High-Fidelity 3D Facial Avatar Synthesis with Controllable Fine-Grained Expressions

Yikang He, Jichao Zhang, Wei Wang, *Member, IEEE*, Nicu Sebe, *Senior Member, IEEE*, and Yao Zhao, *Fellow, IEEE*,

***Abstract*—Facial expression editing methods can be mainly categorized into two types based on their architectures: 2D-based and 3D-based methods. The former lacks 3D face modeling capabilities, making it difficult to edit 3D factors effectively. The latter has demonstrated superior performance in generating high-quality and view-consistent renderings using single-view 2D face images. Although these methods have successfully used animatable models to control facial expressions, they still have limitations in achieving precise control over fine-grained expressions. To address this issue, in this paper, we propose a novel approach by simultaneously refining both the latent code of a pretrained 3D-Aware GAN model for texture editing and the expression code of the driven 3DMM model for mesh editing. Specifically, we introduce a Dual Mappers module, comprising Texture Mapper and Emotion Mapper, to learn the transformations of the given latent code for textures and the expression code for meshes, respectively. To optimize the Dual Mappers, we propose a Text-Guided Optimization method, leveraging a CLIP-based objective function with expression text prompts as targets, while integrating a SubSpace Projection mechanism to project the text embedding to the expression subspace such that we can have more precise control over fine-grained expressions. Extensive experiments and comparative analyses demonstrate the effectiveness and superiority of our proposed method.**

***Index Terms*—3D-Aware Generative Adversarial Networks, Head Avatar Generation, Fine-Grained Expression Editing.**

I. Introduction

FACIAL Expression Editing, aimed at manipulating the expression of a given face image to match a target expression without altering its identity properties, is increasingly prevalent across applications, such as object animation, human-computer interactions, psychological analysis, and augmented reality [1]–[5].

Generative models, including generative adversarial networks (GANs) [6] and diffusion models [7], have revolutionized facial attribute editing by enabling high-quality and flexible image synthesis.

To address this task, early methods [8]–[21] predominantly relied on 2D image translation techniques. However, these approaches often fail to account for 3D facial geometry, limiting their ability to maintain 3D consistency and control pose variations. Consequently, recent advancements have integrated 3D Morphable Face Models (3DMM) [22] with GANs or Neural Radiance Fields (NeRF) to enable finer control.

Despite these improvements, achieving high-fidelity editing with both view consistency and fine-grained expression control remains an open problem. 3D GAN methods, such as AniFaceGAN [23] and Next3D [24], combine volume rendering with mesh-guided deformation. However, they heavily rely on 3D reconstruction priors like DECA [25], which lack rich fine-grained expression variations in their training data, limiting the model's ability to extract and generate subtle expression details.

Furthermore, while diffusion-based methods like DiffusionRig [26] excel in image fidelity, they lack explicit 3D representations, often resulting in compromised multi-view consistency compared to geometry-aware approaches. Although multi-view diffusion methods like Morphable Diffusion [27], incorporating 3D morphable models, can mitigate inconsistency, they typically require extensive multi-view datasets for each individual, restricting their practicality in data-scarce scenarios. Similarly, recent reconstruction-based approaches, such as Gaussian Head Avatars [28] or methods utilizing SDS optimization [29], necessitate training separate models for each individual (per-subject optimization). This requirement makes them computationally inefficient and time-consuming for general applications. In contrast, to address the aforementioned gaps, the primary research problem we tackle in this paper is achieving high-fidelity, fine-grained 3D Facial expression editing from a single image while strictly maintaining multi-view consistency. By leveraging a robust 3D GAN-based framework, we eliminate the need for multi-view data or per-subject training. This approach efficiently ensures view consistency while overcoming the common challenge of uncoordinated texture and geometry updates—where existing methods fail to synergistically match structural mesh deformations with subtle texture changes, often leading to unnatural editing artifacts.

Fine-grained facial expressions, capturing subtle emotional nuances, present challenges in both recognition and generation due to the lack of comprehensive annotations. This scarcity hinders expression editing and generation tasks. To address this, we leverage CLIP [30], a multimodal model trained on image-text pairs. CLIP's ability to generalize across visual and linguistic domains allows it to guide fine-grained expression manipulation through multimodal supervision. By using expression descriptions (*e.g.*, "A person who is raising brow") as supervision via CLIP loss, we can bypass the need for large

Yikang He, Wei Wang, Yao Zhao are with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. E-mail: 23120294@bjtu.edu.cn, wei.wang@bjtu.edu.cn, yzhao@bjtu.edu.cn.

Nicu Sebe is with the Department of Information Engineering and Computer Science (DISI), University of Trento, Italy. E-mail: sebe@disi.unitn.it.

Jichao Zhang is with the School of Computer Science, Ocean University of China, Qingdao, China. E-mail: zhang163220@gmail.com.

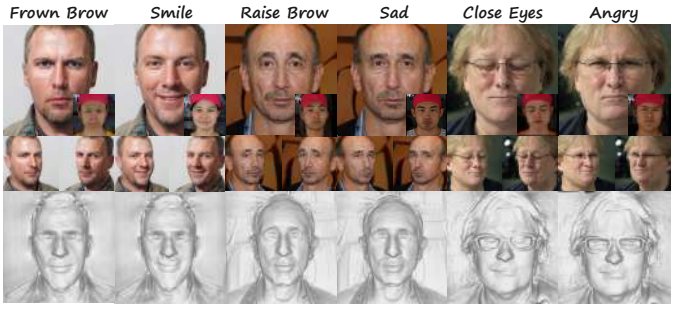


Fig. 1. Our model can generate high-quality, view-consistent face images while enabling fine-grained expression editing. The first row displays the generated images of different facial expression edits (with the reference image located in the lower-right corner), the second row presents different viewpoints, and the third row showcases the generated 3D head meshes.

annotated datasets, enabling the generation of high-quality 3D avatars with precise expressions.

In this paper, to solve the fine-grained expression editing problem, we introduce a novel 3D-aware framework designed to synergistically edit both texture and geometry.

First, given a reference image, we use EMOCA [31], which has a richer expression space, to extract its 3D shape hyper-parameters (*i.e.* FLAME parameters) for initialization. Next, to ensure this coordinated editing, we introduce a **Dual Mappers Module**, comprising a Texture Mapper for refining the UV texture latent code $\boldsymbol{w}$ and an Emotion Mapper for optimizing the expression code $\boldsymbol{\alpha}$ of the mesh. Both components work simultaneously and interact via a cross-attention mechanism. By synchronously updating both spaces, our method ensures that geometric and textural features match coherently, significantly reducing artifacts introduced during the editing process and achieving more accurate expression synthesis.

To train the Dual Mappers, we propose a **Text-Guided Optimization method** that utilizes CLIP-based loss functions, enabling expression refinement based on textual prompts. Note that each textual prompt corresponds to one expression. For each individual expression, we learn one set of specific parameters for Dual Mappers. This method reduces reliance on extensively annotated fine-grained expression datasets by leveraging the generalization capabilities of large pre-trained language-vision models. It effectively bridges the gap between textual descriptions and visual representations, providing a robust mechanism for controlling subtle expression nuances. Furthermore, a Subspace Projection mechanism ensures the preservation of identity attributes, maintaining the fidelity of the synthesized face while allowing fine-grained expression adjustments. As shown in Fig. 1, guided by reference images and target expression texts, our method is capable of generating high-quality multi-view facial images while effectively capturing both the texture and geometric characteristics of the target expression.

By addressing these limitations and leveraging the proposed techniques, our approach advances the field of 3D-aware expression editing and opens new possibilities for realistic and expressive 3D facial synthesis in practical applications.

To sum up, the main contributions are as follows:

- We have proposed a novel Dual Mappers module for face synthesis with fine-grained expression editing which does not require large scale training data.
- The Dual Mappers comprise a Texture Mapper, which refines the latent code of UV textures, and an Emotion Mapper, which learns transformations of the expression parameters of the mesh. Besides, a cross-attention mechanism has been designed for their interactions, and more accurate expression editing can be achieved.
- We have proposed a Text-Guided Optimization for Dual Mappers, utilizing the CLIP-based Loss with the text prompts as the objective function, while integrating a Subspace Projection mechanism to avoid undesired attribute changes.

II. RELATED WORKS

3D-Aware Head-Avatar Generation. Recent advancements in 3D scene representation methods have improved 3D-aware image synthesis for head avatar generation. Many models are trained on 2D images without relying on 3D or multi-view data, using techniques like Explicit Voxel Grids [32], [33], Neural Implicit Representations (NeRF) [34]–[36] or 3D Gaussian Splatting (3DGS) [28], [37], [38]. NeRF-based GANs allow for higher-resolution modeling without visible artifacts. Studies have introduced improvements in articulated object modeling [39], fast rendering [40], and controllability for editing tasks [23].

