Revealing Directions for Text-guided 3D Face Editing

Zhuo Chen, Yichao Yan, Sehngqi Liu, Yuhao Cheng,
Weiming Zhao, Lincheng Li, Mengxiao Bi, Xiaokang Yang

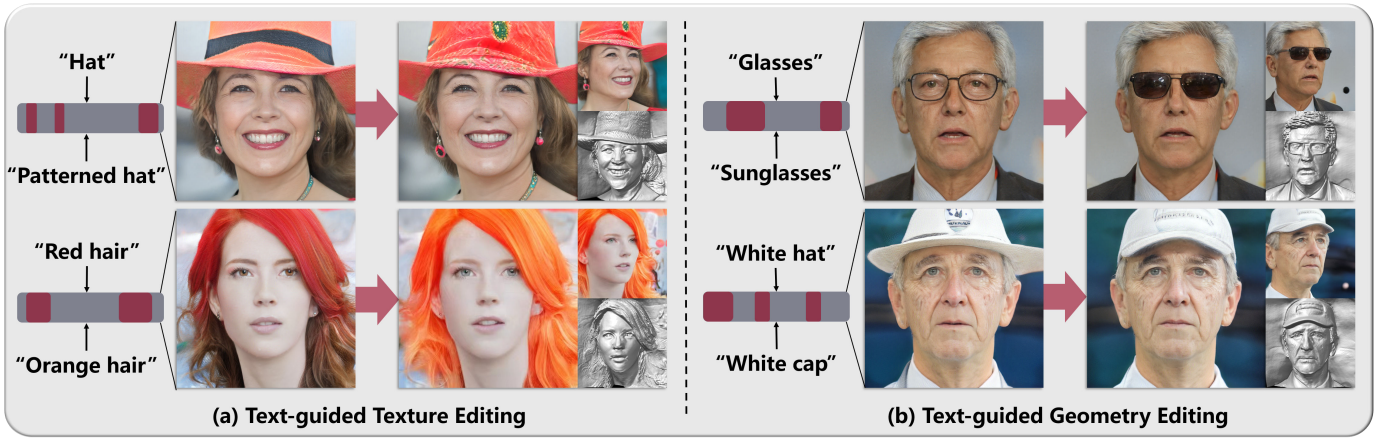

Fig. 1: Examples of our text-guided 3D face editing. Our method links the input text to a mask on latent codes, enabling attribute disentanglement and identity preservation. It has the capacity of both (a) texture editing and (b) geometry manipulation.

***Abstract*—3D face editing is a significant task in multimedia, aimed at the manipulation of 3D face models across various control signals. The success of 3D-aware GAN provides expressive 3D models learned from 2D single-view images only, encouraging researchers to discover semantic editing directions in its latent space. However, previous methods face challenges in balancing quality, efficiency, and generalization. To solve the problem, we explore the possibility of introducing the strength of diffusion model into 3D-aware GANs. In this paper, we present Face Clan, a fast and text-general approach for generating and manipulating 3D faces based on arbitrary attribute descriptions. To achieve disentangled editing, we propose to diffuse on the latent space under a pair of opposite prompts to estimate the mask indicating the region of interest on latent codes. Based on the mask, we then apply denoising to the masked latent codes to reveal the editing direction. Our method offers a precisely controllable manipulation method, allowing users to intuitively customize regions of interest with the text description. Experiments demonstrate the effectiveness and generalization of our Face Clan for various pre-trained GANs. It offers an intuitive and wide application for text-guided face editing that contributes to the landscape of multimedia content creation.**

***Index Terms*—3D face editing, Diffusion model, Latent space, GAN, Text-conditioned**

(Corresponding author: Yichao Yan.)
Zhuo Chen, Yichao Yan, Sehngqi Liu, Yuhao Cheng and Xiaokang Yang are with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China. (email: ningci5252@sjtu.edu.cn, yanyichao@sjtu.edu.cn, lsqlsq@sjtu.edu.cn, chengyuhao@sjtu.edu.cn, xkyang@sjtu.edu.cn)
Weiming Zhao is with the Student Innovation Center, Shanghai Jiao Tong University, Shanghai, China. (email: weiming.zhao@sjtu.edu.cn)
Lincheng Li and Mengxiao Bi are with NetEase Fuxi AI Lab, Hangzhou, China. (email: lilincheng@corp.netease.com, bimengxiao@corp.netease.com)

I. Introduction

3D face editing is an essential task in multimedia, aiming to manipulate face attributes in a 3D representation under various control signals, while preserving the original identity and 3D consistency. In 2D editing, we have witnessed the creative and diverse synthesis achieved by diffusion models. Diffusion models can control multiple modalities and generate diverse images by manipulating the latent or attention features across the time steps. However, existing text-to-image diffusion models lack editing precision, which can produce wonderful changes in a whole part but perform badly in detailed editing. Diffusion model also face challengs to directly lift this talent to 3D generation due to its heavy dependency on the 3D ground-truth data. Fortunately, recent 3D-aware GANs [1]–[14] brings the ability to synthesize 3D representation by learning from unposed single-view 2D images only. A GAN model [15], [16] can learn well-disentangled attributes within its latent space and support an arbitrary combination of attributes, equivalent to compacting the composite 3D face into a light latent code. The light representation and rich distributions encourage researchers to explore the meaningful directions in the latent space, and also potentially empower us to collaborate with the power of mature editing strategy based on the diffusion model.

Previous methods have explored semantic directions in the latent space, but they still face challenges in balancing intuitiveness, generalization, and efficiency. Supervised methods [17]–[21] are time-consuming and incapable of the attribute whose classifiers are not available, while unsupervised methods [22]–[27] are highly sensitive to the analyzed identity and fail to discover arbitrary semantic directions as

desired by the user. Another category of work introduces additional control signals to discover the editing direction. These methods optimize or predict latent codes under the guidance of various conditions, *e.g.*, semantic map [28], [29], 3DMM [30], [31] and dragging points [32]–[34]. Although these conditioned methods have achieved success in shape and expression manipulation, they encounter difficulties when attempting to edit color and texture.

To achieve a more general and intuitive way to manipulate 3D faces, there has been a growing interest on text-guided exploration of semantic directions. Recent text-conditioned methods can be roughly classified into 1) optimization in latent space or parameter space and 2) an encoder that directly projects CLIP features to latent space. Optimization methods [35], [36] guided by CLIP [37] or Stable diffusion [38] can handle general texts, albeit at the cost of prolonged editing times. To improve efficiency, recent works [39]–[43] align the text feature with latent codes via an encoder. However, the one-step determined projection reduces their capacity and leads to a lack of diversity and identity preservation. Thus, there is an inherent need for **an optimization-free model that can balance text generalization, editing quality, and efficiency**.

Inspired by the success of diffusion models in text-to-2D synthesis [38], [44]–[49], we come out with an idea that **diffuses the latent space of GAN** to align text conditions with latent codes instead of 2D images. Diffusion models are good at multi-model controls and fantastic changes, but bad at human detail features and 3D views. GANs perform well in getting and changing human face features, but badly in multi-model controls and style transfer [50]. It can be seen that their characters are complementary, therefore, we choose to combine the diffusion model and GAN into a unified model to leverage both of their advantages.

In contrast to the mapper that directly projects text features into latent spaces, the diffusion model in this work demonstrates superior text-to-latent diversity and consistency, due to its multi-step accumulated bias as a generative model [51]–[53]. Text conditions can be regarded as a direction indicator that leads the sampled noise to the class consistent with the given text. This special characteristic can both enhance text-guided generation quality and facilitate editing direction reveal. Based on the inspiration, we present a fast and text-generalized approach called **Face Clan**, to automatically generate and manipulate the 3D face based on arbitrary attribute descriptions by reconstructing the distribution of latent codes.

A robust method with high quality should take effect on the contents within the area of interest only and preserve the remaining contents as much as possible. We decompose it into **mask** and **direction** in the latent space. **1) To edit the contents**, we need to design a text-guided model that reveals an editing **direction** for latent codes. **2) To preserve the remaining parts**, a direct idea is to apply a mask on the region of interest. However, latent space is visually implicit so that users cannot intuitively add the mask. Therefore, it is supposed to analyze the interest region of given texts and link it to a **mask** on latent codes.

