This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

EmoDiffTalk: Emotion-aware Diffusion for Editable 3D Gaussian Talking Head

Chang Liu[1,2] Tianjiao Jing[1] Chengcheng Ma[1] Xuanqi Zhou[1] Zhengxuan Lian[1]
Qin Jin[2] Hongliang Yuan[3] Shi-Sheng Huang[1*]
[1]Beijing Normal University [2]Renmin University of China [3]Xiaomi Inc

Abstract

*Recent photo-realistic 3D talking head via 3D Gaussian Splatting still has significant shortcoming in emotional expression manipulation, especially for **fine-grained** and **expansive** dynamics emotional editing using multi-modal control. This paper introduces a new editable 3D Gaussian talking head, i.e. EmoDiffTalk. Our key idea is a novel Emotion-aware Gaussian Diffusion, which includes an action unit (AU) prompt Gaussian diffusion process for fine-grained facial animator, and moreover an accurate text-to-AU emotion controller to provide accurate and expansive dynamic emotional editing using text input. Experiments on public EmoTalk3D and RenderMe-360 datasets demonstrate superior emotional subtlety, lip-sync fidelity, and controllability of our EmoDiffTalk over previous works, establishing a principled pathway toward high-quality, diffusion-driven, multimodal editable 3D talking-head synthesis. To our best knowledge, our EmoDiffTalk is one of the first few 3D Gaussian Splatting talking-head generation framework, especially supporting continuous, multimodal emotional editing within the AU-based expression space.Please visit our website for more details and information:https://liuchang883.github.io/EmoDiffTalk/*

1. Introduction

Photo-realistic 3D talking head has achieved remarkable progress using representations such as 3DMM [39, 48, 49, 51, 59], neural radiance fields (NeRF) [2, 18, 20, 29, 40, 52, 63, 69] and 3D Gaussian Splatting (3DGS) [9, 14, 17, 30, 32, 34, 36, 42, 58, 65, 66]. However, most of current 3D talking head focus on photo-realistic portrait rendering quality [12, 46, 53, 60] and audio-lip synchronization [8, 44], which are still limited in semantic-level talking head editing such as emotional expressions [16]. One of the main challenging for emotional talking head manipulation lies in the ambiguous audio-to-emotion and emotion-to-expression mappings, which makes it difficult to conduct accurate emotional editing within the dynamic 3D talking head space [15, 70]. Some early works such as EAMM [25] adopt to perform emotional editing within the implicit latent space using control signals such as reference images. For more better emotional-expression editing, EmoTalk [39] and the subsequent works [19] proposed to disentangle the emotion from speech to rich 3D facial generation. Recent Hallo3 [11] utilize text prompt to the audio-diffusion process, which has achieved impressive emotional 3D talking head generation. However, the quality of emotional editing but not only stylization [51] or personalization [3] is still limited, especially for **fine-grained** and **expansive** emotional editing using multi-modal control.

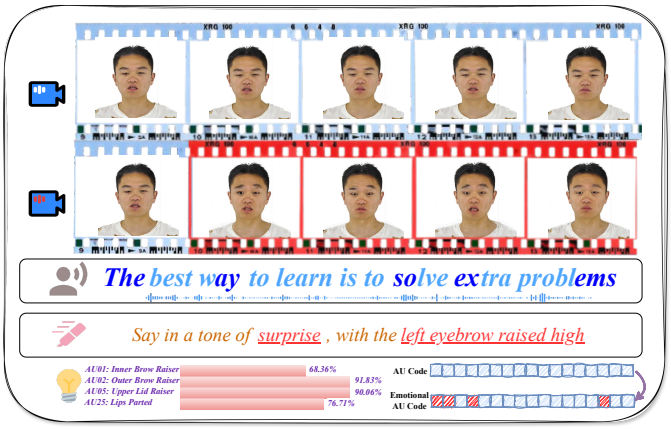


Figure 1. **Text-to-AU based emotion editing results of our EmoDiffTalk.** Our EmoDiffTalk not only supports fine-grained speech-driven (third row) 3DGS talking head generation (first row), but also enabling expansive and accurate text-based emotion editing (second and fourth rows). The AU Code (bottom) are also demonstrated. Please refer to our demo video for more text driven 3D talking head editing results.

To address such issue, this paper proposes a new emotional controllable 3D talking head based on 3D Gaussian splatting [19], called *EmoDiffTalk*, which generates high fidelity dynamic portrait rendering quality while enabling fine-grained emotional editing using cross-modal input such as text. Unlike previous emotion editing talking heads

*corresponding author.

incorporating speech and emotion within implicit feature space [25, 39] or direct diffusion process [11], our key insight is to model the emotion-to-expression mapping via action unit (AU) code space [13], and introduce the AU code as effective emotional embedding to prompt the Gaussian diffusion process to predict the dynamic Gaussian primitive attributions. This formulates as a novel emotional-aware Gaussian diffusion, within which we first build a AU-prompt Gaussian diffusion for fine-grained facial animator, and then introduce a text-to-AU emotion controller to distill the Gaussian diffusion for accurate emotion editing using text input. Benefit from the emotional-aware Gaussian diffusion, our EmoDiffTalk can reconstruct photo-realistic 3D talking head with more better fine-grained expressions, and enables high accurate and easy-to-use emotion editing using texts as shown in Fig. 1.

To evaluate the effectiveness of our EmoDiffTalk, we conducted extensive evaluation on public EmoTalk3D [19] and RenderMe-360 [38] datasets, comparing with previous SOTA approaches such as EAMM[25], SadTalker[69], Real3D-Portrait[64], EmoTalk3D[19], Hallo3[11] and EchoMimic[7]. From the experiments, on average our EmoDiffTalk boosts PSNR (+4.56 dB vs EmoTalk3D) and CPBD (+16.1 % vs Hallo3) on EmoTalk3D, retains best LPIPS and lowers LMD, and on RenderMe-360 cuts LMD by 29.4 % and adds 1.28 dB PSNR over Hallo3; user studies also exceed Hallo3 in video fidelity (+5.3 %), image quality (+4.7 %) on EmoTalk3D, video fidelity (+1.7 %), and emotion control (+2.2 %) on RenderMe-360. To our best knowledge, our approach becomes a new state-of-the-art emotional controllable 3D talking head, especially in simultaneously delivering photo-realistic dynamic reconstruction and precise, text-driven emotional editing with superior expression accuracy, naturalness, and flexibility.