For 3D avatar generation, leveraging a 3D morphable face model [41], [42] within a 3D generative model enhances the controllability of pose and expression in models [23], [24], [43]–[50]. Specifically, DiscoFaceGAN [44] uses imitative-contrastive learning and the 3D Morphable Model (3DMM) for precise attribute control but struggles with pose-inconsistent issues. AniFaceGAN [23] uses expression-driven deformation fields for realistic avatar generation but faces challenges with fine-grained animation, such as gaze, and low-quality generation in regions like the mouth. These issues arise from the under-constrained nature of the 3DMM and deformation fields. Next3D [24] overcomes these challenges by separating dynamic and static components of the head and modeling them separately, using neural textures for facial deformation and an efficient teeth synthesis module. However, it struggles with expression control due to the limitations of the FLAME [51] and DECA [25] estimation methods.

Facial Expression Editing. Several studies have achieved notable facial expression editing using generative adversarial networks (GANs), broadly categorized into 2D-GAN methods [8], [9], [52]–[54], 3DMM-based methods [44], [55], and NeRF-based methods [56], [57]. 2D-GAN methods incorporate an encoder to map source to target domains for facial attribute manipulation, but they are limited by low-resolution outputs. StyleFlow [14] and StyleCLIP [15] enhance visual quality by manipulating pretrained StyleGAN's latent space but lack 3D understanding. 3DMM-based methods [44] use 3D Morphable Model (3DMM) priors for editing, though they struggle with pose control and view-inconsistency. More recently, NeRF-based methods [56], [57] use semantic masks or 3DMM priors for facial expression control, but fine-grained expression control remains difficult due to limitations in capturing subtle variations in NeRF representations.

Recently, diffusion models [7], [58] have demonstrated superior performance compared to GANs, significantly enhancing image fidelity across diverse tasks. DiffusionRig [26] introduces a two-stage training method, utilizing diffusion to learn personalized facial priors on top of generic face priors, enabling the facial expression editing task. However, due to lacking multiple-view data for training, the utilizing diffusion model fails to preserve the view-consistency. Morphable Diffusion [27] introduces a novel method for avatar creation by integrating a 3D morphable model into a multi-view consistent diffusion framework. However, it heavily relies on large amounts of multi-view image training data, making it difficult to generalize to images outside of the dataset.

3D Face Reconstruction and 3DMM. The 3D Morphable Model (3DMM) was first introduced by Blanz and Vetter [22], using PCA to model 3D face shapes and textures from scan data. Paysan et al. [59] further refined this with the Basel Face Model (BFM), employing advanced scanning technology for higher accuracy. In 2017, Gerig et al. [60] enhanced BFM by adding facial expression modeling. Other advancements included multilinear and non-linear representations for 3DMM [61]–[66], while Li et al. [42] introduced the FLAME model, which accurately simulates head poses and eye rotations.

Reconstructing 3D facial shapes from images has long been a challenge. One approach is regressing 3DMM parameters to generate 3D faces [25], [31], [67]–[70], [70], [71], which allows controlled editing of facial meshes. However, it's difficult to obtain large labeled datasets for 3D meshes, making self-supervised methods more appealing. Deep3D [68] regresses BFM parameters for precise 3D face reconstruction from multi-view images, but it lacks full head and neck modeling. To address this, DECA adopted the FLAME model [42] for improved animation and detail, using a two-stage training process to output animated, detailed shapes from a single image. However, DECA's loss functions were insufficient for capturing facial expressions. EMOCA [31], [72] solved this by employing an expression recognition network and a dedicated encoder for more accurate facial expressions. Building upon this, we refined expression codes using CLIP [30] for fine-grained control over facial expressions.

III. METHOD

In this section, we present our novel framework for fine-grained, text-guided 3D facial expression editing. As illustrated in Fig. 2, our approach is built upon a 3D-aware generative backbone and consists of two core modules designed to bridge the gap between semantic text instructions and 3D geometry control: **(a) Dual Mappers** and **(b) Text-Guided Optimization**.

First, to achieve high-fidelity 3D face modeling with precise expression capture, we adopt Next3D [24] as our synthesis network but integrate EMOCA [31] to replace the standard coarse expression estimator. This modification allows for more accurate extraction of fine-grained expression parameters from reference images. Second, we introduce the *Dual Mappers* module designed to explicitly edit both the latent style code $\boldsymbol{w}$ and the expression parameters $\boldsymbol{\alpha}$. By employing a cross-attention mechanism, this module modulates these representations to ensure that both texture and geometric deformations are coherently altered to satisfy target expressions. Finally, to train the Dual Mappers for accurate expression editing, we propose a *Text-Guided Optimization* strategy. Specifically, to interpret textual descriptions (e.g., "A person who is raising brow"), we utilize a subspace projection technique to disentangle expression-related attributes from identity features. This optimization objective guides the Dual Mappers to precisely modify facial expressions according to the text prompts while preserving the subject's identity.

A. Mesh-Guided 3D-aware Head Avatar

1) 3D Face Modeling: In this work, we employ the FLAME [51] model as 3D prior for the generation and manipulation of various geometric shapes and fine-grained facial expressions. Taking a set of identity shape parameters $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$, pose parameters $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$, and facial expression parameters $\boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{\alpha}|}$ as input, FLAME outputs a corresponding mesh and is defined as:

$$M(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \rightarrow (\mathbf{V}, \mathbf{F}) \tag{1}$$

with n_v = 5023 vertices $\mathbf{V} \in \mathbb{R}^{n_v \times 3}$ and n_f = 9976 faces $\mathbf{F} \in \mathbb{R}^{n_f \times 3}$. To reconstruct FLAME parameters from a single image, we employ pretrained EMOCA [31] as the parameter extractor instead of DECA [25] utilized in Next-3D [24]. Trained on Affectnet [73], a large-scale annotated emotion dataset, EMOCA incorporates an additional expression encoder to reconstruct expression parameters $\boldsymbol{\alpha}$. This enhances its capability to capture fine-grained expressions and allows for more precise reconstruction of the required FLAME parameters. Given an image I, EMOCA outputs FLAME geometry parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, expressed as follows:

$$E(I) \rightarrow (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \tag{2}$$

Using the above method enables the 3D face modeling from a single image while simultaneously allowing for precise representation of fine-grained emotions.

2) Synthesis Network: As mentioned above, we adopt the pretrained Next3D [24] as the Synthesis Network, as shown in Fig. 3. Next3D employs a StyleGAN2 CNN generator G_{uv} to generate UV textures. These textures, combined with the 3D mesh obtained from the FLAME model through rasterization rendering, subsequently transform into neural texture tri-planes. Moreover, the final image is generated using volume rendering and a 2D super-resolution module. Specifically, it starts by sampling the latent code $\boldsymbol{z} \sim N(0, 1)$ and then employs a Mapping Network to obtain an intermediate latent code $\boldsymbol{w}$ which is mapped into uv texture features F_{uv} by G_{uv}:

$$F_{uv} = G_{uv}(\boldsymbol{w}) \tag{3}$$

Then, Next3D uses DECA [25] to reconstruct the input image to obtain the 3D mesh M. However, as illustrated in Fig. 2, we employ EMOCA [31] to perform this operation. Employing rasterization techniques R, the frontal, lateral, and top views of this mesh are rendered onto the generated UV texture maps. This leads to the generation of neural texture tri-planes:

$$F_{tri} = R(F_{uv}, M(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha})) \tag{4}$$

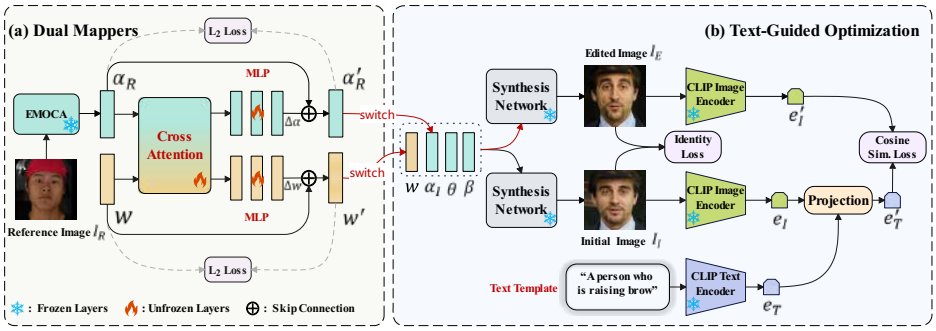


Fig. 2. Overview of the proposed architecture which consists of two main modules: **(a) Dual Mappers** and **(b) Text-Guided Optimization**. In (a), *Dual Mappers* refine the random latent code $\boldsymbol{w}$ and expression code $\boldsymbol{\alpha_R}$ from a reference image $\boldsymbol{I_R}$ using a Cross Attention module and MLP layers, with updates stabilized by skip connections and a L_2 **Loss**. The refined codes $\boldsymbol{w'}$ and $\boldsymbol{\alpha'_R}$ are passed to a frozen synthesis network to generate the edited image. (The red arrow represents replacing $\boldsymbol{w}$,$\boldsymbol{\alpha_I}$ with $\boldsymbol{w'}$,$\boldsymbol{\alpha'_R}$.) In (b), *Text-Guided Optimization* aligns the edited image with a text description (*e.g.*, "A person who is raising brow"). A **Projection** module maps the text embedding $\boldsymbol{e_T}$ to the image embedding space, enhancing compatibility. **Cosine Similarity Loss** aligns the embeddings, while **Identity Loss** preserves facial identity.

