ExpDiff: Generating High-fidelity Dynamic Facial Expressions with BRDF Textures via Diffusion Model

Yuhao Cheng, Xuanchen Li, Xingyu Ren, Zhuo Chen, Xiaokang Yang, *Fellow, IEEE*, Yichao Yan[†]

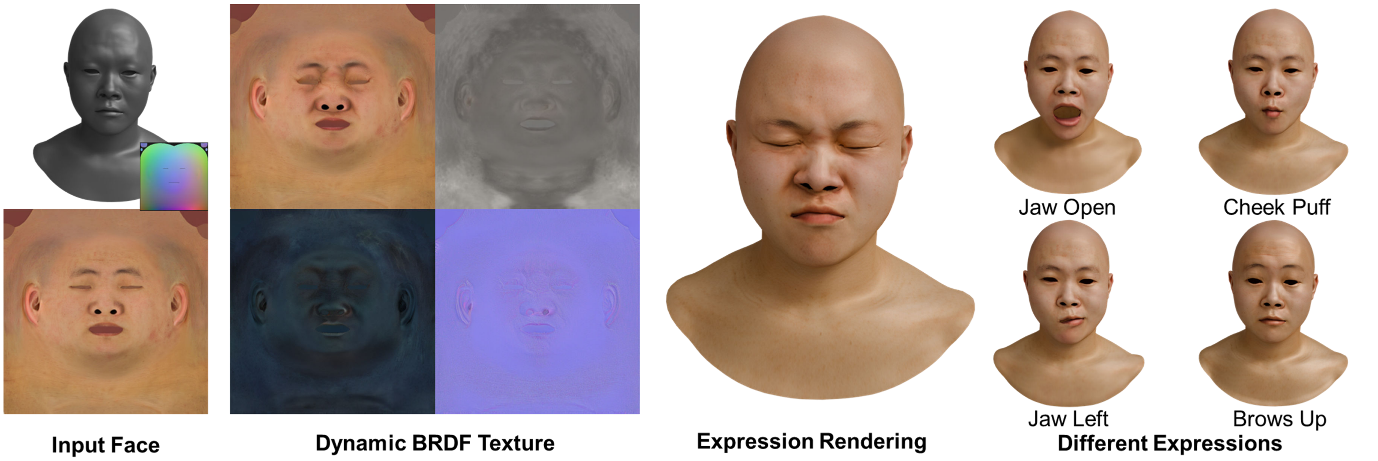


Fig. 1: An example of our dynamic facial expressions generated results. Given a neutral-expression mesh and texture maps as input, our framework enables the generation of FACS-compliant expression meshes with pore-level dynamic BRDF textures through expression text prompts, achieving physically-based photo-realistic rendering.

***Abstract*—3D face generation is a critical task for immersive multimedia applications, where a key challenge is the joint synthesis of expressive geometry and BRDF textures Existing methods often struggle with geometric-textural coherence and corresponding dynamic reflectance modeling. To overcome these limitations, we present ExpDiff, a framework that generates expression meshes and dynamic BRDF textures from a single neutral-expression face. Our method employs an attention-based diffusion model to learn the semantic transition across expressions. To ensure correspondence between geometry and texture, we introduce a unified representation that explicitly models geometric-textural interaction, which is encoded into a shared latent space by models pre-trained on a vast dataset for strong generalization. To achieve semantically coherent and physically consistent generation, we propose to guide the denoising direction with specially designed textual prompts. We further construct two novel dynamic expression datasets, J-Reflectance, for ultra-high-quality assets, and FFHQ-BRDFExp for diverse identities, both of which are publicly released to advance the community. Extensive experiments demonstrate our method's superior performance in photo-realistic facial expression synthesis. Project page: https://cyh-sj.github.io/expdiff/.**

***Index Terms*—Expression generation, Facial mesh, BRDF Textures, Diffusion models**

I. Introduction

3D face generation is an important task in human-centric multimedia applications. As the demand for high-fidelity digital humans continues to rise across immersive virtual environments [1]–[3], including film, VR/AR, and gaming, the need for high-quality facial assets has become increasingly critical. To achieve vivid facial animation and photo-realistic rendering under varying lighting conditions, the detailed **expression meshes** that conform to the Facial Action Coding System (FACS) [4], along with corresponding high-quality **dynamic BRDF textures**, such as albedo, roughness, specular, and normal maps, are typically required.

In industrial pipelines, creating dynamic expressions and textures typically involves capturing multiple facial expressions with the LightStage [5], [6], followed by photometric techniques to produce BRDF assets [7]. These scans are then refined through complex and time-consuming registration processes, during which skilled artists manually process expression meshes and corresponding UV texture maps. This procedure is not only labor-intensive and costly but also difficult to scale. Even though some methods [8]–[10] have been proposed for automatic registration, it is still challenging to acquire film-level dynamic facial assets and inevitable to capture multiple expressions. This motivates the development of fully automated frameworks that can synthesize expression meshes and dynamic BRDF textures directly from a single neutral-expression input, without relying on capture-based pipelines.

However, few works have directly addressed this challenging task. The current approaches primarily focus on two kinds of research: expression mesh generation and BRDF texture generation. **1)** Expression mesh generation typically aims to produce meshes with specific expressions from a neutral facial model. A common approach leverages 3D Morphable Models

[†] Yichao Yan is the Corresponding author.

Yuhao Cheng, Xuanchen Li, Xingyu Ren, Zhuo Chen, Xiaokang Yang, and Yichao Yan are with MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University. Emails: {chengyuhao, lixc6486, rxy_sjtu, ningci5252, xkyang, yanyichao}@sjtu.edu.cn.

(3DMM) [11]–[24] to achieve linear expression generation. However, these approaches lack the representation capacity to capture fine-grained identity-specific variations and detailed expression meshes. To overcome the expression rigidity caused by linear representations, some methods [25]–[28] employ non-linear techniques, such as auto-encoder-decoder and diffusion models, which can generate relatively more diverse and realistic facial animations. Other approaches [29]–[32] use landmarks as guidance for mesh generation, but the overly coupling between landmarks and expressions often limits their ability to preserve personal characteristics across individuals. Moreover, these approaches focus primarily on expression generation on geometry, neglecting the modeling of dynamic BRDF textures. **2)** BRDF texture generation aims to produce multi-channel consistent textures and can be categorized into two main approaches. A kind of approaches [33]–[39] trains transfer networks to infer albedo, normal, roughness, and specular textures from a base texture. However, their performances are often limited by the scale of training data, which impairs generalization to unseen identities. Another kind of approach [40], [41] utilizes inverse rendering for BRDF texture reconstruction, but the incomplete disentanglement of illumination greatly affects the quality of textures. Moreover, both types of methods focus predominantly on the neutral expression, overlooking the dynamic texture changes that arise with expression-dependent deformations.

Certainly, there are a few works that have attempted to jointly synthesize facial expressions and dynamic BRDF textures. TBGAN [42] utilizes GAN to simultaneously generate shapes and textures for dynamic expression generation, while the coupling of expression and identity makes it difficult to generate different expressions while preserving identity. Li et al. [43] first propose generating blendshapes and corresponding BRDF textures from a neutral scan, and Chandran et al. [44] propose a semantically controllable model for expression generation. However, these methods do not fully model the mutual dependencies between geometry and texture, leading to misalignments between texture and geometric deformations in the generated expressions. Besides, these methods ignore the dependencies among different expressions within the FACS, which results in inconsistent local correspondences across different expressions of the same individual, particularly in sensitive facial regions.

