RealDyadic: Synthesizing Realistic Dyadic 3D Dialogue with Neural Appearance Priors

Lei Zhu[1,2‡], Lijian Lin[2], Ye Zhu[2], Jiahao Wu[1,2], Xuehan Hou[1], Yu Li[2], Yunfei Liu[2†✉], and Jie Chen[5,3,1,4✉]

[1] School of Electronic and Computer Engineering, Peking University, Shenzhen,China
[2] International Digital Economy Academy (IDEA)
[3] Pengcheng Laboratory, Shenzhen, China
[4] AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China
[5] School of Intelligence Science and Engineering, Harbin Institute of Technology, Shenzhen, China.

Abstract. Current audio-driven 3D head generation methods primarily focus on single-speaker scenarios, struggling to model natural, bidirectional conversational dynamics. Furthermore, extending these models to dyadic interactions relies heavily on 3D pseudo-labels derived from tracking algorithms. These labels inherently contain severe noise and fail to capture fine-grained facial dynamics, particularly accurate lip closure. To address these limitations, we propose RealDyadic, a two-stage framework that leverages 2D photometric supervision to refine and correct 3D motion priors. In the first stage, a diffusion-based transformer equipped with a dual-audio interaction module generates synchronized 3D facial motions from multi-speaker audio streams. In the second stage, a 3D Gaussian Splatting renderer projects these motions into high-fidelity 2D frames, enabling image-level gradients to backpropagate and penalize the inaccuracies of the initial 3D pseudo-labels via an alternating training strategy. Additionally, we introduce Real-Dialog, a dataset comprising over 50 hours of aligned 2D-3D dyadic conversational data across 500+ identities. Extensive experiments demonstrate that integrating 2D supervision into 3D motion generation significantly outperforms existing baselines in both lip-sync accuracy and conversational realism, explicitly overcoming the performance ceiling dictated by pseudo-label errors. The code is available at `https://github.com/Pixel-Talk/RealDyadic`.

Keywords: Conversational Avatar · Neural Appearance Priors · 2D-to-3D Supervision

1 Introduction

In recent years, considerable research has been dedicated to 3D talking head generation, with a particular emphasis on audio-driven 3D motion synthesis for

✉Corresponding authors. † Project lead.
‡ This work was done during Lei Zhu's internship at IDEA.

applications in virtual reality, film production, education, *etc.* However, current talking head models are typically constrained to either a *speaking* [6,21,47,54] or *listening* [24,39] mode, lacking the ability to transition smoothly and realistically between these two states. Specifically, speaker-only models excel at generating lip movements that are highly synchronized with speech but fall short in producing listening feedback. Conversely, listener-only models can generate attentive responses but are incapable of speaking. Yet in real-world human-computer interaction, both speech generation and responsive listening are equally critical. This highlights the necessity of advancing conversational head generation towards a unified framework capable of bidirectional, dyadic interaction.

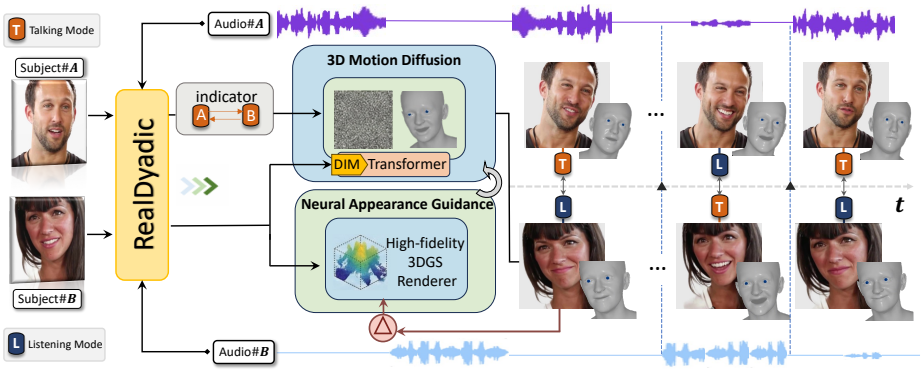

Fig. 1: Overview of RealDyadic. We first generate synchronized 3D motion from multi-speaker audio using a diffusion-based Transformer with a Dual-audio Interaction Module (DIM). Then our method renders these motions into 2D images via 3D Gaussian Splatting, providing photometric supervision to refine 3D facial dynamics through alternate training. The timeline (right) illustrates fluid transitions between Talking (T) and Listening (L) states.

Recently, multi-speaker generation has gained the attention of community. While INFP [57] proposed a 2D end-to-end framework, it lacks fine-grained 3D spatial control over identity and mouth motion. DualTalk [27] introduced a unified 3D dyadic talking head framework, utilizing estimated 3D meshes as ground-truth for supervision. However, in interactive conversational scenarios, lip movements are highly dynamic. We observe that models relying solely on 3D pseudo-labels generated by tracking methods (e.g., Spectre [8]) inherently overfit to tracking noise, establishing an artificial performance ceiling. As shown in Fig. 2, these 3D reconstructions exhibit significant lip misalignment and fail to provide adequate supervision for precise mouth closures. In contrast, original 2D video frames contain absolute, pixel-aligned interaction patterns. Leveraging these 2D signals as **neural appearance priors** can directly compensate for the geometric inaccuracies introduced by 3D motion tracking.

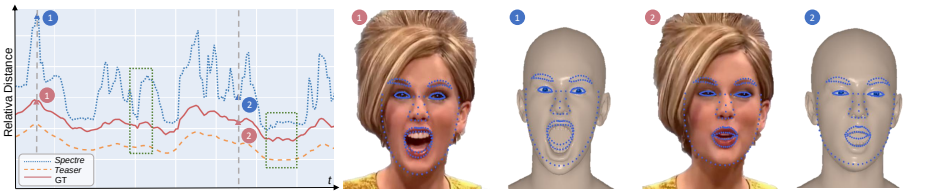

Fig. 2: Existing 3D face reconstruction methods produce either over-smoothed or noisy mouth movements when used for training conversational talking heads. As shown on the right, the actual lip motion (red curve, based on mouth keypoint distances) contrasts with the over-smoothed output from Teaser (orange curve) and the exaggerated, noisy results from Spectre (blue curve, based on projected 3D mesh).

Motivated by the above observations, we propose **RealDyadic**, a novel two-stage framework that leverages 2D photometric supervision to refine and correct 3D motion priors. As illustrated in Fig. 1, in the first stage, our diffusion-based motion generation network generates a sequence of 3D motion parameters from the input audio. To bypass the pseudo-label inaccuracies, the second stage employs a fast 3D Gaussian Splatting (3DGS) renderer. This renderer projects the predicted 3D meshes into high-fidelity 2D images, allowing image-level gradients to back-propagate and optimize the underlying 3D facial dynamics through an alternating joint training strategy. Furthermore, we introduce **Real-Dialog**, a high-quality, temporally-aligned multi-speaker dataset comprising over 50 hours of 2D-3D conversational data across 500+ identities.

