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.

UniLS: End-to-End Audio-Driven Avatars for Unified Listening and Speaking

Xuangeng Chu[1,2*] Ruicong Liu[1,2*†] Yifei Huang[1,2] Yun Liu[1] Yichen Peng[3] Bo Zheng[1]
[1]Shanda AI Research Tokyo [2]The University of Tokyo [3]Institute of Science Tokyo
{xuangeng.chu, ruicong.liu, yifei.huang, yun.liu, bo.zheng}@shanda.com
peng.y.ag@m.titech.ac.jp

Abstract

Generating lifelike conversational avatars requires modeling not just isolated speakers, but the dynamic, reciprocal interaction of speaking and listening. However, modeling the listener is exceptionally challenging: direct audio-driven training fails, producing stiff, static listening motions. This failure stems from a fundamental imbalance: the speaker's motion is strongly driven by speech audio, while the listener's motion primarily follows an internal motion prior and is only loosely guided by external speech. This challenge has led most methods to focus on speak-only generation. The only prior attempt at joint generation relies on extra speaker's motion to produce the listener. This design is not end-to-end, thereby hindering the real-time applicability. To address this limitation, we present UniLS, the first end-to-end framework for generating unified speak-listen expressions, driven by only dual-track audio. Our method introduces a novel two-stage training paradigm. Stage 1 first learns the internal motion prior by training an audio-free autoregressive generator, capturing the spontaneous dynamics of natural facial motion. Stage 2 then introduces the dual-track audio, fine-tuning the generator to modulate the learned motion prior based on external speech cues. Extensive evaluations show UniLS achieves state-of-the-art speaking accuracy. More importantly, it delivers up to 44.1% improvement in listening metrics, generating significantly more diverse and natural listening expressions. This effectively mitigates the stiffness problem and provides a practical, high-fidelity audio-driven solution for interactive digital humans.

1. Introduction

Conversational avatars [15, 20, 25, 31, 44] serve as a bridge between humans and intelligent virtual agents [4, 8, 16, 37], enabling more natural and engaging human–computer interaction [3, 22, 42]. Although recent studies have advanced this field, most systems remain one-way, *i.e.*, either speak-only [23, 24, 33, 40] or listen-only [17, 19, 28, 34]. In real-world interaction, a conversational avatar should not only speak naturally but also listen naturally. When the user is talking, it should show appropriate listening reactions (blinks, nods, subtle expressions), and when it speaks, its mouth and facial expressions should match its own voice.

Recently, the first attempt at building a conversational avatar [26] has emerged, enabling joint speaking and listening facial motion generation. As shown in Fig. 1, it relies on speaker-A's facial sequence to produce speaker-B's speak–listen motions. This dependency prevents the system from being fully end-to-end and requires an additional motion acquisition or generation pipeline. The reliance on external motion input introduces extra processing stages and potential temporal lag, blocking the real-time deployment.

Unlike previous methods, we propose UniLS, an end-to-end framework for unified speaking and listening. As illustrated in Fig. 1, UniLS takes only audio inputs and generates the corresponding motions for both speaker-A and speaker-B. A straightforward strategy is to jointly train an end-to-end model on "speak-listen" data. However, this approach leads to a noticeable issue: the generated listening expressions become stiff and unnatural. We empirically find that this arises from an imbalance in audio–motion correlation. The speaker-A's speech strongly drives its own mouth and facial motion through phoneme and rhythm alignment, but only weakly relates to the speaker-B's face. Consequently, direct joint optimization drives the model to collapse into a safe, static facial prior for the listening branch, yielding expression stiffness.

To address this challenge, our key insight is to reinterpret listening behavior as a combination of internal motion prior and external audio cues. Human listening expressions first follow their own motion patterns a priori (*e.g.*, blink frequency, subtle nods, and muscle coordination), and then to be modulated by external speech. Following this intuition, we introduce a two-stage training paradigm. In stage 1, we remove audio input and train an audio-free generator on un-

*Equal contribution.
†Corresponding author.

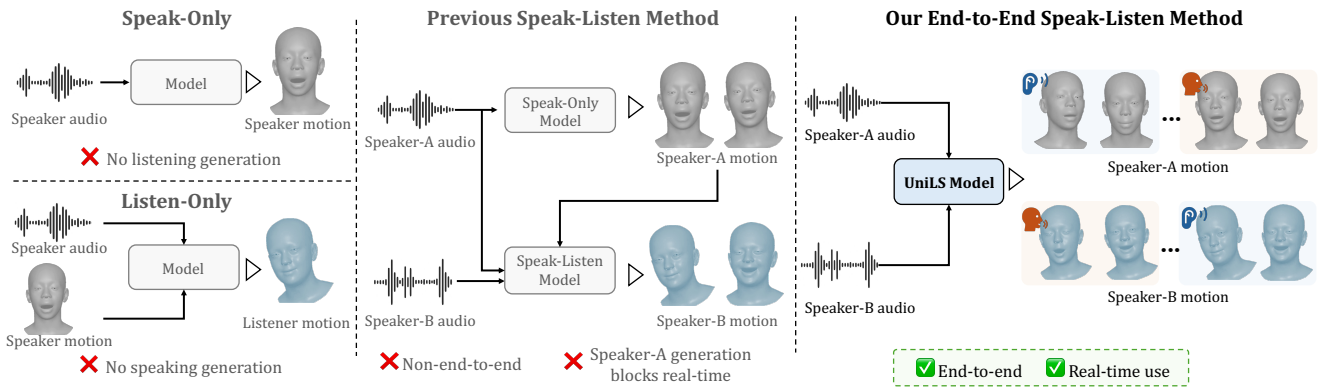


Figure 1. Comparison between previous methods and our proposed approach. Most previous studies remain one-way, *i.e.*, speak-only or listen-only. The previous speak–listen method [26] requires generating speaker A's facial sequence before producing speaker B's motions. The speaker-A generation makes it non-end-to-end and blocks real-time. In contrast, our method provides an end-to-end framework for unified, real-time speak–listen motion generation.

paired multi-scenario data. The generator learns to predict the next facial motion, allowing it to build an internal motion prior. In stage 2, we fine-tune the generator on paired conversational data. The model is conditioned on speaker A and B's speech, allowing the audio to drive expressive facial motion.