Based on the analysis, we argue that modeling the intrinsic correlations between geometric deformations and textural variations can support the development of 3D facial expression generation. In this paper, we propose a unified framework for jointly synthesizing dynamic facial expressions and BRDF textures. As shown in Fig. 1, the framework takes as input a subject's neutral-expression mesh and textures, and outputs corresponding expression meshes with BRDF textures. To facilitate correspondence learning, we unify the representation of geometry and texture by rendering position maps in the same UV space as the texture maps. Both of them are then encoded into a shared latent space using a pre-trained variational autoencoder VAE [45]. To ensure geometric-textural coherence, we propose a hybrid attention-based diffusion architecture that jointly models the bi-directional correspondence between geometric deformations and BRDF spaces, achieving multi-channel consistency in both textures and geometry. Moreover, a textual semantic-guided training paradigm is introduced to enforce local consistency, explicitly capturing cross-region relationships through textual constraints to maintain coherent spatial deformation across distinct facial regions.

Furthermore, a key bottleneck in advancing this field is the lack of publicly available, high-quality datasets that capture dynamic facial expressions with detailed reflectance information. To bridge this gap, we introduce **J-Reflectance**, the first ultra-high-fidelity facial dataset, containing native 8K-resolution BRDF texture maps and high-quality expression meshes, with all assets undergoing manual retopology by professional artists with submillimeter geometric accuracy, to advance facial synthesis research. Additionally, to mitigate identity diversity limitations, we construct **FFHQ-BRDFExp** by extending the FFHQ-UV dataset into a dynamic BRDF format. This enhanced dataset maintains photometric consistency while incorporating realistic expression variations, enabling large-scale perceptual and generative studies. Extensive qualitative and quantitative experiments demonstrate that our method significantly outperforms previous state-of-the-art techniques in both expression synthesis and dynamic BRDF generation.

The main contributions are summarized as follows:

- We present a framework for the consistent facial expression meshes and corresponding dynamic BRDF texture synthesis from a facial model in neutral expression.
- We propose an attention-based diffusion model to capture the relationship between geometry and textures, as well as inner textures, ensuring consistent facial generation.
- We introduce a semantic-aware guidance that leverages text prompts to establish correlations across facial regions, achieving cross-expression consistency.
- We present publicly available, high-quality, large-scale, dynamic BRDF facial datasets.

II. Related Works

A. 3D Expression Generation

3D expression generation is an important task in face modeling, which aims to output facial expression mesh and texture that conforms to the FACS system for subsequent facial binding or blendshape production, bringing vivid facial animation to the face. Early facial expressions were obtained by scanning different facial expressions, such as through camera arrays [27] and lightstage [5], [6] using MVS [7], [46], depth camera [47] reconstruction of point clouds to obtain facial geometry, and then obtaining faces with different expressions through subsequent post-processing such as retopology [12], [48]–[50]. However, such algorithms face the high probability of failure of multiple scans and face retopology under complex and extreme expressions. Automated facial expression generation is imperative to solve the problem of batch facial expression generation. Subsequent facial expression generation can be roughly summarized as parameterized face models and landmark-guided expression generation methods. The earliest

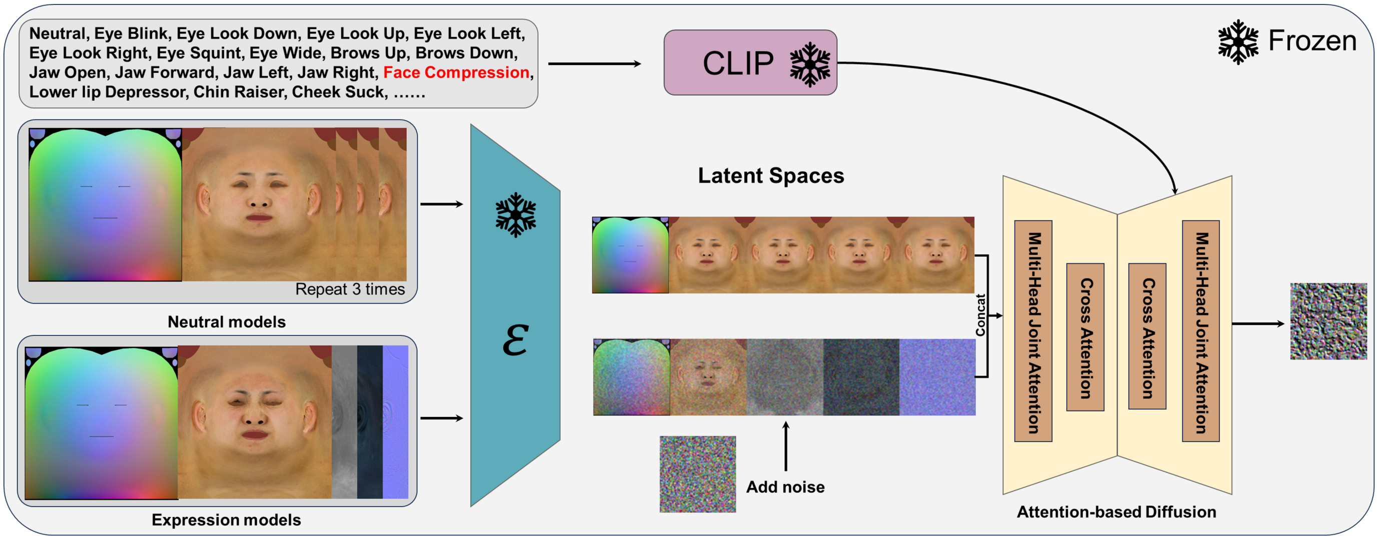

Neutral, Eye Blink, Eye Look Down, Eye Look Up, Eye Look Left, Eye Look Right, Eye Squint, Eye Wide, Brows Up, Brows Down, Jaw Open, Jaw Forward, Jaw Left, Jaw Right, Face Compression, Lower lip Depressor, Chin Raiser, Cheek Suck, ...
CLIP
Frozen
Repeat 3 times
Neutral models
Expression models
$\mathcal{E}$
Latent Spaces
Concat
Add noise
Multi-Head Joint Attention
Cross Attention
Cross Attention
Multi-Head Joint Attention
Attention-based Diffusion

Fig. 2: Overview of our proposed ExpDiff. Given neutral models and expression models, we first represent them in the unified representations and then project them into the latent space through a frozen encoder trained on a large-scale dataset. Then, we propose an attention-based diffusion model for expression asset generation. The textual prompts are encoded by CLIP to obtain semantic information, leading the denoising direction of the diffusion process.

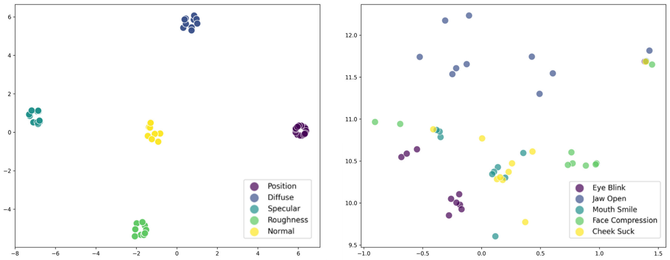
parameterized face model [11]–[16] used PCA to learn low-dimensional features of the face from a large number of face scans, decoupling identity and expression into different facial signals, and was able to use expression parameter changes to achieve facial expression changes. Subsequent methods have also been extended to facial texture and reflection texture learning [51]–[53]. However, limited by linear representation, such methods can only produce fixed expression offsets and cannot represent personalized expressions. Another type of method [29]–[32] uses facial landmarks to guide the generation of facial mesh displacement to achieve expression generation. For example, these works use GAN [54], [55], VAE [45], and diffusion models [56] to achieve facial expression generation, but the acquisition of landmarks is not completely cost-free and is hard to adapt to novel faces. In summary, the above works mainly focus on the structural guidance of facial meshes, but cannot extract high-quality facial expression meshes that meet the FACS to serve the subsequent multimedia development. In addition, they cannot model high-quality dynamic reflection maps. The most similar works to ours are Li et al. [43] and Chandran et al. [44], which can generate scans and high-quality maps of different expressions. However, they only consider the separate modeling of mesh and map, without considering the mutual connection between geometry and texture to promote the generation of facial expressions, nor the correlation and consistency of different expressions. Our work aims to exploit the correlation between geometry and texture to generate expression changes that conform to facial features, and is able to generate expressions with consistent changes in texture geometry.

