, three augmentation strategies are adopted: 1) shuffle along the batch dimension of $\mathcal{F}_w(w_+)$, 2) add random perturbations of Gaussian noise on $\mathcal{F}_w(w_+)$, and 3) combine both of them. By applying such constraints, augmented (or edited) w_+ vectors are forced to have similar non-face regions, thereby disentangling the effect of Δw on the background.

Finally, since the identity condition $\mathcal{F}_w(w_+)$ should only influence the face region, which constitutes a relatively small proportion of the final output, we constrain its impact to be limited by a regularization loss:

$$\mathcal{L}_{reg} = \|M \cdot \epsilon_\theta\left(z_t, t, \tau_{txt}(c_{txt}), \mathcal{F}_w(w_+)\right) - M \cdot \epsilon_\theta\left(z_t, t, \tau_{txt}(c_{txt})\right)\| \tag{7}$$

In this way, a more balanced effect of text and identity condition on the synthesized result is achieved, further mitigating the risk of unwanted noisy information introduced from w_+ and preserving the compatibility with text prompts.

The overall learning objective in Stage II is defined as:

$$\mathcal{L} = \mathcal{L}_{rec} + \gamma_1 \cdot \mathcal{L}_{disen} + \gamma_2 \cdot \mathcal{L}_{reg} \tag{8}$$

where γ_1 and γ_2 denote the trade-off parameters.

4. Experiments

Training Data. In Stage I, we use the pre-trained e4e encoder [32] to obtain the w_+ vector for each face image I_f from FFHQ [15] and StyleGAN2 [16]. FFHQ dataset contains 70,000 images, among which 600 are split for validation while the rest are for training. We also synthesize 70,000 images using StyleGAN2. In Stage II, we use FFHQ in-the-wild images (excluding those used for validation in Stage I) and SHHQ [7] to optimize the $\mathcal{W}_+$ adapter. SHHQ is a human dataset that contains 40,000 high-quality full-body images. BLIP2 [18] is introduced to generate the caption for each in-the-wild image. Face region mask M is obtained based on the FFHQ alignment operation and is eroded and blurred with kernel sizes of 32 and 7, respectively.

Implementation Details. All training is conducted on a server with 8 Tesla V100 GPUs. The batch size is set to 16. We employ the AdamW optimizer [22] with a learning rate of $1e{-}4$ and weight decay of 0.01. In Stage II, we adopt color jittering [44], random rotation, and sampling for in-the-wild images to increase diversity. During training, the cropped images with incomplete or small faces are discarded. It takes around 24 hours to align the $\mathcal{W}_+$ space with SD in Stage I, and nearly 96 hours to train the $\mathcal{W}_+$ adapter in Stage II. In the inference, we adopt DDIM sampler [31] with 50 steps. To enable classifier-free guidance [11], we use the default settings and set the guidance scale to 7.5. We follow IP-Adapter [39] to randomly drop the text c_{txt} and w_+ vector with a probability of 0.05. γ_1 and γ_2 in Eqn. (8) are set to 1.5 and 1.0, respectively. The whole framework is implemented on Diffusers [35]. Our experiments are based on SD v1.5 and can be extended to other SD models (see suppl.).

Table 1. Quantitative comparisons with previous methods. The best and second best results are highlighted by **bold** and underline.

Variants	CLIP Score↑	ID↓	Detection↑
Textual Inversion [8]	.194	.575	.891
Dreambooth [28]	.177	.562	.800
Custom Diffusion [17]	.216	.498	.850
FastComposer [38]	.265	.419	.950
IP-Adapter-Face [39]	.241	**.407**	**.958**
CelebBasis [40]	.253	.448	.916
Ours	**.267**	.418	.950

4.1. Quantitative Comparison

Baselines. The methods we evaluate can be categorized into two distinct groups: general object customization and specific face personalization. For general object customization, we use Dreambooth [28], Textual Inversion [8], and Custom Diffusion [17], implementing these using Diffusers [35]. For specific face personalization, we select closely related works such as FastComposer [38], IP-Adapter-Face [39], and CelebBasis [40]. In our evaluation, we adhere to the settings established by CelebBasis [40] and FastComposer [38]. The metrics used for assessment include the **CLIP Score** [25], which is calculated based on the average image-text similarity using CLIP-L/14, the Face **Detection** Score determined using MTCNN [41], and the Identity Distance (**ID**) measured using ArcFace [5] on the detected facial regions.