To enable this research, we track FLAME [14] parameters for videos from the Seamless Interaction dataset [1], providing high-quality 3D facial motion supervision for both speaking and listening behaviors. Evaluations demonstrate that UniLS achieves state-of-the-art accuracy in generating speaking motions. More importantly, our method delivers substantial gains in listening quality, achieving up to 44.1% improvement in distributional listening metrics over prior methods. This highlights the expressiveness and naturalness of our generated listening behaviors, resulting in more realistic and authentic conversational avatars.

Our contributions are as follows:

- We present the first end-to-end framework that enables a conversational avatar to both speak and listen naturally, driven solely by audio inputs.
- We propose a two-stage training paradigm in response to the issue of listening stiffness.
- Our method achieves state-of-the-art results, generating both high-quality speaking expressions and diverse, natural listening expressions.

2. Related work

2.1. Speech-driven 3D facial animation

Research on speech-driven 3D facial animation has evolved substantially over the past decades. Early works [9, 32, 41] relied on procedural or rule-based approaches, which segmented speech into phonemes and mapped them to visemes using handcrafted rules or constraint-based motion graphs. While these methods provided explicit control, they required laborious parameter tuning, mocap data specific to each subject, and offline processing, making them inflexible and often unable to capture the diversity and nuance of natural speech.

With the rise of deep learning, learning-based methods have largely replaced procedural pipelines by directly learning the mapping from audio to 3D facial motion. End-to-end architectures such as FaceFormer [11], CodeTalker[40], and SelfTalk [23] leverage mesh-based or vertex-level representations to generate expressive and temporally coherent animations. More recent frameworks, including UniTalker [10], ScanTalk [21], and ARTalk [5] further improve cross-identity generalization and enable retargetable animation across diverse facial topologies. Despite these advances, most existing models focus solely on speaking behavior, ignoring speak-listen interactions that occur in natural conversations.

2.2. Speak-listen 3D facial animation

Recent research starts focusing on modeling nonverbal listener behaviors, which are essential for natural and engaging interactions. DualTalk [26] represents the first attempt to generate speak–listen facial motions within a unified framework, producing both speaking and listening behaviors. However, it depends on the interlocutor's precomputed facial sequence, preventing true end-to-end generation and limiting real-time performance. Earlier studies such as Learning2Listen [18] generated only listener reactions (*e.g.*, nods or smiles) based on the speaker's speech and facial mesh. While effective in producing feedback, these methods are limited to single-turn responses and do not support speak-listen interactions. Other methods [19, 37, 43] predicted listener facial expressions given conversational context, but their outputs remained confined to brief, isolated

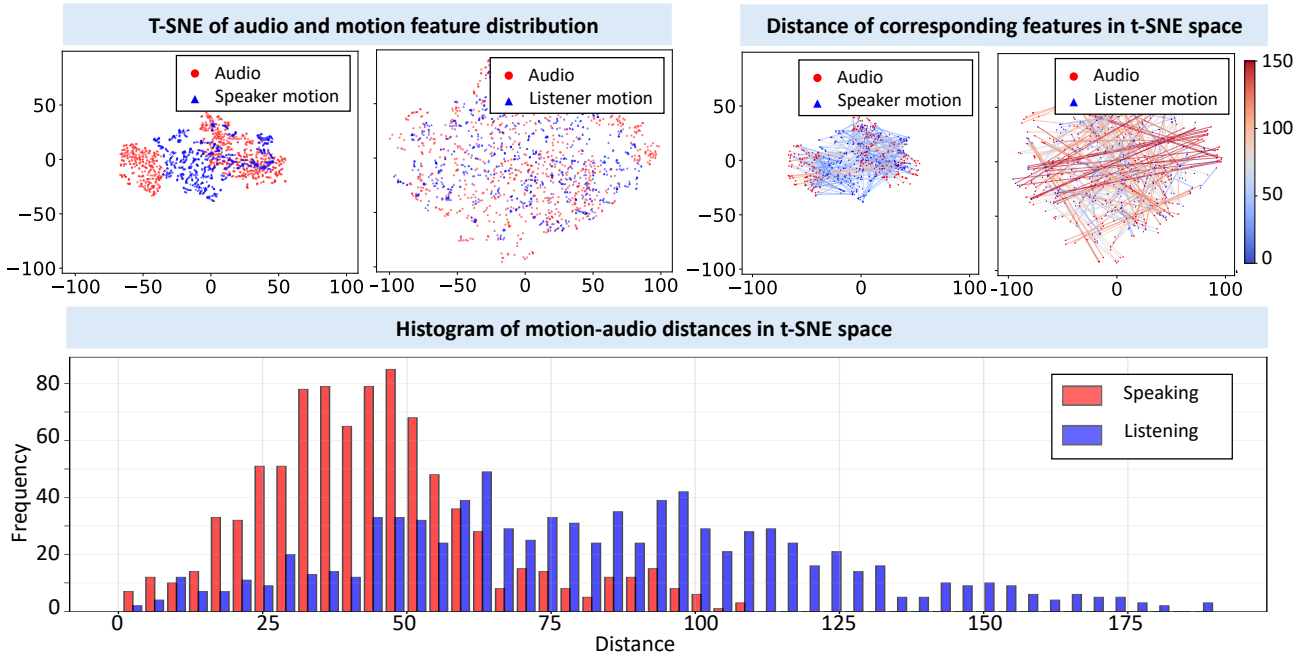


Figure 2. Correlation between facial expression parameters [14] and corresponding audio features [2]. For speaking, the audio is the speaker's own speech. For listening, the audio comes from the other speaker's speech.