2. Related Work

2.1. From 2D Generation to 3D Talking Heads

The pursuit of photorealistic talking heads has evolved along two primary paradigms. **2D-based methods** prioritize generation efficiency and frame-wise realism, achieving significant success in audio-lip synchronization [8, 44] and one-shot video editing [6, 24, 31, 45, 47, 50, 57]. For instance, EAMM [24] enables emotional editing via reference images in a latent space, while Hallo3 [11] and EchoMimic [7] leverage powerful video diffusion models for highly dynamic portraits. However, these methods often struggle with 3D consistency and free-viewpoint synthesis [33]. This limitation has spurred the advancement of **3D-aware approaches**. Early models relied on 3D Morphable Models (3DMM) [27, 39, 48, 59] for explicit control but were constrained by their linear expressivity. The emergence of neural fields, particularly NeRF [18, 20, 21, 52, 63, 69], and more recently, 3D Gaussian Splatting (3DGS) [26] , has enabled high-fidelity novel view synthesis. Recent efforts have significantly pushed the boundaries of 3DGS for high fidelity 3D scene reconstruction [41, 43] with ultrafast rendering [62], physical based deferred rendering [61], and 4D dynamics reconstruction [28], which inspired many subsequent talking head reconstruction [9, 19, 30, 66]. But these talking head approaches mainly focus on photorealism appearance reconstruction. Precise and expansive *semantic-level* control, particularly for fine-grained emotional expressions, remains a critical challenge in 3D talking heads [16].

2.2. Emotion Control and AU-Based Synthesis

A central challenge in controllable talking head generation is the ambiguous mapping from audio or semantic prompts to expressive facial dynamics [15, 70]. Research has explored various control signals. While methods like EmoTalk [39] and its 3DGS extension [19] disentangle emotion from speech, and diffusion-based approaches like Diffposetalk [51] and Hallo3 [11] use text prompts for stylization, they often operate on holistic expressions or implicit features, lacking fine-grained anatomical grounding [54]. In contrast, the Facial Action Coding System (FACS) [1, 13], which defines Action Units (AUs) corresponding to specific facial muscle movements, provides a foundational, interpretable framework for precise expression modeling. Although widely used in analysis and 3DMM-based animation [32, 39], the integration of AUs as a central control mechanism for *editable* 3D Gaussian-based avatars is novel. Our work bridges this gap by introducing an explicit AU code space to mediate between multi-modal inputs (speech, text) and the Gaussian diffusion process, enabling anatomically-aware fine-grained emotional editing.

3. Method

Given a subject's multi-view images $\mathcal{I} = \{\mathbf{I}_i\}$ accompanied with a facial template $\mathcal{T}_f$ as like EmoTalk3D [19], our goal is to reconstruct the dynamic 3DGS primitives representation $\mathcal{G} = \{g_t^i\}$, which can be driven with fine-grained facial expression by any speech input, and more importantly supports emotion editing using text input. As show in Fig. 2, we first perform a canonical Gaussian rig reconstruction (Sec. 3.1) to get a basic canonical 3DGS rig along with high accurate Gaussian color fetching. Then we set up a novel emotion-aware Gaussian diffusion, by which we can drive the canonical 3DGS rig using any speech input and perform text-based emotion editing simultaneously.

Our emotion-aware Gaussian diffusion consists of two key components: (1)AU-prompt Gaussian Diffusion (Sec. 3.2), which encodes audio signals into Action Unit (AU) codes and subsequently decodes them to predict Gaussian position along with all other fine-grained Gaussian splatting attributes, and (2)Text-to-AU Emotion Controller

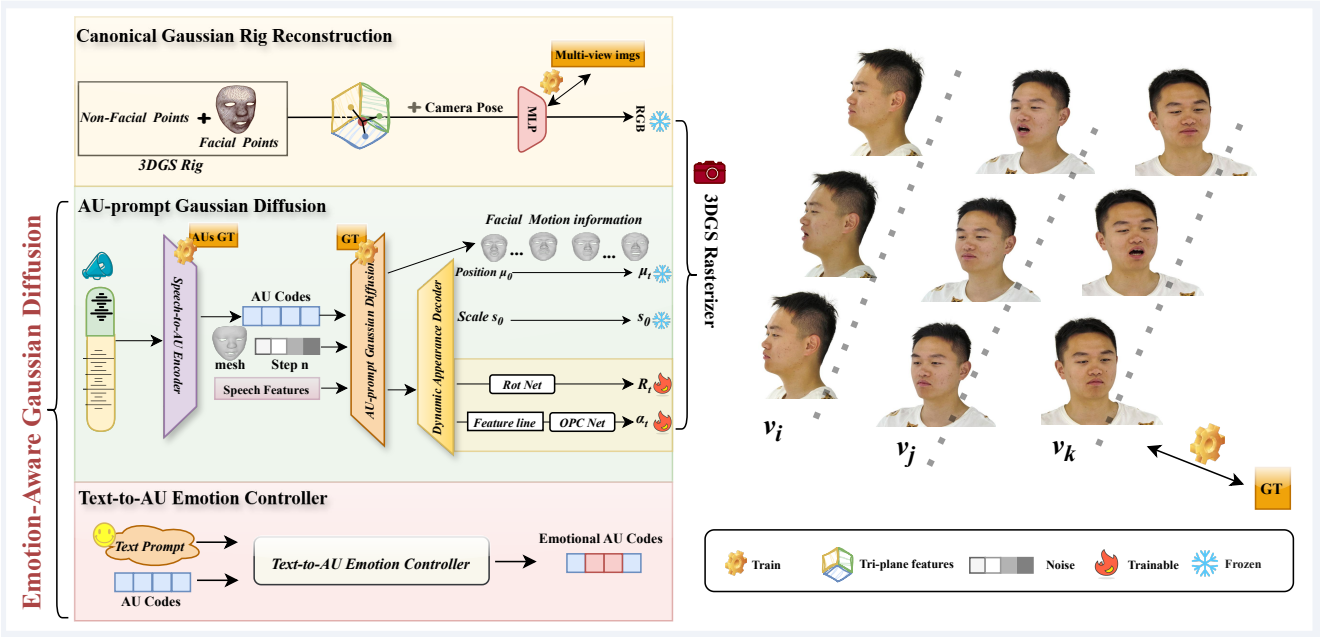


Figure 2. **Overall Pipeline of EmoDiffTalk.** We first reconstruct the canonical Gaussian rig from a subject's multi-view images, and then perform an emotion-aware Gaussian diffusion, including a AU-prompt Gaussian Diffusion process and Text-to-AU Emotion Controller, to animate the canonical Gaussian rig with any speech and text-based emotion prompt input, ultimately enabling free-viewpoint rendering via the 3DGS splatting.