Given the extensive fine-tuning requirements for each identity in methods like Dreambooth, Textual Inversion, Custom Diffusion, and CelebBasis – which can be notably time-consuming (for instance, Textual Inversion demands about an hour on a single V100 GPU server[1]) – we encounter constraints in evaluating a substantial number of test instances. To ensure a thorough and equitable quantitative analysis, we carefully choose 120 identities from the CelebA-HQ dataset [14], using one reference image per identity for consistent comparisons. In addition, we randomly select 20 text prompts that describe various aspects such as clothing, styles, and backgrounds, to provide detailed characterizations of each individual.

Table 1 shows that optimization-based methods operating in the text embedding space, such as Textual Inversion,

[1] https://huggingface.co/docs/diffusers/training/text_inversion

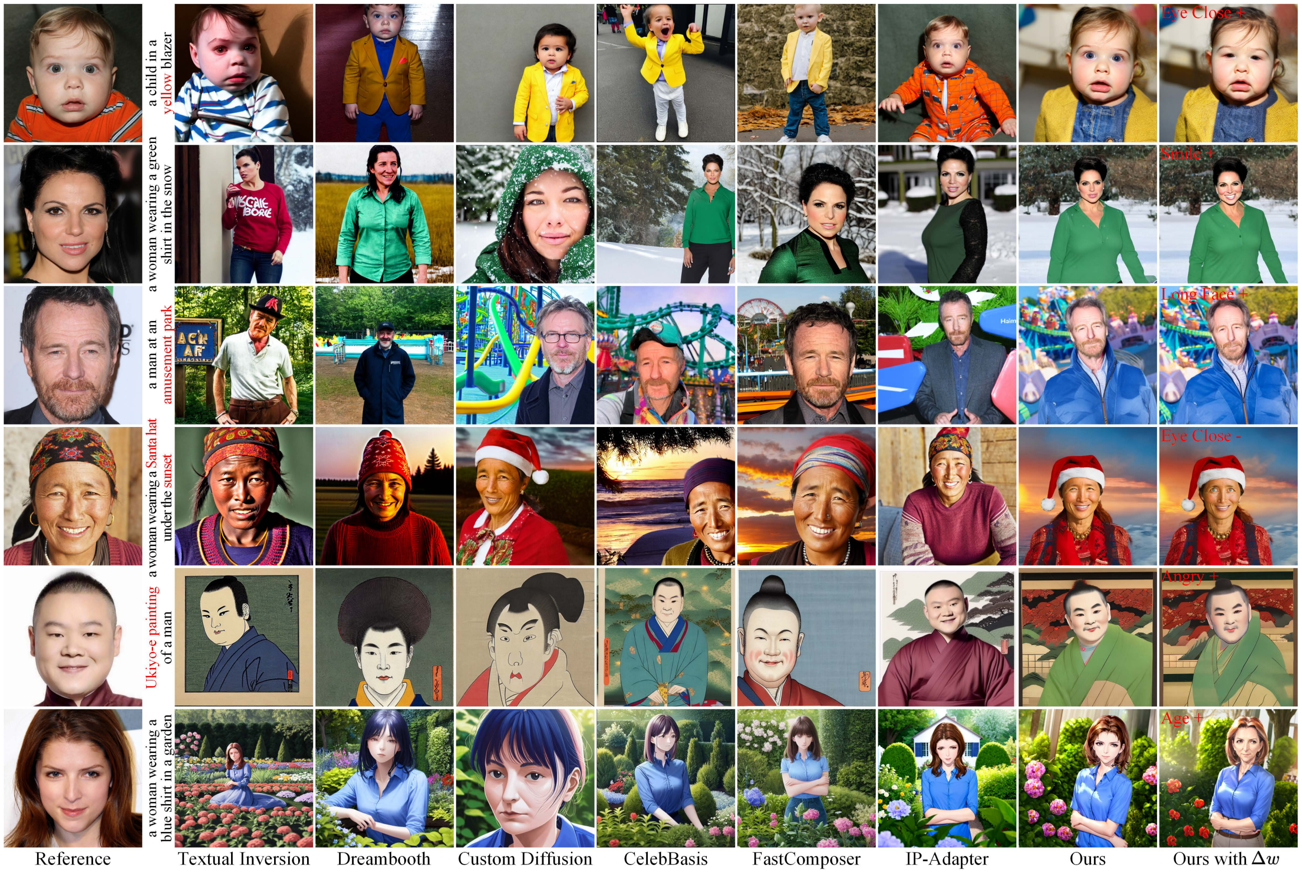

a child in a yellow blazer
a woman wearing a green shirt in the snow
a man at an amusement park
a woman wearing a Santa hat under the sunset
Ukiyo-e painting of a man
a woman wearing a blue shirt in a garden
Eye Close +
Smile +
Long Face +
Eye Close -
Angry +
Age +
Reference
Textual Inversion
Dreambooth
Custom Diffusion
CelebBasis
FastComposer
IP-Adapter
Ours
Ours with Δw

Figure 3. Visual comparisons with baselines on different scenarios of T2I generation. Best view by zooming in.

Dreambooth, and Custom Diffusion, fall short in accurately aligning with text prompts and preserving identity features. On the contrary, our method shows comparable performance to FastComposer and IP-Adapter-Face in terms of both CLIP Score and ID metrics. However, it is important to note that IP-Adapter has been fine-tuned on the large-scale LAION-2B dataset, whereas our model is trained on the relatively smaller FFHQ-wild and SHHQ datasets, comprising around 110,000 images. Both FastComposer and IP-Adapter directly process the reference image as input, while our approach first maps it to the $\mathcal{W}_+$ space before embedding it into the SD model. This additional mapping step could potentially introduce minor discrepancies