Face Clan first trains a diffusion model to align the distribution of the text with the learned latent manifolds of pre-trained 3D-aware GAN in a self-supervised way. Given a set of synthesized data from the pre-trained GAN, it is capable of learning to map a randomly sampled Gaussian noise to a text-consistent latent code with identity diversity. Modulated by the latent code, the 3D-aware generator can further produce a 3D-consistent face according to the textual description. Based on the trained text-to-3D-face model, we are empowered to precisely edit face attributes. During editing, we initially estimate a text-relevant mask on the latent code by measuring the principal difference between two predicted noises under a pair of opposite descriptions, *e.g.*, *"hat"* and *"cap"* as despite in Fig. 1. With the mask, a denosing procedure can be performed on the masked region of noisy latent codes, while keeping the unmasked region as usual. Consequently, it is flexible that users can customize the desired regions under input prompts. Extensive experimental results show that our Face Clan achieves precise controllability and strong robustness in various 3D-aware GANs. The main contributions are summarized as follows:

- We design a fast and generalized text-guided face editing pipeline based on a self-supervised diffusion model that aligns distributions of texts and latent manifolds in the pre-trained GAN.
- We propose a directional mask estimation to link the text to the region of interest in latent codes, achieving robust attribute disentanglement.
- Our editing approach can handle identity-specific attributes out of the common face attribute domain, *e.g.*, hat style as despite in Fig. 1.

II. Related Works

A. Semantic Direction Exploration in GANs

The latent space of GANs contains a wealth of semantic features, facilitating the exploration of diverse semantic directions for image editing. Face editing methods in the latent space can be broadly categorized as **condition-guided direction discovery** and **unconditioned direction discovery** Unconditioned direction discovery can be further subdivided into supervised and unsupervised exploration. Supervised methods [17]–[20] involve labeling image data with the target binary attribute and training a classifier to reveal the inherent hyperplane. The normal of this hyperplane indicates the editing direction between the binary attribute. While this approach is stable and effective, it is laborious to discover an editing direction due to the need for labeled data and additional classifier training for each attribute. Some unsupervised methods [22]–[27], [54] analyze the principle feature of the latent space to discover meaningful editing directions. For real face editing, StyleRes [55] addresses the trade-off between reconstruction fidelity and editing quality by learning residual features in higher latent codes and transforming them. StyleFeatureEditor [56] enables editing in both latent space and feature space, achieving finer image details and preserving them during editing. However, these methods lack the intuitiveness of semantic directions, as these directions in editing images often need to be manually discerned. Besides, it is hard to generalize to diverse attributes, limited to the principal face attributes. To

achieve more intuitive editing, some methods explore semantic directions through explicit conditions, *e.g.*, 3DMM [30], [31], semantic maps [28], [29], and point dragging [32]–[34]. These methods incorporate multi-modality into facial editing of multimedia, *e.g.*, image, video, and 3D representations.

Specifically, Zhu *et al.* [57] separate the face into the overall shape and detailed regions, utilizing 3DMM to guide shape deformation and semantic mask to guide region manipulation. Yang *et al.* [58] also take region masks to edit the latent code for facial editing. Unlike previous works, GlassesCLIP [59] focuses on the specific glasses region, which is challenging in facial editing. Despite their effectiveness and intuitiveness, these methods are constrained to geometry manipulation due to the representative ability of these conditions. Failure in texture editing restricts their applications. Among the vast control signals, natural language is the primary means of human expression, containing both geometry and texture semantics that can align with image attributes and realize more diverse editing on targets.

B. Text-guided Editing in GANs

In the realm of multimedia innovation, some researchers delve into the intricate interplay between textual descriptions and image manipulation, offering an intuitive approach to face editing. Some earlier approaches [60]–[63] utilize encoder-decoder structures to construct the relationship between images and text for editing. To enable semantic editing, some approaches [19], [35], [39]–[41] propose editing within GANs, leveraging pre-trained large-scale multi-modal models, *e.g.*, CLIP [37]. StyleCLIP [19] introduces a paradigm to optimize the latent code of GANs. StyleGAN-Nada [35] proposes to fine-tune the generator of GANs to improve a specific type of editing. MorphNeRF [64] proposes a learnable network that morphs the 3D geometry of faces toward the text descriptions via NeRF-based GAN and CLIP. Despite high-quality editing results, they suffer from inefficiency due to the time-consuming optimization. Aiming at efficiency improvement, the following works [39]–[41], [65]–[67] directly project the text feature extracted by the CLIP model to the latent space. TextFace [43] introduces text-to-style mapping based on text-image similarity matching and a face-caption alignment, achieving high-fidelity face generation and manipulation. However, the one-step determined projection limits their capacity, leading to conflict between diversity and identity preservation. Therefore, it is essential to propose a method that can balance the quality and efficiency of editing. Our proposed Face Clan leverages a lightweight diffusion model to map texts to latent codes without optimizing the generators, achieving efficient and wide-ranging face editing.

C. Diffusion-based Face Edit

The success of diffusion in 2D image editing has been widely witnessed. Early methods either focus on textural inversion [48], [49] or redistribute the image via noising and denoising [44]–[47]. To enable controllable editing, similar to the development of GAN, further approaches explore the cooperation of the diffusion model and different modalities, *e.g.*textual instructions [68]–[72], masks [73], [74], segmentation maps [75], and point-dragging interfaces [76]. Specifically, MMGInpainting [77] incorporates an Anchored Stripe Attention mechanism that utilizes anchor points to model global contextual dependencies, effectively integrating the semantic information into the target region of faces. Despite the desired modifications, these methods often fall short in precise and disentangled editing. This means that they always change the identity or other irrelevant facial attributes when editing the target region. Sketch-based editing approaches [78]–[80] offer an interactive brush to constrain the painting region. However, the modality of sketch or painting still faces the limitation of identity shift when editing overall attributes, *e.g.*, make-up and complexion. It is also hard to precisely control the details in the editing areas, such as accidentally changing eye color when adding glasses. Recent Flux-based models can be seen as an improved alternation of the base model. They have better performance for customized generation, but editing methods [81] based on them still have the problem of editing precision. In contrast, our focus lies in the latent space of GANs, where the latent codes, unlike 2D images, are well-distributed and lightweight. This characteristic means that the editing direction of an attribute is nearly unified across all codes, and revealing the direction from a group of codes is fast and efficient. The advantages of the latent code and the diffusion model have motivated researchers to explore the possibility of their combination. Recent works [12], [42], [82]–[85] connect a diffusion model to the latent space of GANs, enabling text-controllable synthesis. However, these approaches have not addressed the preservation of identities, thereby failing in precise editing. The work most closely related to ours is InstructPix2NeRF [86] (ICLR 2024), which employs a conditional 3D diffusion to lift 2D editing to 3D space through the acquisition of the correlation between the image and the instruction. Despite diverse and generalized editing, it still falls short in disentanglement and identity preservation. To enhance the editing precision, we design a method for mask estimation on latent codes to realize more controllable editing.

III. Methods

In this section, we begin by reviewing the 3D-aware GAN and the diffusion model that respectively produces high-quality 3D faces and learns text-conditioned latent distribution (Sec. III-A). Benefiting from the rich semantics learned in the latent space of GANs, we design a self-supervised diffusion model that aligns the distributions of text prompts and latent codes, to achieve text-guided face synthesis. (Sec. III-B). Based on this trained text-to-3D-face model, we further propose a text-guided editing pipeline, which consists of two specific steps, *i.e.*,mask estimation (Sec. III-C) and target direction discovery (Sec. III-D) on latent codes. The overview of our method is shown in Fig. 2.

A. Preliminaries

Classic diffusion models, such as stable diffusion, are built on the latent space of a pre-trained generative VAE model.

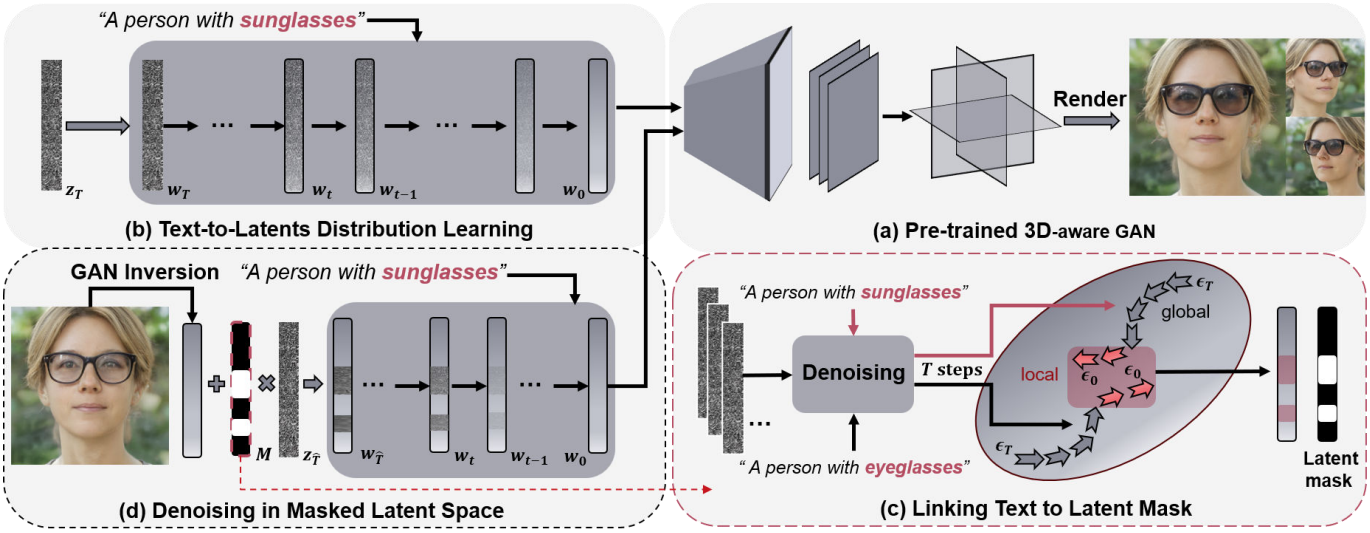

Fig. 2: Overview of our proposed method. (a) The architecture of our based 3D-aware generator, EG3D. (b) The inference pipeline to redistribute the latent code for text-guided synthesis. (c) The illustration of linking text to the region of interest on latent codes. (d) Apply denoising to masked latent codes for disentangled face editing.