In summary, our contributions are as follows:

(1) We propose RealDyadic, a unified dyadic 3D talking head framework. By decoupling a diffusion based 3D motion generator with 2D photometric supervision via a 3DGS renderer, we effectively mitigate the inherent noise of 3D tracking pseudo-labels.

(2) We design the Dual-audio Interaction Module (DIM) to model complex inter-speaker motion distributions from overlapping audio, enabling natural transitions between listening and speaking states.

(3) We construct Real-Dialog, a large-scale, photometrically-aligned conversational dataset. Extensive experiments demonstrate that our method outperforms existing approaches in multi-speaker 3D conversation tasks, particularly in terms of mouth movement accuracy and visual alignment with 2D target videos.

2 Related Work

2.1 3D Talking Head Generation

The field of speech-driven 3D talking head generation [10, 31, 32, 40, 46, 49, 50, 53] continues to garner significant research attention. Existing methodologies in this domain typically follow two distinct approaches. The first approach [6, 40, 47] utilizes acoustic features, such as MFCCs or representations derived from

pretrained speech models [1,11,15], mapping them to either 3D Morphable Model (3DMM) parameters [2, 19, 26] or a 3D mesh [5, 13, 37] to achieve decoupled expression and motion control. However, a major limitation lies in the lack of true 3D ground-truth labels, forcing methods to rely on 3D pseudo-labels generated by 3D reconstruction techniques [7, 22, 36, 44], whose accuracy is consequently constrained by the absence of genuine 3D supervision. Alternatively, the second category of methods [29, 34, 43, 53] employs a pure 3D Gaussian pipeline, where offsets of Gaussian attributes are predicted through spatial-audio interaction to render the desired deformation effects. Nevertheless, a critical drawback of these techniques is their current restriction to a single person per model, hindering their generalization ability to arbitrary identities. Additionally, due to the highly disentangled and controllable nature of 3D parameters, many 2D facial animation methods also utilize 3D motion parameters to model facial movements. [35,51,54] learn the mapping from audio to 3D parameters and use these parameters as control signals to generate talking heads, thereby achieving highly controllable results.

2.2 Multi-speaker Audio-driven Motion Generation

With the rapid development of audio-driven digital human generation [12, 30, 45, 48, 55, 56], audio-driven motion generation for multi-speaker interaction has emerged as an important research direction. Most existing methods support only a single function (speaking or listening), failing to achieve natural transitions between these two states. [57] pioneered a speech-driven 2D talking head generation method for multi-speaker interaction by mapping speech to visual latent codes containing conversational semantics to animate a static image. However, its latent space does not offer explicit control, and any fine-tuning requires re-inference from the audio input. [52] employed Stable Diffusion and Appearance Reference Net, using speech as condition along with LLM-generated labels to produce three states (listening, speaking, and communicating). However, this approach struggles to capture inter-speech relationships, resulting in insufficient representation of interactive semantics in generated videos. [17] implemented multi-stream audio injection through Label Rotary Position Embedding (L-RoPE), computing cross-attention between video latent space and speech embeddings using DIT architecture. While achieving realistic results, it demands excessive computational resources for training and inference.

In the field of 3D interactive generation, DualTalk [27] first realized speech-driven 3D digital human generation for multi-speaker interaction with natural state transitions, but its mouth movement modeling lacks precision due to reliance solely on 3D pseudo-label supervision. [33] achieved 3D human motion modeling via a temporal interactive module, [9] extracted motion token sequences from speech through Hierarchical Masked Modeling, and [23] performed multi-speaker interaction synthesis using diffusion models conditioned on multimodal inputs like text and speech. These methods [18, 28] demonstrate the importance of interaction-aware temporal modeling, yet they mainly focus on motion-level synthesis and do not explicitly exploit image-level cues to refine

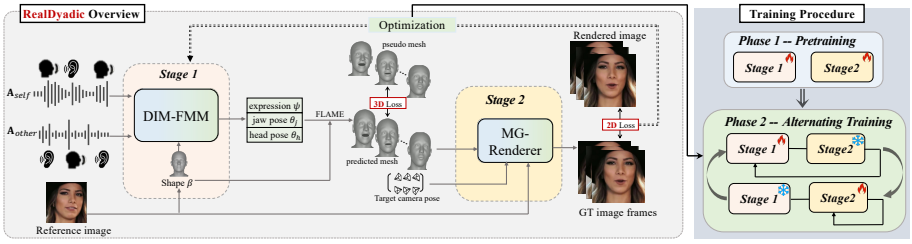

Fig. 3: The overall pipeline of RealDyadic. Our method first generates 3D facial motions from speech via **DIM-FMM** (stage1: audio-to-mesh), and then synthesizes images with **MG-Renderer** (stage2: mesh-to-image), where 2D supervision further refines the 3D motion.

facial details. In particular, fine-grained mouth shapes and subtle expression variations are difficult to supervise using noisy 3D pseudo-labels alone. None of these methods addressed the alignment between 3D and 2D generation. Our work is the first to incorporate 2D supervision in 3D conversational head generation, significantly improving the accuracy of 3D motion generation.

3 Method

3.1 Overview

As illustrated in Fig. 3, the proposed method consists of two stages.

Specifically, given an agent audio $\mathbf{A}_{self}$, an equal-length conversation partner audio $\mathbf{A}_{other}$, the shape parameter β of the agent speaker, and an agent speaker indicator $\mathbb{I}_{self}$ indicating whether the current speaker is speaking. The proposed first-stage diffusion model (**DIM-FMM**) generates the corresponding motion parameter sequences $\mathbf{X} = \{\psi, \theta_j, \theta_h\}$ under the above conditions. Here $\{\psi, \theta_j, \theta_h\}$ indicate the expression, jaw pose, and head rotation parameters, respectively. Then we can generate the facial 3D geometry through a 3D morphable model [19], *i.e.*,

$$\mathbf{V} = \text{FLAME}(\beta, \psi, \theta_j, \theta_h), \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{w \times 5023 \times 3}$, w is the length of generated frames. Subsequently, based on a reference image I, 3D facial geometries $\mathbf{V}$, and a pre-defined camera pose $\mathbf{R}$, we adopt a 3DGS renderer (**MG-Renderer**) to generate the final 2D video.