reactions without modeling temporal coherence.

2.3. Autoregressive 3D facial animation

Autoregressive approaches model facial motion as a temporal sequence, predicting each frame step-by-step conditioned on past motion and audio context. These methods often utilize pre-trained speech representation models such as wav2vec [2], EnCodec [7], or HuBERT [12] to extract expressive audio features, which are then mapped to 3D mesh vertices or facial model parameters. FaceFormer [11] autoregressively predicts continuous facial motion parameters to ensure smooth temporal coherence, while MeshTalk [27] models discrete expression tokens to capture diverse articulation patterns. CodeTalker [40] and MultiTalk [30] adopt a VQVAE-based [35] tokenization of facial motion and employ transformer decoders to predict future motion tokens from audio cues. Learn2Talk [45] enhances FaceFormer with auxiliary lip-reading and lip-sync losses for improved alignment between phonemes and lip movements. MMHead [38] introduces textual conditioning for semantically guided motion control. More recently, ARTalk [5] advances this field by combining autoregressive modeling with style priors, achieving high-fidelity and temporally stable 3D facial animation from speech.

3. Motivation

Observation: stiffness in listening motion. When training a conversational avatar in an end-to-end manner, we observe a clear issue: the listening expressions become stiff whenever their own audio is silent. For example, when speaker-A is listening while speaker-B is talking, speaker-A's audio contains only silence. The model learns to generate minimal or no motion, causing speaker-A's listening behavior to degrade into a poker-face expression with little blinking, negligible micro-expressions, and almost no head movement. In contrast, a natural conversational avatar should exhibit active listening cues, such as subtle nods, timely blinks, shifts in eye gaze, and soft facial movements that reflect attention and engagement.

Analysis: low correlation between audio and listening motion. To understand why naive end-to-end training produces stiff listening behaviors, we analyze the correlation between audio features [2] and expression parameters [14] for both speaking and listening. As shown in Fig. 2, the t-SNE distributions reveal a clear pattern: audio features cluster closely with speaking motions but show much weaker alignment with listening motions. In the speaking case, audio and motion embeddings occupy nearby regions, with consistently small pairwise distances. In contrast, listening motions are scattered farther from the corresponding audio features. This indicates that listening behavior is only weakly correlated with the audio of the interlocutor, as many aspects of natural listening, such as blinking or micro-expressions, occur independently of the speech signal. This imbalance explains the listening stiffness: the network easily learns strong audio-driven mappings for speaking but re-

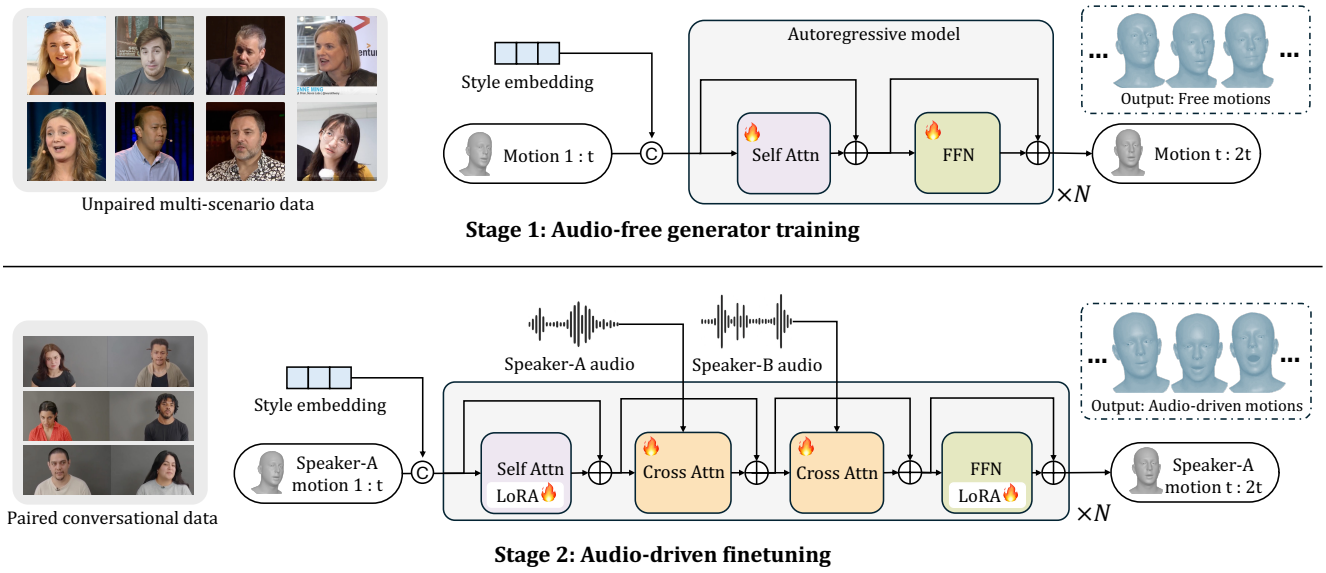


Figure 3. Overview of our two-stage training strategy. Stage 1 trains an audio-free generator on unpaired multi-scenario video data without using audio. Given past motions and a style embedding, the model predicts future free motion chunks. Stage 2 finetunes the generator on paired conversational clips by conditioning on speaker-A and speaker-B's audios through cross-attention, producing audio-driven speak–listen motions.

ceives insufficient guidance for listening, causing the listening branch to collapse into safe, low-variance motions.