(Sec. 3.3), which establishes a direct mapping from a textual emotion prompt to an Activation vector enabling freely editable emotional adjustment.

3.1. Canonical Gaussian Rig Reconstruction

Before conducting speech driven talking head generation, we first reconstruct the subject's *canonical* 3DGS rig from the multi-view image input as like EmoTalk3D [19], which divides the 3D head into facial and non-facial regions and then creates a motion-binding between such two regions. Using the canonical 3DGS rig, we can focus on the facial dynamics during the thereafter talking head animation, since the non-facial region can be driven along with the facial region automatically and efficiently.

Different from previous canonical 3DGS rig [19] that utilizes Spherical Harmonics (SH) to fetch view-dependent Gaussian's color appearance, we employ a triplane-based Gaussian color prediction, which can achieve better color attribution fetching especially in the Gaussian attribution decoding processing of our AU-Prompt Gaussian diffusion (Sec. 3.2) thereafter. Specifically, the color c for each point is generated by:

$$c = \mathcal{M}(F_{xy}(x, y) \oplus F_{xz}(x, z) \oplus F_{yz}(y, z)), \quad (1)$$

where $\mathcal{M}$ is an MLP decoder and F_{xy}, F_{xz}, F_{yz} are triplane feature maps.

We denote the canonical 3DGS $G_0 = \{\mu_0, S_0, R_0, \alpha_0, c_0\}$, and during the speech driven animation, the color attribution c_0 will be updated by fetching the triplane based color prediction, while leave the other attributions by the AU-Prompt Gaussian diffusion.

3.2. AU-prompt Gaussian Diffusion

Based on the canonical 3DGS rig, we propose to map the speech feature to dynamic 3DGS attribution using an AU-prompt Gaussian diffusion on the facial region, which mainly consists of three stages: (1) **Speech-to-AU Encoder**, which maps the speech feature to AU Codes, (2) **AU-prompt Gaussian Diffusion**, serving as speech-to-facial motion prompted with the AU Codes, and (3) **Dynamic Appearance Decoder**, which decodes the dynamic attribution offsets relative to the canonical 3DGS rig, ultimately generating a high-quality 3D talking head.

Speech-to-AU Encoder. As shown in Fig. 2, we first extract self-supervised HuBERT [23] features from raw audio, which encode richer phonetic/prosodic information than spectrograms. The feature sequence is denoted as:

$$\boldsymbol{A}_{0:T-1} = \{\boldsymbol{A}_t \mid t = 0, \ldots, T-1\}, \quad \boldsymbol{A}_t \in \mathbb{R}^{768}. \quad (2)$$

Then we use a multi-layer Transformer $\mathrm{Enc}(\cdot)$ processing $\boldsymbol{A}_{0:T-1}$ to predict AU Codes:

$$\boldsymbol{E}_{0:T-1} = \mathrm{Enc}(\boldsymbol{A}_{0:T-1}; \theta), \quad (3)$$

where lower layers use constrained attention to capture rapid articulatory changes (e.g., lip closure), while upper layers model slower prosodic variations.

AU-prompt Gaussian Diffusion. To incorporate the AU Codes obtained from the first stage into facial motion modeling, we further build an AU-prompt Gaussian diffusion model, which is inspired from Diffposetalk [51]. Specifically, we first utilize AU Codess $\boldsymbol{E}_0 ... \boldsymbol{E}_{T-1}$ extracted from audio features as one input to guide the denoising process, rather than style information extracted from 2D video. Then we employ mesh point positions $\boldsymbol{P}$ as the initial template, with the network learning the offsets of all mesh points $\Delta \boldsymbol{P}_t$ no longer represents the 3DMM coefficients $\boldsymbol{\beta}$. Naturally, learning point offsets presents greater challenges than predicting a priori 3DMM coefficients. Therefore, we designed a series of losses to constrain facial structural changes and conducted ablation experiments comparing our approach with simpler GRU networks predicting point offsets without diffusion strategies. The denoising process of our AU-prompt Gaussian diffusion model is formulated as:

$$\hat{\boldsymbol{x}}^0_{0:T} = D_\theta(\boldsymbol{x}^n_{0:T}, \boldsymbol{P}, \boldsymbol{E}_{0:T}, \boldsymbol{A}_{0:T}, n), \tag{4}$$

which establishes a fine-grained binding between AU Codes dimensions and specific facial point motions, effectively learning the relationship between Action Units and facial movements.

Dynamic Appearance Decoder. After the AU-prompt Gaussian diffusion, we then decode the dynamic appearance attributions of the 3DGS primitives. To effectively decode the fine-grained 3DGS dynamic attributions, we set up separate attribution decoders with the guidance of AU Codes, including rotation and opacity respectively.

For rotation decoder, we introduce a RotNet, which consists of a 3-layer MLP that predicts rotation parameters for each Gaussian point by integrating motion cues with AU information:

$$R_t = \mathcal{N}_{\text{Rot}}(R_0, E_t, \mu_t), \tag{5}$$

where μ_t represents the deformed Gaussian position at timestep t derived from the diffusion output.

For opacity decoder, we designed a compact Feature Line whose initial purpose was to store implicit features regarding variations in flame expression coefficients [55]. Here, we repurpose it to store fine-grained features of AU Codes changes related to opacity. The learnable feature line $\mathcal{F} \in \mathbb{R}^{17\times Q\times 16}$ captures AU-specific opacity patterns, where Q is the number of facial Gaussian points. This continuous AU-based representation enables smooth expression interpolation and infinite blending possibilities. Then we use an OPCNet, another 3-layer MLP, processes the aggregated features to predict opacity changes conditioned on both motion information and AU Codes:

$$\Delta\alpha^i_t = \mathcal{N}_{\text{OPC}}(\mathbf{f}^i_t, \mathbf{E}_t, \mu_t), \tag{6}$$

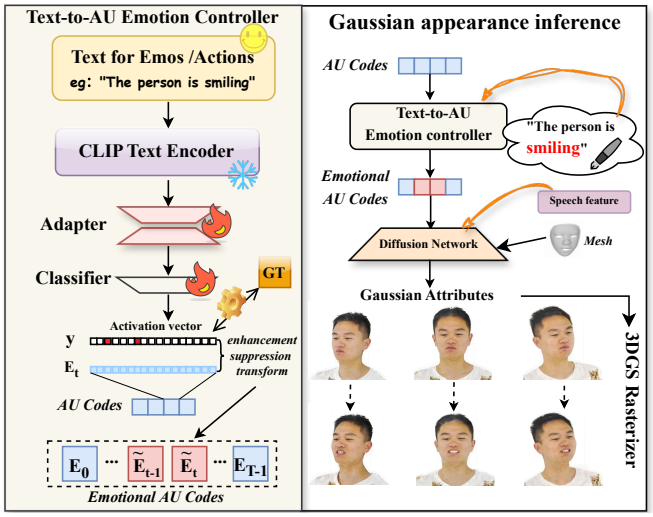


Figure 3. Text-to-AU Emotion Controller pipeline (left) and AU-based emotion editing for Gaussian appearance inference (right).

where $\mathbf{f}^i_t$ is the feature combination weighted by AU intensities. Please refer to our supplementary materials for more details.