3.2 Motion Generation with Conversation Audio

Dual-audio Interaction module (DIM). To learn the joint representation of the agent ($\mathbf{A}_{self}$) and partner ($\mathbf{A}_{other}$), our **DIM** (Fig. 4, left) first extracts their features using a pre-trained encoder [15] $\mathcal{E}$:

$$\mathbf{H}_{self} = \mathcal{E}(\mathbf{A}_{self}), \quad \mathbf{H}_{other} = \mathcal{E}(\mathbf{A}_{other}), \tag{2}$$

where $\mathbf{H}_{self}$ and $\mathbf{H}_{other}$ are the audio features of the agent speaker and the other speaker, respectively. The audio features $\mathbf{H}_{self}$ and $\mathbf{H}_{other}$ are then fed to a Transformer encoder to capture long-range dependencies and intricate interaction patterns between those two speakers. However, incorporating the other speaker's audio may weaken the contribution of the speaking agent's audio cues. To preserve motion-speech synchronization for the agent speaker, we adopt a residual connection by adding $\mathbf{H}_{self}$ back to the fused representation. The final dual audio feature $\mathbf{H}_{dual}$ is generated by:

$$\mathbf{H}_{dual} = \text{Transformer}(\mathbf{H}_{self}, \mathbf{H}_{other}) \oplus \mathbf{H}_{self}. \tag{3}$$

The fused audio features are then concatenated with an agent speaker indicator $\mathbb{I}_{self}$, which is a binary vector indicating whether the agent speaker is speaking. This indicator is processed offline using TalkNet [41]. As detailed in the supplementary material, our approach remains robust to potential noise within these labels. The concatenated features are then fed into a linear layer to obtain the fused audio feature representation $\mathbf{H}_{fuse}$.

Fused-audio Motion Generation Model (FMM). Based on $\mathbf{H}_{fuse}$, we design a diffusion-based model (FMM) to predict the synchronized motion sequences. The fused-audio feature $\mathbf{H}_{fuse}$ is fed into FMM to generate the corresponding motion parameter sequences $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, where $\mathbf{x}_i \in \mathbb{R}^{56}$. In the forward diffusion process, Gaussian noise $\mathbf{z}$ is added to the initial data sample $\mathbf{X}^0 = \{\mathbf{x}_0^0, \mathbf{x}_1^0, ..., \mathbf{x}_T^0\}$ according to a variance schedule.

Eventually, the data distribution converges to a standard normal distribution, which can be represented as $q(\mathbf{X}^N|\mathbf{X}^0)$, where N indicates the total number of timesteps. In the reverse process, we use the distribution $q(\mathbf{X}^{n-1}|\mathbf{X}^n)$ to gradually recover the original sample from noise. To predict this distribution $q(\mathbf{X}^{n-1}|\mathbf{X}^n)$, we employ a denoising network to directly predict the clean sample from the noisy sample, rather than stepwise predicting the noise at each step. This approach enables the introduction of geometric loss, providing more precise constraints for facial motion and significantly benefiting motion generation [40, 42]. Specifically, our **FMM**, as shown in the right part Fig. 4, performs the denoising process as follows: for the predicted fused-audio feature $\mathbf{H}_{fuse}$, we divide it into frame blocks of w frames, and use $\mathbf{H}_{0:w}$ as the input for each block. For each window, at diffusion step n, the denoising network takes as input the previous and current fused audio features $\mathbf{H}_{-w_p:w}$, the real previous motion sequence $\mathbf{X}^0_{-w_p:0}$, and the current noisy sequence $\mathbf{X}^n_{0:w}$ sampled from $q(\mathbf{X}^n_{0:w}|\mathbf{X}^0_{0:w})$. In addition to the fused-audio feature, we incorporate the FLAME shape parameter β shared across all windows. Therefore, the final input and output format of the denoising network can be represented as:

$$\hat{\mathbf{X}}^0_{-w_p:w} = \text{DiT}(\mathbf{H}_{-w_p:w}, \mathbf{X}^0_{-w_p:0}, \mathbf{X}^n_{0:w}, n, \beta). \tag{4}$$

The loss for the first stage consists of three parts:

1) **Parameter Loss.** We calculate the L2 loss between the denoised motion and the ground-truth motion at each step, directly constraining the the predicted

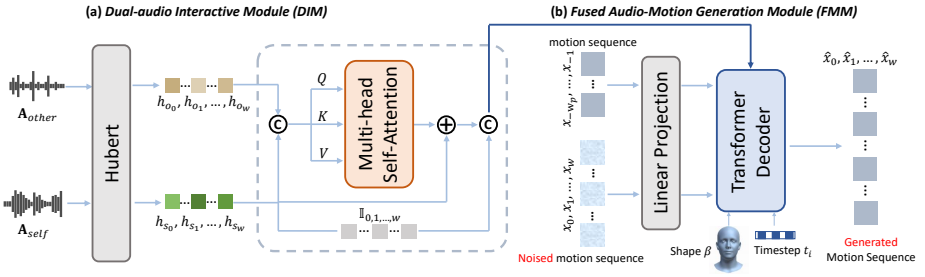

Fig. 4: Dual-audio interactive module (**DIM**) and fused audio motion generation module (**FMM**). The speech signals $\mathbf{A}_{self}$ and $\mathbf{A}_{other}$ are fed into Hubert for semantic features $\mathbf{h}^s_{0:w}$ and $\mathbf{h}^o_{0:w}$. These features are concatenated and fed into a multi-head self-attention module, which is then combined with $\mathbf{h}^s_{0:w}$ via residual connection, followed by concatenation with the agent speaker indicator $\mathbb{I}$. These features are then fed into **FMM** for motion $\hat{\mathbf{X}}_{0:w}$ generation.

parameter distribution to be close to the real distribution:

$$\mathcal{L}_{param} = ||\hat{\mathbf{X}}^0_{-w_p:w} - \mathbf{X}^0_{-w_p:w}||^2_2, \tag{5}$$

since jaw pose parameters θ_j are especially important for lip movements, we also additionally add a loss on jaw pose parameters:

$$\mathcal{L}_{jaw} = ||\hat{\theta}_{j,-w_p:w} - \theta_{j,-w_p:w}||^2_2, \tag{6}$$

2) **3D Loss.** We convert the parameters into zero-head-posed 3D mesh vertices, then we get $\mathbf{V}_{-w_p:w_p} = \text{FLAME}(\beta, \mathbf{X}^0_{-w_p:w_p})$ and $\hat{\mathbf{V}}_{-w_p:w_p} = \text{FLAME}(\beta, \hat{\mathbf{X}}^0_{-w_p:w_p})$ and compute geometric loss directly in the 3D space.
3) **Stability Loss.** We further use velocity loss to improve temporal consistency, and smooth loss to regularize excessive motion amplitudes and prevent abrupt changes. More details will be presented in the supplementary material. The overall loss for the first stage is:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{param}} + \lambda_{\text{jaw}}\mathcal{L}_{\text{jaw}} + \lambda_{\text{vert}}\mathcal{L}_{\text{vert}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}. \tag{7}$$