B. BRDF Texture Generation

To achieve physically based rendering, the industry typically employs LightStages to acquire facial reflectance properties through spectral separation of specular/albedo components. Recent research efforts towards cost-effective acquisition focus on neutral BRDF estimation. AlbedoMM [51]–[53] first used 3DMM to predict albedo, though it remained limited in quality due to its linear representation. Subsequently, AvatarMe [33] and AvatarMe++ [34] leveraged texture transfer networks for BRDF parameter regression, FitMe [40] and ID2Albedo [36] adopted StyleGAN-encoded [55] texture priors for identity-consistent mapping, while Relightify [35] advanced texture inpainting through diffusion models. Recent extensions like MoSAR [41] incorporated semi-supervised learning for auxiliary maps (ambient occlusion/translucency) synthesis, and ID2Reflectance [38] introduced identity-aware fusion networks for multi-channel reflectance generation. In contrast to these expression-static approaches, our framework explicitly models dynamic reflectance induced by expression-dependent wrinkle formation and micro-geometric pore compression effects, establishing novel correlations between geometric deformation fields and non-linear reflectance variations.

C. Diffusion Models for 3D Facial Models

The diffusion model [57] has emerged as one of the most powerful generative models in recent years, achieving target generation by iteratively denoising Gaussian noise. While their multi-step sampling procedure inherently suffers from computational latency, subsequent research has focused on optimizing the noise prediction schedule to accelerate inference [58]–[60]. Owing to their exceptional synthesis quality, the diffusion model has been widely applied to facial domains for tasks including 3D face reconstruction [61]–[63], texture synthesis [37], [64], facial animation [65], [66], and stylization [67], [68]. These existing implementations typically process isolated modalities, *e.g.*, latent codes [65], [66], texture maps [37], [64], or 3DMM coefficients [69], guided by auxiliary signals to generate target facial attributes. However, these approaches are typically limited to a single domain,



(a) The distribution of different channels

(b) The distribution of different expressions

Fig. 3: The distribution visualization of our representations.

without explicitly modeling the bidirectional dependencies between geometric deformations and reflectance variations, which can result in inconsistencies. Our work pioneers a joint diffusion framework that simultaneously models dynamic facial mesh deformations and expression-correlated BRDF material transitions, achieving photo-realistic and geometry-texture corresponding expression synthesis.

III. Methods

Our proposed framework **ExpDiff**, as illustrated in Fig. 2, aims to generate N dynamic expression models $\{\mathbf{E}_n\}_{n=1}^{N}$ with corresponding BRDF textures (albedo maps, specular maps, roughness maps, and normal maps) $\{\mathbf{A}_n, \mathbf{S}_n, \mathbf{R}_n, \mathbf{N}_n\}_{n=1}^{N}$. The input to the framework is a neutral-expression facial model $\mathbf{M} = \{(V, F) \mid V \in \mathbb{R}^{n_v \times 3}\}$ and its corresponding texture $\mathbf{T}$, where F is the pre-defined fixed topology of facial models and n_v is the number of vertexes. A textual expression code $\mathbf{e}$ is also provided as the condition. To this end, we first review the diffusion model used for high-quality generation (III-A). To effectively learn the correspondence between geometry and texture, we explain how we represent facial models and BRDF texture maps in similar forms (III-B). We then introduce our attention-based diffusion model with the guidance of textual prompts to generate dynamic expressions (III-C). Finally, we present the method of high-quality facial assets extraction (III-D).

A. Preliminary on Diffusion

The diffusion probabilistic model has recently gained popularity for image generation, which defines a forward Markov process to learn the Gaussian noises added to the input data in T steps. Specifically, the random Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is sampled and added to the input $\mathbf{x}_0$ at the t steps to get the final noisy target $\mathbf{x}_T$:

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

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right), \tag{2}$$

where β_t is used to control the quality of noise. To gradually denoise the random noisy target, the diffusion model is to learn the reverse Markov process by a network θ:

$$p_\theta\left(\mathbf{x}_{0:T}\right) = p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right). \tag{3}$$

The optimization target is to minimize the learned noise $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ and the added noise $\boldsymbol{\epsilon}$ in every step t by:

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{t,\epsilon}\left[w(t)\left\|\epsilon_\theta\left(\mathbf{x}_t; t\right) - \epsilon\right\|_2^2\right], \tag{4}$$

where $w(t)$ is a weighting function. After training the diffusion model, a random noise input can be progressively denoised to generate meaningful images. In this paper, we leverage the powerful capabilities of diffusion models to generate high-quality facial expressions and dynamic BRDF textures.

B. Facial Model Representation

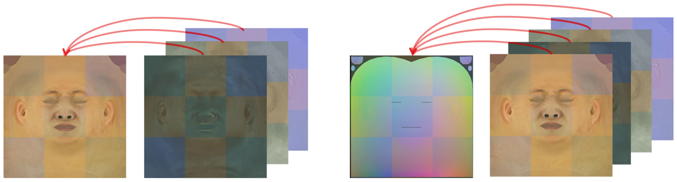
3D facial models primarily represent geometry through meshes, while reflection properties are defined with UV BRDF texture maps. However, meshes rely on discrete vertex coordinates and topological face connections, whereas textures are image-based representations, making it challenging to learn relationships between geometric deformations and reflectance variations. To address this fundamental issue, we propose a unified representation that encodes both dynamic geometric displacements and texture changes in a consistent format, facilitating joint learning of their relationships under expression variation.

First, we establish a unified representation of textures and geometry. Instead of sampling textures onto vertices, which severely degrades texture resolution, our approach encodes meshes as image-based structures according to their topology. Following prior works [66], [70], we construct position maps $\mathbf{P}$ by projecting per-vertex relative displacements through rendering. Considering the high dimensionality of images, it is challenging to directly train a diffusion model on an image-level dataset. Furthermore, the correlations between reflectance and geometric properties with Vanilla position maps and texture maps are not explicit. Therefore, we introduce a pre-trained autoencoder-decoder that maps them into a shared latent space to better explore the implicit relation between geometry and reflectance. Specically, we leverage the VAE $\mathcal{E}$ in Stable Diffusion [71] to process both geometry and texture data, achieving a unified representation as:

$$\mathbf{z}^{\mathbf{P}}, \mathbf{z}^{\mathbf{T}}, \mathbf{z}^{\mathbf{A}}, \mathbf{z}^{\mathbf{S}}, \mathbf{z}^{\mathbf{R}}, \mathbf{z}^{\mathbf{N}} = \mathcal{E}(\mathbf{P}, \mathbf{T}, \mathbf{A}, \mathbf{S}, \mathbf{R}, \mathbf{N}), \tag{5}$$

where $\mathcal{E}$ is trained on large datasets and can capture both high-frequency and low-frequency image details accurately in the latent space. It should be noted that although reflectance texture in UV-space is fundamentally different from natural images, the encoder $\mathcal{E}$ pre-trained by a large-scale dataset demonstrates sufficient representational capacity to faithfully reconstruct the textures without a fine-tuning process.

To evaluate the effectiveness of the representation method, as shown in Fig. 3, we cluster the latent codes with T-SNE for different channels of BRDF textures and various expressions across 10 target individuals, respectively. Fig. 3 (a) reveals distinct distribution patterns of reflectance texture channels in the latent space, indicating effective learning of their inter-channel relationships. Fig. 3 (b) demonstrates that expressions with distinct semantics form separate clusters, while semantically similar expressions show closer proximity in the latent space, confirming the representation's capacity to capture



(a) Joint attention

(b) Cross attention

Fig. 4: Overview of our attention module. The joint attention ensures cross-channel consistency among BRDF textures, while cross-attention establishes cross-modal correspondence between geometric deformations and texture variations.

expression-level correlations, demonstrating the effectiveness of our proposed consistent representation.