when integrating the reference image into the SD model, slightly affecting the ID results. Nevertheless, this slight trade-off in ID preservation is balanced by our method's ability to flexibly edit facial attributes while ensuring the background remains consistent. This unique balance of attribute editability and background consistency is a novel contribution that sets our method apart from previous approaches.

4.2. Qualitative Comparison

Figure 3 depicts visual comparisons with baselines across various scenarios. Since Textual Inversion, Dreambooth, and Custom Diffusion are not specifically designed for facial images, we select results with the best facial quality from their generated set of 40 images. For the remaining methods, we choose from a set of 10. It is observed that Textual Inversion, Dreambooth, and Custom Diffusion encounter difficulties in capturing and maintaining identity details with a single reference, leading to sub-optimal performance in this task. CelebBasis and FastComposer face challenges in striking a balance between text compatibility and identity preservation. Though IP-Adapter shows improved identity retention, it tends to ignore text conditions in certain instances (see the $1st$, $3\sim5th$ rows).

Leveraging our $\mathcal{W}_+$ adapter, our approach successfully generates images that are not only compatible with text descriptions but also more effectively retain the target identity. Additionally, our method allows for the editing of facial attributes along the Δw direction, causing only minor alterations in the non-facial regions (illustrated in the last column). Furthermore, our approach can be seamlessly adapted to other pre-trained SD models without the need for additional fine-tuning, while retaining its editing capabilities. This versatility is exemplified in the last row of Fig. 3, which showcases our method's effectiveness with the dreamlike-anime model[2].

[2] https://huggingface.co/dreamlike-art/dreamlike-anime-1.0

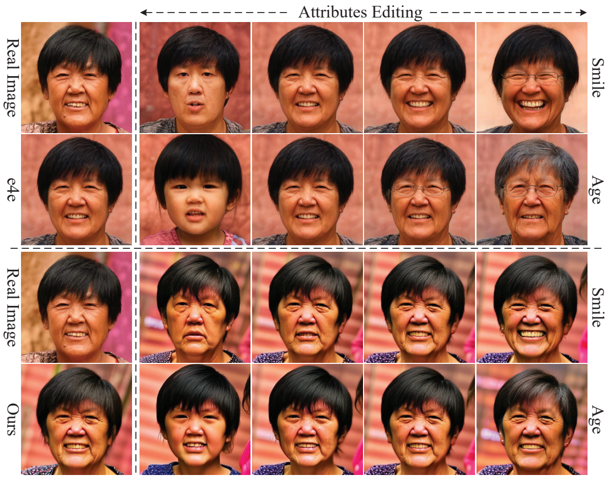


Figure 4. Inversion and editing comparisons between e4e and Ours in Stage I. They have the same w_+ and attributes editing Δw.