3.3. Text-to-AU Emotion Controller

Finally, to support accurate text-based emotion editing, we further introduce a Text-to-AU Emotion Controller [4] .As shown in Fig. 3, this module maps a textual emotion prompt (e.g., "the person is smiling", "the person is sad") to a binary AU activation vector $\mathbf{y} \in \{0,1\}^K$, where $y_k = 1$ denotes AU k should be emotionally up-regulated, and $y_k = 0$ implies suppression. Then We apply a lightweight enhancement–suppression transform to those AU Codes $\mathbf{E}_t$ selected by user for activation by the binary mask $\mathbf{y}$. Each resulting Emotional AU Code $\tilde{\mathbf{E}}_t$ is computed as

$$\tilde{\mathbf{E}}_t = \mathbf{E}_t \odot (1 + \alpha\mathbf{y}) - \beta(1 - \mathbf{y}) \odot \mathbf{E}_t, \tag{7}$$

where $\alpha, \beta > 0$ are enhancement and suppression coefficients, respectively, and $\odot$ denotes element-wise multiplication. Intuitively, activated AUs ($y_k = 1$) are amplified by a factor $1 + \alpha$, while inactive AUs ($y_k = 0$) are attenuated by $1 - \beta$. After modulation, we obtain the emotional AU Codes

$$\tilde{\mathbf{E}}_{0:T-1} = \{\tilde{\mathbf{E}}_t\}^{T-1}_{t=0}, \tag{8}$$

which replaces $\mathbf{E}_{0:T-1}$ as the conditioning input to the AU-aware Diffusion Transformer. This guides the prediction of dynamic Gaussian positions and related attributes, injecting controlled affect while preserving articulation fidelity inherited from the original speech-driven codes.

3.4. Training

Our training pipeline employs a coherent four-stage optimization strategy that systematically builds upon the core

components of our EmoDiffTalk framework. Below we outline the macro-level training stages corresponding to Sec. 3.1- 3.3, with detailed implementations available in our supplementary material.

Stage 1: Speech-to-AU Encoder Training. We first train the transformer-based encoder to map audio features to AU Codes, establishing the foundation for emotional semantics. The training utilizes a combination of regression loss($\mathcal{L}_{reg}$) for AU intensity prediction and temporal consistency loss ($\mathcal{L}_{temp}$) to maintain coherent emotional dynamics across frames.

$$\mathcal{L}_{\text{AU}} = \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} \tag{9}$$

Stage 2: AU-prompt Gaussian Diffusion Training. we condition the diffusion model on the AU Codes produced in Stage 1 to model temporal position offsets of facial Gaussians. Instead of a generic noise prediction loss, we employ a unified geometric objective combining global vertex reconstruction, velocity/acceleration temporal coherence, deformation regularization, and fine-grained lip fidelity:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{vertex}}\mathcal{L}_{\text{vertex}} + \lambda_{\text{motion}}\mathcal{L}_{\text{motion}} + \lambda_{\text{deform}}\mathcal{L}_{\text{deform}} + \lambda_{\text{lip}}\mathcal{L}_{\text{lip}}. \tag{10}$$

Stage 3: Dynamic Appearance Decoder Training. This Stage centers on two lightweight networks:RotNet for stable primitive updates via a simple reconstruction loss and OPCNet for expressive refinement through reconstruction plus motion-aware regularization.
Specifically, we apply four terms in OPCNet training: mixed reconstruction, motion–amplitude coupling, sparsity + temporal smoothness, and displacement limiting:

$$\mathcal{L}_{\text{OPC}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{opcmotion}} + \mathcal{L}_{\text{dist}}. \tag{11}$$

Stage 4: Text-to-AU Emotion Controller Training. Finally, we train the Text-to-AU Emotion Controller to map textual emotion prompt to AU activation vector. This module employs a multi-objective learning strategy:

$$\mathcal{L}_{\text{control}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{infoNCE}}\mathcal{L}_{\text{infoNCE}} \tag{12}$$

where binary cross-entropy loss ensures accurate AU activation prediction, and InfoNCE loss aligns text embeddings with AU semantics.

4. Experiment

4.1. Datasets

We conduct experiments on two widely-used talking face datasets to evaluate our method: EmoTalk3D [19] and RenderMe-360 [38].

EmoTalk3D Dataset is a high-quality 3D emotional talking face dataset. It contains multi-view video recordings with synchronized audio and 3D facial annotations. The dataset covers multiple subjects expressing various emotions including neutral, happy, sad, angry, surprised, and disgusted. Each sequence includes accurate 3D ground truth, making it suitable for training and evaluating 3D-aware talking head generation with emotion control.

RenderMe-360 Dataset is a large-scale high-fidelity multi-view human head avatar dataset. It provides synchronized 60-view 2K video captures with audio and rich 3D/semantic annotations. The collection spans 500 diverse subjects performing speaking, facial expressions, and hair motion under varied appearance conditions. Each sequence bundles calibrated cameras, per-frame images, and structured labels, enabling training and evaluation of controllable high-fidelity talking head, view synthesis, expression editing, and hair-aware avatar generation methods.

4.2. Implementation Details

Training details. Our model is trained on a single NVIDIA RTX 5090 GPU (32GB) for approximately 3 days, with the training time carefully allocated to optimize different components: one day is dedicated to the Canonical Gaussian Rig Reconstruction, one day to the combined training of the Speech-to-AU Encoder and AU-prompt Gaussian Diffusion, and another day to the Dynamic Appearance Decoder. The Text-to-AU Emotion Controller requires less than an hour due to its lightweight design. This structured approach ensures efficient learning of both static and dynamic Gaussian attributes.The learning rate is set to 1e-4 with cosine annealing schedule. Batch size is 4 for canonical reconstruction, 8 for diffusion training, and 16 for emotion controller training. We use AdamW [35] optimizer with gradient clipping. Input images are resized to 512×512. More details are provided in the supplementary materials.