Actually, the diffusion model can be applied to the latent space of most generative models, where GAN is one of them. Our training process also follows classic diffusion models in that we first train a generative model with its latent space which is a 3D-aware GAN in our pipeline, and then we generate latent-label pairs to train a diffusion model with DiT structure. The diffusion model learns how to predict the noise from a noisy latent conditioned by the corresponding label. After that, we can use this diffusion model to manipulate the latent of our trained 3D-GAN.

3D-aware GAN. Considering the generation quality, our framework is constructed based on EG3D [7], as shown in Fig. 2 (a). Similar to other conditioned GANs, EG3D contains a mapping network $f(\cdot)$ responsible for projecting random latent code z (together with camera parameters) to an intermediate latent code $\mathbf{w} = f(z) \in \mathbb{R}^{512}$, which subsequently modulates a synthesis network. This latent code $\mathbf{w}$ plays a crucial role in face editing, as it learns disentangled attributes and aligns with the data distribution. Conditioned on $\mathbf{w}$, the generator of EG3D further produces a tri-plane feature $\mathbf{F} \in \mathbb{R}^{3\times32\times256\times256}$ as a 3D representation. Subsequently, a shallow MLP decoder projects the tri-plane feature $\mathbf{F}$ into volume density σ and color feature c, which are further rendered into high-resolution images via volume rendering and a super-resolution module. The entire pipeline can be simplified as $I_0 = \mathcal{G}(\mathbf{w_0}) = \mathcal{G}(f(z_0))$, where $\mathcal{G}$ represents the generator of EG3D. Compared with the inherent 3D representation $\mathbf{F} \in \mathbb{R}^{3\times32\times256\times256}$, a latent code $\mathbf{w} \in \mathbb{R}^{512}$ is heavily compacted, while it still contains most information of a 3D face. Moving along a semantic direction within the latent space allows for easy face editing,

$$I_0' = \mathcal{G}(\mathbf{w_0} + \Delta\mathbf{w}).$$

Latent Diffusion Model. The diffusion model has shown its strength in diverse generations. It incorporates a forward process, given by the equation:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}),$$

which add noise to the latent representation of the source image $\mathbf{x}_0$, and a reverse process, given by the equation:

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t),$$

which utilizes a parameterized denoising network θ to gradually denoise the target latent representation. To ensure that the forward process is approximately equal to the reverse process, the latent diffusion model is trained by minimizing weighted evidence lower bound:

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{t,\epsilon}\left[w(t)\|\epsilon_\theta(\mathbf{x}_t;t) - \epsilon\|_2^2\right], \quad (1)$$

where $w(t)$ is a weighting function, time step $t \sim \mathcal{U}(0,1)$, random noise $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, and θ denotes the parameters of the denoising network. After the training, we can randomly sample a Gaussian noise input $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and apply denoising on it to generate the target x_0 conforming to data distribution.

B. Text-to-Latents Distribution Learning

The diffusion model possesses a strong capability to project Gaussian noise onto a sample from the data distribution, making it well-suited for the mapping network $f(z)$ in GANs. Consequently, we introduce a diffusion model as an alternative mapping network into the latent space of GAN, to learn the mapping from Z space to W space, as shown in Fig. 2 (b). Following the preliminaries, we train a diffusion model whose θ is parameterized by DiT [87] for the latent space of GANs. To further enable explicit-conditioned generation based on a pre-trained generator, we add a cross-attention layer [88] to introduce conditions y and train the model by an objective [38],

$$L_{LDM} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),y,t}\left[\|\epsilon - \epsilon_\theta(w_t, t, \tau_\theta(y))\|_2^2\right]. \quad (2)$$

Although training a latent diffusion model typically raises concerns about the requirement for a large amount of paired image-text data, we can solve it through self-supervised data generation. Following previous works [89], we can synthesize pairs of the condition and the latent code infinitely by the synthesized images from the 3D generator and a feature extractor for the target condition. While our focus in this work is on the text condition, it is worth noting that automatic caption models [90] exhibit poor performance and limited diversity when applied to human faces. Fortunately, the aligned space of CLIP [37] allows us to utilize CLIP image embedding as the condition during training, while using text embedding during inference, as in previous methods [85], [89], [91]. With the CLIP feature extractor $C(\cdot)$, the training objective can be written as,

$$L_{LDM} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),C(\mathcal{G}(w_0)),t}\left[\|\epsilon - \epsilon_\theta\left(w_t, t, C\left(\mathcal{G}\left(w_0\right)\right)\right)\|_2^2\right]. \tag{3}$$

To improve training efficiency, both data generation and feature extraction are off-the-shelf.

C. Linking Text to Latent Mask

Common 2D image editing encourages applying a mask to keep irrelevant pixels the same within the region of interest. Different from direct 2D image editing, there is no visually explicit semantics in latent codes, which makes it hard to directly apply masks as users' intuitiveness. To intuitively control local semantics in the output 3D face, it leads to a demand that links the explicit meaningful text to a mask on the implicit latent code. Inspired by the supervised methods that collect labeled samples with binary attributes to classify, we come out with an idea that we can treat our diffusion model as a natural classifier to both generate samples and find directions between them, as shown in Fig. 3 and Fig. 2 (a).

Specifically, to reveal a collective direction of a specific attribute, we first sample several Gaussian noises $\mathbf{Z}_T = \{\mathbf{z}_T^0, \mathbf{z_T}^1, \cdots, \mathbf{z}_T^n\}$ and further denoise under the condition of paired descriptions with opposite attributes y_{src} and y_{tgt}, *e.g.*, *"a person with sunglasses"* and *"a person without sunglasses"*. The detailed procedure of each step is represented as,

$$\mathbf{Z}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{Z}_t - \sqrt{1-\alpha_t}\cdot\epsilon_t}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\cdot\epsilon_t, \tag{4}$$

$$\epsilon_t = \epsilon_\theta\left(\mathbf{Z}_t, t, \tau_\theta\left(y\right)\right), \tag{5}$$

where noise schedule $\alpha_t \in (0,1)$, $\sqrt{1-\alpha_{t-1}}\cdot\epsilon_t$ indicates the direction to the target Z_0. To simplify the representation, we donate the full denoising procedure D as,

$$\mathbf{W}_{\text{src}} = \mathbf{Z}_0^{src} = D\left(\mathbf{Z}_T, y_{src}\right), \tag{6}$$

$$\mathbf{W}_{\text{tgt}} = \mathbf{Z}_0^{tgt} = D\left(\mathbf{Z}_T, y_{tgt}\right), \tag{7}$$

where $\mathbf{W}_{\text{src}}$ and $\mathbf{W}_{\text{tgt}}$ are the denoised latent codes of the pre-trained model under opposite prompts y_{src} and y_{tgt}.

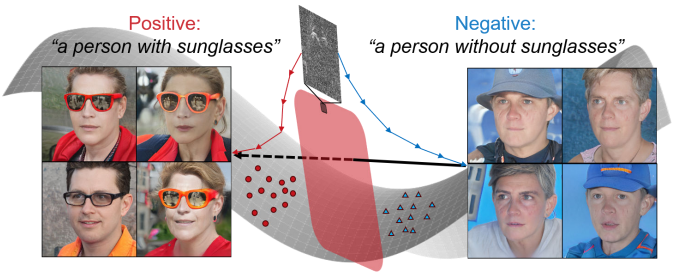

Fig. 3: The illustration of the inspiration that adopts diffusion model as both a controllable data generator and an attribute classifier.