Table 2. Quantitative comparisons of face inversion and editing. Editing (ID↓) is the ID metric on results of w_+ and $w_+ - 3 \cdot \Delta w$.

Methods	Inversion			Editing (ID↓)	
	ID↓	FID↓	LPIPS↓	Age	Smile
e4e [32]	**.431**	32.26	**.205**	.409	.312
Ours (Stage I)	.435	**31.47**	.263	**.393**	**.275**

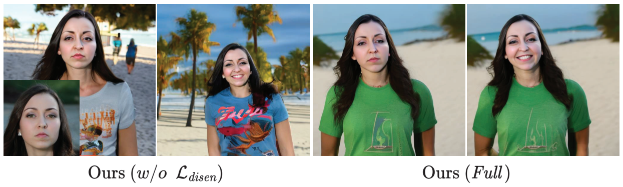
4.3. Ablation Study

Analyses of Aligning $\mathcal{W}_+$ in Stage I. We examine if the StyleGAN's w_+ embedding is well aligned to the latent space of SD during Stage I training. Fig. 4 offers a qualitative comparison of inversion and attribute editing outcomes between e4e [32] and our approach, applied to a real-world image. Both methods use the identical w_+ embedding generated by the e4e encoder. As depicted in Fig. 4, our method yields inversion results on par with e4e, as seen in the left column, using the same w_+ vector. Furthermore, the alignment process we implement in this phase preserves the editability inherent to the $\mathcal{W}_+$ space. By introducing attribute editing directions Δw (such as smile and age) from InterFaceGAN [29], our method can semantically adjust these attributes while maintaining the individual's identity. This is illustrated in the right columns, where the edited embedding, denoted as $w_+ + \alpha \cdot \Delta w$, reflects these changes. The parameter α is used to control the extent of attribute modification.

To quantitatively evaluate the alignment performance, we randomly select 1,000 images from CelebA-HQ [14] and measure their inversion results with metrics of ID, FID, and LPIPS [43][3]. As indicated in the Inversion column of Table 2, our method exhibits comparable performance to e4e in the image inversion task. For assessing attribute editing capabilities, we focus on two key attributes, namely smile, and age, applying a constant scale α (*i.e.*, $w_+ - 3 \cdot \Delta w$). The Editing column in Table 2 demonstrates that our approach surpasses e4e in maintaining identity during the editing process. These findings confirm that our method has successfully aligned the $\mathcal{W}_+$ space with the SD model, preserving both the inversion accuracy and the robust editability of attributes.

Table 3. Quantitative comparisons of different variants.

Variants	CLIP Score↑	ID↓	Detection↑
Ours (*PCA*)	.243	.526	.936
Ours (*w/o* $\mathcal{L}_{reg}$)	.136	**.461**	**.951**
Ours ($\mathcal{L}_{reg}^{0.5}$)	.192	.497	.946
Ours (*Full*)	**.267**	.516	.943



Ours (w/o $\mathcal{L}_{disen}$)

Ours (*Full*)

Figure 5. Visual comparisons of smile attribute editing. The prompt is "a woman wearing a t-shirt on the beach".

Analyses of Variants. We examine different variants to assess each component of our $\mathcal{W}_+$ adapter: 1) Ours (*PCA*), which substitutes our residual cross-attention with parallel cross-attention as in [39], and 2) Ours (*w/o* $\mathcal{L}_{reg}$), which omits the $\mathcal{L}_{reg}$ objective. The outcomes of these variants are presented in Table 3. It is observed that the residual cross-attention module outperforms the parallel cross-attention design. We conjecture that in tasks focused on integrating local facial imagery into broader, in-the-wild scenarios, directly injecting facial embeddings into the original hidden state f_z might adversely impact non-facial regions. Conversely, our residual cross-attention determines fusion weights by calculating attention activations between the text-embedded hidden state f'_z and our facial embedding. This results in a more precise and effective fusion within the facial region.

Besides, the results suggest that the regularization loss $\mathcal{L}_{reg}$ is crucial in maintaining semantic consistency between text prompts and images. In the absence of $\mathcal{L}_{reg}$, the variant Ours (*w/o* $\mathcal{L}_{reg}$) tends to overemphasize facial regions while neglecting the accompanying text descriptions, leading to a reduced CLIP score. Incorporating $\mathcal{L}_{reg}$ with a weight of $\gamma_2 = 0.5$, as in Ours ($\mathcal{L}_{reg}^{0.5}$), enhances the CLIP score but slightly compromises the ID metric. In comparison, Ours (*Full*) achieves a satisfactory trade-off in balancing the performance of ID and CLIP score.