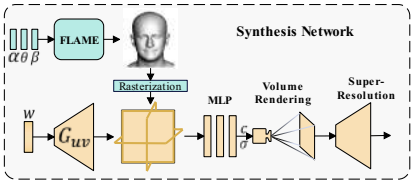


Fig. 3. The *Synthesis Network* utilizes Tri-Plane and Volume Rendering to achieve 3D-aware head avatar generation. The generator $\boldsymbol{G_{uv}}$ produces a Tri-Plane from the latent code $\boldsymbol{w}$, while FLAME accepts $\boldsymbol{\alpha}$,$\boldsymbol{\theta}$,$\boldsymbol{\beta}$ as input and provides a coarse head mesh aligned via rasterization. An MLP predicts color $\boldsymbol{c}$ and density $\boldsymbol{\sigma}$, which are used for **Volume Rendering**. Finally, a **Super-Resolution** module enhances the output, creating realistic, editable 3D avatars.

Next, by sampling features from the tri-planes F_{tri}, the density σ and color c can be obtained with a mapping network. To conserve computational resources, Next3D first utilizes volume rendering with these two parameters to generate a low-resolution image, which is then processed through a super-resolution module to generate the final RGB image.

B. Dual Mappers

The latent space $\mathcal{W}$ of the pretrained StyleGAN2 generator serves as a rich representation that encapsulates diverse facial attributes and texture features. By modifying the latent code $\boldsymbol{w}$, it is possible to alter the identity, facial attributes, and texture features of the generated 3D head. Simultaneously, the facial expression parameters $\boldsymbol{\alpha}$ control the geometric configuration of the corresponding 3D mesh, enabling precise adjustments to facial expressions. Consequently, we devise a dual-mapping structure employing cross-attention mechanisms to predict transformations for the input $\boldsymbol{w}$ and $\boldsymbol{\alpha}$, thereby enabling fine-grained control over facial expression editing.

As shown in Fig. 2 (a), the Dual Mappers consist of a Texture Mapper $\mathcal{M}_T$ and an Emotion Mapper $\mathcal{M}_E$, along with a Cross Attention module. Owing to the aggregation of latent codes $\boldsymbol{w}$ and $\boldsymbol{\alpha}$ within the synthesis network through rasterization rendering on the neural texture tri-planes, we employ cross-attention to guide the editing of $\boldsymbol{\alpha}$ and $\boldsymbol{w}$ with $\boldsymbol{w}$ and $\boldsymbol{\alpha}$ as conditions, respectively. In addition, due to the significant impact that subtle numerical changes in the latent code can significantly influence the subsequent generation results, we refrain from directly predicting the edited parameters. Instead, we predict the discrepancy information between the edits pre- and post-editing. The Texture Mapper $\mathcal{M}_T$ is employed to predict $\Delta\boldsymbol{w}$, while the Emotion Mapper $\mathcal{M}_E$ predicts $\Delta\boldsymbol{\alpha}$.

Specifically, as shown in Fig. 2 (a), we first extract the facial expression parameters $\boldsymbol{\alpha}_R$ of the reference image I_R using EMOCA [31]. Then, to derive the transformation $\Delta\boldsymbol{w}$, we utilize $\boldsymbol{\alpha}_R$ as the query and $\boldsymbol{w}$ as the key and value, resulting in the attention map A_T. Subsequently, $\Delta\boldsymbol{w}$ is obtained through mapping by $\mathcal{M}_T$, formulated as:

$$A_T = \text{softmax}(\frac{Q_{\boldsymbol{\alpha}} K_{\boldsymbol{w}}^\top}{\sqrt{d}}), \qquad \Delta\boldsymbol{w} = \mathcal{M}_T(A_T V_{\boldsymbol{w}}), \tag{5}$$

where $Q_{\boldsymbol{\alpha}}$ is mapped from $\boldsymbol{\alpha}_R$, $K_{\boldsymbol{w}}$ and $V_{\boldsymbol{w}}$ are mapped from $\boldsymbol{w}$, d denotes the length of the key and query features. Then the target latent code $\boldsymbol{w'}$ can be derived with $\Delta\boldsymbol{w} + \boldsymbol{w}$. By a similar approach but in reverse, utilizing $\boldsymbol{w}$ as the condition, we can derive $\Delta\boldsymbol{\alpha}$ and $\boldsymbol{\alpha'}$ using $\mathcal{M}_E$.

Moreover, to prevent the edited parameters from deviating significantly from their initial values, thereby causing changes in other facial attributes or increasing artifacts in the generated images, we apply an L_2 loss term to constrain the Dual Mappers as follows:

$$\mathcal{L}_{\text{M}} = ||\boldsymbol{w} - \boldsymbol{w'}||_2 + ||\boldsymbol{\alpha} - \boldsymbol{\alpha'}||_2 \tag{6}$$

C. Text-Guided Optimization

Recent studies [15], [74]–[76] have increasingly showcased the efficacy of editing latent space $\mathcal{W}$ from StyleGAN under CLIP guidance, enabling high-fidelity image generation and precise control over editing directions. Therefore, we leverage the rich semantic information embedded in CLIP to guide the editing directions of the Dual Mappers outlined in Section III-B through textual cues. Additionally, some scholars [77] have noted that the text embedding obtained from CLIP may not always align well with the image embedding, possibly resulting in undesired attribute alterations during text-guided image editing and increased artifact presence in generated images. Inspired by their work, we abstain from directly optimizing our model through cosine similarity computation between the edited image embedding and the text description embedding. Instead, as shown in Fig. 2 (b), guided by text embedding $\boldsymbol{e}_T$, we project the initial image embedding $\boldsymbol{e}_I$ onto a subspace $\mathfrak{T}$ composed of some basis vectors corresponding to a set of predefined text prompts. Then we calculate the cosine similarity between $\boldsymbol{e}'_T$ which is obtained from the projection operation and the edited image embedding $\boldsymbol{e}'_I$ to optimize our Dual Mappers.

Specifically, since our primary target attribute for editing is facial expression, we choose some text descriptions (e.g., "A person who is raising brow.") related to expressions as basis vectors $\{\boldsymbol{b}_k\}_{k=1}^N$ to construct the subspace $\mathfrak{T}$. We then project the initial image embedding $\boldsymbol{e}_I$ into $\mathfrak{T}$ while preserving a residual vector $\boldsymbol{r}$, formulated as:

$$\boldsymbol{e}_P = \mathcal{P}_{\mathfrak{T}}(\boldsymbol{e}_I), \qquad \boldsymbol{r} = \boldsymbol{e}_I - \boldsymbol{e}_P, \tag{7}$$

where $\mathcal{P}$ is the projection operation on $\mathfrak{T}$, $\boldsymbol{e}_P$ is the projected embedding. After that, we augment the impact of the text prompt t on the projected embedding $\boldsymbol{e}_P$, and subsequently, reintroduce the residual vector $\boldsymbol{r}$ to form the final embedding:

$$\boldsymbol{e}'_T = \mathcal{A}(\boldsymbol{e}_P, \boldsymbol{e}_T, \gamma) + \boldsymbol{r}, \tag{8}$$

where $\mathcal{A}$ is the augmentation operation performed on $\boldsymbol{e}_P$ under the guidance of $\boldsymbol{e}_T$, $\gamma \in \mathbb{R}^+$ is the augmentation power. The principle is to weaken attributes in $\boldsymbol{e}_P$ that are unrelated to $\boldsymbol{e}_T$, thereby enhancing the influence of the attributes represented by $\boldsymbol{e}_T$. Specifically, $\mathcal{A}$ is defined as:

$$\mathcal{A}(\boldsymbol{e}_P, \boldsymbol{e}_T, \gamma) := \sum_{k=1}^{N} \left(c_k - \gamma \left|c_k\right|\right) \boldsymbol{b}_k + \frac{\gamma \sum_{k=1}^{N} \left|c_k\right|}{\sum_{k=1}^{N} d_k} \boldsymbol{e}_T, \tag{9}$$

where $c_k = \boldsymbol{e}_P^T \boldsymbol{b}_k$, $d_k = \boldsymbol{e}_T^T \boldsymbol{b}_k$., and $\boldsymbol{b}_k$ is a set of basis vectors, each describing an expression.

In summary, after projecting $\boldsymbol{e}_I$ into the expression subspace $\mathfrak{T}$, text-guided facial editing focuses only on attributes related to expressions, ensuring that other attributes remain unchanged. By optimizing the cosine similarity between the final embedding $\boldsymbol{e}'_T$ and the edited image embedding $\boldsymbol{e}'_I$, fine-grained expression editing under text guidance can be achieved. Additionally, we introduce an identity loss to constrain identity changes during editing, formulated as:

$$\mathcal{L}_{\text{CLIP}} = -\left\langle \boldsymbol{e}'_I, \boldsymbol{e}'_T \right\rangle \quad \mathcal{L}_{\text{ID}} = 1 - \left\langle FR(I_I), FR(I_E) \right\rangle \tag{10}$$

where FR represents a pretrained ArcFace [78] network utilized for face recognition, and $\langle \cdot, \cdot \rangle$ calculates the cosine similarity between its arguments. The final optimization objective is given by:

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{CLIP}} \mathcal{L}_{\text{CLIP}} + \lambda_{\text{M}} \mathcal{L}_{\text{M}} + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}} \tag{11}$$

where the hyperparameters controlling the relative weights of each loss term are empirically set to $\lambda_{\text{CLIP}} = 1$, $\lambda_{\text{M}} = 0.05$, and $\lambda_{\text{ID}} = 0.2$.