C. Expression Generation Diffusion Model

Training. After encoding geometry and reflectance attributes in a unified latent space, our goal is to generate target expression models $\{\mathbf{E}_n\}_{n=1}^N$ along with their corresponding textures $\{\mathbf{A}_n, \mathbf{S}_n, \mathbf{R}_n, \mathbf{N}_n\}_{n=1}^N$. Notably, facial expression variations adaptively manifest through both geometric deformations and textural variations, involving both shared and modality-specific patterns. For example, jaw movement causes large-scale shape displacement with little change in texture, while wrinkle formation appears mostly in texture maps due to limitations in low-poly meshes. Moreover, to ensure physical plausibility, the generated BRDF channels (albedo, roughness, specular, normal) must remain semantically consistent throughout the expression sequence. To address these challenges, we propose an attention-based diffusion model that captures inter-modal correspondences and channel-wise dependency for dynamic 3D facial asset synthesis.

As illustrated in Fig. 4, our framework adopts a dual-attention architecture composed of two modules: a joint attention module and a cross-attention module. The joint attention module enforces consistency across BRDF channels by modeling inter dependencies among reflectance textures. The cross-attention module aligns geometric deformation with expression-driven texture variations, facilitating coordinated synthesis. This dual attention design enables the generation of photometrically consistent BRDF texture maps and corresponding mesh deformations under expression dynamics. Additionally, to capture the semantic interplay between different expressions, we employ a CLIP [72]-guided approach for model denoising. Each expression is annotated with a corresponding text description, which is encoded into a semantic embedding using CLIP's pretrained text encoder $\mathcal{C}$. This embedding conditions the denoising model, encouraging the predicted noise to align with the expression semantics.

In the training process, the texture latent $\mathbf{z}^{\mathbf{T}}$ is repeated 3 times, and then they are concated with the position latent $\mathbf{z}^{\mathbf{P}}$ as neutral latents. The expression latents $\{\mathbf{z}^{\mathbf{A}}, \mathbf{z}^{\mathbf{S}}, \mathbf{z}^{\mathbf{R}}, \mathbf{z}^{\mathbf{N}}\}$ are added random Gaussian noise as Eq. (2). Following, the neutral latents and the noisy expression latents are concated as the input of the attention-based diffusion model. The training objective can be formulated as follows:

$$L = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),\mathcal{C}(\mathbf{e}),t}\left[\left\|\epsilon - \epsilon_\theta\left(w_t, t, \mathcal{C}\left(\mathbf{e}\right)\right)\right\|_2^2\right]. \tag{6}$$

Dataset	BRDF.	#ID	Exp.	Res.	Acqu.	Regis.	Avai.
FaceScape [74]	✗	847	20	4K	Capt.	Auto	✓
Faceverse [27]		128	21	2K	Capt.	None	✓
MimicMe [75]		4700	20	1K	Capt.	Auto	✓
Multiface [76]		13	65	1K	Capt.	Auto	✓
RealFaceDB [33]	✓	200	7	4K	Capt.	Auto	✗
ICT-FaceKit [26]		99	26	4K	Capt.	Auto	✗
FFHQ-UV-Intrinsics [41]		10K	1	1K	Gene.	Auto	✓
J-Reflectance (Ours)		100	**56**	**8K**	**Capt.**	**Manual**	✓
FFHQ-BRDFExp (Ours)		10K	**56**	**4K**	**Mixed.**	**Auto**	✓

TABLE I: Existing 3D face datasets. The proposed J-Reflectance dataset has the highest quality and is publicly available, surpassing existing 3D expression datasets.

This approach offers two advantages: 1) It effectively guides the model to learn coordinated variations between expressions and textural details through semantic alignment. 2) By leveraging the semantic space of CLIP [72], it enables the generation of novel expressions beyond training prompts.

Inference. At inference time, the neutral-expression position map and texture map are first injected with noise as Eq. (2) and then iteratively denoised under the guidance of expression-specific textual prompts, producing expression-aware latent codes. $\{\mathbf{z}_n^{\mathbf{P}}, \mathbf{z}_n^{\mathbf{A}}, \mathbf{z}_n^{\mathbf{S}}, \mathbf{z}_n^{\mathbf{R}}, \mathbf{z}_n^{\mathbf{N}}\}_{n=1}^N$. These latent codes are subsequently decoded by the VAE decoder $\mathcal{D}$ to reconstruct the corresponding position maps and BRDF textures as:

$$\mathbf{P}_n, \mathbf{A}_n, \mathbf{S}_n, \mathbf{R}_n, \mathbf{N}_n = \mathcal{D}(\mathbf{z}_n^{\mathbf{P}}, \mathbf{z}_n^{\mathbf{A}}, \mathbf{z}_n^{\mathbf{S}}, \mathbf{z}_n^{\mathbf{R}}, \mathbf{z}_n^{\mathbf{N}}), \tag{7}$$

where we can sample mesh $\mathbf{E}_n$ from the position map $\mathbf{P}_n$ and directly obtain the low-resolution BRDF maps.

D. High-quality Assets Generation

Due to the low resolution of Stable Diffusion [71], it is unable to directly generate pore-level detail, which limits photorealistic rendering. To enhance texture fidelity, we incorporate a super-resolution module based on Real-ESRGAN [73] to upsample the generated BRDF maps. However, direct upsampling often oversmooths dynamic features such as wrinkles. To preserve these high-frequency details, we adopt the difference map as input to enhance the perception of wrinkles. Take albedo A_n as an example, the difference map $\mathbf{W}_n$ can be obtained by interpolating between the generated expression $\mathbf{D}_n$ and the input texture $\mathbf{D}_0$ as follows:

$$\mathbf{W}_n = 1/(1 + \exp(-\mathbf{D}_n/(\mathbf{D}_0 + 1^{-8}))). \tag{8}$$

Besides, the diffusion model will introduce geometric bias into the position maps that cannot faithfully invert expressions. Therefore, we propose a directional mesh reconstruction strategy. That is, we adopt the difference between the target position map P_n and that in the "Neutral Expression" $P_{\mathbf{Neutral}}$ as guidance, the final dynamic expression meshes can be extracted via:

$$E_n' = M + f(P_n - P_{\mathbf{Neutral}}|F), \tag{9}$$

where $f(\cdot|F)$ is a mesh recovery method with topology F.

IV. Dataset

A. Existing 3D facial expression Dataset

High-quality expression datasets are essential for dynamic face generation. As the comparisons illustrated in Tab. I,

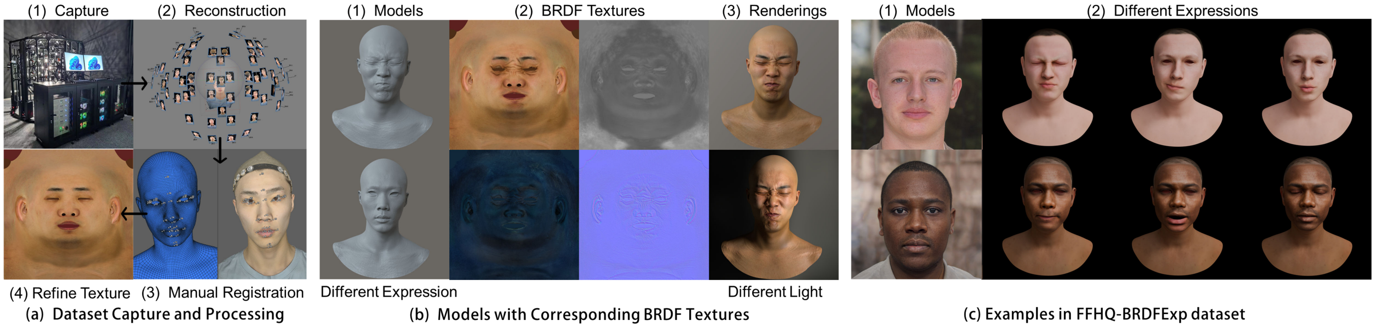


Fig. 5: Overview of our capture pipeline and datasets. (a) We utilize a Light Stage system to capture high-resolution facial images and employ skilled artists to meticulously process the data to ensure film-quality assets. (b) We showcase an identity's expression results, showing muscle-level geometry and BRDF textures with pore-level details, which can integrate seamlessly with the relighting applications. (c) We display the examples of our FFHQ-BRDFExp dataset.