The impact of our disentanglement loss $\mathcal{L}_{disen}$ (Eqn. (6)) is visually demonstrated in Fig. 5. This example shows that omitting $\mathcal{L}_{disen}$ leads to unintended alterations in non-facial regions when the smile editing direction Δw is applied to w_+. This indicates the importance of $\mathcal{L}_{disen}$ in effectively separating identity-relevant and irrelevant information. By combining our residual cross-attention, regularization loss, and disentanglement loss, Ours (*Full*) preserves identity and ensures compatibility with text prompts, even amidst changes to the face embedding.

[3] https://github.com/chaofengc/IQA-PyTorch

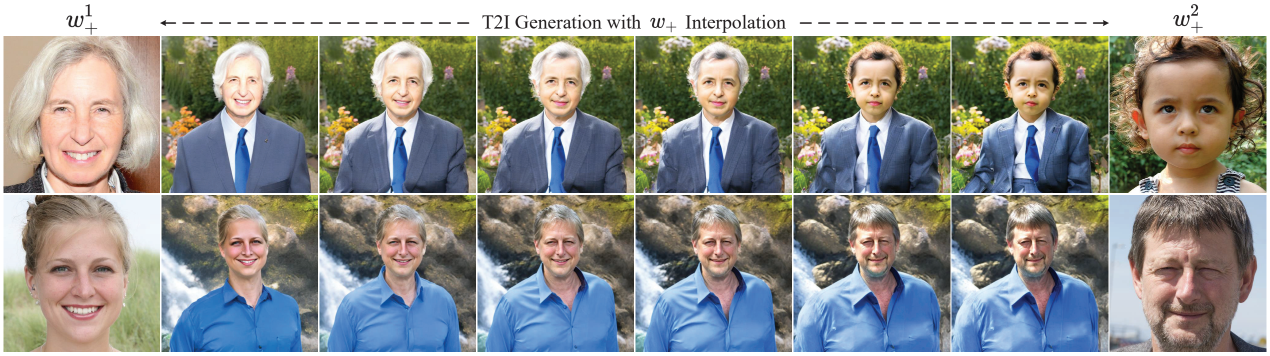


Figure 6. Visual results of w_+ embeddings interpolation during inference. The prompts are “one person wearing suit and tie in a garden” and “one person wearing a blue shirt by a secluded waterfall”, respectively.

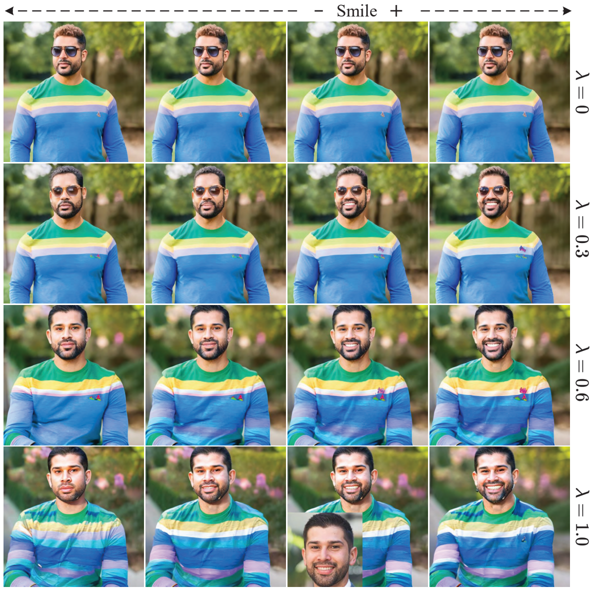


Figure 7. Visual results of using different λ during inference. The prompt is “a man wearing a rainbow shirt in a garden”.

Analyses of λ in Cross-attention. The parameter λ determines the extent to which the w_+ vector influences the hidden state f'_z. Fig. 7 shows the effects of varying λ during inference. At $\lambda = 0$, the outcome aligns with that of the original SD model, unaffected by any editing direction. However, as λ increases, the generated identity more closely resembles the reference image. Interestingly, at relatively low values of λ (for example, $\lambda = 0.3$), even though the identity does not closely match the reference, our model effectively edits attributes in the specified direction. This observation suggests that our method proficiently leverages the $\mathcal{W}_+$ space from StyleGAN within the SD framework.