Mask estimated from ϵ or $\mathbf{W}$. With two groups of latent codes, a naive idea is to estimate the mask M by the difference as,

$$M[i] = \begin{cases} 1, Norm\left(|\mathbf{W}_{tgt} - \mathbf{W}_{src}|\right)[i] > threshold, \\ 0, otherwise, \end{cases} \tag{8}$$

where i is the ith dimension in the latent code. It is effective but suffers from attribute entanglement. The reason is that the difference between paired latent codes can be seen as accumulating the bias of full-step paired noises. However, similar to 2D image diffusion models, predicted noises in different steps indicate directions for different-level attributes [92]–[94]. The endpoint estimate $\hat{x_0}$ travels on the latent manifold, initiating from the center of the distribution, moving first along the high variance axes, and then the lower variance axes. Consequently, low-level attributes are produced in the early stage, while finer details tend to emerge in the last few steps. Therefore, the noise difference accumulated in latent codes contains more unexpected global changes besides the local target attribute. Fig. 10 in our ablation study also supports this theory. According to the above analysis, we turn to estimate the mask by the difference between paired predicted noises in the last few steps. It avoids the influence of global and fused changes led by the early stage. The procedure is represented as,

$$\epsilon_t^{\text{src}} = \epsilon_\theta\left(\mathbf{Z}_t, t, \tau_\theta\left(y_{\text{src}}\right)\right), \tag{9}$$

$$\epsilon_t^{\text{tgt}} = \epsilon_\theta\left(\mathbf{Z}_t, t, \tau_\theta\left(y_{\text{tgt}}\right)\right), \tag{10}$$

$$M[i] = \begin{cases} 1, Norm\left(\left|\epsilon_t^{\text{tgt}} - \epsilon_t^{\text{src}}\right|\right)[i] > threshold, \\ 0, otherwise. \end{cases} \tag{11}$$

With the estimated mask from noise direction, a simple idea for face editing is to directly swap the $\mathbf{W}_{\text{tgt}}$ with $\mathbf{W}_{\text{src}}$ in the masked region as the editing direction, which is represented as,

$$\Delta\mathbf{w} = M \odot \left(\bar{\mathbf{W}}_{\text{tgt}} - \bar{\mathbf{W}}_{\text{src}}\right), \tag{12}$$

$$I_{\text{edit}} = \mathcal{G}\left(\mathbf{w}_{\text{input}} + \alpha\Delta\mathbf{w}\right). \tag{13}$$

Unfortunately, a precise latent mask can constrain the region dimensions in the latent code, but can not accurately indicate the change direction. It neither achieves a large-scale change for the target attribute nor preserves identity during editing towards this direction. Despite a scalable α to manipulate

the aptitude, a large scale α with the misdirection $\Delta\mathbf{w}$ leads to corner samples in the distribution. We have conducted an ablation study in Fig. 10 to support the claim. Therefore, we need an additional step to ascertain the efficient direction in the masked latent space.

D. Denoising in Masked Latent Space

Given an input image $\mathbf{I}_e$, we first invert it into a latent code $\mathbf{w}_e$,

$$\mathbf{z}_0^e = \mathbf{w}_e = \mathcal{E}(\mathbf{I}_e). \tag{14}$$

With the estimated mask M, we can perform text-conditioned diffusion on the masked region of the latent code $\mathbf{z}_0^e$ to reveal the efficient direction. Compared to 2D image editing, latent codes are much less sensitive on the edge between masked and unmasked regions, obviating concerns regarding edge inconsistency. Specifically, we add noise to the latent code $\mathbf{z}_0^e$ instead of sampling several Gaussian noises,

$$q\left(\mathbf{z}_t^e \mid \mathbf{z}_{t-1}^e, \mathbf{z}_0^e\right) = \mathcal{N}\left(\mathbf{z}_t^e; \sqrt{1-\beta_t}\mathbf{z}_{t-1}^e, \beta_t\mathbf{I}\right), \tag{15}$$

where $\mathcal{N}(\cdot)$ is a Gaussian distribution, and $\mathbf{I}$ is an identity matrix. Here, rather than completely destruct the original distribution, we get a partially noisy latent code $\mathbf{z}_{T'}^e$, where $0 < T' < T$. It can help to avoid the early steps that move along the high variance axes and focus more on the local target attribute, similar to the discussion about mask estimation. Another reason is that a smaller step can boost the editing speed. For clarity, the latent code $\mathbf{z}_{T'}^e$ is denoted as $\mathbf{w}_{T'}^e$ during denoising, while retaining the representation $\mathbf{z}_{T'}^e$ during noise addition. Different from the mask estimation, we replaced the masked region of $\mathbf{w}_t^e$ with $\mathbf{z}_t^e$ in each denoising step,

$$\mathbf{w'}_t^e = M \odot \mathbf{w}_t^e + (1-M) \odot \mathbf{z}_t^e. \tag{16}$$

The full pipeline is simply represented as,

$$\mathbf{w}_{\text{edit}} = D\left(\mathbf{w}_{T'}^e, y_{tgt}, M\right). \tag{17}$$

The editing direction discovered in a specific instance can also be generalized to other instances and produce similar attribute editing,

$$\Delta\hat{\mathbf{w}} = \mathbf{w}_{\text{edit}} - \mathbf{w}_{\text{e}}, \tag{18}$$

$$I_{\text{edit}} = \mathcal{G}\left(\mathbf{w}_{\text{input}} + \alpha\Delta\hat{\mathbf{w}}\right). \tag{19}$$

IV. Experiments

A. Implementation Details

Training Details. To train our diffusion model, we perform offline synthesis of 1,000,000 images using $\mathbf{w}$ codes obtained from the pre-trained EG3D [7]. Subsequently, we extract the features of these images using CLIP [37]. Similar to previous works [85], [91], we incorporate classifier-free guidance and pseudo-text embeddings during training to mitigate overfitting and enhance diversity. The training procedure for the diffusion model requires approximately 3 days on FFHQ dataset [15] and 2 days on AFHQ dataset [95]. The implementation is executed on 1 Nvidia A6000 GPU.

3D GAN Inversion. To perform editing on real-world images, we employ both an encoder-based method GOAE [96], and an optimization method PTI [97] for GAN inversion. We choose the superior outcome from these two inversion methods for each specific case.

Baselines. We compare our proposed method with seven face editing approaches that utilize text guidance. **Diffusion-based:** InstructPix2Pix [68], MagicQuill [78], (Flux-based) FlowEdit [81] **GAN-based:** StyleGAN-Nada [35], StyleFeatureEdit [56] and StyleRes [55]. **Hybrid model-based:** InstructPix2NeRF [86]. Among these, InstructPix2NeRF is the most closely related to our method, as it combines a diffusion model with a 3D-aware GAN. StyleGAN-Nada is a fine-tuning-based method that optimizes the parameters of a pre-trained generator supervised by the CLIP directional loss. Besides, InstructPix2Pix [68], MagicQuill [78], (Flux-based) FlowEdit [81], StyleFeatureEdit [56] and StyleRes [55] are 2D editing methods. We incorporate the same 3D GAN Inversion [96], [97] for these methods to compare with other 3D methods. Here we list detailed parameters of diffusion-based methods. **MagicQuill**, Sampler Name: Euler ancestral, Steps: 20 and CFG: 5. **(Flux-based) FlowEdit**, Steps: 28, source CFG: 1.5 and target CFG: 5.5.

B. Qualitative Evaluation

Appearance Comparison. In this section, we conduct comparison experiments between our methods and several text-guided editing methods. Based on experience, we sequentially present four attributes, each representing a different level of editing difficulty: beard for Easy, sunglasses for Normal, hair for Hard, and cap for Extreme. Among these methods, StyleGAN-Nada and InstructPix2NeRF are pure 3D editing, and others are 2D editing mixing 3D GAN inversion. The combination of StyleGAN-Nada [35] with EG3D enables color editing, while it encounters challenges in geometry manipulation. Besides, the overall face color slightly changes. Although InstructPix2NeRF [86] can edit most attributes, it still fails on the most challenging attribute and identity preservation. In contrast, our method achieves text-consistent and natural editing results under arbitrary text prompts while maintaining identity and disentanglement. For 2D editing methods + 3D Inversion, it is significant to clarify the difference between direct manipulation on 3D space and 2D editing + 3D inversion. 2D editing mainly focuses on appearance and ignores 3D geometry. As shown in Fig. 4, although some 2D editing methods such as InstrcutPix2Pix [68] and MagicQuill [78] can produce photo-realistic and text-consistent results in the frontal view, they cannot provide a plausible 3D shape. It results in the loss of details and ambiguity in appearance, *e.g.*, blurry blue hair instead of a blue cap. The following 3D GAN inversion exacerbates the ambiguity and produces bad results from side views. StyleFeatureEdit [56] is a StyleGAN-based method that contributes to a faithful reconstruction for complex faces, while its editing is based on the pre-defined directions of the e4e encoder [98]. Therefore, it cannot handle special attributes such as sunglasses and caps.