early datasets primarily focus on increasing the variety of identities and expressions. However, they lack facial BRDF assets that meet physically based rendering standards, such as albedo, specular, roughness, and normal maps, to achieve photo-realistic rendering under varied lighting conditions. Several datasets, *e.g.*, RealFaceDB [33] and ICT-FaceKit [26], use Light Stages to capture BRDF textures. However, these datasets employ automated registration strategies, resulting in relatively lower quality, especially in extreme expressions. Moreover, due to the high cost of capture systems, these datasets have not been completely open-released, limiting advancements in expression generation. MoSAR [41] proposes a BRDF dataset in static expression, while lacking dynamic variations in both meshes and textures. To overcome these limitations, we propose the J-Reflectance dataset. Compared to existing works, our dataset has several advantages: **1)** 8K high-resolution textures, **2)** Artist-crafted meshes and textures, **3)** Diverse identities balanced in gender and age, **4)** Publicly available. These advantages ensure high quality and diversity of the dataset, supporting the advancement of the field.

B. Data Acquisition and Registration

Capture. We capture 3D facial expressions with a Light Stage. Our system consists of 128 individually controlled lights, capable of synchronized polarized lighting within 10ms. The capture system comprises 53 calibrated cameras that can capture 6K-resolution raw images. To ensure consistent topology, we mark subtle points on faces where non-grid ICP often struggles to maintain geometric precision, such as cheeks, eyes, and mouth. Actors are instructed to perform 32 expressions with maximum intensity to show facial muscles and wrinkles. Each expression is illuminated with 15 types of polarized lighting to calculate BRDF textures.

Processing. First, high-resolution facial scans exceeding 10 million polygons and 8K-resolution textures are produced with Agisoft Metashape [77]. Then, we employ manual retopology crafted by artists with Wrap4D [78] for ultra-high-quality mesh acquisition. In this pipeline, artists first create 32 AUs for the same template. To ensure topological consistency among different expressions, the markers are aligned with the template face to avoid vertex and UV drift, allowing our registration precision within 0.1mm. Subsequently, the 32 expressions will be processed into 56 FACS-compliant expression units manually. For UV textures, we first merge facial texture with templates, then manually remove high-frequency noise from the eyes, mouth, and nostrils. Through these steps, our facial models and BRDF textures achieve film-level precision, as shown in Fig. 5.

C. Extend Dynamic Expression to FFHQ-UV

Our dataset boasts extremely high precision, yet due to computer ethics and identity confidentiality requirements, it is essential to fully review the applicants' qualifications before releasing data to them, which somewhat constrains subsequent research. To foster advancements of community, we are inspired by FFHQ-UV [79] and MoSAR [41] to construct a high-quality dynamic dataset with the publicly available data. First, we align FFHQ-UV with our topology automatically. Since FFHQ-UV possesses consistent topology and UV mapping in the neutral expression, it can be accurately registered to our model. We then blend the aligned textures with our textures. Initially, we apply a super-resolution approach to these textures, followed by automatic rendering based on skin tone and identity, following ID2Reflectance. Blending algorithms are employed to enhance pore-level skin details while preserving identity characteristics. Similar techniques are applied to normal, roughness, and specular maps. Finally, we process these models through our ExpDiff to synthesize dynamic BRDF textures and meshes to construct the dynamic expression dataset, referred to as FFHQ-BRDFExp.

V. Experiments

In the experiments, we first introduce the implementation details. Then, we showcase our generated assets and rendering results under different lights. Afterwards, we compare our ExpDiff with the current SOTAs qualitatively and quantitatively. Finally, we conduct ablation studies to evaluate the effectiveness of our proposed modules.

A. Implementation Details

In our experiments, we use 90 identities in J-Reflectance for training, and the rest 10 identities for evaluation. All facial

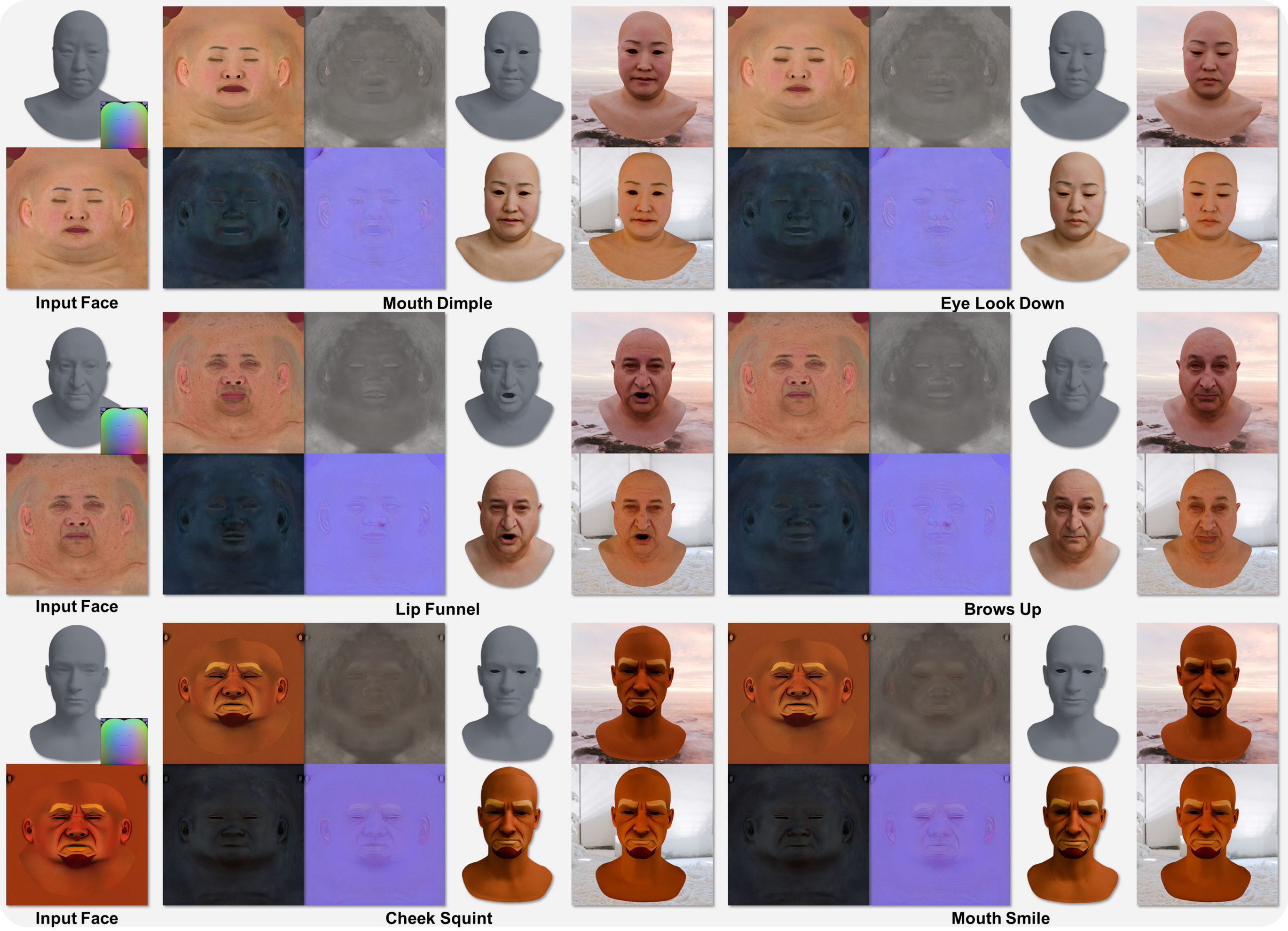


Fig. 6: We showcase some dynamic BRDF textures and rendering results of faces from different datasets, which are from J-Reflectance, 3DScanStore, and generated from Dreamface. **Please Zoom In for detailed observation.**