IV. IMPLEMENTATION

A. Experimental Setup

Training Datasets and Details. During training, we utilized EMOCA [31] to extract corresponding FLAME [51] parameters from 2D images containing various expressions in FaceScape [79], serving as the initial 3D priors. Specifically, we extract six expressions to demonstrate the editing ability of our model, i.e., "Smile", "Angry", "Sad", "Close Eyes", "Raise Row", and "Frown Brow". We train separately for each expression. Taking the expression "Smile" as an example, we randomly selected 5,000 latent codes from the z space of the StyleGAN generator pre-trained in Next3D [24] to generate neural texture tri-planes. Simultaneously, previously obtained FLAME parameters for the "Smile" expression were randomly chosen as the driving information for generating initial images. Finally, through the optimization process guided by the text prompt "A person who is smiling", we obtain the final edited images.

Metrics. We employ Fréchet Inception Distance (FID) [81] and Kernel Inception Distance (KID) [82] to measure the quality of generated images. To explicitly evaluate the accuracy of fine-grained expression editing—which we define as facial modifications that capture subtle, localized muscle movements beyond coarse categorical emotion labels—we use Action Units Accuracy (AU Acc) and the CLIP Score. Unlike traditional categorical metrics that evaluate a single global emotion, AU analysis decomposes facial expressions into fundamental muscular movements based on the Facial Action Coding System (FACS). By quantitatively measuring the presence of specific Action Units (AUs) using OpenFace [83], [84],we can rigorously benchmark the structural fidelity of nuanced expressions at a micro-level.

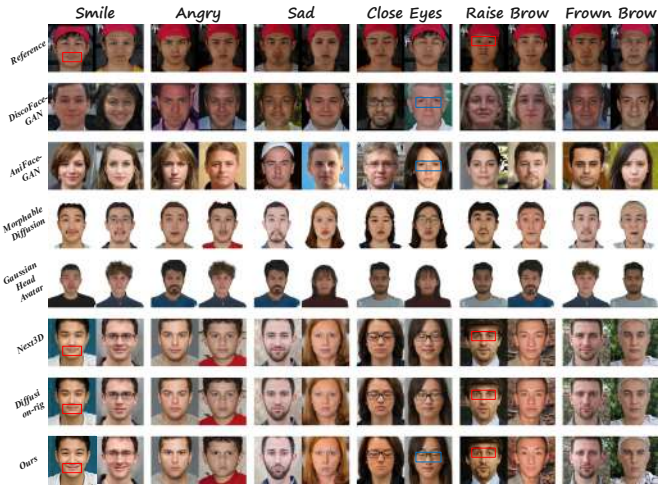


Fig. 4. Fine-grained expression editing comparison with state-of-the-art animatable 3D image synthesis methods. To ensure a strict and direct comparison, methods supporting reference-guided editing within the same domain (Next3D, Diffusion-rig, and Ours) are evaluated using the exact same identity. For unconditional generative models (DiscoFaceGAN, AniFaceGAN) and models trained on entirely disparate datasets (Morphable Diffusion on FaceScape [79], Gaussian Head Avatar on NeRSemble [80]), forcing a specific target identity would require error-prone inversion or cross-domain adaptation. To avoid disadvantaging these baselines with reconstruction artifacts, we display high-quality native identities sampled directly from their respective distributions. Fig. 5(a) illustrates a zoomed comparison of the red-box region.

In practice, for each target expression, we randomly sample latent codes and reference images to generate 1,000 sample pairs. To ensure a fair evaluation across fundamentally different architectures, unconditional baselines and models trained on disparate datasets (e.g., FaceScape [79] and NeRSemble [80]) are evaluated by sampling native identities from their respective distributions. This deliberately avoids reconstruction artifacts associated with forced GAN inversion or cross-domain alignment. Crucially, to strictly rule out any selection bias caused by evaluating different visual identities, our quantitative evaluation relies entirely on identity-agnostic metrics.

Specifically, we employ the following metrics to provide a rigorous and objective benchmark:

Action Units Accuracy (AU Acc): We utilize OpenFace to detect localized facial muscle movements. An edit is considered successful if the generated image possesses all AUs associated with the target expression. Specifically, "Smile" corresponds to AU_06, AU_12; "Angry" to $AU\_04, AU\_05, AU\_07, AU\_23$; "Sad" to AU_01, AU_04, AU_15; "Raise Brow" to AU_02; and "Frown Brow" to AU_04. Note that OpenFace cannot detect the corresponding AU for "Close Eyes".

TABLE I

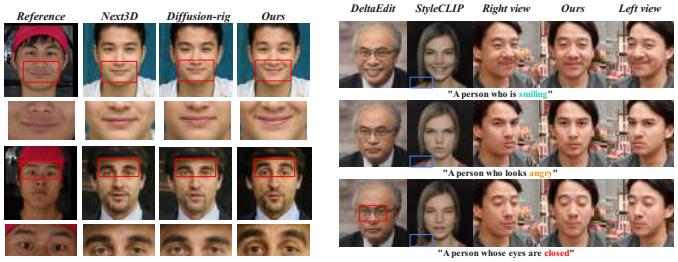
QUANTITATIVE RESULTS ON THE EXPRESSION EDITING RESULTS USING TWO METRICS: ACTION UNITS ACCURACY (AU ACC) AND CLIP SCORE. THE USED EXPRESSIONS ARE SMILE, ANGRY, SAD, CLOSE EYES (CE), RAISE BROW (RB), AND FROWN BROW (FB). SINCE OPENFACE CANNOT DETECT THE AU CORRESPONDING TO "CLOSE EYES", WE DO NOT PROVIDE THE AU ACCURACY FOR THIS EXPRESSION.

Method	AU Acc ↑							CLIP Score ↑						
	Smile	Angry	Sad	CE	RB	FB	Avg	Smile	Angry	Sad	CE	RB	FB	Avg
DiscoFaceGAN [44]	0.87	0.48	0.33	N/A	0.35	0.23	0.45	25.10	23.89	23.02	23.82	24.04	23.53	23.90
AniFaceGAN [23]	0.57	0.43	0.38	N/A	0.60	0.46	0.49	24.01	25.94	24.64	23.98	24.50	25.06	24.69
Morphable Diffusion [27]	0.46	0.39	0.52	N/A	0.48	0.48	0.49	24.28	25.11	24.29	25.22	24.75	25.26	24.82
Gaussian Head Avatar [28]	0.34	0.44	**0.54**	N/A	0.50	**0.52**	0.47	23.83	26.26	**25.28**	24.63	24.26	24.84	24.85
Next3D [24]	0.65	0.41	0.42	N/A	0.49	0.38	0.47	24.56	24.61	23.35	25.93	24.63	24.61	24.62
Diffusion-rig [26]	0.65	0.38	**0.54**	N/A	0.66	**0.52**	0.55	24.18	25.76	24.59	**26.30**	24.69	**25.40**	25.15
Ours	**0.91**	**0.49**	0.48	N/A	**0.68**	0.43	**0.60**	**25.36**	**26.52**	24.21	26.13	**24.76**	24.68	**25.28**

TABLE II

QUANTITATIVE RESULTS OF IMAGE QUALITY USING FID AND KID.

Method	FID ↓	KID ↓
StyleCLIP [15]	35.53	0.021
DeltaEdit [87]	45.78	0.024
DiscoFaceGAN [44]	58.95	0.040
AniFaceGAN [23]	51.00	0.035
Next3D [24]	32.95	0.019
Diffusion-rig [26]	31.26	0.024
Ours	**30.71**	**0.018**



(a) Zoomed-in details. (b) Comparison with 2D methods.

Fig. 5. **(a)** In contrast to Next3D and Diffusion-rig, our approach excels in capturing fine-grained expression facial details, such as the curvature of the mouth corners during smiling and the positioning of the eyebrows during frowning, bringing generated facial features closer to the reference image and target emotion text descriptions. **(b)** Comparison with the recent 2D text-guided face editing methods. We perform three expression edits on the same individual. The final three columns demonstrate view-consistent editing capabilities from our method, a feature lacking in both 2D methods, DeltaEdit and StyleCLIP, which struggle to control views.

CLIP Score: To evaluate semantic image-text faithfulness, we compute the cosine similarity between the embeddings of the generated image and the expression's text description in the CLIP space [85]. Because both AU Acc and CLIP Score are fundamentally independent of the subject's underlying identity, they ensure objective measurement of fine-grained expression fidelity.

FID and KID: To assess overall image quality, we combine the generated samples from all expressions and compare them against a reference set of 5,000 randomly selected images from the FFHQ [86] dataset.

We evaluate our method against several state-of-the-art baselines for 3D-aware fine-grained expression editing. Our primary comparison includes 3D-aware models compatible with single-view data: DiscoFaceGAN [44], AniFaceGAN [23], Next3D [24], and Diffusion-rig [26]. We also compare against 2D text-guided editing methods, specifically StyleCLIP [15] and DeltaEdit [87].