4.3. Evaluation

Baseline and Metrics. To comprehensively assess the effectiveness of our approach, we select representative methods from two mainstream paradigms that cover key recent advances as baselines: 2D video generation methods include EAMM [24], Hallo3 [10], and EchoMimic [6]; 3D-aware methods include SadTalker [68], Real3D-Portrait [64], and EmoTalk3D [19].

To quantitatively evaluate the talking head rendering quality, we use PSNR [22] and SSIM [56] to quantify overall image fidelity, LPIPS[67] to measure perceptual similarity to the ground-truth videos, Landmark Distance (LMD) [5] to evaluate the alignment between lip movements and speech, and the Cumulative Probability of Blur Detection (CPBD) [37] to assess image sharpness.

Besides,, we also conducted a user study to ask participant score the talking head rendering results across four dimensions: (1) S–V Sync. (Speech–Visual Synchronization), measuring the temporal consistency between audio and lip movements; (2) Video Fidelity, assessing the overall realism and temporal coherence of the video; (3) Image Qual-

Table 1. **Quantitative comparison** results evaluated on the Emotalk3D and RenderMe-360 datasets using different comparing approaches respectively.

	EmoTalk3D					RenderMe-360				
Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**LMD↓**	**CPBD↑**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**LMD↓**	**CPBD↑**
EAMM [24]	9.82	0.57	0.40	20.25	0.12	10.96	0.57	0.48	24.79	0.09
SadTalker [68]	9.90	0.64	0.47	43.49	0.11	12.20	0.65	0.41	26.80	0.19
Real3D-Portrait [64]	13.53	0.72	0.30	24.11	0.26	16.42	0.74	0.26	15.35	0.18
EmoTalk3D [19]	21.22	0.83	**0.12**	3.62	0.30	18.44	0.79	0.19	9.98	0.19
Hallo3 [10]	18.30	0.78	0.24	18.31	0.31	20.13	**0.83**	**0.14**	9.33	**0.30**
EchoMimic [6]	13.80	0.73	0.35	26.06	0.21	17.13	0.71	0.33	18.57	0.26
Ours	**25.78**	**0.86**	**0.12**	**3.56**	**0.36**	**21.41**	**0.83**	0.15	**6.59**	0.26

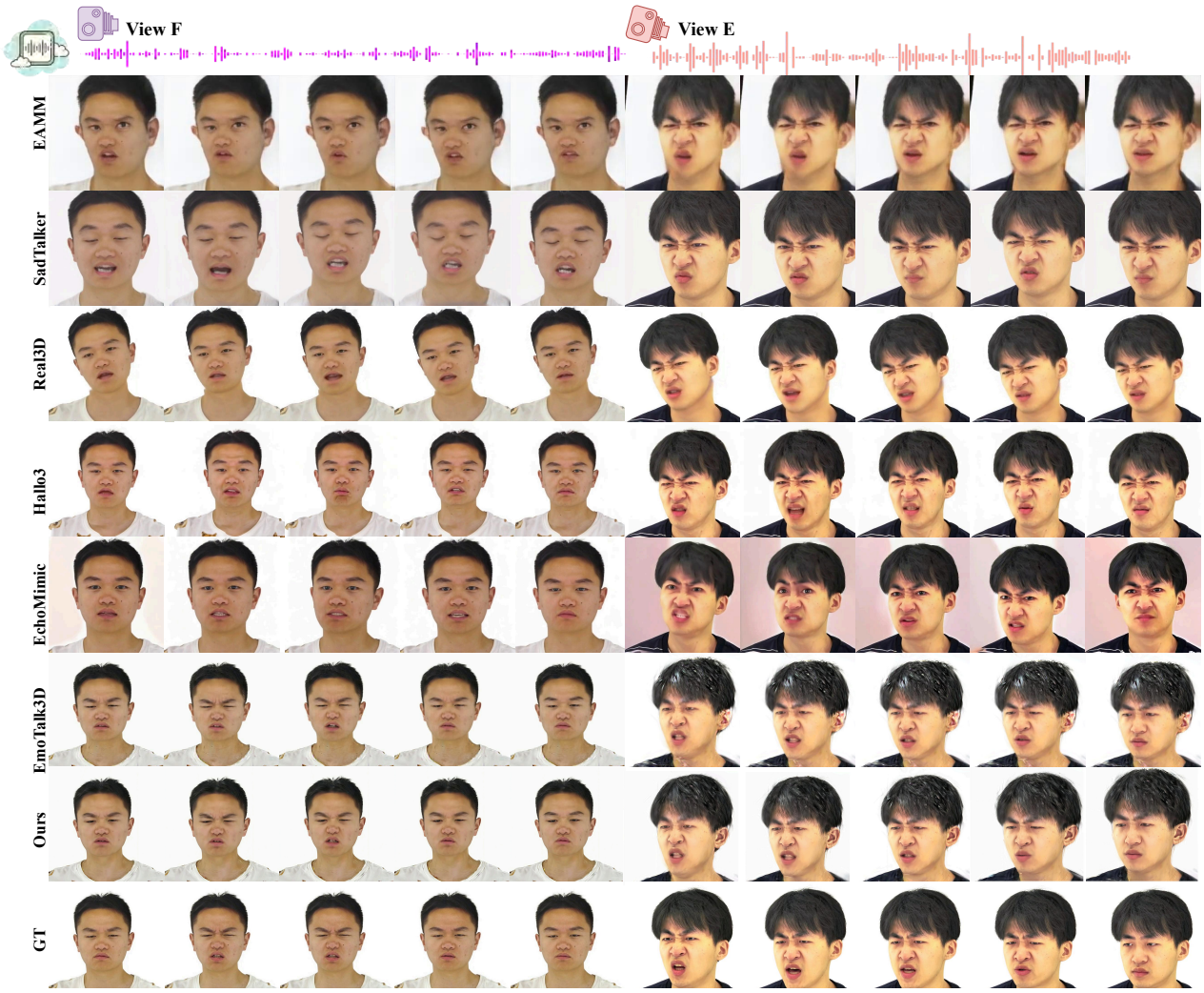


Figure 4. **Qualitative Comparison on the Emotalk3D Dataset.** Our approach outperforms existing approaches in both lip-sync accuracy and facial reconstruction detail than those previous SOTA talking head approaches.

ity, reflecting participants' ratings of the generated video's quality; (4) Emotion Control, assessing the accuracy and naturalness of facial expression regulation. For these user study metrics, we collected ratings using a 1–5 point scale (higher scores indicate better performance) and report the mean for each metric.