Insight. According to the analysis, listening behavior should not be learned as a direct audio-to-motion mapping. Instead, natural listening arises from two components: 1) an internal motion prior that reflects spontaneous behaviors such as blinks, nods, and micro-expressions, and 2) external audio cues that modulate these intrinsic dynamics in response to conversational context. We design a two-stage training framework that separately learns these two components. In Stage 1, we train an audio-free generator to model the internal dynamics of facial behavior. In Stage 2, we finetune this generator by conditioning on dual-track audios, allowing external speech signals to modulate the facial expression.

4. Methods

An overview of our method is presented in Fig. 3. We train a chunk-based autoregressive model to generate facial motions, conditioned on both speakers' audio. The following sections provide further details: Sec. 4.1 describes the problem definitions, Sec. 4.2 explains the multi-scale codec model, Secs. 4.3 and 4.4 introduce the two-stage training of our speak-listen model.

4.1. Preliminaries

We represent facial motion using the FLAME 3D deformable model [14], a widely adopted parametric head model. FLAME encodes dynamic facial behavior through expression parameters ψ, poses parameters θ (including head pose, jaw pose and eye gaze), and shape parameters β. Given these parameters, FLAME reconstructs a full 3D face mesh with 5,023 vertices V using standard blendshape deformation and articulated pose transformations. For each motion chunk, we form a motion matrix $M \in \mathbb{R}^{T\times D}$ by concatenating the expression ψ and pose θ over T consecutive frames. Our goal is to learn an autoregressive generator $\mathcal{G}$ that takes audios as input and produces unified speak–listen facial motions coherently and naturally.

4.2. Multi-scale codec

Following previous work [5], we adopt a temporal multi-scale VQ codec to obtain a compact yet expressive discrete representation of 3D facial motion. This codec has been proven to achieve high-fidelity synthesis with strong temporal coherence.

Given an input motion chunk M, a transformer encoder first extracts a temporal feature $\mathbf{f}$ into a latent space. This latent feature is then quantized using a multi-scale codebook. Specifically, the codebook has a sequence of scales $[k_1, k_2, ..., k_L]$, where each level progressively refines the temporal resolution. At the l-th scale, the feature $\mathbf{f}^{(l)}$ are quantized and interpolated to motion code $\mathbf{c}^{(l+1)}$, following:

$$\mathbf{c}^{(l+1)} = \text{Interp}(\text{Quant}(\mathbf{f}^{(l)}), k_l),$$
$$\mathbf{f}^{(l+1)} = \mathbf{f}^{(l)} - \mathbf{c}^{(l+1)}, \quad (1)$$

where feature $\mathbf{f}^{(0)}$ is from the original encoder output. Function $\text{Quant}(\cdot)$ maps features to the nearest codebook entries, while $\text{Interp}(\cdot, k_l)$ adjusts temporal resolutions to ensure smooth transitions. We accumulate motion code $\mathbf{c}^{(l)}$ across scales to represent the original motion. These multi-scale discrete motion codes serve as the supervision targets for training our autoregressive generator. For clarity in the following subsections, we directly describe the model using the continuous motion matrix M, while the underlying codec operates in the discrete code space.

4.3. Stage 1: Audio-free generator training

In the first stage, we train an audio-free generator to learn internal motion priors that capture natural facial dynamics independent of speech. This generator is trained on unpaired multi-scenario data, including diverse video sources such as news broadcasts, interviews, streaming content, and casual talking videos, providing diverse facial behaviors across identities and environments.

As shown in Fig. 3, the input contains motion chunk M with style embedding $\mathbf{s}$. This style embedding is learned to encode speaker-specific motion characteristics, following [5]. This embedding is concatenated with the input chunk and then fed them into our autoregressive model $\mathcal{G}$. Our model is transformer-based, composed of stacked self-attention and feed-forward blocks. At each time step t, the generator predicts the next-chunk motion based solely on past motions and the style embedding, *i.e.*:

$$\hat{M}_{t:2t} = \mathcal{G}(M_{1:t}, \mathbf{s}). \quad (2)$$

The model is trained with an autoregressive reconstruction loss over each chunk:

$$\mathcal{L} = \sum_{t=1}^{T} ||\hat{M}_{t:2t} - M_{t:2t}||. \quad (3)$$

Through this process, the model learns to produce free motions, which reflects the internal motion prior such as blinking, subtle head movements, and micro-expressions.

4.4. Stage 2: Audio-driven finetuning

In the second stage, we finetune the generator to produce audio-driven conversational motions, enabling both speaking and listening behaviors. This stage uses paired conversational data, where synchronized videos and audios from speaker-A and speaker-B provide the appropriate dynamics required for natural dialogue modeling.