3.3 Image Synthesis with Meta Gaussian Renderer

To enhance the synchronization and fidelity of the generated results, we propose a 3D meta Gaussian renderer (**MG-Renderer**) to render the above generated motion sequences into high-fidelity 2D images, which was then supervised using ground-truth images. Specifically, given $\hat{\mathbf{X}}^0_{0:w}$, we randomly sample n frames from this window sequence to get $S = \{\hat{\mathbf{x}}_0, \hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_n\}$. Through FLAME, we obtain the vertex sequence $V = \{\hat{\mathbf{v}}_0, \hat{\mathbf{v}}_1, ...\hat{\mathbf{v}}_n\}$. We randomly sample an image of the agent speaker as a reference image I_r, which is fed into Dinov2 [25] for the appearance feature. Besides, we use target camera parameters R which is a

4×4 matrix including translation and rotation scale, to project the vertices into 2D image plane. For each $\hat{\mathbf{x}}_i$, the process through our 3D Gaussian Renderer $\mathcal{R}$ can be described as:

$$\hat{\mathbf{I}}_i = \mathcal{R}(\hat{\mathbf{v}}_i, I_r, R). \tag{8}$$

Meta Gaussian Renderer. In our Gaussian Renderer $\mathcal{R}$, we model faces in a canonical space using 3D Gaussians [16]. Each Gaussian is defined by its position, rotation, scale, opacity, and a latent appearance vector: $G = \{\mu, r, s, \alpha, c\}$. The meta Gaussians are constructed from two components: **(1)** Template Gaussians derived from the FLAME model G_T which are relatively sparse and responsible for representing the overall texture and geometry; and **(2)** UV Gaussians attached to the triangulated mesh G_{UV}, which are used to encode fine-grained details. The detailed construction of each Gaussian and Renderer $\mathcal{R}$ are provided in the supplementary material.

Image-level Supervision. After the above processing, we obtain all the Gaussians of the reference image I_r. Then we replace the positions of the Gaussians corresponding to I_r with those of $\hat{\mathbf{v}}_i$ to generate the animated Gaussians, while keeping all other attributes unchanged. During rendering, we splat the animated Gaussians to produce a coarse feature map F_{raw}, where the first three channels correspond to a coarse RGB image $\hat{I}_{raw}$. To enhance texture details, we feed this feature map F_{raw} into a StyleUNet-based refiner, ultimately producing the final image $\hat{\mathbf{I}}_i$ with enhanced features such as detailed teeth textures.

Loss Function. For the sequence $\hat{\mathbf{X}}^0_{0:w}$ generated in the first stage, we randomly sample n frames and obtain $\hat{\mathbf{I}}_{0:n} = \{\hat{\mathbf{I}}_0, \hat{\mathbf{I}}_1, ..., \hat{\mathbf{I}}_n\}$ through the rendering process described above. Similar to previous methods, we compute image-level losses for each frame in $\mathcal{R}$, mainly including photometric loss $\mathcal{L}_{pho}$ and VGG perceptual loss $\mathcal{L}_{per}$, to ensure consistency between the rendered images and the ground truth, which is:

$$\mathcal{L}_{stage2} = \lambda_{pho}\mathcal{L}_{pho} + \lambda_{per}\mathcal{L}_{per}. \tag{9}$$

Two-phase Training Strategy. Our total training process contains two phases:
1) Phase 1. To ensure that we can learn fully decoupled FLAME parameters from the audio, we first train the first stage with $\mathcal{L}_{stage1}$. Additionally, we also train the second stage separately with $\mathcal{L}_{stage2}$ to prevent the potential catastrophic impact that from-scratch training have on the first stage.
2) Phase 2. In order to introduce the image-level loss and reduce the effect of the gaussian renderer compensating between these two stages [22], we alternate the training between the first and second stages. At this point, the loss function for the first stage is:

$$\mathcal{L}_J = \mathcal{L}_{stage1} + \mathcal{L}_{stage2}, \tag{10}$$

where $\mathcal{L}_{stage2}$ provides a more accurate optimization path for the first stage. After one iteration, we still apply $\mathcal{L}_{stage2}$ to optimize the second stage for better converging.

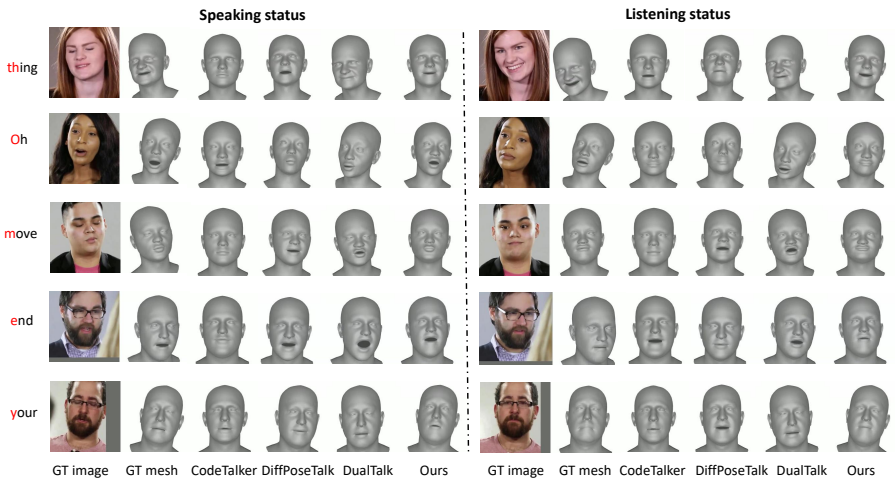

Fig. 5: Visual comparison of the 3D conversational talking head generation results with SOTA methods on our RealDyadic-Dialog testset.

3.4 Real-Dialog Dataset

To support the training of our multi-speaker framework, we constructed a dataset of dual-speaker dialogues (Real-Dialog). For diversity and authenticity, we collected dialogue videos over diffrent daily conversation scenarios, such as emotional communication, casual dialogue, interview connections, etc. Real-Dialog contains more than 5,000 dialogue clips ranging from 30 seconds to 2 minutes, with a total duration of 50 hours and covering 500 speakers. The clip length setting of 30 seconds to 2 minutes ensures the inclusion of listening-speaking state transitions. Each dialogue clip ensures that both people appear on screen simultaneously, allowing us to obtain any facial changes of both participants. Our conversational video processing pipeline includes three steps: (1) We using TackNet [41] to separate speech segments and assign them to the corresponding speaker; (2) applying Spectre [8] for 3D FLAME motion parameter tracking; and (3) refining camera parameters via keypoint-based alignment optimization.

Finally, all clips are divided into three parts: training, validation, and test sets, containing 4,300 clips, 400 clips, and 200 clips, respectively. To evaluate the model's generalization ability, the identities in the testset were unseen during training. We used the testset for subsequent metric evaluation.