Comparison with existing methods. As shown in Fig. 4 and Fig. 5, our method achieves leading performance on both datasets for the previous 2D image-based and 3D-based talking heads, demonstrating superiority particularly in fa-

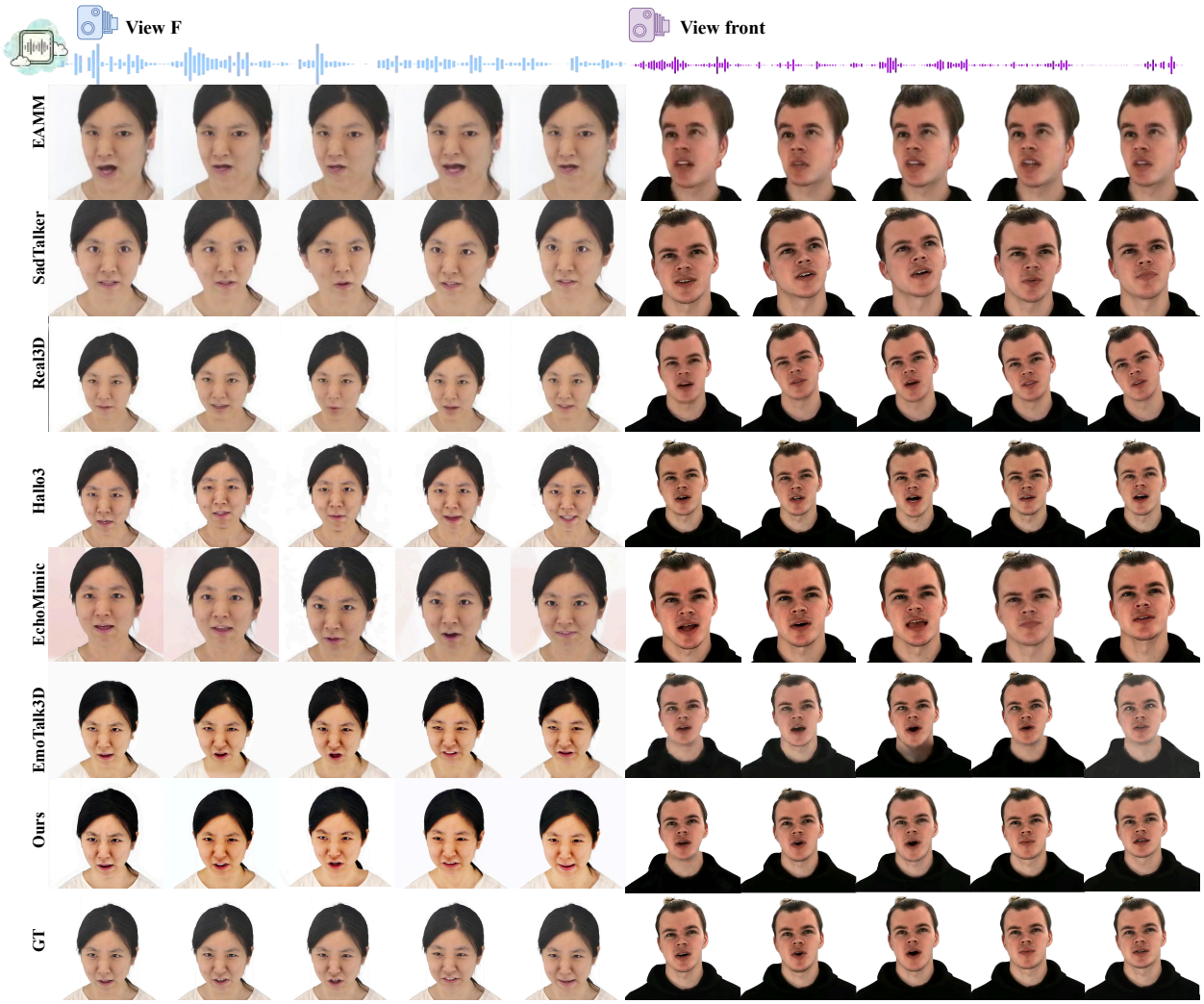


Figure 5. Additional experimental results on the Emotalk3D dataset and comparative evaluation results on the RenderMe-360 benchmark.

Table 2. The quantitative comparison of talking head generation quality user study using different comparing approaches respectively.

Method	EmoTalk3D				RenderMe-360			
	S-V Sync.	Video Fidelity	Image Quality	Emotion Control	S-V Sync.	Video Fidelity	Image Quality	Emotion Control
EAMM [24]	1.86	1.17	1.30	-	1.80	1.27	1.17	-
SadTalker [68]	3.15	2.20	3.80	-	3.10	2.57	3.92	-
Real3D-Portrait [64]	3.75	3.95	3.50	-	3.70	3.67	3.55	-
EmoTalk3D [19]	4.05	4.20	4.11	-	3.91	4.31	4.15	-
Hallo3 [10]	**4.70**	4.51	4.30	3.75	4.55	4.63	**4.55**	3.65
EchoMimic [6]	3.85	3.45	3.75	-	3.60	3.30	4.00	-
Ours	4.65	**4.75**	**4.50**	**3.77**	**4.62**	**4.71**	4.35	**3.73**

cial detail and emotional expression. As shown in Tab. 1, on the EmoTalk3D dataset, our method matches or surpasses existing approaches across all metrics. On the RenderMe-360 dataset, we achieve higher PSNR/SSIM and lower LMD. This demonstrates that compared to prior methods, our reconstructed speaker heads enhance rendering fidelity and

lip-sync accuracy while preserving fine-grained emotional cues.

User Study. We conducted a user study to compare real videos with results generated by various methods, evaluating the subjective quality of synthetic talking heads. Specifically, we selected five video sequences from EmoTalk3D and two from RenderMe-360 featuring the same speech content but different identities as evaluation materials. Participants score the videos across four dimensions: speech-visual synchronization, video fidelity, image quality, and emotion control. Since only Hallo3 and our model support text-based emotion control, other methods were excluded from this dimension's evaluation. Results are shown in Tab. 2: On EmoTalk3D, our method achieved the highest scores across all metrics; on RenderMe-360, we scored higher in speech-visual synchronization, video fidelity, and emotion control. These findings demonstrate our method's superior performance in subjective user experience.