Analyses of w_+ Interpolation. We select two facial images and acquire their respective w_+^1 and w_+^2 embeddings from e4e encoder. The interpolated w_+ embedding is then obtained through $(1-\kappa) \cdot w_+^1 + \kappa \cdot w_+^2$, where $\kappa \in [0, 1]$. As shown in Fig. 6, this T2I generation process results in a smooth transition in the facial regions of the generated images, while maintaining a similar and consistent layout throughout the interpolation. This result confirms that the w_+ embedding in our method not only preserves the editability characteristic of the original StyleGAN but also effectively distinguishes between facial and background regions.

4.4. Limitation

Our work aims at integrating StyleGAN’s editable $\mathcal{W}_+$ space into the SD model. We notice a challenge in this integration: the process of converting real images to w_+ vectors in StyleGAN often leads to a loss of detail, impacting the preservation of identity features. Despite employing a substantial number of training pairs $\{I_f, w_+\}$ to establish the mapping, we still observe limitations in maintaining identity fidelity and challenges in editing certain attributes (*e.g.*, pose and glasses). The current framework is designed to generate and edit images with only a single face. Our future work aims to explore the potential of applying localized injections of distinct w_+ to address multiple human subjects.

5. Conclusion

We have presented the first attempt to embed the $\mathcal{W}_+$ space of StyleGAN into the SD model. We showed that both the mapping network and the residual cross-attention module play crucial roles in facilitating the injection of w_+ embedding into the SD model, balancing between text prompt influence and identity conditions. Our experiments demonstrate that the $\mathcal{W}_+$ space, as used in our approach, not only enables personalized text-to-image generation but also allows for precise editing of facial attributes. We envision this capability being highly beneficial in various practical applications, such as portrait customization with seamless attribute modifications. Furthermore, our methodology holds the potential for application across other object domains that possess distinct prior spaces.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *ICCV*, 2019. 2

[2] Yufei Cai, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Hu Han, and Wangmeng Zuo. Decoupled textual embeddings for customized image generation. In *AAAI*, 2024. 2

[3] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. DisenBooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023. 2

[4] 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. 2

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

[7] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A data-centric odyssey of human generation. In *ECCV*, 2022. 5

[8] 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. In *ICLR*, 2022. 2, 5

[9] 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. *arXiv preprint arXiv:2302.12228*, 2023. 2

[10] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. In *NeurIPS*, 2020. 2

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2022. 5

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[13] Xuhui Jia, Yang Zhao, Kelvin C. K. 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. 2

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 5, 7

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3, 5

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 3, 5

[17] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2, 5

[18] 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 *ICML*, 2023. 4, 5

[19] Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *CVPR*, 2023. 2

[20] Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. A survey on leveraging pre-trained generative adversarial networks for image editing and restoration. *arXiv preprint arXiv:2207.10309*, 2022. 2

[21] Yang Liu, Cheng Yu, Lei Shang, Ziheng Wu, Xingjun Wang, Yuze Zhao, Lin Zhu, Chen Cheng, Weitao Chen, Chao Xu, Haoyu Xie, Yuan Yao, Wenmeng Zhou, Yingda Chen, Xuansong Xie, and Baigui Sun. FaceChain: A playground for identity-preserving portrait generation. *arXiv preprint arXiv:2308.14256*, 2023. 3

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[24] Hamza Pehlivan, Yusuf Dalva, and Aysegul Dundar. StyleRes: Transforming the residuals for real image editing with StyleGAN. In *CVPR*, 2022. 2

[25] 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*, 2021. 3, 5

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[28] 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*, 2023. 2, 5

[29] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, 2020. 2, 7

[30] 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. 2

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[32] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. In *ACM TOG*, 2021. 2, 3, 5, 7

[33] Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. *arXiv preprint arXiv:2306.06638*, 2023. 3

[34] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 3

[35] 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. 5

[36] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 2, 4

[37] 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. 2

[38] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2, 5

[39] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3, 4, 5, 7

[40] 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

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. In *IEEE signal processing letters*, 2016. 5

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[44] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *ECCV*, 2020. 5