4 Experiment

4.1 Experimental Setup

Evaluation Metrics. In our experiments, we mainly evaluate our method from two aspects: 3D mesh modeling and 2D image quality, to comprehensively as-

sess the performance of our approach in dual-speaker dialogue scenarios. For 3D mesh modeling, we use key metrics such as lip vertex error (**LVE**) [38] and mean vertex error (**MVE**) to measure the frame-wise differences between mouth vertices and overall facial vertices compared to tracking results, use upper face dynamics deviation (**FDD**) [47] to evaluate the accuracy of inter-frame vertex motion and use **MOD** to represent the speech-synchronized fidelity of lip opening/closing movements. Notably, we also employ the recently proposed metrics from [3]: Mean Temporal Misalignment (**MTM**) and Speech and Lip Intensity Correlation Coefficient (**SLCC**), which are used to evaluate the temporal consistency of 3D meshes and the correlation between lip and speech intensity, respectively.

For 2D image generation, we evaluate visual quality and facial fidelity using complementary metrics. Fréchet Inception Distance (**FID**) [14] is adopted to measure distribution-level realism, identity cosine similarity (**CSIM**) is used to assess identity preservation, while Average Expression Distance (**AED**) and Average Pose Distance (**APD**) [35] are used to evaluate expression accuracy and head-pose consistency, respectively. For image-level evaluation of lip synchronization and mouth shape, we adopt perceptual metrics from Wav2Lip [20], including the distance score (**LSE-D**) and confidence score (**LSE-C**). To comprehensively evaluate the interaction capability, following DualTalk, we utilize **FD** to measure the realism of the listener and speaker states separately, and use **SID** to evaluate the diversity of facial expressions and head movements. The details of all metrics and our implementation details are provided in the supplementary material.

Table 1: Quantitative comparisons of the SOTA methods on 2D image generation in our Real-Dialog testsets.

Methods	Visual Quality				Lip Sync	
	FID↓	CSIM↑	AED↓	APD↓	LSE-C↑	LSE-D↓
ARTalk	26.13	0.641	0.341	0.123	4.923	7.632
SadTalker	24.35	0.723	0.408	0.105	5.136	7.427
AniTalker	28.67	0.594	0.642	0.156	3.279	9.362
Ours	**21.31**	**0.832**	**0.257**	**0.064**	**5.382**	**7.296**

Compared SOTA Methods. We compare our method with several SOTA methods: single-speaker 3D facial animation methods FaceFormer [6], CodeTalker [47], DiffPoseTalk [40], ARTalk [4] , and the recent dual-speaker 3D facial animation method DualTalk [27]. We also compare with single-speaker 2D talking head generation methods SadTalker [54] and AniTalker [21], to evaluate our 2D generation capability. For the single-speaker methods, we mutes the speech of the other conversational partner.

4.2 Quantitative Results

We evaluated our 3D performance on our Real-Dialog testset and the only available 3D conversion dataset: DualTalk. As shown in Tab. 2, our method outperforms all baselines across all metrics on these two datasets, especially in LVE and MVE, indicating more accurate modeling of frame-wise facial vertices, particularly mouth vertices. This advantage mainly stems from our dual-audio fusion module, which effectively captures the mapping between speech semantics and

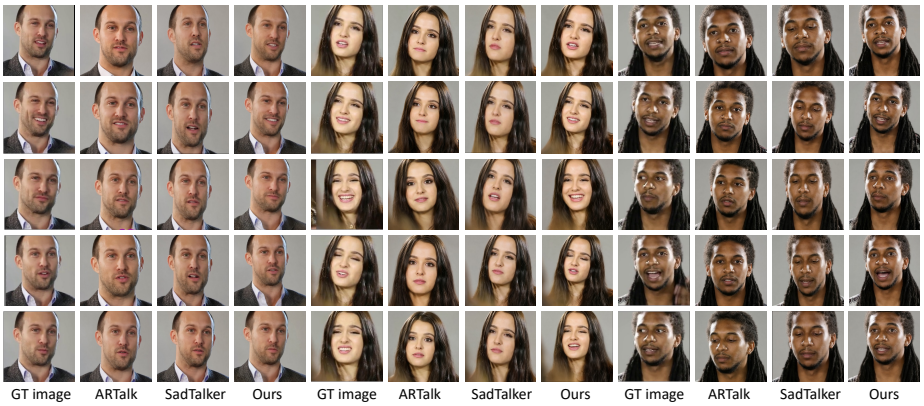

Fig. 6: Visual comparison of the 2D conversational talking head generation results with SOTA methods on the Real-Dialog testset.

Table 2: Quantitative comparisons of the comparative methods on mesh accuracy in our Real-Dialog testset and Dualtalk testset.

Version	Real-Dialog Testset					DualTalk Testset				
	LVE↓	MVE↓	MOD↓	MTM↓	SLCC↑	LVE↓	MVE↓	MOD↓	MTM↓	SLCC↑
FaceFormer	3.276	1.754	1.463	4.812	0.623	3.186	1.685	1.442	4.769	0.632
CodeTalker	3.445	1.638	1.477	4.793	0.638	3.258	1.612	1.456	4.703	0.641
DiffPoseTalk	2.694	1.542	1.391	4.781	0.645	2.574	1.503	1.347	4.692	0.649
ARTalk	2.452	1.521	1.304	4.532	0.662	2.368	1.425	1.289	4.487	0.671
DualTalk	2.083	1.382	1.228	4.321	0.707	2.021	1.226	**1.145**	4.297	0.721
Ours	**1.741**	**1.225**	**1.096**	**4.015**	**0.791**	**1.894**	**1.182**	1.162	**4.024**	**0.764**

mouth movements, as well as the joint training of the two stages, which provides stronger supervision for motion generation. The lower MOD and MTM further confirms the improved lip-sync accuracy of our approach, while a higher SLCC indicates stronger correlation between the mesh generated by our method and vocal intensity.

Due to the lack of publicly available 2D conversation datasets, we conducted the 2D evaluation solely on the testset of Real-Dialog, and the results are shown in Tab. 1. The experimental results demonstrate that our method outperforms other approaches in terms of both visual quality and lip-sync accuracy of the generated videos. This finding further corroborates that our 3D mesh achieves the highest alignment precision with the ground truth data.

Furthermore, as shown in Tab. 5, the optimal FD scores for both the speaker and listener segments indicate that our method demonstrates superior realism in expression, jaw, and pose compared to the baseline methods. Simultaneously, these optimal SID metrics also show that our method achieves richer and more

Table 3: Ablation study on our Real-Dialog, including 3D-mesh and 2D-image metrics.