Here we describe the process for generating speaker-A's motion as an example. As illustrated in Fig. 3, the input consists of three components: the motion chunk M, style embedding $\mathbf{s}$, and the audios $\mathbf{a}^A, \mathbf{a}^B$ from speaker-A and speaker-B, respectively. To incorporate audio guidance, we extend the stage 1 architecture by adding two cross-attention layers to each transformer block. Specifically, one attends to speaker-A's audio (for speaking behavior) and the other attends to speaker-B's audio (for listening behavior). The newly added cross-attention layers are trained from scratch, while the backbone weights inherited from stage 1 are finetuned with LoRA [13]. This design ensures allows the model to adapt efficiently to audio conditioning without overwriting the learned internal motion priors. Formally, the generating process of stage 2 is expressed as:

$$\hat{M}_{t:2t} = \mathcal{G}(M_{1:t}, \mathbf{a}^A_{1:t}, \mathbf{a}^B_{1:t}, \mathbf{s}). \quad (4)$$

The training objective remains a chunk-wise autoregressive reconstruction loss, which is the same as Eq. (3). Generating motions for speaker-B follows the same procedure, with the two audios simply exchanged in their roles. Through this finetuning, the generator combines internal motion priors learned in stage 1 with dual-track audio guidance, producing smooth and expressive speak–listen motions.

Implementation Details. Our codec model's codebook consists of 256 entries, each with a code dimension of 64. The time window size is 100 frames (4 seconds), and the multi-scale levels are [1, 5, 25, 50, 100]. We used the AdamW optimizer with a learning rate of 1.0e-4 for training the codec, with a total batch size of 64 for 100,000 iterations. In the two-stage training, we train the autoregressive model using the AdamW optimizer with the same learning rate of 1.0e-4 and a batch size of 128 for 200,000 iterations. During this training, we employ a frozen wav2vec audio encoder [2]. All training was conducted on four NVIDIA H200 GPUs, requiring a total of approximately 40 GPU hours (10 GPU hours for the stage 1 and 30 GPU hours for stage 2).

5. Experiments

In this section, we first introduce the details of the dataset, provide an overview of the implementation of our method, describe the metrics used, and present the baseline methods. Subsequently, we compare our method with existing approaches across a range of evaluation metrics.

5.1. Experiment setting

Datasets. For paired conversational data, we use the Seamless Interaction dataset [1], which offers large-scale dyadic conversational videos. For multi-scenario data, we additionally adopt four large-scale video datasets: CelebV, TalkingHead-1KH [36], TEDTalk [6], and VFHQ [39]. To enable 3D facial motion supervision, we apply a carefully designed tracking pipeline to extract per-frame FLAME

Table 1. Evaluation of speaking and listening facial motions on the Seamless Interaction [1] test split. We use colors to denote the first and second places, respectively.

Method	Speak					Listen				
	LVE↓	MHD↓	FDD↓	PDD↓	JDD↓	FDD↓	PDD↓	JDD↓	F-FID↓	P-FID↓
DiffPoseTalk [29]	9.48	2.96	32.66	7.89	1.40	-	-	-	-	-
ARTalk [5]	7.46	2.12	31.64	7.66	1.19	-	-	-	-	-
ARTalk* [5]	6.79	2.02	27.41	8.55	0.81	30.62	9.52	1.53	10.779	0.072
DualTalk [26]	6.35	1.95	37.46	9.70	1.02	43.58	10.71	2.02	13.143	0.079
UniLS (ours)	5.83	1.89	18.41	4.67	0.71	17.12	4.75	0.98	4.304	0.038

[14] parameters, including detailed eye-gaze and head pose annotations. After filtering, we obtain 657.5 hours of conversational data from Seamless Interaction, and 546.5 hours of multi-scenario data from other datasets. The conversational data includes 251.5 hours of speaking motions comprising 22.6M frames, and 406.0 hours of listening motions comprising 36.5M frames. For the conversational dataset, we use 622.5 hours for training, 4.8 hours for validation, and 30.2 hours for testing.

Evaluation Metrics. Our metrics evaluate both speaking accuracy and listening naturalness.

a) Speaking. To assess lip synchronization quality, we use Lip Vertex Error (LVE) [27], which measures the maximum per-frame lip-vertex error. To assess overall facial accuracy, we use Mean Head Distance (MHD), the average distance across all head vertices. To assess upper-face motion consistency, we use Upper-face Dynamic Deviation (FDD) [40]. To assess temporal pose dynamics, we use Pose Dynamic Deviation for head pose (PDD) and jaw pose (JDD).

b) Listening. For listening behaviors, we compute FDD, PDD, and JDD on listening segments to assess whether the generated motion dynamics follow the ground-truth distribution. We additionally report FID on FLAME expression and pose parameters to assess distributional naturalness.

5.2. Quantitative results

To evaluate our method's performance, we compare UniLS with representative baselines:

- *DiffPoseTalk* [29]: A diffusion-based transformer model to generate speaking expressions and poses.
- *ARTalk* [5]: A autoregressive model to generate speaking expressions and poses.
- *ARTalk**: We adapt the original ARTalk for speak-listen generation by adding additional audio input.
- *DualTalk* [26]: A non-end-to-end speak-listen model.

Tab. 1 reports an evaluation of speaking and listening facial motions. Our method shows clear improvements in lip-sync accuracy (LVE, MHD) and speech-style alignment (FDD, PDD, JDD). These results indicate that our model not only tracks phoneme-to-motion correspondence precisely but also captures the characteristic dynamics of speech, such as upper-face involvement and coordinated head-jaw movement. For listening, our approach shows large improvements in distributional measures (FDD, PDD, JDD, F-FID, P-FID). This indicates that our model generates diverse expressions and head movements instead of collapsing into a neutral or static listening pose. Together, these results validate our two-stage design, showing that the model successfully learns to produce both natural listening reactions and accurate speaking expressions within a unified framework.