4.4. Ablation Study

To comprehensively validate the effectiveness of each key module, we designed three sets of ablation experiments below. Note that Codes4P injects AU Codes into diffusion-based Gaussian offset predictors to conditionally modulate Gaussian point shifts based on affect perception; Codes4O injects AU Codes into OPCNet to endow its opacity predictions with affect-related modulation capabilities. Comparison results are shown in Fig. 6 and Tab. 3.

· **(A) w/o Codes4P.** By removing AU Codes from the diffusion Transformer, we ensure that facial Gaussian point offsets are restored solely by speech features, mesh priors, and diffusion time steps, while all other settings remain unchanged. Tab. 3 and Fig. 6 show a significant performance drop after removing Codes4P. While the model maintains basic speech-lip sync, facial structure and dynamic details are noticeably absent. Because AU Codes provides interpretable, region-specific constraints that align emotional semantics with local facial dynamics; without it, the diffusion struggles to reliably recover fine-grained expression changes based solely on audio.

· **(B) w/o Codes4O.** We retain the guiding role of AU Codes in Gaussian offset prediction, but remove AU Codes input in OPCNet, relying solely on predicted position offsets and triplane features to regress Gaussian point opacity. As shown in Fig. 6, this variant generates plausible facial geometry but exhibits significant deficiencies in appearance details: dynamic wrinkles (e.g., crow's feet and forehead lines) are markedly weakened, muscle tension variations are indistinct, and overall expressions tend toward neutrality and flatness. This indicates that without the emotional semantic constraints provided by AU Codes, OPCNet struggles to accurately model emotion-related local appearance features. This validates its effectiveness in modeling dynamic emotional appearances.

· **(C) w/o Diffusion** In Section 3.2, we proposed multi-step diffusion denoising via a Transformer-based temporal decoder to progressively restore the positional offsets of Gaussian points on the face. Here, we replace the Transformer-based diffusion denoising network with a GRU of comparable parameters. Consistent with Fig. 6, this variant exhibits amplitude shrinkage and conservatism at extreme expressions, resulting in an overall appearance that tends toward "mean-shifting." This demonstrates that diffusion is crucial for preserving detail intensity and diversity.



Figure 6. **Visual results of Ablation Study** for the key modules within different system variants including 'w/o Codes4P', 'w/o Codes4O', 'w/o Diffusion' and our 'FULL' respectively.

Table 3. **Quantitative Comparison of Ablation Study** for the key modules within different system variants including 'w/o Codes4P', 'w/o Codes4O', 'w/o Diffusion' and our 'FULL' respectively.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	LMD ↓	CPBD ↑
w/o Codes4P	20.12	0.72	0.21	6.25	0.22
w/o Codes4O	22.43	0.75	0.14	4.75	0.26
w/o Diffusion	24.96	0.82	0.21	4.51	**0.36**
FULL	**25.78**	**0.86**	**0.12**	**3.56**	**0.36**

5. Conclusion and Limitations

This paper proposes EmoDiffTalk, a novel emotion-aware diffusion framework for editable 3D talking head synthesis via 3D Gaussian Splatting. Our key innovations include an AU-prompted Gaussian diffusion process for fine-grained facial animation and a text-to-AU emotion controller enabling expansive multimodal emotional editing. Experimental results on EmoTalk3D and RenderMe-360 datasets validate superior performance in emotional subtlety, lip-sync accuracy, and controllability over state-of-the-art methods. However, one main limitation comes from the substantial computational overhead involving multiple pre-trained networks within our emotion-aware Gaussian diffusion. Besides, we also have limitation for very exaggerated expression, where our current activation-suppression-based text-to-AU emotion controller for intricate facial expressions would fail. But this remains challenging for all current emotion editing based 3D talking head approaches. We hope our work could inspire subsequent efforts for more advanced multimodal-based 3D talking head editing, with high-fidelity rendering quality and expansive manipulation.

Acknowledgements. We thank the anonymous reviewers for the constructive comments. This work is supported by the Natural Science Foundation of China (No. 62572060), Beijing Natural Science Foundation (25QY0304). We would also like to thank the AIM3 Lab (Renmin University of China) and the 3D Vision Lab (NJU-3DV) for their valuable academic support.

References

[1] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20, 2016. 2

[2] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21263–21273, 2024. 1

[3] Shivangi Aneja, Artem Sevastopolsky, Tobias Kirschstein, Justus Thies, Angela Dai, and Matthias Nießner. Gaussianspeech: Audio-driven personalized 3d gaussian avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13065–13075, 2025. 1

[4] Hejia Chen, Haoxian Zhang, Shoulong Zhang, Xiaoqiang Liu, Sisi Zhuang, Yuan Zhang, Pengfei Wan, Di Zhang, and Shuai Li. Cafe-talk: Generating 3d talking face animation with multimodal coarse- and fine-grained control, 2025. 4

[5] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. *CoRR*, abs/1803.10404, 2018. 5

[6] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions, 2024. 2, 5, 6, 7

[7] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2403–2410, 2025. 2

[8] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1, 2

[9] Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time talking head synthesis with 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10985–10994, 2024. 1, 2

[10] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer, 2024. 5, 6, 7

[11] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21086–21095, 2025. 1, 2

[12] Zijian Dong, Longteng Duan, Jie Song, Michael J Black, and Andreas Geiger. Moga: 3d generative avatar prior for monocular gaussian avatar reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13304–13314, 2025. 1

[13] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2

[14] Zhen Fan, Shiguang Huang, Yichi Zhang, Dachao Shang, Juyong Zhang, Yudong Guo, and Hua Huang. Rgavatar: Relightable 4d gaussian avatar from monocular videos. *IEEE Transactions on Visualization and Computer Graphics*, 31: 7421–7435, 2025. 1

[15] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22634–22645, 2023. 1, 2

[16] Xuan Gao, Haiyao Xiao, Chenglai Zhong, Shimin Hu, Yudong Guo, and Juyong Zhang. Portrait video editing empowered by multimodal generative priors. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 2

[17] Shengjie Gong, Haojie Li, Jiapeng Tang, Dongming Hu, Shuangping Huang, Hao Chen, Tianshui Chen, and Zhuoman Liu. Monocular and generalizable gaussian talking head animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5523–5534, 2025. 1