Additionally, we include Morphable Diffusion [27] and Gaussian Head Avatar [28] in our comparison. Morphable Diffusion is based on a multi-view diffusion model trained on FaceScape [79], while Gaussian Head Avatar utilizes 3D Gaussian Splatting (3DGS) for reconstruction on the NeRSemble [80] dataset. Since these methods rely on extensive multi-view data and are difficult to adapt to the single-view FFHQ [86] dataset, we focus solely on comparing their expression editing performance using their officially released code and pretrained models.

For the remaining baselines, all results are attained using their released code and pretrained models on the FFHQ dataset, except for StyleCLIP, which required retraining using default hyperparameters as no pretrained model was available. For Diffusion-rig, we further fine-tuned the model for each identity using Next3D-generated images with diverse expressions to support the required image generation.

B. Qualitative Evaluation

1) 3D Methods: Fig. 4 illustrates the comparative results between our approach and the 3D-aware animatable baselines. Overall, our method more accurately captures the fine-grained facial expression features in the reference images and generates high-quality images that are consistent with the target facial expression text descriptions. Specifically, Next3D can tolerate the changes in topology in the expression control by adapting surface deformation into a continuous volumetric representation. Diffusion-rig can extract the expression coefficients of the reference image using a DECA [25], then generate the corresponding mesh, which is subsequently rendered into an albedo map used as a condition for the diffusion model.

Consequently, ours, Diffusion-rig and Next3D can successfully generate the "Close Eyes" expression. In contrast, DiscoFaceGAN and AniFaceGAN struggle to match this challenging expression, as depicted in Fig. 4 (Blue Box). Similarly, although Morphable Diffusion and Gaussian Head Avatar align well with the majority of target expressions, they perform poorly on challenging expressions such as "Close Eyes". Specifically, generating challenging expressions like "Close Eyes" exposes two primary bottlenecks in existing methods: data bias in standard 3D priors (e.g., DECA), which are dominated by open-eyed subjects, and uncoordinated updates between eyelid geometry and eye textures, leading to 'ghosting' artifacts. Our approach overcomes these issues by leveraging EMOCA for a more accurate geometric initialization. Crucially, our Dual Mappers synergistically coordinate topological mesh deformations with UV texture updates, while the CLIP loss provides semantic compensation for the scarcity of closed-eye 3D data. This ensures robust, natural closed-eye generation without texture-mesh mismatches.

Furthermore, since DiscoFaceGAN and Diffusion-rig do not explicitly utilize 3DMM information to model the entire 3D representation corresponding to the generated images, they encounter view-inconsistency issues. Although Next3D and Diffusion-rig can match most target expressions, their ability

TABLE III
QUANTITATIVE RESULTS ON THE FINE-GRAINED EXPRESSION EDITING RESULTS USING CLIP SCORE. THE USED EXPRESSIONS ARE SMILE, ANGRY, SAD, CLOSE EYES (CE), RAISE BROW (RB), AND FROWN BROW (FB).

Method	CLIP Score ↑						
	Smile	Angry	Sad	CE	RB	FB	Avg
DeltaEdit [87]	**25.81**	23.69	23.26	23.41	23.51	22.68	23.73
StyleCLIP [15]	25.15	**26.48**	**25.69**	23.59	23.67	23.27	24.64
Ours	24.70	26.27	23.96	**26.16**	**24.66**	**25.06**	**25.14**

to capture fine-grained expressions is limited. Specifically, as shown in Fig. 4 (Red box) and zoomed results in Fig. 5(a), Next3D and Diffusion-rig do not align well with both the reference image and the given expression category simultaneously. For example, they may exhibit "Smile" characteristics in "Sad" expressions, shown in Fig. 4 "Sad", or they may fail to represent some expressions like "Raise brow", shown in Fig. 5(a).

Overall, our method not only achieves "challenging expression" editing by successfully modeling topology changes (e.g., "Close eyes") but also accurately generates the fine-grained expression features in the reference images and aligns more closely with the textual description of the target expression.

2) 2D Methods: To assess the alignment between the edited expressions and the given text prompts achieved by our method, we compare the editing quality with StyleCLIP [15] and DeltaEdit [87] under the given text descriptions, such as "A person who is smiling". As shown in Fig. 5(b), generated images from StyleCLIP can match the corresponding text descriptions well, but they have poor background preservation (the clothes color has been changed in the blue box). On the other hand, although DeltaEdit can maintain good background preservation, it does not match well with some expressions such as "eye closed" (Red box). In contrast, our method not only achieves better preservation of background but also produces edited expressions closer to the corresponding text descriptions. Moreover, our method can generate high-quality images from novel views and maintain good 3D consistency. More results can be found in the Demo Video.

C. Quantitative Evaluation

Table I presents the comparison results of our method with 3D-aware animatable baselines in terms of AU Acc and CLIP Score. Our method achieves the best scores in most expressions. It is observed that DiscoFaceGAN performs well on the "Smile" expression due to its training set containing many "Smile"-related images, but struggles with other expressions. Its limited ability to extract and generate fine-grained expression features leads to generating mostly "Smile"-related images. For other expressions, our method achieves higher average AU Acc and CLIP Score, indicating a better ability to extract fine-grained expression features from the reference image and maintain consistency with the expression text description. Additionally, for the "Close Eyes" expression, OpenFace [83], [84] fails to detect the corresponding AU, but the CLIP Score shows that our method can effectively handle this more challenging expression editing task.

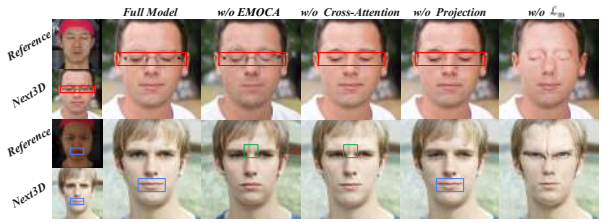


Fig. 6. The qualitative ablation experiments examine the main components of our method, including the removal of identity loss (*w/o* $\mathcal{L}_{\text{ID}}$), Subspace Projection (*w/o Projection*), Cross-Attention (*w/o Cross-Attention*) and EMOCA (*w/o EMOCA*). We present the results of expression editing on the "Close Eyes" and "Angry" expressions, for each ablation.

Table II presents the visual quality evaluation results of our method compared to all baselines. Our method demonstrates significant improvements in FID and KID compared to 3D methods AniFaceGAN and DiscoFaceGAN, as well as 2D methods StyleCLIP and DeltaEdit. It is noted that, except for Next3D, all 3D methods generate results at 256^2 resolution, while the 2D methods generate results at 1024^2 resolution. Therefore, we first select images at 1024^2 resolution from the FFHQ dataset and then resize them to 256^2 and 512^2 resolutions for comparison. Our method also exhibits certain improvements compared to Next3D and Diffusion-rig, indicating that our proposed modules and optimization methods do not compromise the generation quality. We note that Morphable Diffusion [27] and Gaussian Head Avatar [28] are excluded from Table II. Since they utilize FaceScape and NeRSemble, respectively, the inherent dataset discrepancy heavily impacts FID and KID scores. Consequently, we limit the comparison with these methods to AU Accuracy and CLIP Score to ensure a fair evaluation.

In Table III, we compare the text-guided 2D face editing methods DeltaEdit [87] and StyleCLIP [15] in terms of CLIP Score under given textual descriptions. Our method performs the best on most expressions, especially on challenging ones like "Close eyes". This indicates that our method outperforms 2D approaches by better aligning with corresponding textual descriptions.

D. Ablation Study

We also conducted an evaluation to assess the impact of various design choices within our method, including the identity loss $\mathcal{L}_{\text{ID}}$, Subspace Projection, Cross-Attention Module, and EMOCA. The qualitative results depicted in Fig. 6 illustrate that our full model achieves greater precision in fine-grained expression editing compared to all ablations. Notably, the altered face identity in the absence of $\mathcal{L}_{\text{ID}}$ underscores the effectiveness of $\mathcal{L}_{\text{ID}}$ in preserving identity. Similarly, the absence of EMOCA leads to a noticeable reduction in the accuracy of expression editing, as highlighted in Fig. 6 (Red Box). Without EMOCA, the fine-grained facial movements are not accurately captured, resulting in less expressive and misaligned outputs. Moreover, both the absence of Subspace Projection and the Cross-Attention Module result in excessive editing problems. Particularly, the absence of Subspace Projection causes unintended attribute alterations, as seen in Fig. 6 (Red Box). The blue box highlights the improved alignment of subtle facial features, such as the curvature of the