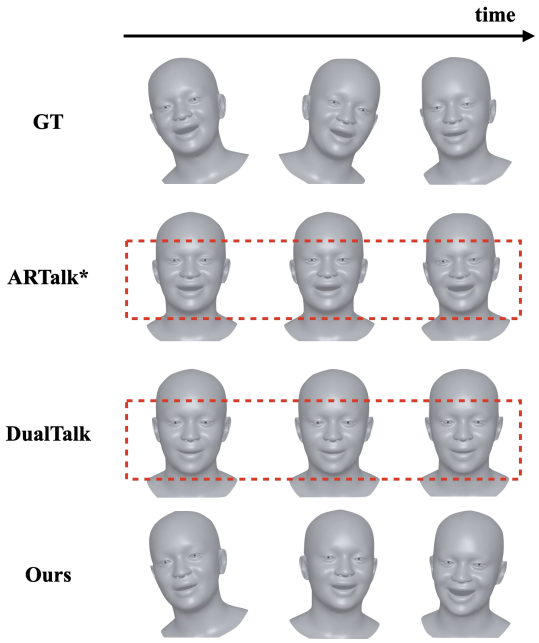


Figure 4. Qualitative comparison on listening motions. Red rectangles highlight motion stiffness over time. Additional qualitative evaluation results are available in the supplementary materials.

5.3. Qualitative results

In Figs. 4 and 5, we qualitatively compare our method with other baseline methods for listening and speaking motions,

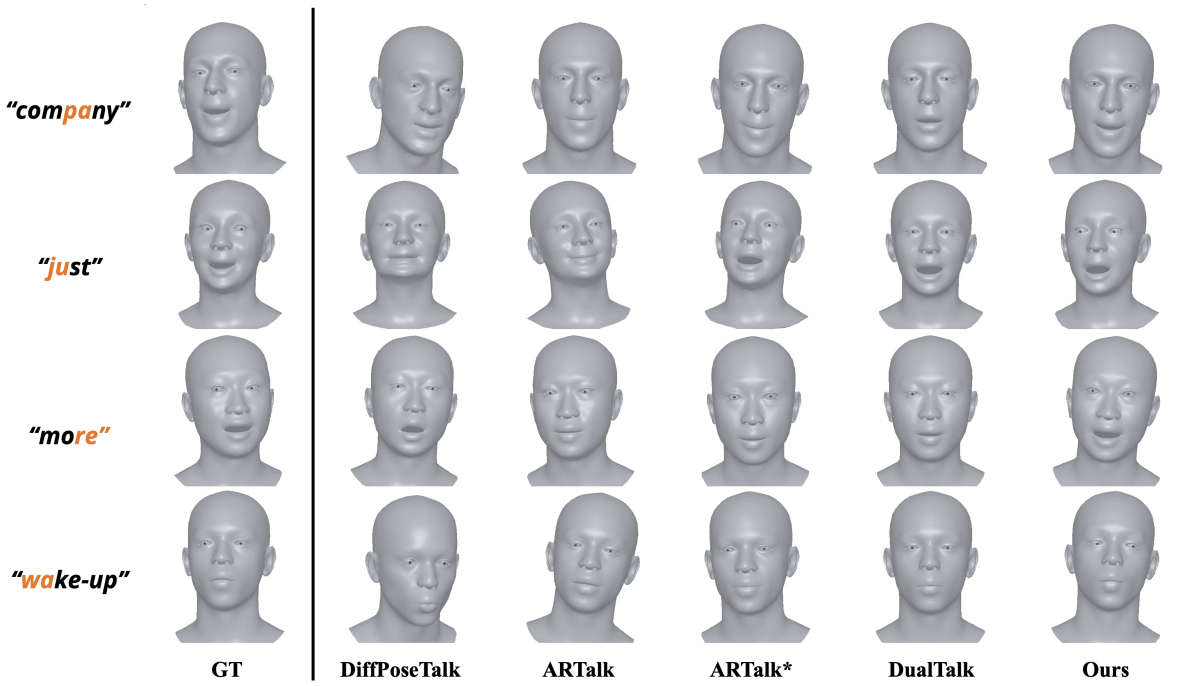


Figure 5. Qualitative comparison on speaking motions. Our method shows better alignment with the ground truth in expression style and lip synchronization. Additional qualitative evaluation results are available in the supplementary materials.

Table 2. User study results with 25 participants. Numbers (%) indicate the proportion of users who prefer our method over each baseline. "Sync" measures lip synchronization, while "Exp", "React", and "Pose" assess the naturalness of facial expressions, listening reactions, and head pose, respectively.

Method	Sync	Exp	React	Pose
vs. DiffPoseTalk [29]	55.34	61.65	-	58.25
vs. ARTalk [5]	75.36	75.36	-	71.98
vs. ARTalk* [5]	76.92	77.88	79.80	74.52
vs. DualTalk [26]	86.06	90.38	91.35	89.42

respectively. As shown in Fig. 4, we show the ability of our method on listening motion generation. ARTalk* and DualTalk exhibit noticeably stiff and low-variance facial motions, as highlighted by the red dashed boxes. In contrast, our method produces vivid, expressive, and temporally diverse expressions. The generated faces exhibit natural head dynamics, mouth shapes, and micro-expressions, demonstrating the effectiveness of our two-stage training in capturing realistic listening behavior. In Fig. 5, our method demonstrates excellent lip synchronization, accurately capturing various phonetic elements. Furthermore, the generated results exhibit realistic facial behavior, such as micro-expressions and head movements. Additional qualitative results are available in supplementary materials.