Version	3D-mesh Metrics						2D-image Metrics			
	MVE↓	LVE↓	FDD↓	MOD↓	MTM↓	SLCC↑	PSNR↑	SSIM↑	LPIPS↓	MAE↓
baseline	1.542	0.269	1.607	1.391	4.781	0.645	18.97	0.745	0.352	0.076
+audio fusion	1.401	0.193	1.603	1.281	4.242	0.661	20.17	0.783	0.278	0.067
+audio res	1.454	0.233	1.601	1.230	4.371	0.683	20.53	0.789	0.257	0.063
+indicator	1.513	0.235	1.601	1.247	4.526	0.698	20.42	0.782	0.238	0.064
+jaw pose	1.507	0.235	**1.579**	1.221	4.147	**0.802**	21.20	0.794	0.246	0.061
Ours(+two stage)	**1.225**	**0.174**	1.593	**1.096**	**4.015**	0.791	**23.25**	**0.821**	**0.213**	**0.054**

Table 4: User study results evaluating animation realism, expressiveness, and interaction correlation. Rating is on a scale of 1-5; higher is better.

Methods	Realism and Expressiveness							Interaction Correlation		
	L-Visual Quality↑	L-Exp Richness↑	L-Pose Naturalness↑	S-Lip Sync Accuracy↑	S-Visual Quality↑	S-Exp Richness↑	S-Pose Naturalness↑	Temporal Coherence↑	Contextual Appropriateness↑	Pose Naturalness↑
CodeTalker	2.6	2.6	2.8	2.1	1.8	2.2	1.3	1.2	1.5	2.8
DiffPoseTalk	2.3	3.4	3.5	**4.4**	3.8	3.8	3.9	1.7	2.1	3.5
DualTalk	3.5	3.8	3.2	4.0	3.5	3.6	3.7	2.9	3.0	3.2
Ours	**3.9**	**3.9**	**4.0**	4.3	**4.1**	**3.9**	**4.0**	**3.7**	**3.3**	**3.7**

diverse motion patterns. These experiments collectively validate the effectiveness of our diffusion-based 3D motion generation and 2D-level supervision.

4.3 Qualitative Results

To intuitively demonstrate the accuracy and naturalness of our method's results, we conduct qualitative comparisons with SOTA methods, as shown in Fig. 5. We present results for both "listening" and "speaking" states. It can be observed that CodeTalker [47] captures some mouth movements, but the amplitude of mouth changes is small; in speaking scenarios where the mouth should be open, it sometimes remains closed (e.g., column 3, row 1 and 4), and in listening states, the mouth sometimes fails to close (e.g., column 9, row 4). DiffPoseTalk [40] exhibits larger motion amplitudes, but still shows inconsistencies between mouth opening/closing and the ground truth (e.g., column 4, row 2 and 3). While DualTalk [27] yields rich dynamics, it suffers from exaggerated mouth openings and fails to achieve proper lip closure in silent states (e.g., col. 5, row 4; col. 10, rows 2-4). Crucially, our generated meshes often exhibit better visual alignment with the ground-truth images than the tracking pseudo-labels themselves, which inherently contain excessive opening or incomplete closure errors (e.g., col. 2, row 2). This confirms that our two-stage joint training effectively corrects 3D tracking noise via 2D-lifted alignment.

Meanwhile, we present a comparison of the 2D effects of our method with existing SOTA methods, as illustrated in Fig. 6. It is clearly evident that our 2D results outperform the comparative methods in terms of motion accuracy, synchronization, and identity preservation when compared to the ground truth.

Table 5: Quantitative Evaluation on Realism and Diversity. 'L' and 'S' denote Listener and Speaker segments, respectively.

Method	S-FD (exp)↓	S-FD (jaw)↓	S-FD (pose)↓	L-FD (exp)↓	L-FD (jaw)↓	L-FD (pose)↓	SID (pose)↑	SID (exp)↑	SID (jaw)↑
DiffPoseTalk	23.86	2.89	3.61	18.39	3.56	4.79	1.47	1.96	1.73
DualTalk	**21.91**	3.06	3.83	12.94	2.12	3.23	1.48	2.17	1.98
Ours	22.37	**2.75**	**3.54**	**11.93**	**1.99**	**2.78**	**1.53**	**2.23**	**2.26**

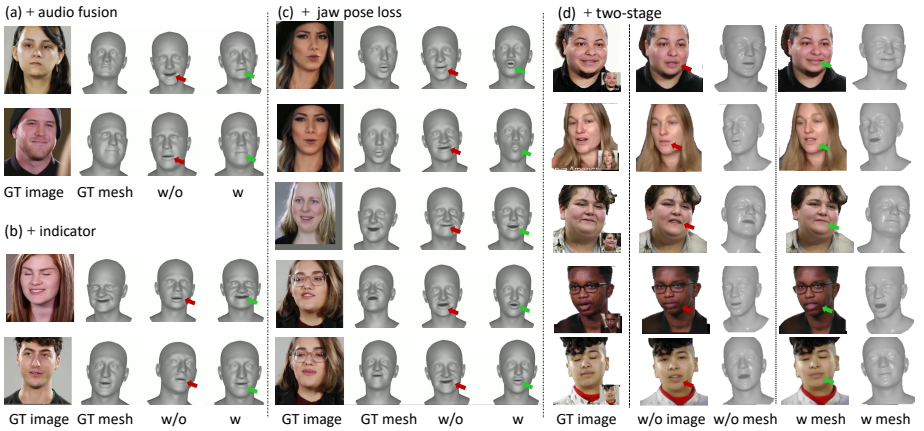

Fig. 7: Visual results of ablation study in the 3D-level strategy in stage1 (**a,b,c**) and the two-stage joint training strategy (**d**).

More notably, our approach is capable of capturing subtle dynamics in conversations, such as producing a natural smile in conversation (e.g. column 4, raw 4 and column 8, raw 5), without appearing rigid or unnatural.

To further validate these observations, we conducted a user study in which we invited 15 participants to rate the realism and expressiveness of the generated animations. Utilizing the Mean Opinion Score (MOS) protocol, participants rated the realism of the test set's listening and speaking segments across four dimensions: pose naturalness, expression richness, visual quality, and audio-lip synchronization [27]. Additionally, we asked users to evaluate the correlation between the speaker's speech and the listening behavior across three dimensions: temporal synchronization, action appropriateness, and head pose naturalness. The results, presented in Tab. 4, indicate that our method achieved higher average scores across virtually all dimensions compared to the baseline methods.

4.4 Ablation Study

Effectiveness of the DIM. To validate the effectiveness of the audio fusion module, we conducted an ablation study using DiffPoseTalk [40] as the baseline.