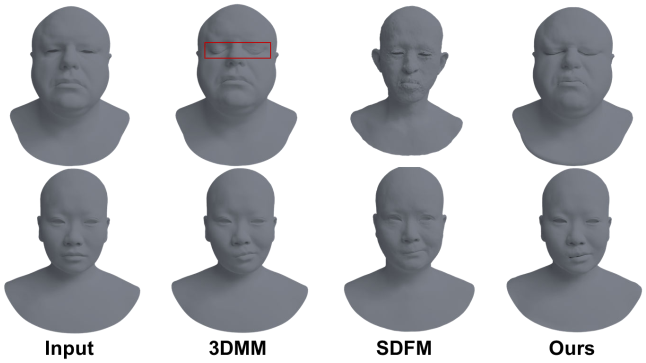


Fig. 7: Qualitative comparisons for 3D expression mesh generation. The meshes in the 1st row are from 3DScanStore, and those in the 2nd row are from J-Reflectance.

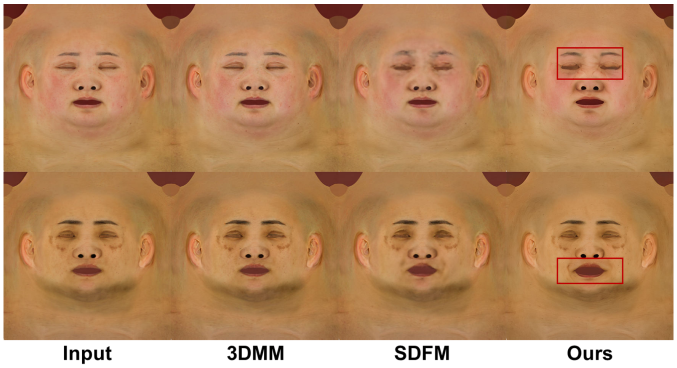


Fig. 8: Qualitative comparisons with current SOTAs for dynamic texture generation.

meshes are aligned to the template mesh and normalized to the range of [-1, 1], maintaining variability of face shape. In the diffusion training process, we use images with a resolution of 512 × 512 and set the size of the latent embedding to 1024. The model is optimized using the Adam optimizer [80] with a learning rate of $1e-5$, trained for 50000 steps with a total batch size of 128. All the implementations are based on PyTorch and set up on 4 Nvidia A6000 GPUs.

B. Generated results

Fig. 6 presents generated expression assets with rendering results under varied illumination conditions. The input faces

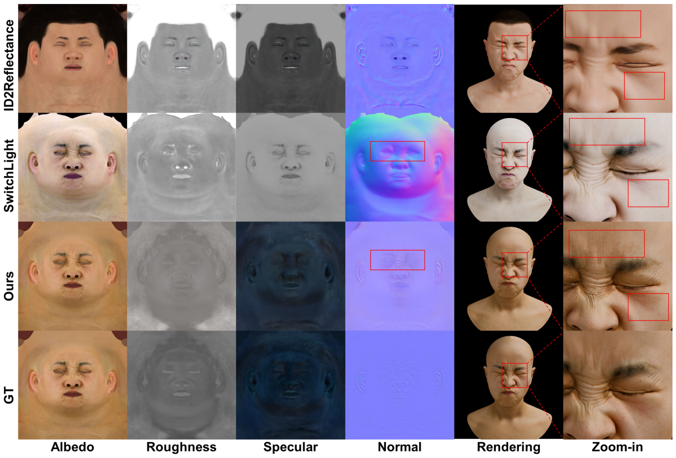


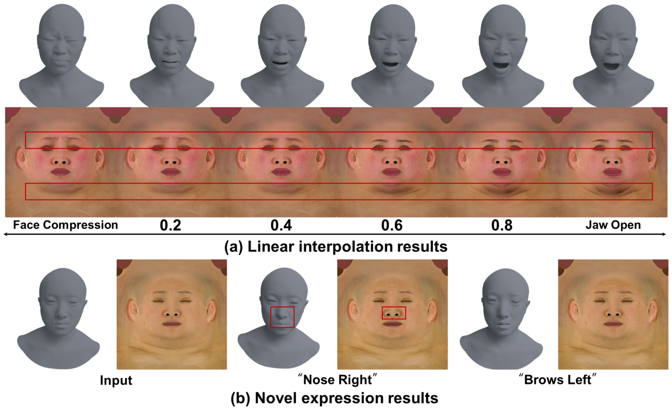
Fig. 9: Qualitative comparisons for BRDF texture generation. Please zoom-in for detailed observation.

Method	Geometry	Texture		
	FVD↓	PSNR↑	SSIM↑	LPIPS↓
3DMM	0.255	37.607	0.973	0.040
SDFM [44]	0.473	37.980	0.962	0.070
Ours	**0.214**	**39.057**	**0.983**	**0.032**

TABLE II: Quantitative comparisons towards geometry and texture with several expression generation methods. The best results are labeled in bold.

Method	Shape	Texture		
	FVD↓	PSNR↑	SSIM↑	LPIPS↓
w/o Attention	0.240	38.095	0.966	0.040
w/ Onehot	0.322	37.397	0.969	0.035
Ours	**0.214**	**39.057**	**0.983**	**0.032**

TABLE III: Ablation studies on attention modules and the guidance of textual prompts.



(a) Linear interpolation results

(b) Novel expression results

Fig. 10: The evaluation of semantic learning. (a) The interpolated generation results between "Face Compression" and "Jaw Open". (b) Novel expression results, which do not appear in training. Both results show the generalization of our ExpDiff.

are drawn from three distinct sources: the test set of J-Reflectance, the 3DScanStore dataset, and synthetic identities generated by DreamFace [37]. Our ExpDiff can synthesize expressive expressions with high-fidelity multi-channel consistent BRDF textures, which are beneficial for physically-based rendering, also demonstrating our method's generalization to unseen identities.

C. Comparisons

Qualitative Comparisons. We conduct qualitative comparisons between our proposed ExpDiff and current SOTAS, *i.e.*, 3DMM and SDFM [44], in terms of both expression geometry and dynamic textures. We retrain the models by their released code on our dataset for fair comparisons. As shown in Fig. 7, 3DMM is capable of producing generic facial expressions, but fails to capture identity-specific expression nuances and introduces interpenetration artifacts. Although SDFM can better preserve identity-specific expressions, it fails to generate semantically plausible results for facial geometries deviating from training distributions. In contrast, our method leverages a carefully designed facial representation, enabling robust generalization across diverse identities while preserving semantic expression fidelity.

As shown in the texture comparisons of Fig. 8, 3DMM fails to generate expression-dependent textures and often produces overly smooth or blurry appearance. While SDFM can generate dynamic wrinkles, it overlooks the mutual dependencies between geometric deformations and texture variations, leading to insufficient expression-specific details in critical regions such as the corners of the mouth and the bridge of the nose. In contrast, our framework achieves semantically consistent facial expression synthesis with wrinkle-level textural fidelity. The qualitative comparisons effectively demonstrate the superior performance of our proposed ExpDiff over existing SOTAs.