5.4. User study

User studies are recognized as an essential and reliable method for evaluating the performance of generative models. To thoroughly assess our method, we conducted a user study examining four key aspects of conversational facial motion: lip synchronization, facial expression naturalness, listening reaction naturalness, and head pose naturalness. For baseline methods that do not generate listening behaviors, the reaction naturalness category is omitted. We adopt a pairwise comparison protocol. For each trial, videos generated by our method and a baseline model are shown side by side in a randomized order. After viewing the video pairs, participants select the result they find more natural and realistic. We then compute the percentage of users who prefer our method in each category.

As summarized in Tab. 2, our method is consistently preferred over all baselines across all aspects. Among the 25 participants in our study, the most notable improvement appears in listening reactions, where over 90% of participants prefer our results compared to DualTalk. This overwhelming preference highlights the strength of our two-stage design: our model produces listening motions that are significantly more expressive, responsive, and human-like than existing methods.

Table 3. Ablation study on different components in our framework.

Method	Speak					Listen				
	LVE↓	MHD↓	FDD↓	PDD↓	JDD↓	FDD↓	PDD↓	JDD↓	F-FID↓	P-FID↓
single cross-attn	11.48	3.00	31.35	5.83	0.92	27.46	7.01	1.61	5.97	0.059
w/o stage 1	6.32	2.02	25.10	5.76	0.89	25.64	6.66	1.42	5.97	0.055
w/o multi-scenario data	6.26	1.99	18.77	4.81	0.76	17.81	4.85	1.04	4.62	0.040
Ours	**5.83**	**1.89**	**18.41**	**4.67**	**0.71**	**17.12**	**4.75**	**0.98**	**4.30**	**0.038**

Table 4. Comparison of real-time performance and computational costs. "RT" indicates whether real-time use is supported.

Method	RT	FPS	GFLOPS	#Params.
ARTalk* [5]	✓	357.7	720.7	489.5M
DualTalk [26]	✗	N/A	348.1	647.3M
UniLS (ours)	✓	560.6	739.4	421.3M

5.5. Ablation study

To better understand the contribution of each component in our framework, we conduct a series of ablation experiments.

Single cross-attention. To validate the importance of utilizing two cross-attention layers, we replaces our dual cross-attention design with a single cross-attention layer that receives a mixed audio stream (speaker A + speaker B). As shown in Tab. 3, this leads to a substantial degradation in speaking performance, especially in lip-synchronization metrics such as LVE and MHD. This decline is due to the difficulty in differentiating between the speech features when processing the mixed audio stream.

Without stage 1. This variant removes the audio-free training stage and directly trains the model using Stage 2 only. Without stage 1, the model struggles to generate natural listening behaviors, resulting in a clear drop in listening performance across FDD, PDD, JDD, and FID. These results confirm that learning the internal motion prior plays a vital role for natural listening reactions.

Without multi-scenario data. In this variant, we train the audio-free generator in stage 1 using only the conversational data from stage 2, removing the diverse multi-scenario dataset. As shown in Tab. 3, this leads to consistent drops across both speaking and listening metrics. Speaking performance becomes less stable, with higher LVE and MHD. Listening performance also declines noticeably, reflected by increased dynamic deviations and worse FID. These results indicate that multi-scenario data enables the model to learn a stronger internal motion prior, which generalizes better to both expressive speaking and listening behaviors.

5.6. Real-time performance

Tab. 4 compares our method with ARTalk [5] and DualTalk [26] in terms of computational cost and real-time capability. All experiments are conducted on a single NVIDIA GeForce RTX 5090 GPU. Although UniLS requires higher GFLOPS, it achieves the highest frame rate at 560.6 FPS, significantly outperforming both ARTalk and DualTalk (N/A due to its non–end-to-end design). Overall, UniLS provides the best trade-off: real-time generation, high throughput, and a smaller model size, making it well-suited for practical applications.

6. Conclusion and discussion

In this paper, we present UniLS, the first end-to-end audio-driven framework for generating unified speak–listen facial motions. Through a detailed analysis of audio–motion correlation, we identify the fundamental cause of listening stiffness in conventional end-to-end training. Building on this insight, we introduce a two-stage training paradigm that separates the learning of internal motion priors from audio-driven modulation. Extensive experiments on large-scale conversational datasets demonstrate that UniLS achieves state-of-the-art speaking accuracy, the most diverse and natural listening expressions, and real-time performance. Overall, UniLS establishes a foundation for conversational digital humans and opens new possibilities for interactive, lifelike human–AI communication.

Limitations and future work. While UniLS demonstrates strong capability in jointly generating speaking and listening motions, it has several limitations. First, the model lacks semantic understanding, so listener behaviors are primarily driven by acoustic cues (*e.g.*, rhythm, emphasis) rather than conversational meaning. As a result, it cannot yet produce semantically grounded reactions, such as nodding to agree or shaking the head to disagree. In addition, our chunk-based formulation limits fully continuous motion generation, occasionally causing subtle discontinuities. We view these challenges as promising directions for future work. Incorporating semantic information and developing continuous long-horizon models may enable more context-aware and smoother conversational avatars.