Comparison on Real Persons. We provide a comparison on real persons with several 2D editing methods, includ-

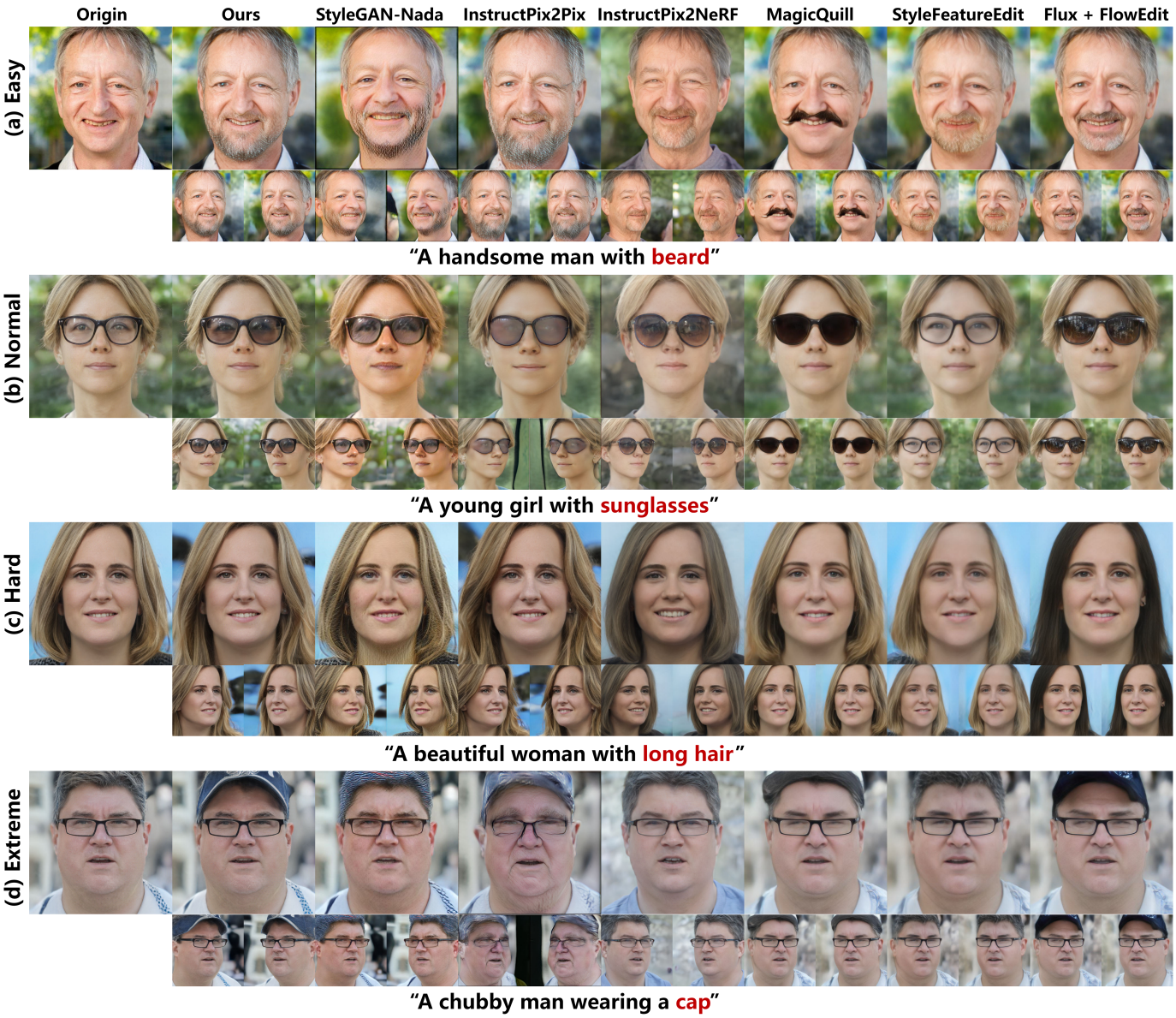

Fig. 4: Qualitative comparisons with text-guided face editing methods. (a), (b), (c), and (d) are the cases with empirically different difficulties, *i.e.*,easy for beard, normal for sunglasses, hard for hair, and extreme for cap. Our method achieves better results in most cases compared to the other six methods , with higher text consistency and stronger identity preservation.

ing diffusion-based MagicQuill [78], GAN-based StyleFeatureEditor [56], Flux-based FlowEdit [81] and GAN-based StyleRes [55]. As shown in Fig. 5, our method can also achieve better editing results than other methods. GAN-based methods [55], [56] can well handle those principle attributes, *e.g.*"wearing glasses", but fail on the attributes out of principle scopes, *e.g.*"black lips". The Flux-based method [81] shows identity shifts in some cases, such as "girl with red hair" and "with glasses". The latest diffusion method MagicQuill [78] achieves good attributes disentanglement and wide editing scope, but it shows artifacts in some details, such as eyebrows for the man wearing glasses in the fourth case. Besides, the flux-based and diffusion-based methods are unable to manipulate the shape such as a fatter face. However, it must be said that real-world faces are sensitive to the inversion method which is not our main focus, and it shows some identity shifts during inversion in real-world cases. Particularly, it shows severe identity shifts in the fifth case with a large side pose. It owes to the GAN inversion that can hardly recover the front face from a large side pose.

Geometry Comparison. Moreover, an additional comparison of geometric manipulation is made between our work and one of the SOTA 3D dragging methods, FaceEdit3D [34] (CVPR2024). As shown in Fig. 7, we can achieve more natural results in terms of geometry manipulation compared with FaceEdit3D. Our method also maintains better identity consistency, *e.g.*, the case of "Open mouth". It is important to note that FaceEdit3D is a warping-based method that

Fig. 5: Qualitative comparisons for real persons. Our method achieves better results in most cases compared to the other three methods, with higher text consistency and stronger identity preservation.

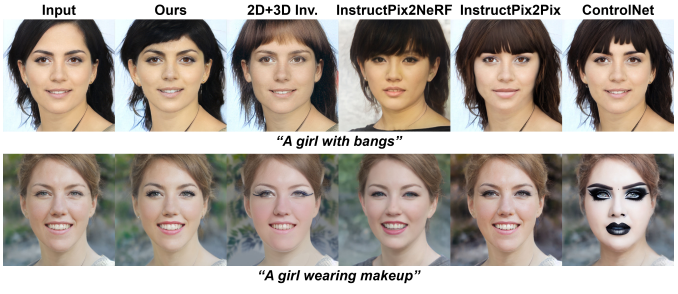

Fig. 6: Qualitative comparisons with additional 2D methods.

only focuses on geometry manipulation and does not support texture editing. In contrast, our method has the capacity to edit both the geometry and texture with arbitrary attribute descriptions. To demonstrate the geometric changes, we provide two synthetic cases with before-and-after editing results for visualization in Fig. 8. The geometric difference between eyeglasses and sunglasses can provide a better understanding of the 3D role, *i.e.*,the consistent change between appearance and geometry.

C. Quantitative Evaluation

In quantitative experiments, we choose 10 editing directions, each of which samples 5 identities, to evaluate the efficiency, text consistency, ID consistency, and attribute disentanglement. We adopt the editing time as the metric for the efficiency measurement. For real-world image editing, we have subtracted the time spent on the GAN inversion for all methods. As shown in Tab. I, we compare CLIP similarity and ID similarity to measure the effectiveness of text-guided editing. The highest scores achieved by our methods in both metrics demonstrate our ability to balance text consistency and ID preservation, which are the most essential aspects of the text-guided editing task. We further adopt the pixel-wise MSE outside the target region to evaluate the attribute disentanglement. Our method also achieves the lowest error in irrelevant regions. The best performance across all these evaluation metrics proves the superiority of our methods. For editing time, two GAN-based methods achieve the fastest speed, but their editing scopes are limited to the pre-defined principle attributes and they fail to handle some corner attributes such as sunglasses and cap.

As it is hard to directly evaluate the geometric results during

TABLE I: Quantitative comparison of efficiency and quality. We also conduct an ablation study to analyze the effect of ϵ mask.

Methods	Editing Time ↓	CLIP Similarity ↑	ID Similarity ↑	MSE_o ↓
MagicQuill [78] (November, 2024)	5s	0.270	0.766	0.035
StyleFeatureEditor [56]	**1s**	0.235	-	0.031
StyleRes [56]	**1s**	0.228	-	0.039
Instructpix2pix [68]	92s	0.259	0.739	0.043
StyleGAN-Nada [35] (SIGGRAPH2023)	107s + 1s	0.258	0.754	0.056
InstructPix2NeRF [86] (ICLR2024)	45s	0.264	0.764	0.037
Ours (ϵ mask)	30 × 0.3s	**0.273**	**0.830**	**0.031**
Ours (W mask)	10s	0.262	0.788	0.041
Ours ($\mathbf{w}$ swap)	5s	0.251	0.735	0.043
Ours (w/o mask)	5s	0.270	0.598	0.068

TABLE II: Quantitative comparison of multi-view results.