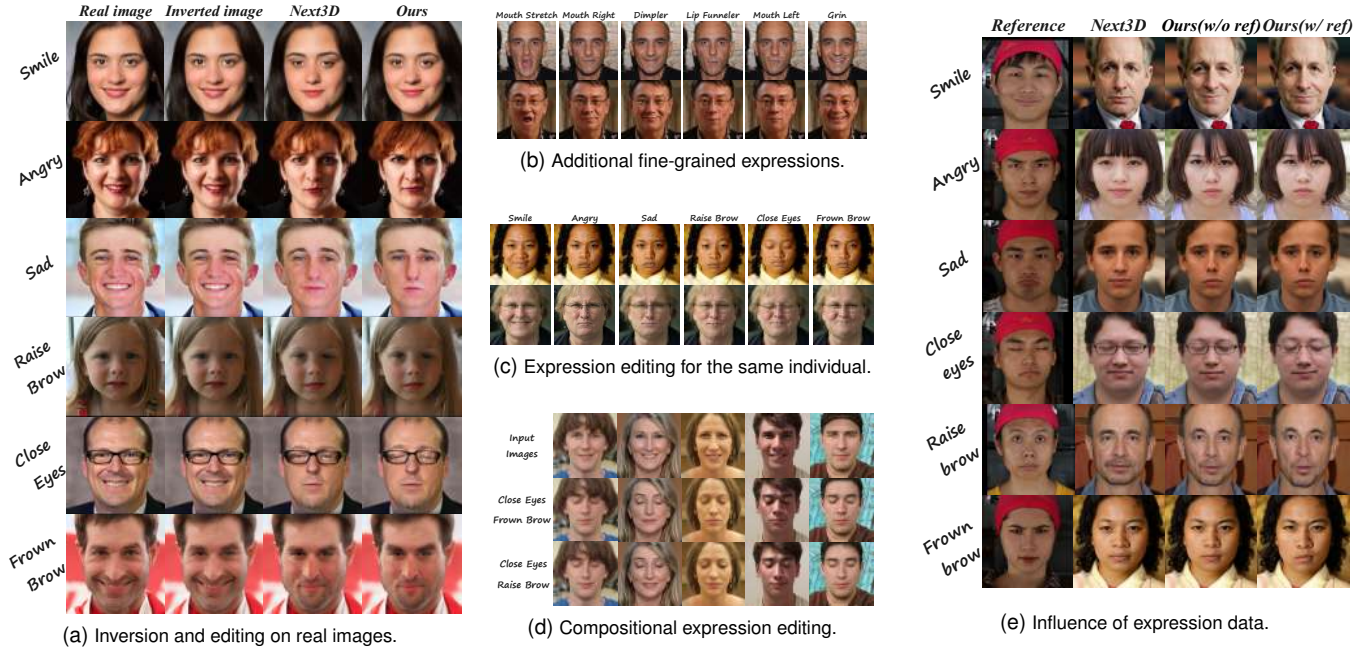
This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3674337

TABLE IV
ABLATION STUDY ON FINE-GRAINED EXPRESSION EDITING QUALITY AND IMAGE QUALITY.

Method	AU Acc ↑							CLIP Score ↑							FID&KID ↓	
	Smile	Angry	Sad	CE	RB	FB	Avg	Smile	Angry	Sad	CE	RB	FB	Avg	FID	KID
w/o $\mathcal{L}_{\text{ID}}$	0.99	0.48	0.81	N/A	0.92	0.21	0.68	27.30	28.80	26.07	27.94	26.96	29.29	27.73	43.42	0.027
w/o Projection	0.75	0.46	0.49	N/A	0.68	0.67	0.61	24.89	25.77	23.98	26.18	24.85	25.62	25.22	30.38	0.018
w/o Cross-Attention	0.83	0.56	0.48	N/A	0.68	0.62	0.63	25.34	26.43	23.94	26.28	24.76	25.56	25.38	30.93	0.018
w/o EMOCA	0.70	0.48	0.52	N/A	0.64	0.47	0.56	24.79	26.44	24.44	26.49	24.80	25.52	25.41	33.47	0.019
Full Model	0.90	0.49	0.56	N/A	0.73	0.56	0.65	25.12	26.38	24.67	26.10	24.72	24.96	25.33	30.13	0.017

TABLE V
QUANTITATIVE RESULTS FOR BACKGROUND PRESERVATION.

Method	SSIM ↑
DiscoFaceGAN [44]	0.8457
AniFaceGAN [23]	0.9996
Morphable Diffusion [27]	0.9911
Gaussian Head Avatar [28]	0.9998
Next3D [24]	0.9998
Diffusion-rig [26]	0.9998
Ours	**0.9996**



(a) Inversion and editing on real images.
(b) Additional fine-grained expressions.
(c) Expression editing for the same individual.
(d) Compositional expression editing.
(e) Influence of expression data.

Fig. 7. **(a)** Comparison results of fine-grained expression editing on real images: real images, inverted images, Next3D outputs, and results from the proposed method across six expressions ("Smile", "Angry", "Sad", "Raise Brow", "Close Eyes", and "Frown Brow"). **(b)** Illustration of new expression editing results using the proposed method, including "Mouth Stretch", "Mouth Right", "Dimpler", "Lip Funneler", "Mouth Left" and "Grin". The edited images demonstrate the model's ability to generate diverse fine-grained expressions while maintaining high visual quality. **(c)** Expression editing results for the same individual. Each row corresponds to a different subject, and each column represents a target expression, including "Smile", "Angry", "Sad", "Raise Brow", "Close Eyes", and "Frown Brow". The results demonstrate the ability of the proposed method to modify expressions while preserving the identity and overall image quality. **(d)** Qualitative results of compositional expression editing demonstrate our model's ability to seamlessly combine multiple distinct facial actions, such as "Close Eyes" alongside "Frown Brow" or "Raise Brow", applied to various input images. **(e)** Comparison results under different training setups: training without target expression data versus training initialized with target expression codes, evaluated across six expressions ("Smile", "Angry", "Sad", "Raise Brow", "Close Eyes", and "Frown Brow").

mouth corners, achieved by incorporating Subspace Projection. Additionally, the green box demonstrates that the inclusion of the Cross-Attention Module reduces the generation of artifacts, ensuring smoother and more coherent editing results.

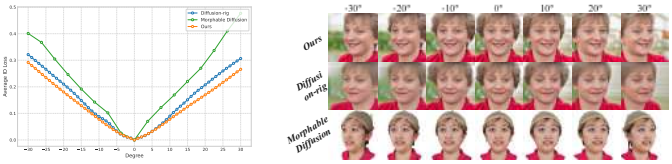
Table IV demonstrates that our method outperforms all other ablations in terms of AU Acc, except for the *w/o* $\mathcal{L}_{\text{ID}}$ variant. In the absence of $\mathcal{L}_{\text{ID}}$, the model excessively edits the expression under CLIP guidance, resulting in generated images that exhibit stronger features of the target expression, but struggle to maintain identity consistency and image quality. This is reflected in both the FID and KID scores. In terms of CLIP Score, the full model performs similarly to the other ablations, with *w/o* $\mathcal{L}_{\text{ID}}$ still achieving the best results. For FID and KID, the Full model achieves the best performance. In conclusion, when considering both the accuracy of expression editing and the quality of generated images, the Full Model exhibits the best performance.

V. IDENTITY CONSISTENCY MAINTENANCE AND RESULTS FOR ADDITIONAL EXPRESSIONS.

We provide more expression editing results in Fig. 7(b) and (c). Fig. 7(b) shows new expression editing which demonstrates that our method improves the editing precision of Next3D for subtle expressions while supporting common expressions. Both figures illustrate multiple expression editing for the same identity. Notably, the identity remains well-preserved. Moreover, Table V provides a quantitative evaluation of the preservation of non-target regions (backgrounds), with the high scores highlighting the effectiveness of our method in maintaining identity.

VI. INVERSION AND EDITING RESULTS

To demonstrate the effectiveness of our method on real-world images, we randomly selected several images from the FFHQ dataset [86] and utilized the 3D GAN inversion technique described in [88]. The inverted images were then edited



(a) Quantitative evaluation. (b) Qualitative comparison.

Fig. 8. **(a)** The view consistency across different viewpoints is evaluated by calculating $\mathcal{L}_{\text{ID}}$ between images generated by rotating the head model by 1-30 degrees from the frontal view. Our method demonstrates a smoother trend compared to Diffusion-rig and Morphable Diffusion (evaluated via uniform sampling from $-30°$ to $30°$ due to its 16-image generation limit). **(b)** The images generated from different viewpoints demonstrate that our method maintains superior view consistency, particularly in the generation of hair, compared to Diffusion-rig.

to modify their facial expressions, with the results presented in Fig. 7(a). As shown, the inverted images maintain a high level of fidelity, closely resembling the original real images. By applying our expression editing approach, we successfully generated images that not only retain high visual quality but also depict facial expressions that align more accurately with the corresponding text descriptions. This highlights the robustness of our method in handling real-image scenarios and demonstrates its ability to preserve essential image details while achieving expressive edits.

VII. Influence of Expression Data

In practical applications, obtaining large-scale, high-quality facial image datasets with diverse fine-grained expressions poses significant challenges. This makes it impractical to adopt methods that require training with a large number of images matching the target expressions for initializing expression codes. In our method, we use target expression images to initialize expression parameters and CLIP loss for optimization, leveraging textual descriptions to guide the generated expressions and align them with the intended semantics. Specifically, we present the results in Fig. 7(e), comparing the training initialized with neutral expressions (*w/o ref*) against those using reference images of the target expressions (*w/ ref*). While the *w/o ref* method is not as effective as the *w/ ref* approach in terms of editing performance, it still achieves noticeable editing results, particularly for expressions like Smile, Angle, and Sad. This improvement can be attributed to the CLIP loss, which plays a pivotal role in aligning the generated expressions with the intended textual descriptions.