Our framework is capable of generating expression-dependent BRDF textures given a dynamic facial input and its corresponding textual description. To validate the effectiveness of the capability on BRDF generation, we conduct comparisons with SOTA methods (SwitchLight [81] and ID2Reflectance [38]. As the results show in Fig. 9, ID2Reflectance produces over-smoothed results, where facial details are poorly preserved. SwitchLight can maintain corresponding dynamic wrinkles in albedo, specular, and roughness maps, but it lacks high-frequency details in the normal map, limiting its ability for pore-level high-fidelity rendering. In contrast, our method maintains sharper micro-details (e.g., wrinkles and pores) across albedo and reflectance maps, achieving superior photo-realistic rendering results, demonstrating our effectiveness in high-fidelity BRDF texture synthesis.

Quantitative Comparisons. We also evaluate the effectiveness of ExpDiff quantitatively, with the geometric metric of Facial Vertex Distance (FVD), which represents the L2 distance between corresponding meshes, and the widely-used textural metrics, *i.e.*, PSNR, SSIM, and LPIPS [82], as shown in Tab. II. It shows that our proposed ExpDiff can generate Identity-specific expression meshes and dynamic textures, achieving superior generation results compared to traditional 3DMM and current SDFM.

D. Ablation Study

We investigate the effectiveness of attention modules and textual prompts, as depicted in Tab. III. The "w/o attention" removes the attention mechanisms between shape-texture and inter-texture relations during training. This leads to quality degradation due to insufficient cross-modal correlation learning. The "w/ one-hot" indicates that we replace CLIP's semantic space with randomly generated embeddings for conditioning, resulting in a significant performance decrease in both texture and geometry synthesis. The combination of them achieves the highest quantitative metrics, demonstrating the effectiveness of the proposed components.

Moreover, we conduct experiments to evaluate the effectiveness of semantic interpolation. Leveraging CLIP's semantically structured text space, our method enables seamless linear interpolation between geometric and textural attributes. As shown in Fig. 10 (a), ExpDiff achieves smooth transitions from the "facial compression" expression to "jaw open" expression, where wrinkles on the nose and chin vary corresponding to the geometric deformations. Furthermore, CLIP's semantic space permits controllable extrapolation as demonstrated in Fig. 10 (b), where we illustrate expressions absent from the training data like "nose right" and "brows left".

VI. Conclusion

In this paper, we propose ExpDiff, a generalized and effective framework for generating expression-specific facial meshes and dynamic BRDF textures. Unlike previous approaches, ExpDiff better captures both inter- and intra-relations between geometry and BRDF textures, thereby enabling the consistent generation of high-fidelity facial assets. Our framework leverages an attention-based diffusion model conditioned on textual prompts to jointly synthesize geometry and textures for diverse facial expressions, ensuring semantic and structural consistency. To enhance alignment and generalization, we introduce a unified representation for meshes and textures, which is further encoded through a VAE trained on a large-scale dataset. Furthermore, we incorporate a super-resolution module and a post-processing pipeline to refine the generated assets, improving overall visual quality. Extensive experiments demonstrate the effectiveness and generalization of our proposed method, showing our superior performance over other SOTAs. In addition, we release two high-quality expression datasets to facilitate the research community. Overall, ExpDiff offers a robust solution for text-driven 3D facial expression generation, contributing to the development of multimedia applications.

References

[1] G. Fyffe, A. Jones, O. Alexander, R. Ichikari, and P. Debevec, "Driving high-resolution facial scans with video performance capture," *TOG*, pp. 1–14, 2014. 1

[2] W. Yang, Y. Zhao, B. Yang, and J. Shen, "Learning 3d face reconstruction from the cycle-consistency of dynamic faces," *TMM*, pp. 3663–3675, 2023. 1

[3] L. Liu, X. Liu, J. Sun, C. Gao, and J. Chen, "Seif: semantic-constrained deep implicit function for single-image 3d head reconstruction," *TMM*, 2024. 1

[4] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978. 1

[5] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *CVIT*, 2000. 1, 2

[6] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec, "Multiview face capture using polarized spherical gradient illumination," *ACM TOG*, 2011. 1, 2

[7] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, P. E. Debevec *et al.*, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination." *Rendering Techniques*, p. 2, 2007. 1, 2

[8] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen, "Production-level facial performance capture using deep convolutional neural networks," in *SCA*, 2017, pp. 1–10. 1

[9] H. Li, T. Weise, and M. Pauly, "Example-based facial rigging," *TOG*, pp. 1–6, 2010. 1

[10] P. Gotardo, J. Riviere, D. Bradley, A. Ghosh, and T. Beeler, "Practical dynamic facial appearance modeling and acquisition," *TOG*, pp. 1–13, 2018. 1

[11] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH*, 1999, pp. 187–194. 2, 3

[12] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *TOG*, pp. 194–1, 2017. 2, 3

[13] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in *CVPR*, 2016, pp. 5543–5552. 2, 3

[14] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *AVSS*, 2009, pp. 296–301. 2, 3

[15] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, "3d face morphable models" in-the-wild"," in *CVPR*, 2017, pp. 48–57. 2, 3

[16] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter, "Morphable face models-an open framework," in *FG*, 2018, pp. 75–82. 2, 3

[17] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3d face morphable model," in *CVPR*, 2019, pp. 1126–1135. 2

[18] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE TVCG*, pp. 413–425, 2013. 2

[19] A. Patel and W. A. Smith, "3d morphable face models revisited," in *CVPR*, 2009, pp. 1327–1334. 2

[20] S. Ploumpis, E. Ververas, E. O'Sullivan, S. Moschoglou, H. Wang, N. Pears, W. A. Smith, B. Gecer, and S. Zafeiriou, "Towards a complete 3d morphable model of the human head," *TPAMI*, pp. 4142–4160, 2020. 2

[21] H. Dai, N. Pears, W. A. Smith, and C. Duncan, "A 3d morphable model of craniofacial shape and texture variation," in *ICCV*, 2017, pp. 3085–3093. 2

[22] S. Giebenhain, T. Kirschstein, M. Georgopoulos, M. Rünz, L. Agapito, and M. Nießner, "Learning neural parametric head models," *arXiv preprint arXiv:2212.02761*, 2022. 2

[23] S. Wu, Y. Yan, Y. Li, Y. Cheng, W. Zhu, K. Gao, X. Li, and G. Zhai, "Ganhead: Towards generative animatable neural head avatars," in *CVPR*, 2023, pp. 437–447. 2

[24] X. Fan, S. Cheng, K. Huyan, M. Hou, R. Liu, and Z. Luo, "Dual neural networks coupling data regression with explicit priors for monocular 3d face reconstruction," *TMM*, pp. 1252–1263, 2020. 2

[25] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3d faces using convolutional mesh autoencoders," in *ECCV*, 2018, pp. 704–720. 2

[26] 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 *CVPR*, 2020, pp. 3410–3419. 2, 5, 6

[27] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, and Y. Liu, "Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset," in *CVPR*, 2022, pp. 20 333–20 342. 2, 5

[28] R. A. Potamias, A. Lattas, S. Moschoglou, S. Ploumpis, and S. Zafeiriou, "Animateme: 4d facial expressions via diffusion models." 2

[29] N. Otberdout, C. Ferrari, M. Daoudi, S. Berretti, and A. Del Bimbo, "Sparse to dense dynamic 3d facial expression generation," in *CVPR*, 2022, pp. 20 385–20 394. 2, 3

[30] X. Lu, Z. Lu, Y. Wang, and J. Xiao, "Landmark guided 4d facial expression generation," in *SIGGRAPH Asia*, 2023, pp. 1–2. 2, 3

[31] K. Zou, S. Faisan, B. Yu, S. Valette, and H. Seo, "4d facial expression diffusion model," *TOMM*, 2023. 2, 3

[32] X. Lu, C. Zhuang, Z. Lu, Y. Wang, and J. Xiao, "Fc-4dfs: Frequency-controlled flexible 4d facial expression synthesizing," in *ACM MM*, 2024, pp. 10 882–10 890. 2, 3

[33] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou, "Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild"," in *CVPR*, 2020, pp. 760–769. 2, 3, 5, 6