Methods	3D CLIP Score ↑	Multi-view ID Sim. ↑
2D + 3D Inv.	0.252	0.517
StyleGAN-Nada	0.238	0.533
InstructPix2NeRF	0.249	0.468
Ours	**0.255**	**0.560**

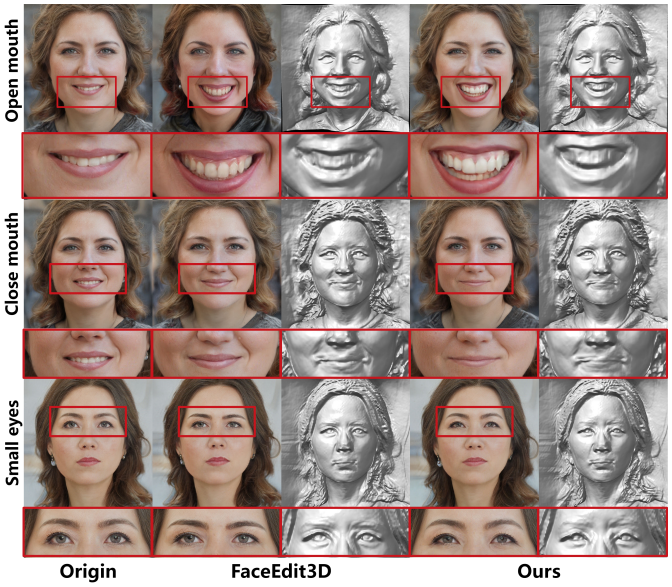

Fig. 7: Qualitative comparisons with FaceEdit3D [34] on geometry manipulation.

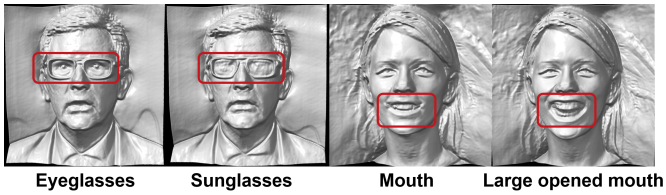

Fig. 8: Geometric changes after editing face attributes.

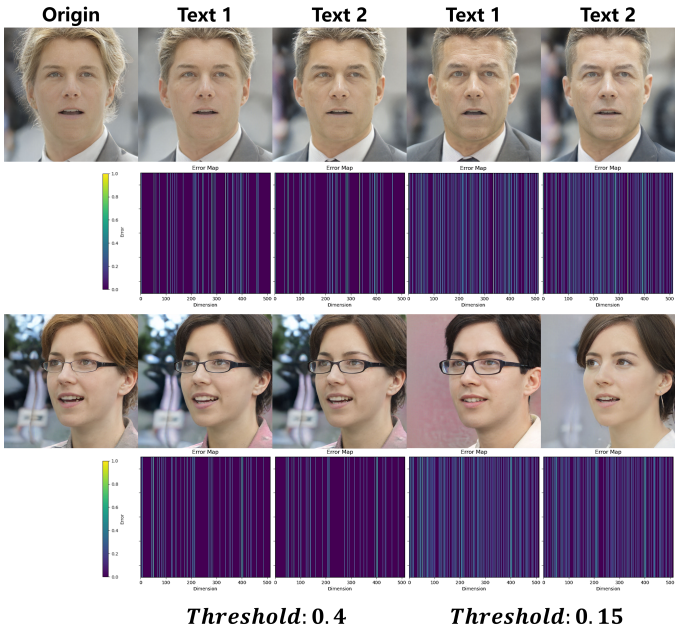

Fig. 9: Mask estimation from different text pairs. The paired texts 1 in the first example are *"a person with short hair"* and *"a person with long hair"*. The paired texts 2 in the first example are *"a long-haired man"* and *"a short-haired man"*. The paired texts 1 in the second example are *"a smart girl with blonde hair"* and *"a smart girl with black hair"*. The paired texts 2 in the second example are *"a blond person"* and *"a black-haired person"*. Please zoom in for detailed observation on mask distribution.

editing, here we additionally provide the 3D quantitative results under multi-view metrics. The metrics are similar to those in single-view quantitative results, except that we render images from two random side views and measure the similarity of those two views. For 2D editing methods, their geometry results are always determined by the 3D inversion, so we only adopt the best results of these 2D methods to compare. As shown in Tab. II, our method also outperforms others in multi-view evaluations, demonstrating the superiority of our methods in 3D space.

D. Ablation Study

Text-guided Mask Estimation. To investigate the robustness of our text-guided mask estimation, we conduct an ablation study to demonstrate the variation of an attribute's mask on latent code when using different text pairs with similar meanings. For example in Fig. 9, we use *"a person with short hair"* and *"a short-haired man"* as different target text prompts to identify the latent mask for editing hair length manipulation. We visually represent the latent masks as error maps together with their corresponding edited faces. As illustrated in Fig. 9, they exhibit nearly the same latent masks and editing results with different target prompts within

Fig. 10: The ablation study of the latent mask.

an appropriate threshold. It proves that the relative direction of paired texts is stable and effective in estimating the mask. The experiment also reveals an interesting phenomenon that the results in low threshold include obvious changes in age besides hair length. This can be attributed to our generative mapper's tendency to associate certain attributes according to data distribution, *e.g.*, sparse hair and old age, thick hair and youthfulness. The experiment shows that our estimated mask serves as a filter, eliminating these associated attributes while preserving the primary attribute. A similar trend is observed in the hair color editing.

Mask-guided Direction Discovery. We further conduct an ablation study on the effect of different ways to apply the mask. Compared with mask-guided ways, results without mask guidance expose inconsistencies in identity and other irrelevant attributes, *e.g.*, hairstyles. Directly applying the first-step $\mathbf{w}$ direction to the masked region proves to be effective, benefiting from the accurate estimation of the mask. However, to precisely filter out text-irrelevant attributes, the magnitude of change in the primary attribute is constrained, as shown in the *"Age"* case in Fig. 10. Despite the scalability of the editing direction, a large scale leads to the interference of other attributes due to the incomplete disentanglement of this direction, as shown in the *"Beard"* case in Fig. 10. Finally, we compare the masks estimated from the paired predicted noises with those from the paired latent codes. The difference between paired latent codes can be understood as the cumulative bias of full-step noises. It contains more global changes, as the initial steps primarily contribute to global synthesis, while the subsequent steps primarily contribute to local generation. Compared to masks from the paired latent codes, masks from the last few noises show better disentanglement. We additionally conduct a quantitative comparison in Tab. I to support the analysis.

E. Applications

Multiple-attributes Editing. Our method supports both continuous single-attribute editing and simultaneous multiple-attribute editing. As shown in Fig. 11, we gradually edit the original face with *"without eyeglasses"*, *"serious"*, and *"blond hair"*. The results demonstrate our capacity for ID preservation and attribute disentanglement during continuous editing. We conduct a comparison using a multi-concept prompt to directly edit multiple attributes, *i.e.*,*"A serious girl without eyeglasses"*

Fig. 11: The illustration of continuous single-attribute editing and simultaneous multi-attribute editing.

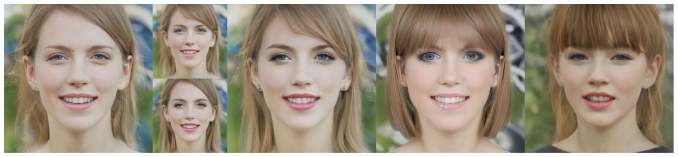

Fig. 12: Comparison of multi-attribute editing.

and *"A serious blonde girl without eyeglasses"*. Despite minimal entanglement, the natural results prove that our method can support the simultaneous editing of three attribute targets in a text prompt. We also provide a comparison of multi-attribute editing with other 3D methods. As shown in Fig. 12, our method better keeps the identity during editing multiple attributes than other methods.

Generalization to Other Generators. Our method can be generalized to other generators, *e.g.*, cat faces or full heads. As shown in Fig. 13, the generated samples exhibit diversity and photo-realism, while edited samples are identity-consistent. The experiment demonstrates that our architecture can generate and manipulate full heads and cat faces on other pre-trained generators, consistent with the user-provided text guidance.

V. Conclusion

In this paper, we propose Face Clan, a fast and generalized method for text-guided 3D face editing. Compared to previous text-guided 3D face editing, our method does not require optimization but can handle arbitrary attribute descriptions. Leveraging a pre-trained 3D-aware GAN, we construct a self-supervised diffusion model to align the text distribution with the latent manifolds of the GAN. It serves as a mapping network that projects noises to semantic meaningful latent codes, while its classifier capability empowers us to reveal the editing direction in the latent space. To enhance the disentanglement, we propose estimating a semantic mask in latent space based on pairs of opposite descriptions. The mask significantly preserves the original identity and irrelevant attributes. Experiments demonstrate the efficiency, generalization, and effectiveness of our methods in both synthetic and real-world face editing. As we conclude our exploration into

Fig. 13: Generalization to cat faces and full heads.

text-guided image editing, our study well incorporates textual guidance to image manipulation, advancing the frontiers of multimedia innovation and content creation.