[18] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5784–5794, 2021. 1, 2

[19] Qianyun He, Xinya Ji, Yicheng Gong, Yuanxun Lu, Zhengyu Diao, Linjia Huang, Yao Yao, Siyu Zhu, Zhan Ma, Songchen Xu, Xiaofei Wu, Zixiao Zhang, Xun Cao, and Hao Zhu. Emotalk3d: High-fidelity free-view synthesis of emotional 3d talking head. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 5, 6, 7

[20] Yuxiao He, Yiyu Zhuang, Yanwen Wang, Yao Yao, Siyu Zhu, Xiaoyu Li, Qi Zhang, Xun Cao, and Hao Zhu. Head360: Learning a parametric 3d full-head for free-view synthesis in 360. In *European Conference on Computer Vision*, pages 254–272. Springer, 2024. 1, 2

[21] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2

[22] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010. 5

[23] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *CoRR*, abs/2106.07447, 2021. 3

[24] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2, 5, 6, 7

[25] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 1, 2

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[27] Jisoo Kim, Jungbin Cho, Joonho Park, Soonmin Hwang, Da Eun Kim, Geon Kim, and Youngjae Yu. Deeptalk: Dynamic emotion embedding for probabilistic speech-driven 3d face animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4275–4283, 2025. 2

[28] Deqi Li, Shiguang Huang, Zhiyuan Lu, Xinran Duan, and Hua Huang. St-4dgs: Spatial-temporally consistent 4d gaussian splatting for efficient dynamic scene rendering. *ACM SIGGRAPH 2024 Conference Papers*, 2024. 2

[29] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023. 1

[30] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In *European Conference on Computer Vision*, pages 127–145. Springer, 2024. 1, 2

[31] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Jun Zhou, and Lin Gu. Instag: Learning personalized 3d talking head from few-second video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10690–10700, 2025. 2

[32] Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10747–10757, 2025. 1, 2

[33] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023. 2

[34] Xiaoxi Liang, Yanbo Fan, Qiya Yang, Xuan Wang, Wei Gao, and Ge Li. Dgtalker: Disentangled generative latent space learning for audio-driven gaussian talking heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11079–11088, 2025. 1

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[36] Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 3d gaussian blendshapes for head avatar animation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers 24*, page 1–10. ACM, 2024. 1

[37] Niranjan D. Narvekar and Lina J. Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience*, pages 87–91, 2009. 5

[38] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: Large digital asset library and benchmark towards high-fidelity head avatars". In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 5

[39] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20687–20697, 2023. 1, 2

[40] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024. 1

[41] Zhexi Peng, Tianjia Shao, Yong Liu, Jingke Zhou, Yin Yang, Jingdong Wang, and Kun Zhou. Rtg-slam: Real-time 3d reconstruction at scale using gaussian splatting. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 2

[42] Ziqiao Peng, Wentao Hu, Junyuan Ma, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Hui Tian, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk++: High-fidelity and efficient synchronized talking heads synthesis using gaussian splatting. *arXiv preprint arXiv:2506.14742*, 2025. 1

[43] Zhexi Peng, Kun Zhou, and Tianjia Shao. Gaussian-plus-sdf slam: High-fidelity 3d reconstruction at 150+ fps. *Computational Visual Media*, 2025. 2

[44] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 1, 2

[45] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 128(3):698–713, 2020. 2

[46] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5020–5030, 2024. 1

[47] Hao-Pan Ren, Wei Duan, Wan-Yu Li, Yi Liu, Yu-Dong Guo, Shi-Sheng Huang, Ju-Yong Zhang, and Hua Huang. Egavatar: Efficient gan inversion for generalizable head avatar from few-shot images. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[48] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1173–1182, 2021. 1, 2

[49] Wenfeng Song, Xuan Wang, Shi Zheng, Shuai Li, Aimin Hao, and Xia Hou. Talkingstyle: personalized speech-driven 3d facial animation with style preservation. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[50] Xusen Sun, Longhao Zhang, Hao Zhu, Peng Zhang, Bang Zhang, Xinya Ji, Kangneng Zhou, Daiheng Gao, Liefeng Bo, and Xun Cao. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. In *2025 International Conference on 3D Vision (3DV)*, pages 713–722. IEEE, 2025. 2

[51] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–9, 2024. 1, 2, 4

[52] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Ziwei Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *International Journal of Computer Vision*, pages 1–12, 2025. 1, 2

[53] Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. Gaussianheads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations. *ACM Transactions on Graphics (TOG)*, 43(6):1–12, 2024. 1

[54] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024. 2

[55] Yating Wang, Xuan Wang, Ran Yi, Yanbo Fan, Jichen Hu, Jingcheng Zhu, and Lizhuang Ma. 3d gaussian head avatars with expressive dynamic appearances by compact tensorial representations, 2025. 4

[56] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 5

[57] Zhongjian Wang, Peng Zhang, Jinwei Qi, Guangyuan Wang, Chaonan Ji, Sheng Xu, Bang Zhang, and Liefeng Bo. Omnitalker: One-shot real-time text-driven talking audio-video generation with multimodal style mimicking. *arXiv preprint arXiv:2504.02433*, 2025. 2

[58] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4): 613–642, 2024. 1

[59] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 1, 2

[60] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin 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*, pages 1931–1941, 2024. 1

[61] Keyang Ye, Qiming Hou, and Kun Zhou. 3d gaussian splatting with deferred reflection. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers 24*, page 1–10. ACM, 2024. 2

[62] Keyang Ye, Tianjia Shao, and Kun Zhou. When gaussian meets surfel: Ultra-fast high-fidelity radiance field rendering. *ACM Transactions on Graphics (TOG)*, 44(4):1–15, 2025. 2

[63] Zhenhui Ye, Tianyun Zhong, Yi Ren, Ziyue Jiang, Jiawei Huang, Rongjie Huang, Jinglin Liu, Jinzheng He, Chen Zhang, Zehan Wang, et al. Mimictalk: Mimicking a personalized and expressive 3d talking face in minutes. *Advances in neural information processing systems*, 37:1829–1853, 2024. 1, 2

[64] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024. 2, 5, 6, 7

[65] Youyi Zhan, Tianjia Shao, Yin Yang, and Kun Zhou. Real-time high-fidelity gaussian human avatars with position-based interpolation of spatially distributed mlps. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26297–26307, 2025. 1

[66] Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5535–5545, 2025. 1, 2

[67] {Richard Yi} Zhang, Phillip Isola, {Alexei A.} Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings - 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 586–595. IEEE Computer Society, 2018. 5

[68] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. *arXiv preprint arXiv:2211.12194*, 2022. 5, 6, 7

[69] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023. 1, 2

[70] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 1, 2