[34] A. Lattas, S. Moschoglou, S. Ploumpis, B. Gecer, A. Ghosh, and S. Zafeiriou, "Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans," *TPAMI*, pp. 9269–9284, 2021. 2, 3

[35] F. P. Papantoniou, A. Lattas, S. Moschoglou, and S. Zafeiriou, "Relightify: Relightable 3d faces from a single image via diffusion models," in *ICCV*, 2023, pp. 8806–8817. 2, 3

[36] X. Ren, J. Deng, Y. Cheng, J. Guo, C. Ma, Y. Yan, W. Zhu, and X. Yang, "Monocular identity-conditioned facial reflectance reconstruction," in *CVPR*, 2024, pp. 885–895. 2, 3

[37] L. Zhang, Q. Qiu, H. Lin, Q. Zhang, C. Shi, W. Yang, Y. Shi, S. Yang, L. Xu, and J. Yu, "Dreamface: Progressive generation of animatable 3d faces under text guidance," *TOG*, pp. 1–16, 2023. 2, 3, 8

[38] X. Ren, J. Deng, Y. Cheng, J. Guo, C. Ma, Y. Yan, W. Zhu, and X. Yang, "Monocular identity-conditioned facial reflectance reconstruction," in *CVPR*, 2024. 2, 3, 8

[39] R. Liu, Y. Cheng, S. Huang, C. Li, and X. Cheng, "Transformer-based high-fidelity facial displacement completion for detailed 3d face reconstruction," *TMM*, pp. 799–810, 2023. 2

[40] A. Lattas, S. Moschoglou, S. Ploumpis, B. Gecer, J. Deng, and S. Zafeiriou, "Fitme: Deep photorealistic 3d morphable model avatars," in *CVPR*, 2023, pp. 8629–8640. 2, 3

[41] A. Dib, L. G. Hafemann, E. Got, T. Anderson, A. Fadaeinejad, R. M. Cruz, and M.-A. Carbonneau, "Mosar: Monocular semi-supervised model for avatar reconstruction using differentiable shading," in *CVPR*, 2024, pp. 1770–1780. 2, 3, 5, 6

[42] B. Gecer, A. Lattas, S. Ploumpis, J. Deng, A. Papaioannou, S. Moschoglou, and S. Zafeiriou, "Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks," in *ECCV*, 2020, pp. 415–433. 2

[43] J. Li, Z. Kuang, Y. Zhao, M. He, K. Bladin, and H. Li, "Dynamic facial asset and rig generation from a single scan." *ACM Trans. Graph.*, pp. 215–1, 2020. 2, 3

[44] P. Chandran, D. Bradley, M. Gross, and T. Beeler, "Semantic deep face models," in *3DV*, 2020, pp. 345–354. 2, 3, 8

[45] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014. 2, 3

[46] M. Goesele, B. Curless, and S. M. Seitz, "Multi-view stereo revisited," in *CVPR*, 2006, pp. 2402–2409. 2

[47] H. Wu, J. Jia, J. Xing, H. Xu, X. Wang, and J. Wang, "Mmface4d: a large-scale multi-modal 4d face dataset for audio-driven 3d face animation," *arXiv preprint arXiv:2303.09797*, 2023. 2

[48] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, 1992, pp. 586–606. 2

[49] P. Ji, H. Li, L. Jiang, and X. Liu, "Light-weight multi-view topology consistent facial geometry and reflectance capture," in *CGI*, 2021, pp. 139–150. 2

[50] G. Fyffe, K. Nagano, L. Huynh, S. Saito, J. Busch, A. Jones, H. Li, and P. Debevec, "Multi-view stereo on consistent face topology," in *CGF*, 2017, pp. 295–309. 2

[51] W. A. Smith, A. Seck, H. Dee, B. Tiddeman, J. B. Tenenbaum, and B. Egger, "A morphable face albedo model," in *CVPR*, 2020, pp. 5011–5020. 3

[52] A. Tewari, H.-P. Seidel, M. Elgharib, C. Theobalt *et al.*, "Learning complete 3d morphable face models from images and videos," in *CVPR*, 2021, pp. 3361–3371. 3

[53] A. Dib, C. Thebault, J. Ahn, P.-H. Gosselin, C. Theobalt, and L. Chevallier, "Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing," in *ICCV*, 2021, pp. 12 819–12 829. 3

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, pp. 139–144, 2020. 3

[55] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410. 3

[56] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *NeurIPS*, pp. 21 696–21 707, 2021. 3

[57] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *TPAMI*, pp. 10 850–10 869, 2023. 3

[58] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. 3

[59] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *ICML*. PMLR, 2023, pp. 32 211–32 252. 3

[60] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, "One-step diffusion with distribution matching distillation," in *CVPR*, 2024, pp. 6613–6623. 3

[61] 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

[62] T. Kirschstein, S. Giebenhain, and M. Nießner, "Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars," in *CVPR*, 2024, pp. 5481–5492. 3

[63] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou, "Arc2face: A foundation model for id-consistent human faces," in *ECCV*, 2024, pp. 241–261. 3

[64] M. Zhou, R. Hyder, Z. Xuan, and G. Qi, "Ultravatar: A realistic animatable 3d avatar diffusion model with authenticity guided textures," in *CVPR*, 2024, pp. 1238–1248. 3

[65] Q. Zhao, P. Long, Q. Zhang, D. Qin, H. Liang, L. Zhang, Y. Zhang, J. Yu, and L. Xu, "Media2face: Co-speech facial animation generation with multi-modality guidance," in *SIGGRAPH*, 2024, pp. 1–13. 3

[66] X. Li, J. Wang, Y. Cheng, Y. Zeng, X. Ren, W. Zhu, W. Zhao, and Y. Yan, "Towards high-fidelity 3d talking avatar with personalized dynamic texture," *arXiv preprint arXiv:2503.00495*, 2025. 3, 4

[67] H. Wu, M. Zhao, Z. Hu, C. Fan, L. Li, W. Chen, R. Zhao, and X. Yu, "Ice: Interactive 3d game character facial editing via dialogue," *TMM*, 2025. 3

[68] Z. Chen, Y. Yan, S. Liu, Y. Cheng, W. Zhao, L. Li, M. Bi, and X. Yang, "Revealing directions for text-guided 3d face editing," *TMM*, 2025. 3

[69] Z. Sun, T. Lv, S. Ye, M. Lin, J. Sheng, Y.-H. Wen, M. Yu, and Y.-j. Liu, "Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models," *TOG*, pp. 1–9, 2024. 3

[70] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018, pp. 534–551. 4

[71] 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. 4, 5

[72] 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. 5

[73] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *ICCV*, 2021, pp. 1905–1914. 5

[74] 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 *CVPR*, 2020, pp. 601–610. 5

[75] A. Papaioannou, B. Gecer, S. Cheng, G. Chrysos, J. Deng, E. Fotiadou, C. Kampouris, D. Kollias, S. Moschoglou, K. Songsri-In *et al.*, "Mimicme: A large scale diverse 4d database for facial expression analysis," in *ECCV*, 2022, pp. 467–484. 5

[76] C.-h. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, E. Brockmeyer, L. Evans, T. Godisart, H. Ha, X. Huang *et al.*, "Multiface: A dataset for neural face rendering," *arXiv preprint arXiv:2207.11243*, 2022. 5

[77] "Agisoft metashape: Agisoft metashape," https://www.agisoft.com/. 6

[78] "Wrap4d - faceform," https://faceform.com/wrap4d/. 6

[79] H. Bai, D. Kang, H. Zhang, J. Pan, and L. Bao, "Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction," in *CVPR*, 2023, pp. 362–371. 6

[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2015. 7

[81] H. Kim, M. Jang, W. Yoon, J. Lee, D. Na, and S. Woo, "Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting," in *CVPR*, 2024, pp. 25 096–25 106. 8

[82] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595. 8