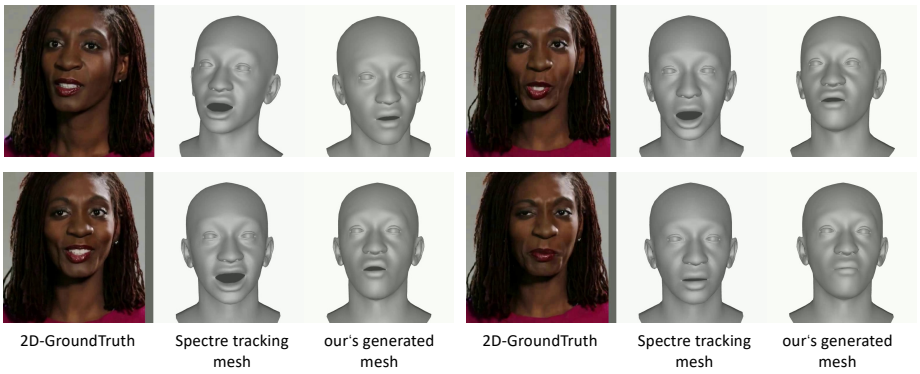

Fig. 8: Comparison of Spectre tracked mesh and our generated mesh. The Spectre mesh exhibits an overly exaggerated mouth opening, whereas our mesh is visually better aligned with the 2D ground truth image.

As shown in Fig. 7 **(a)**, the audio fusion module offers two main advantages: (1) it effectively distinguishes between the speaker and the listener, preventing mouth movements associated with speaking from appearing during listening states (see the upper row of Fig. 7 **(a)**); (2) it establishes a correlation between the speaker and listener's audio, such that when the other person is speaking with a smile, the model also exhibits a corresponding smile (see the lower row of Fig. 7 **(a)**). As shown in Tab. 3, incorporating the audio fusion module leads to significant improvements in both 3D and 2D metrics.

Effectiveness of Speaking Indicator. Experimental results show that incorporating the speaking indicator leads to richer facial expressions and a significant improvement in overall perceptual quality (as shown in Fig. 7 (**b**)). We attribute this to the indicator's ability to help the model accurately distinguish between the speaker and the listener, enabling more targeted interaction modeling.

Effect of the Second Stage. Introducing the second stage, as in the last row of Tab. 3, all metrics improve significantly, which fully demonstrates the positive impact of the second stage on the first stage. The notable improvement in 2D metrics indicates that the motion after passing through the 3D GS Renderer becomes more accurate, validating our hypothesis that 2D loss can influence 3D performance, thereby enhancing the accuracy of 3D motion and making the rendered 2D images closer to the ground truth. As shown in Fig. 7 **(d)**, before introducing the second stage, many motion sequences and their rendered images cannot be perfectly aligned with the target, especially for mouth dynamics; after introducing the second stage, both the motion and rendered images are well aligned with the target image. This demonstrates that relying solely on 3D pseudo-loss constraints is far from sufficient, and 2D supervision provides a more accurate optimization for motion learning. More ablation study results are shown in the supplementary. To further demonstrate the effectiveness of our proposed

two-stage framework and the benefits of joint training, we provide additional case studies and quantitative evaluations. As shown in Fig. 8, we compare results from different stages: ground-truth image, mesh without stage-2, and mesh after stage-1 and stage-2 joint training. It can be observed that incorporating the second stage significantly improves mesh alignment with the ground-truth image, particularly in mouth opening and closing. Without stage-2, many cases that should depict closed mouths fail to do so.

To quantitatively validate this observation, we compare our method with Spectre with manually annotated ground-truth keypoints ($\mathbf{X}$gt). We obtain 2D keypoints from Spectre's reconstructed meshes ($\mathbf{X}s$) and our meshes ($\mathbf{X}m$), and compute the Lip Distance Metric (D_{Lip}) and Mean Absolute Error (MAE_{Lip}). Spectre-Reconstruction yields an MAE_{Lip} of $\mathbf{9.68}$, while our method achieves a significantly lower error of $\mathbf{6.23}$. These results confirm our hypothesis: RealDyadic achieves superior accuracy over Spectre in specific cases, particularly when Spectre produces exaggerated or non-physical mouth shapes. Visual comparisons are provided in the supplementary material for further illustration.

5 Conclusion

In this paper, we present RealDyadic, a method for multi-speaker 3D talking head generation. RealDyadic fully explores the mapping relationship between speech semantics and lip movements by introducing a diffusion-based dual-audio fusion module. In addition, we incorporate a 3D Gaussian renderer to synthesize images under the supervision of real 2D images. Through a two-stage joint training strategy, our method achieves 2D-aligned 3D mesh generation. We conducted extensive experiments on our self-built dual-speaker conversation dataset, and the results show that our method outperforms existing approaches in both 3D mesh accuracy and 2D image quality. Despite these advances, our method still shows minor artifacts in some non-facial areas. We leave these improvements to future work.

Acknowledgements

Jie Chen and Yunfei Liu are the corresponding authors. This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122702), the Shenzhen Medical Research Funds in China (No. B2302037), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465), and AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China, and the Shenzhen Loop Area Institute under grant FPF10120250014.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

2. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2013)
3. Chae-Yeon, L., Hyun-Bin, O., EunGi, H., Sung-Bin, K., Nam, S., Oh, T.H.: Perceptually accurate 3d talking head generation: New definitions, speech-mesh representation, and evaluation metrics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21065–21074 (2025)
4. Chu, X., Goswami, N., Cui, Z., Wang, H., Harada, T.: Artalk: Speech-driven 3d head animation via autoregressive model. arXiv preprint arXiv:2502.20323 (2025)
5. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10101–10111 (2019)
6. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18770–18780 (2022)
7. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021)
8. Filntisis, P.P., Retsinas, G., Paraperas-Papantoniou, F., Katsamanis, A., Roussos, A., Maragos, P.: Visual speech-aware perceptual 3d facial expression reconstruction from videos. arXiv preprint arXiv:2207.11094 (2022)
9. Ghosh, A., Zhou, B., Dabral, R., Wang, J., Golyanik, V., Theobalt, C., Slusallek, P., Guo, C.: Duetgen: Music driven two-person dance generation via hierarchical masked modeling. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
10. Guan, J., Yang, Q., Wang, K., Zhou, H., He, S., Xu, Z., Feng, H., Ding, E., Wang, J., Xie, H., et al.: Talk-act: Enhance textural-awareness for 2d speaking avatar reenactment with diffusion model. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
11. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
12. Hao, L., Cao, H., Feng, B., Shao, D., Tang, R., Yan, Z., Tian, Y., Yuan, L., Li, Y.: Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. Advances in Neural Information Processing Systems **38** (2026)
13. Haque, K.I., Yumak, Z.: Facexhubert: Text-less speech-driven e (x) pressive 3d facial animation synthesis using self-supervised speech representation learning. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 282–291 (2023)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
15. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing **29**, 3451–3460 (2021)
16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)