VIII. View consistency analysis of Diffusion based and 3D-aware GAN based methods

As shown in Fig. 8(a) and (b), diffusion-based methods lack explicit 3D head modeling, leading to artifacts in view consistency. For instance, Diffusion-rig relies on conditional image generation, and Morphable Diffusion, despite employing a multi-view diffusion architecture, still generates novel views without a consistent 3D geometry. Consequently, they exhibit inconsistencies when the viewing angle changes. In contrast, 3D-aware GAN-based methods like ours model the head utilizing neural radiance fields, ensuring strict geometric consistency across different rendered viewpoints. As a result, our method maintains better view consistency.

TABLE VI
Comparison of inference time (s) per run.

Methods	Time (s)
DiscoFaceGAN [44]	0.62
AnifaceGAN [23]	0.95
Next3D [24]	0.52
Gaussian Head Avatar [28]	0.42
Diffusion-rig [26]	1.58
Morphable Diffusion [27]	1.26
Ours	**0.71**

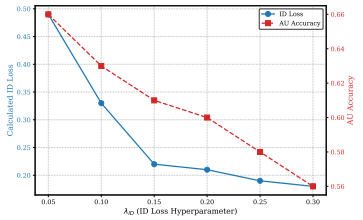


Fig. 9. Sensitivity Analysis of the Identity Regularization Weight (λ_{ID}).

IX. Computational Complexity Analysis

We mainly evaluate the computational efficiency of our proposed method in terms inference time. All experiments were conducted on a workstation equipped with a single NVIDIA RTX 4090 GPU (24GB VRAM).

Our approach trains a dedicated *Dual Mapper* for each specific target expression. The training process for refining a Dual Mapper typically requires approximately 14 hours. While our method involves an optimization-based training strategy, which is computationally intensive, this is a one-time cost. Once the Dual Mapper is trained, no further optimization is required during the inference stage.

Unlike optimization-based editing methods that require time-consuming iterations during generation, our method achieves optimization-free inference. As shown in Table VI, we compare the inference time of our method against several approaches. While the 3DGS-based reconstruction method, Gaussian Head Avatar, achieves the fastest inference speed, it requires two days of training for each individual subject. Our method requires only 0.71 seconds per run. This is significantly faster than diffusion-based editing methods such as Diffusion-rig (1.58s) and Morphable Diffusion (1.26s), and is comparable to GAN-based approaches like DiscoFaceGAN (0.62s) and Next3D (0.52s), striking a favorable balance between generation quality and computational efficiency.

X. Sensitivity Analysis of λ_{ID}:

We evaluated the network's performance while varying λ_{ID} from 0.05 to 0.30 in intervals of 0.05. The analysis tracks two key metrics: Calculated ID Loss (measuring identity preservation, where lower is better) and AU Accuracy (measuring expression editing precision). As illustrated in Fig. 9, increasing λ_{ID} effectively reduces the Calculated ID Loss, indicating stronger identity preservation. However, this comes at the cost of expression expressiveness, evidenced by the steady decline in AU Accuracy as the parameter increases. We observe a sharp improvement in identity preservation (a steep drop in ID loss) as λ_{ID} increases from 0.05 to 0.15. Beyond this point, the identity loss begins to stabilize. The range between 0.15 and 0.20 represents a "sweet spot" where identity features are robustly maintained without overly sacrificing the target expression's accuracy. To ensure the high-fidelity retention of the source identity while maintaining competitive expression generation capabilities, we empirically selected $\lambda_{\text{ID}} = 0.20$ as our final hyperparameter setting for all primary experiments.

XI. Compositional Expression Editing

To further evaluate our model's capabilities, we investigate its performance in compositional expression generation by combining multiple distinct facial actions. As illustrated in Fig. 7(d), we successfully combined the action "Close Eyes" with either "Frown Brow" or "Raise Brow", applying them simultaneously to various input images. The visual results clearly demonstrate that our model handles these composite edits seamlessly. The distinct facial actions—such as the closing of the eyelids and the vertical or inward movement of the eyebrows—are synthesized together with high fidelity. The geometric deformations in these adjacent regions do not conflict or cause unnatural artifacts, confirming that our framework can effectively merge multiple, distinct expression instructions into a single, cohesive output.

XII. Limitations and Discussion

Our proposed editing method may not be effective for all generated images. For instance, when the initial expression of some images differs significantly from the reference image, or when the reference image does not clearly exhibit the expression corresponding to the text description, artifacts may appear in the generated images, or they may fail to reflect the intended expression. Additionally, challenging expressions such as "Close eyes" may introduce more artifacts. Fine-tuning the pretrained Next3D model with a more diverse dataset of expressions, and using more accurate reference images for driving, may reduce the occurrence of these situations. Large Generative Models are one type of method requiring intensive computational resources in general. Compared to the methods training from scratch, our method takes a pretrained 3DGAN as the backbone, thus saving largely computational resources in the training stage. What we want to emphasize is that our method focuses on the expression editing problem. The speed of image editing mainly depends on the rendering progress of the NeRF module used, as our dual mappers are small networks. In addition, using better 3D representations, such as 3D Gaussian instead of NeRF, may improve rendering speed in the future.

XIII. Conclusion

We have introduced a method that directly utilizes the pretrained 3D-Aware Animitable GAN model for fine-grained expression editing tasks. Our approach involves the incorporation of the Dual Mappers module into the pretrained model, allowing for the simultaneous refinement of the latent space of UV texture and the expression space of the driven mesh. Additionally, we have proposed a Text-Guided Optimization method, which employs a CLIP-based objective function with expression text prompts as targets. Furthermore, we have integrated a SubSpace Projection mechanism to provide more precise control over fine-grained expressions. These advancements pave the way for more effective and nuanced expression editing capabilities within the realm of computer graphics and animation. In future work, we aim to explore additional techniques for further enhancing the expressiveness and realism of expression editing. This may involve investigating novel approaches for refining the alignment between text prompts and edited expressions, as well as exploring the integration of multimodal inputs, such as audio cues or physiological signals, to enrich the editing process.

References

[1] A. Siarohin, W. Menapace, I. Skorokhodov, K. Olszewski, J. Ren, H.-Y. Lee, M. Chai, and S. Tulyakov, "Unsupervised volumetric animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4658–4669.

[2] J. Li, J. Wu, and C. K. Liu, "Object motion guided human motion synthesis," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–11, 2023.

[3] T. Kosch, J. Karolus, J. Zagermann, H. Reiterer, A. Schmidt, and P. W. Woźniak, "A survey on measuring cognitive workload in human-computer interaction," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023.

[4] D. Borsboom, M. K. Deserno, M. Rhemtulla, S. Epskamp, E. I. Fried, R. J. McNally, D. J. Robinaugh, M. Perugini, J. Dalege, G. Costantini *et al.*, "Network analysis of multivariate data in psychological science," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 58, 2021.

[5] S. Dargan, S. Bansal, M. Kumar, A. Mittal, and K. Kumar, "Augmented reality: A comprehensive review," *Archives of Computational Methods in Engineering*, vol. 30, no. 2, pp. 1057–1080, 2023.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[7] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[8] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 818–833.

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[10] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade ef-gan: Progressive facial expression editing with local focuses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5021–5030.

[11] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 627–635.

[12] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE transactions on image processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[13] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[14] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 3, pp. 1–21, 2021.

[15] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2085–2094.

[16] X. Hu, N. Aldausari, and G. Mohammadi, "2cet-gan: Pixel-level gan model for human facial expression transfer," in *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, 2023, pp. 49–56.

[17] W. Huang, W. Luo, X. Cao, and J. Huang, "Interactive generative adversarial networks with high-frequency compensation for facial attribute editing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8215–8229, 2024.

[18] Z. Li, Z. Zhang, P. Liu, Q. Liu, and X. Sun, "Toward open-world text-driven face generation and manipulation via stylegan3," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 12, pp. 12 862–12 872, 2024.

[19] Y. Wang, W. Zhou, J. Bao, W. Wang, L. Li, and H. Li, "Clip2gan: Toward bridging text with the latent space of gans," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6847–6859, 2024.

[20] 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," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 9997–10 010, 2024.

[21] K. Zhang, M. Sun, J. Sun, K. Zhang, Z. Sun, and T. Tan, "Open-vocabulary text-driven human image generation," *International Journal of Computer Vision*, vol. 132, no. 10, pp. 4379–4397, 2024.

[22] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[23] Y. Wu, Y. Deng, J. Yang, F. Wei, Q. Chen, and X. Tong, "Anifacegan: Animatable 3d-aware face image generation for video avatars," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 188–36 201, 2022.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 991–21 002.

[25] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," vol. 40, no. 8, 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459936

[26] Z. Ding, X. Zhang, Z. Xia, L. Jebe, Z. Tu, and X. Zhang, "Diffusionrig: Learning personalized priors for facial appearance editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12 736–12 746.

[27] X. Chen, M. Mihajlovic, S. Wang, S. Prokudin, and S. Tang, "Morphable diffusion: 3d-consistent diffusion for single-image avatar creation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 359–10 370.

[28] Y. Xu, B. Chen, Z. Li, H. Zhang, L. Wang, Z. Zheng, and Y. Liu, "Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 1931–1941.