Limitations and Future Work. Despite natural and text-consistent results, our method also encounters several limitations. Due to the curved trajectory of DDIM or DDPM sampling, the mask estimation needs to measure the difference of predicted noise in the last few steps which is inefficient. Besides, it leads to up to 5 seconds during denoising, still a long time for users. To solve the problem, we are exploring the possibility of introducing Rectified Flow [99], [100] whose trajectory is a straight line. The straight trajectory may improve the editing speed and stability. Besides, it is hard for our method to manipulate some corner attributes, *e.g.*, wearing a headset and eyepatch, due to the limited capacity of pre-trained 3D-aware GANs.

VI. Acknowledgment

This work was supported in part by NSFC (62201342), and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). The authors would like to thank the Student Innovation Center of SJTU for providing GPUs.

References

[1] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "Graf: Generative radiance fields for 3d-aware image synthesis," in *NIPS*, 2020. 1

[2] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *CVPR*, 2021, pp. 5799–5809. 1

[3] M. Niemeyer and A. Geiger, "Giraffe: Representing scenes as compositional generative neural feature fields," in *CVPR*, 2021, pp. 11 453–11 464. 1

[4] J. Gu, L. Liu, P. Wang, and C. Theobalt, "Stylenerf: A style-based 3d aware generator for high-resolution image synthesis," in *ICLR*, 2021. 1

[5] P. Zhou, L. Xie, B. Ni, and Q. Tian, "Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis," *arXiv preprint arXiv:2110.09788*, 2021. 1

[6] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman, "Stylesdf: High-resolution 3d-consistent image and geometry generation," in *CVPR*, 2022, pp. 13 503–13 513. 1

[7] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *CVPR*, 2022, pp. 16 123–16 133. 1, 4, 6

[8] Y. Xu, S. Peng, C. Yang, Y. Shen, and B. Zhou, "3d-aware image synthesis via learning structural and textural representations," in *CVPR*, 2022. 1

[9] Y. Deng, J. Yang, J. Xiang, and X. Tong, "Gram: Generative radiance manifolds for 3d-aware image generation," in *CVPR*, 2022, pp. 10 673–10 683. 1

[10] J. Xiang, J. Yang, Y. Deng, and X. Tong, "Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds," in *ICCV*, 2023, pp. 2195–2205. 1

[11] I. Skorokhodov, S. Tulyakov, Y. Wang, and P. Wonka, "Epigraf: Rethinking training of 3d gans," *NeurIPS*, pp. 24 487–24 501, 2022. 1

[12] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen *et al.*, "Rodin: A generative model for sculpting 3d digital avatars using diffusion," in *CVPR*, 2023, pp. 4563–4573. 1, 3

[13] S. An, H. Xu, Y. Shi, G. Song, U. Y. Ogras, and L. Luo, "Panohead: Geometry-aware 3d full-head synthesis in 360deg," in *CVPR*, 2023, pp. 20 950–20 959. 1

[14] Y. Yichao, C. Yuhao, C. Zhuo, P. Yicong, W. Sijing, Z. Weitian, L. Junjie, L. Yixuan, G. Jingnan, Z. Weixia, Z. Guangtao, and Y. Xiaokang, "A survey on generative 3d digital humans based on neural networks: representation, rendering, and learning," *SCIENTIA SINICA Informationis*, pp. 1858–, 2023. 1

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410. 1, 6

[16] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*, 2020, pp. 8110–8119. 1

[17] Y. Shen, C. Yang, X. Tang, and B. Zhou, "Interfacegan: Interpreting the disentangled face representation learned by gans," *TPAMI*, pp. 2004–2018, 2020. 1, 2

[18] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows," *TOG*, pp. 1–21, 2021. 1, 2

[19] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *CVPR*, 2021, pp. 2085–2094. 1, 2, 3

[20] E. Simsar, A. Tonioni, E. P. Ornek, and F. Tombari, "Latentswap3d: Semantic edits on 3d image gans," in *ICCV*, 2023, pp. 2899–2909. 1, 2

[21] Q. Song, J. Li, S. Wu, and H.-S. Wong, "A graph-based discriminator architecture for multi-attribute facial image editing," *IEEE Transactions on Multimedia*, vol. 26, pp. 436–446, 2024. 1

[22] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," in *ICML*, 2020, pp. 9786–9796. 1, 2

[23] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," *NeurIPS*, pp. 9841–9850, 2020. 1, 2

[24] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," in *CVPR*, 2021, pp. 1532–1540. 1, 2

[25] J. Zhu, R. Feng, Y. Shen, D. Zhao, Z.-J. Zha, J. Zhou, and Q. Chen, "Low-rank subspaces in gans," *NeurIPS*, pp. 16 648–16 658, 2021. 1, 2

[26] J. Zhu, Y. Shen, Y. Xu, D. Zhao, and Q. Chen, "Region-based semantic factorization in gans," in *ICML*, 2022, pp. 27 612–27 632. 1, 2

[27] J. Zhu, C. Yang, Y. Shen, Z. Shi, B. Dai, D. Zhao, and Q. Chen, "Linkgan: Linking gan latents to pixels for controllable image synthesis," in *ICCV*, 2023, pp. 7656–7666. 1, 2

[28] J. Sun, X. Wang, Y. Shi, L. Wang, J. Wang, and Y. Liu, "Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis," *ToG*, pp. 1–10, 2022. 2, 3

[29] K. Jiang, S.-Y. Chen, F.-L. Liu, H. Fu, and L. Gao, "Nerffaceediting: Disentangled face editing in neural radiance fields," in *SIGGRAPH Asia*, 2022, pp. 1–9. 2, 3

[30] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Stylerig: Rigging stylegan for 3d control over portrait images," in *CVPR*, 2020, pp. 6142–6151. 2, 3

[31] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu, "Next3d: Generative neural texture rasterization for 3d-aware head avatars," in *CVPR*, 2023, pp. 20 991–21 002. 2, 3

[32] Y. Endo, "User-controllable latent transformer for stylegan image layout editing," in *Computer Graphics Forum*, 2022, pp. 395–406. 2, 3

[33] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your gan: Interactive point-based manipulation on the generative image manifold," in *ASIGGRAPH*, 2023, pp. 1–11. 2, 3

[34] C. Yuhao, C. Zhuo, R. Xingyu, Z. Wenhan, X. Zhengqin, X. Di, Y. Changpeng, and Y. Yichao, "3d-aware face editing via warping-guided latent direction learning," in *CVPR (CVPR)*, 2024. 2, 3, 7, 9

[35] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *TOG*, pp. 1–13, 2022. 2, 3, 6, 9

[36] Z. Chen, X. Xu, Y. Yan, Y. Pan, W. Zhu, W. Wu, B. Dai, and X. Yang, "Hyperstyle3d: Text-guided 3d portrait stylization via hypernetworks," *arXiv preprint arXiv:2304.09463*, 2023. 2

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763. 2, 3, 5, 6

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695. 2, 4

[39] B. Li, Y.-k. Li, Z.-f. He, B. Liu, and Y.-K. Lai, "3d-aware image generation and editing with multi-modal conditions," *arXiv preprint arXiv:2403.06470*, 2024. 2, 3

[40] J. Zhang, Y. Zhou, Q. Zheng, X. Du, G. Luo, J. Peng, X. Sun, and R. Ji, "Fast text-to-3d-aware face generation and manipulation via direct cross-modal mapping and geometric regularization," *arXiv preprint arXiv:2403.06702*, 2024. 2, 3

[41] C. Yu, G. Lu, Y. Zeng, J. Sun, X. Liang, H. Li, Z. Xu, S. Xu, W. Zhang, and H. Xu, "Towards high-fidelity text-guided 3d face generation and manipulation using only images," in *ICCV*, 2023, pp. 15 326–15 337. 2, 3

[42] C. Zhang, Y. Chen, Y. Fu, Z. Zhou, G. Yu, B. Wang, B. Fu, T. Chen, G. Lin, and C. Shen, "Styleavatar3d: Leveraging image-text diffusion models for high-fidelity 3d avatar generation," *arXiv preprint arXiv:2305.19012*, 2023. 2, 3

[43] X. Hou, X. Zhang, Y. Li, and L. Shen, "Textface: Text-to-style mapping based face generation and manipulation," *IEEE Transactions on Multimedia*, vol. 25, pp. 3409–3419, 2023. 2, 3

[44] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021. 2, 3

[45] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," *arXiv preprint arXiv:2210.11427*, 2022. 2, 3

[46] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *CVPR*, 2022, pp. 18 208–18 218. 2, 3

[47] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *CVPR*, 2022, pp. 11 461–11 471. 2, 3

[48] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022. 2, 3

[49] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *CVPR*, 2023, pp. 6007–6017. 2, 3

[50] J. Kim, C. Oh, H. Do, S. Kim, and K. Sohn, "Diffusion-driven gan inversion for multi-modal face image generation," in *CVPR*, 2024, pp. 10 403–10 412. 2