17. Kong, Z., Gao, F., Zhang, Y., Kang, Z., Wei, X., Cai, X., Chen, G., Luo, W.: Let them talk: Audio-driven multi-person conversational video generation. arXiv preprint arXiv:2505.22647 (2025)
18. Li, H., Huang, J., Jin, P., Song, G., Wu, Q., Chen, J.: Weakly-supervised 3d spatial reasoning for text-based visual question answering. IEEE Transactions on Image Processing **32**, 3367–3382 (2023)
19. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
20. Liang, C., Wang, Q., Chen, Y., Tang, M.: Wav2lip-hr: Synthesising clear high-resolution talking head in the wild. Computer Animation and Virtual Worlds **35**(1), e2226 (2024)
21. Liu, T., Chen, F., Fan, S., Du, C., Chen, Q., Chen, X., Yu, K.: Anitalker: animate vivid and diverse talking faces through identity-decoupled facial motion encoding. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 6696–6705 (2024)
22. Liu, Y., Zhu, L., Lin, L., Zhu, Y., Zhang, A., Li, Y.: Teaser: Token enhanced spatial modeling for expressions reconstruction (2025)
23. Mughal, M.H., Dabral, R., Habibie, I., Donatelli, L., Habermann, M., Theobalt, C.: Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1388–1398 (2024)
24. Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., Ginosar, S.: Learning to listen: Modeling non-deterministic dyadic facial motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20395–20405 (2022)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
26. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301. Ieee (2009)
27. Peng, Z., Fan, Y., Wu, H., Wang, X., Liu, H., He, J., Fan, Z.: Dualtalk: Dual-speaker interaction for 3d talking head conversations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21055–21064 (2025)
28. Peng, Z., Hu, W., Ma, J., Zhu, X., Zhang, X., Zhao, H., Tian, H., He, J., Liu, H., Fan, Z.: Synctalk++: High-fidelity and efficient synchronized talking heads synthesis using gaussian splatting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
29. Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., Zhao, H., He, J., Liu, H., Fan, Z.: Synctalk: The devil is in the synchronization for talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 666–676 (2024)
30. Peng, Z., Liu, J., Zhang, H., Liu, X., Tang, S., Wan, P., Zhang, D., Liu, H., He, J.: Omnisync: Towards universal lip synchronization via diffusion transformers. Advances in Neural Information Processing Systems **38**, 11405–11424 (2026)
31. Peng, Z., Luo, Y., Shi, Y., Xu, H., Zhu, X., Liu, H., He, J., Fan, Z.: Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5292–5301 (2023)

32. Peng, Z., Wu, H., Song, Z., Xu, H., Zhu, X., He, J., Liu, H., Fan, Z.: Emotalk: Speech-driven emotional disentanglement for 3d face animation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20687–20697 (2023)
33. Qi, X., Wang, Y., Zhang, H., Pan, J., Xue, W., Zhang, S., Luo, W., Liu, Q., Guo, Y.: Co^3 Gesture: Towards coherent concurrent co-speech 3d gesture generation with interactive diffusion. arXiv preprint arXiv:2505.01746 (2025)
34. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024)
35. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13759–13768 (2021)
36. Retsinas, G., Filntisis, P.P., Danecek, R., Abrevaya, V.F., Roussos, A., Bolkart, T., Maragos, P.: Smirk: 3d facial expressions through analysis-by-neural-synthesis. arXiv preprint arXiv:2404.04104 (2024)
37. Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1173–1182 (2021)
38. Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1173–1182 (2021)
39. Song, L., Yin, G., Jin, Z., Dong, X., Xu, C.: Emotional listener portrait: Neural listener head generation with emotion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20839–20849 (2023)
40. Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y.H., Yu, M., Liu, Y.j.: Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. ACM Transactions on Graphics (TOG) **43**(4), 1–9 (2024)
41. Tao, R., Pan, Z., Das, R.K., Qian, X., Shou, M.Z., Li, H.: Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3927–3935 (2021)
42. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
43. Wang, J., Xie, J.C., Li, X., Xu, F., Pun, C.M., Gao, H.: Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. IEEE Transactions on Visualization and Computer Graphics (2025)
44. Wang, Z., Zhu, X., Zhang, T., Wang, B., Lei, Z.: 3d face reconstruction with the geometric guidance of facial part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1672–1682 (2024)
45. Wu, J., Liu, Y., Lin, L., Zhu, Y., Zhu, L., Li, J., Li, Y.: Pear: Pixel-aligned expressive human mesh recovery. arXiv preprint arXiv:2601.22693 (2026)
46. Wu, J., Peng, R., Wang, Z., Xiao, L., Tang, L., Yan, J., Xiong, K., Wang, R.: Swift4d: Adaptive divide-and-conquer gaussian splatting for compact and efficient reconstruction of dynamic scene. arXiv preprint arXiv:2503.12307 (2025)
47. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12780–12790 (2023)

48. Xiong, L., Cheng, X., Tan, J., Wu, X., Li, X., Zhu, L., Ma, F., Li, M., Xu, H., Hu, Z.: Segtalker: Segmentation-based talking face generation with mask-guided local editing. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3170–3179 (2024)
49. Yang, Q., Guan, J., Wang, K., Yu, L., Chu, W., Zhou, H., Feng, Z., Feng, H., Ding, E., Wang, J., et al.: Showmaker: Creating high-fidelity 2d human video via fine-grained diffusion modeling. Advances in Neural Information Processing Systems **37**, 51039–51062 (2024)
50. Yang, Q., Huang, L., Wang, K., Guan, J., He, S., Li, F., Zhou, H., Yu, L., Li, Y., Feng, H., et al.: Gesturehydra: Semantic co-speech gesture synthesis via hybrid modality diffusion transformer and cascaded-synchronized retrieval-augmented generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12615–12625 (2025)
51. Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: European conference on computer vision. pp. 85–101. Springer (2022)
52. Yu, H., Wang, Z., Pan, Y., Cheng, M., Yang, H., Wang, C., Xie, T., Xu, X., Wei, X., Cai, X.: Llia–enabling low-latency interactive avatars: Real-time audio-driven portrait video generation with diffusion models. arXiv preprint arXiv:2506.05806 (2025)
53. Yu, H., Qu, Z., Yu, Q., Chen, J., Jiang, Z., Chen, Z., Zhang, S., Xu, J., Wu, F., Lv, C., et al.: Gaussiantalker: Speaker-specific talking head synthesis via 3d gaussian splatting. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3548–3557 (2024)
54. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: 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. pp. 8652–8661 (2023)
55. Zhu, L., Cai, X., Chen, Y., Li, Y., Yang, B., Liu, H., Chen, J., Li, C., LYu, J.: Omnihuman: A large-scale dataset and benchmark for human-centric video generation. arXiv preprint arXiv:2604.18326 (2026)
56. Zhu, L., Chen, Y., Liu, X., Li, T.H., Li, G.: Learning semantic facial descriptors for accurate face animation. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
57. Zhu, Y., Zhang, L., Rong, Z., Hu, T., Liang, S., Ge, Z.: Infp: Audio-driven interactive head generation in dyadic conversations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10667–10677 (2025)