[29] Y. Cheng, F. Yin, X. Huang, X. Yu, J. Liu, S. Feng, Y. Yang, and Y. Tang, "Efficient text-guided 3d-aware generation with score distillation on 3d distribution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 10, pp. 9865–9877, 2025.

[30] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[31] R. Danecek, M. J. Black, and T. Bolkart, "EMOCA: Emotion driven monocular face capture and animation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 311–20 322.

[32] M. Gadelha, S. Maji, and R. Wang, "3d shape induction from 2d views of multiple objects," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 402–411.

[33] P. Henzler, N. J. Mitra, and T. Ritschel, "Escaping plato's cave: 3d shape from adversarial rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9984–9993.

[34] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "Graf: Generative radiance fields for 3d-aware image synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 154–20 166, 2020.

[35] M. Niemeyer and A. Geiger, "Giraffe: Representing scenes as compositional generative neural feature fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 453–11 464.

[36] X. Zhu, J. Zhou, L. You, X. Yang, J. Chang, J. J. Zhang, and D. Zeng, "Dfie3d: 3d-aware disentangled face inversion and editing via facial-contrastive learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8310–8326, 2024.

[37] R. Hu, X. Wang, Y. Yan, and C. Zhao, "Tgavatar: Reconstructing 3d gaussian avatars with transformer-based tri-plane," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[38] J. Zhang, X. Li, H. Zhong, Q. Zhang, Y. Cao, Y. Shan, and J. Liao, "Humanref-gs: Image-to-3d human generation with reference-guided diffusion and 3d gaussian splatting," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[39] J. Zhang, E. Sangineto, H. Tang, A. Siarohin, Z. Zhong, N. Sebe, and W. Wang, "3d-aware semantic-guided generative model for human synthesis," in *European Conference on Computer Vision*. Springer, 2022, pp. 339–356.

[40] K. Schwarz, A. Sauer, M. Niemeyer, Y. Liao, and A. Geiger, "Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 999–34 011, 2022.

[41] S. Galanakis, B. Gecer, A. Lattas, and S. Zafeiriou, "3dmm-rf: Convolutional radiance fields for 3d face modeling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3536–3547.

[42] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[43] J. Tang, B. Zhang, B. Yang, T. Zhang, D. Chen, L. Ma, and F. Wen, "Explicitly controllable 3d-aware portrait generation," *arXiv preprint arXiv:2209.05434*, 2022.

[44] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and controllable face image generation via 3d imitative-contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5154–5163.

[45] M. Yu, D. Pang, Z. Kang, Z. Sun, T. Lv, J. Sheng, R. Yi, Y.-H. Wen, and Y.-J. Liu, "Ecavatar: 3d avatar facial animation with controllable identity and emotion," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 468–10 476.

[46] L.-A. Zeng, G. Wu, A. Wu, J.-F. Hu, and W.-S. Zheng, "Progressive human motion generation based on text and few motion frames," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 9, pp. 9205–9217, 2025.

[47] J. Hu, H. Zhang, Y. Wang, M. Ren, and Z. Sun, "Personalized graph generation for monocular 3d human pose and shape estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2399–2413, 2024.

[48] G. Xia, D. Luo, Z. Zhang, Y. Sun, and Q. Liu, "3d information guided motion transfer via sequential image based human model refinement and face-attention gan," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3270–3283, 2023.

[49] J. Liu, X. Wang, X. Fu, Y. Chai, C. Yu, J. Dai, and J. Han, "Osm-net: One-to-many one-shot talking head generation with spontaneous head motions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6888–6900, 2024.

[50] S. Li, X. Qi, B. Yang, C. Weile, Z. Tian, M. Sun, Q. Liu, M. Zhang, and Z. Sun, "Vividlistener: Expressive and controllable listener dynamics modeling for multi-modal responsive interaction," *arXiv preprint arXiv:2504.21718*, 2025.

[51] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, pp. 194:1–194:17, 2017. [Online]. Available: https://doi.org/10.1145/3130800.3130813

[52] H. Li, W. Wang, C. Yu, and S. Zhang, "Swapinpaint: Identity-specific face inpainting with identity swapping," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4271–4281, 2022.

[53] Y. Liu, Q. Li, Q. Deng, and Z. Sun, "Towards spatially disentangled manipulation of face images with pre-trained stylegans," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1725–1739, 2023.

[54] J. Sun, Q. Deng, Q. Li, M. Sun, Y. Liu, and Z. Sun, "Anyface++: A unified framework for free-style text-to-face synthesis and manipulation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 47, no. 11, pp. 9438–9453, 2024.

[55] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9821–9830.

[56] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, and J. Wang, "Fenerf: Face editing in neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7672–7682.

[57] 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 Conference Papers*, 2022, pp. 1–9.

[58] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[59] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.

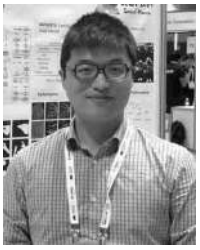
[60] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter, "Morphable face models-an open framework," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 75–82.

[61] R. Li, K. Bladin, Y. Zhao, C. Chinara, O. Ingraham, P. Xiang, X. Ren, P. Prasad, B. Kishore, J. Xing *et al.*, "Learning formation of physically-

based face attributes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3410–3419.
[62] H. Li, T. Weise, and M. Pauly, "Example-based facial rigging," *Acm transactions on graphics (tog)*, vol. 29, no. 4, pp. 1–6, 2010.
[63] T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, and C. Theobalt, "Sparse localized deformation components," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 1–10, 2013.
[64] A. Brunton, T. Bolkart, and S. Wuhrer, "Multilinear wavelets: A statistical shape space for human faces," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 297–312.
[65] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3d face morphable model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1126–1135.
[66] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7346–7355.
[67] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.
[68] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.
[69] H. Wei, S. Liang, and Y. Wei, "3d dense face alignment via graph convolution networks," *arXiv preprint arXiv:1904.05562*, 2019.
[70] Q. Deng, B. H. Le, A. Jin, and Z. Deng, "End-to-end 3d face reconstruction with expressions and specular albedos from single in-the-wild images," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4694–4703.
[71] Z. Chen, Y. Wang, T. Guan, L. Xu, and W. Liu, "Transformer-based 3d face reconstruction with end-to-end shape-preserved domain transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8383–8393, 2022.
[72] P. P. Filntisis, G. Retsinas, F. Paraperas-Papantoniou, A. Katsamanis, A. Roussos, and P. Maragos, "Visual speech-aware perceptual 3d facial expression reconstruction from videos," *arXiv preprint arXiv:2207.11094*, 2022.
[73] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
[74] U. Kocasari, A. Dirik, M. Tiftikci, and P. Yanardag, "Stylemc: multi-channel based fast text-guided image generation and manipulation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 895–904.
[75] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2256–2265.
[76] S. Aneja, J. Thies, A. Dai, and M. Nießner, "Clipface: Text-guided editing of textured 3d morphable models," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.
[77] C. Zhou, F. Zhong, and C. Öztireli, "Clip-pae: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–9.
[78] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
[79] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[80] T. Kirschstein, S. Qian, S. Giebenhain, T. Walter, and M. Nießner, "Nersemble: Multi-view radiance field reconstruction of human heads," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, 2023.
[81] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
[82] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.
[83] B. Tadas, Z. Amir, L. Y. Chong, and M. Louis-Philippe, "Openface 2.0: Facial behavior analysis toolkit," in *13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018.
[84] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–6.
[85] S. Zhengwentai, "clip-score: CLIP Score for PyTorch," https://github.com/taited/clip-score, March 2023, version 0.1.1.
[86] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.
[87] Y. Lyu, T. Lin, F. Li, D. He, J. Dong, and T. Tan, "Deltaedit: Exploring text-free training for text-driven image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6894–6903.
[88] J. Ko, K. Cho, D. Choi, K. Ryoo, and S. Kim, "3d gan inversion with pose optimization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2967–2976.

Yikang He received the B.S. degree from the School of Computer Science & Technology, Beijing Jiaotong University. He is currently a master student at Beijing Jiaotong University. His research interests include 3D face generation and editing.

Jichao Zhang is an Assistant Professor with Ocean University of China. He received the Ph.D. degree from Multimedia and Human Understanding Group at the University of Trento. His research interests include deep generative model, 2D and 3D image generation and editing.



Wei Wang (Member, IEEE) received the PhD degree from the University of Trento, in 2018. He is a full Professor at the Institute of Information Science at Beijing Jiaotong University. His research interests include machine learning and its application to computer vision and multimedia analysis.

Nicu Sebe (Senior Member, IEEE) is Professor with the University of Trento, Italy, leading the Multimedia and Human Understanding Group. He was the General Co-Chair of ACM Multimedia 2013 and 2022, and the Program Chair of ACM Multimedia 2007 and 2011, ICCV 2017, ECCV 2016 and ICPR 2020. He is a fellow of the International Association for Pattern Recognition.

Yao Zhao (Fellow, IEEE) is currently the Director of the Institute of Information Science, Beijing Jiaotong University. His current research interests include image/video coding and video analysis and understanding. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010 and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013.