[51] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020. 2

[52] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2020. 2

[53] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*, 2021, pp. 8162–8171. 2

[54] J. Huang, J. Liao, and S. Kwong, "Unsupervised image-to-image translation via pre-trained stylegan2 network," *IEEE Transactions on Multimedia*, vol. 24, pp. 1435–1448, 2022. 2

[55] H. Pehlivan, Y. Dalva, and A. Dundar, "Styleres: Transforming the residuals for real image editing with stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1828–1837. 2, 6, 7

[56] D. Bobkov, V. Titov, A. Alanov, and D. Vetrov, "The devil is in the details: Stylefeatureeditor for detail-rich stylegan inversion and high quality image editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9337–9346. 2, 6, 7, 9

[57] Y. Zhu, W. Zhao, Y. Tang, Y. Rao, J. Zhou, and J. Lu, "Stableswap: Stable face swapping in a shared and controllable latent space," *IEEE Transactions on Multimedia*, vol. 26, pp. 7594–7607, 2024. 3

[58] J. Yang, C. Cheng, S. Xiao, G. Lan, and J. Wen, "High fidelity face-swapping with style convtransformer and latent space selection," *IEEE Transactions on Multimedia*, vol. 26, pp. 3604–3615, 2024. 3

[59] J. Wang, P. Liu, J. Liu, and W. Xu, "Text-guided eyeglasses manipulation with spatial constraints," *IEEE Transactions on Multimedia*, vol. 26, pp. 4375–4388, 2024. 3

[60] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in *ICCV*, 2017, pp. 5706–5714. 3

[61] S. Nam, Y. Kim, and S. J. Kim, "Text-adaptive generative adversarial networks: manipulating images with natural language," *NeurIPS*, vol. 31, 2018. 3

[62] Y. Liu, M. De Nadai, D. Cai, H. Li, X. Alameda-Pineda, N. Sebe, and B. Lepri, "Describe what to change: A text-guided unsupervised image-to-image translation approach," in *ACM MM*, 2020, pp. 1357–1365. 3

[63] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr, "Manigan: Text-guided image manipulation," in *CVPR*, 2020, pp. 7880–7889. 3

[64] Y. Yu, R. Wu, Y. Men, S. Lu, M. Cui, X. Xie, and C. Miao, "Morphnerf: Text-guided 3d-aware editing via morphing generative neural radiance fields," *IEEE Transactions on Multimedia*, vol. 26, pp. 8516–8528, 2024. 3

[65] X. Du, J. Peng, Y. Zhou, J. Zhang, S. Chen, G. Jiang, X. Sun, and R. Ji, "Pixelface+: Towards controllable face generation and manipulation with text descriptions and segmentation masks," in *ACM MM*, 2023, pp. 4666–4677. 3

[66] J. Peng, X. Du, Y. Zhou, J. He, Y. Shen, X. Sun, and R. Ji, "Learning dynamic prior knowledge for text-to-face pixel synthesis," in *ACM MM*, 2022, pp. 5132–5141. 3

[67] J. Peng, H. Pan, Y. Zhou, J. He, X. Sun, Y. Wang, Y. Wu, and R. Ji, "Towards open-ended text-to-face generation, combination and manipulation," in *ACM MM*, 2022, pp. 5045–5054. 3

[68] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023, pp. 18 392–18 402. 3, 6, 9

[69] K. Feng, Y. Ma, B. Wang, C. Qi, H. Chen, Q. Chen, and Z. Wang, "Dit4edit: Diffusion transformer for image editing," *arXiv preprint arXiv:2411.03286*, 2024. 3

[70] Z. Geng, B. Yang, T. Hang, C. Li, S. Gu, T. Zhang, J. Bao, Z. Zhang, H. Li, H. Hu *et al.*, "Instructdiffusion: A generalist modeling interface for vision tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 709–12 720. 3

[71] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "Magicbrush: A manually annotated dataset for instruction-guided image editing," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 3

[72] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman, "Emu edit: Precise image editing via recognition and generation tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8871–8879. 3

[73] Y. Huang, L. Xie, X. Wang, Z. Yuan, X. Cun, Y. Ge, J. Zhou, C. Dong, R. Huang, R. Zhang *et al.*, "Smartedit: Exploring complex instruction-based image editing with multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8362–8371. 3

[74] J. Singh, J. Zhang, Q. Liu, C. Smith, Z. Lin, and L. Zheng, "Smartmask: Context aware high-fidelity mask generation for fine-grained object insertion and layout control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6497–6506. 3

[75] Y. Yang, H. Peng, Y. Shen, Y. Yang, H. Hu, L. Qiu, H. Koike *et al.*, "Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 3

[76] C. Mou, X. Wang, J. Song, Y. Shan, and J. Zhang, "Dragondiffusion: Enabling drag-style manipulation on diffusion models," *arXiv preprint arXiv:2307.02421*, 2023. 3

[77] C. Zhang, W. Yang, X. Li, and H. Han, "Mmginpainting: Multi-modality guided image inpainting based on diffusion models," *IEEE Transactions on Multimedia*, vol. 26, pp. 8811–8823, 2024. 3

[78] Z. Liu, Y. Yu, H. Ouyang, Q. Wang, K. L. Cheng, W. Wang, Z. Liu, Q. Chen, and Y. Shen, "Magicquill: An intelligent interactive image editing system," *arXiv preprint arXiv:2411.09703*, 2024. 3, 6, 7, 9

[79] W. Mao, B. Han, and Z. Wang, "Sketchffusion: Sketch-guided image editing with diffusion model," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 790–794. 3

[80] C. Xiao and H. Fu, "Customsketching: Sketch concept extraction for sketch-based image synthesis and editing," in *Computer Graphics Forum*. Wiley Online Library, 2024, p. e15247. 3

[81] V. Kulikov, M. Kleiner, I. Huberman-Spiegelglas, and T. Michaeli, "Flowedit: Inversion-free text-based editing using pre-trained flow models," *arXiv preprint arXiv:2412.08629*, 2024. 3, 6, 7

[82] X. Shen, J. Ma, C. Zhou, and Z. Yang, "Controllable 3d face generation with conditional style code diffusion," in *AAAI*, 2024, pp. 4811–4819. 3

[83] B. Lei, K. Yu, M. Feng, M. Cui, and X. Xie, "Diffusiongan3d: Boosting text-guided 3d generation and domain adaption by combining 3d gans and diffusion priors," *arXiv preprint arXiv:2312.16837*, 2023. 3

[84] T. Kirschstein, S. Giebenhain, and M. Nießner, "Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars," *arXiv preprint arXiv:2311.18635*, 2023. 3

[85] Y.-J. Li, T. Xu, J. Hou, B. Wu, X. Dai, A. Pumarola, P. Zhang, P. Vajda, and K. Kitani, "3d-clfusion: Fast text-to-3d rendering with contrastive latent diffusion," *arXiv preprint arXiv:2303.11938*, 2023. 3, 5, 6

[86] J. Li, S. Liu, Z. Liu, Y. Wang, K. Zheng, J. Xu, J. Li, and J. Zhu, "Instructpix2nerf: Instructed 3d portrait editing from a single image," *arXiv preprint arXiv:2311.02826*, 2023. 3, 6, 9

[87] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, pp. 1–39, 2023. 4

[88] J. Chen, Y. Jincheng, G. Chongjian, L. Yao, E. Xie, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li, "Pixart: Fast training of diffusion transformer for photorealistic text-to-image synthesis," in *ICLR*, 2023. 4

[89] J. Gu, Q. Gao, S. Zhai, B. Chen, L. Liu, and J. Susskind, "Learning controllable 3d diffusion models from single-view images," *arXiv preprint arXiv:2304.06700*, 2023. 5

[90] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023, pp. 19 730–19 742. 5

[91] J. N. Pinkney and C. Li, "clip2latent: Text driven sampling of a pre-trained stylegan using denoising diffusion and clip," *arXiv preprint arXiv:2210.02347*, 2022. 5, 6

[92] B. Wang and J. J. Vastola, "Diffusion models generate images like painters: an analytical theory of outline first, details later," *arXiv preprint arXiv:2303.02490*, 2023. 5

[93] Z. Huang, T. Wu, Y. Jiang, K. C. Chan, and Z. Liu, "Reversion: Diffusion-based relation inversion from images," *arXiv preprint arXiv:2303.13495*, 2023. 5

[94] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang, "Svdiff: Compact parameter space for diffusion fine-tuning," in *ICCV*, 2023, pp. 7323–7334. 5

[95] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *CVPR*, 2020, pp. 8188–8197. 6

[96] Z. Yuan, Y. Zhu, Y. Li, H. Liu, and C. Yuan, "Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding," in *ICCV*, 2023, pp. 2437–2447. 6

[97] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *TOG*, 2021. 6

[98] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021. 6

[99] X. Liu, C. Gong *et al.*, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *ICLR*, 2022. 11

[100] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *ICLR*, 2022. 11