References

[1] Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, Sahir Gomez, Hongyu Gong, Srivathsan Govindarajan, Brandon Han, Sen He, Denise Hernandez, Yordan Hristov, Rongjie Huang, Hirofumi Inaguma, Somya Jain, Raj Janardhan, Qingyao Jia, Christopher Klaiber, Dejan Kovachev, Moneish Kumar, Hang Li, Yilei Li, Pavel Litvin, Wei Liu, Guangyao Ma, Jing Ma, Martin Ma, Xutai Ma, Lucas Mantovani, Sagar Miglani, Sreyas Mohan, Louis-Philippe Morency, Evonne Ng, Kam-Woh Ng, Tu Anh Nguyen, Amia Oberai, Benjamin Peloquin, Juan Pino, Jovan Popovic, Omid Poursaeed, Fabian Prada, Alice Rakotoarison, Alexander Richard, Christophe Ropers, Safiyyah Saleem, Vasu Sharma, Alex Shcherbyna, Jia Shen, Jie Shen, Anastasis Stathopoulos, Anna Sun, Paden Tomasello, Tuan Tran, Arina Turkatenko, Bo Wan, Chao Wang, Jeff Wang, Mary Williamson, Carleigh Wood, Tao Xiang, Yilin Yang, Zhiyuan Yao, Chen Zhang, Jiemin Zhang, Xinyue Zhang, Jason Zheng, Pavlo Zhyzheria, Jan Zikes, and Michael Zollhoefer. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. 2025. 2, 5, 6

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 3, 5

[3] Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, and Ying Shan. High-fidelity facial avatar reconstruction from monocular video with generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4541–4551, 2023. 1

[4] Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, and Hao Yan. Human conversation as a system framework: Designing embodied conversational agents. embodied conversational agents. j. cassell, j. sullivan, s. prevost and e. churchill, 2000. 1

[5] Xuangeng Chu, Nabarun Goswami, Ziteng Cui, Hanqin Wang, and Tatsuya Harada. Artalk: Speech-driven 3d head animation via autoregressive model. In *SIGGRAPH Asia 2025 Conference Papers*. Association for Computing Machinery, 2025. 2, 3, 4, 5, 6, 7, 8

[6] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016. 5

[7] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024. 3

[8] Stephan Diederich, Alfred Benedikt Brendel, Stefan Morana, and Lutz Kolbe. On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, 23(1):96–138, 2022. 1

[9] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11, 2016. 2

[10] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*, pages 204–221. Springer, 2024. 2

[11] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[12] 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. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460, 2021. 3

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[14] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, pages 194:1–194:17, 2017. 2, 3, 4, 6

[15] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17978, 2023. 1

[16] Lizi Liao, Grace Hui Yang, and Chirag Shah. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 3452–3455, 2023. 1

[17] Xi Liu, Ying Guo, Cheng Zhen, Tong Li, Yingying Ao, and Pengfei Yan. Customlistener: Text-guided responsive interaction for user-friendly listening head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2415–2424, 2024. 1

[18] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022. 2

[19] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10083–10093, 2023. 1, 2

[20] Arthur Niswar, Ee Ping Ong, Hong Thai Nguyen, and Zhiyong Huang. Real-time 3d talking head from a synthetic viseme dataset. In *Proceedings of the 8th International Con-*

ference on Virtual Reality Continuum and its Applications in Industry, pages 29–33, 2009. 1

[21] F. Nocentini, T. Besnier, C. Ferrari, S. Arguillere, S. Berretti, and M. Daoudi. Scantalk: 3d talking heads from unregistered scans. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[22] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2023. 1

[23] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023. 1, 2

[24] 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

[25] 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

[26] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21055–21064, 2025. 1, 2, 6, 7, 8

[27] 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. 3, 6

[28] Siyang Song, Micol Spitale, Cheng Luo, German Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth Andre, et al. React2023: The first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9620–9624, 2023. 1

[29] 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), 2024. 6, 7

[30] Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, Oh Hyun-Bin, Janghoon Ju, Suekyeong Nam, and Tae-Hyun Oh. Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. *arXiv preprint arXiv:2406.14272*, 2024. 3

[31] Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. Laughtalk: Expressive 3d talking head generation with laughter. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6404–6413, 2024. 1

[32] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284, 2012. 2

[33] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20621–20631, 2023. 1

[34] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dim: Dyadic interaction modeling for social behavior generation. In *European Conference on Computer Vision*, pages 484–503. Springer, 2024. 1

[35] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[36] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 5

[37] Haoyu Wu, Ziqiao Peng, Xukun Zhou, Yunfei Cheng, Jun He, Hongyan Liu, and Zhaoxin Fan. Vgg-tex: A vivid geometry-guided facial texture estimation model for high fidelity monocular 3d face reconstruction. *arXiv preprint arXiv:2409.09740*, 2024. 1, 2

[38] Sijing Wu, Yunhao Li, Yichao Yan, Huiyu Duan, Ziwei Liu, and Guangtao Zhai. Mmhead: Towards fine-grained multi-modal 3d facial animation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7966–7975, 2024. 3

[39] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 5

[40] 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, 3, 6

[41] Yuyu Xu, Andrew W Feng, Stacy Marsella, and Ari Shapiro. A practical and configurable lip sync method for games. In *Proceedings of Motion on Games*, pages 131–140. 2013. 2

[42] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, et al. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–12, 2023. 1

[43] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: a benchmark dataset and baseline. In *European conference on computer vision*, pages 124–142. Springer, 2022. 2

[44] Xukun Zhou, Fengxin Li, Ziqiao Peng, Kejian Wu, Jun He, Biao Qin, Zhaoxin Fan, and Hongyan Liu. Meta-learning

empowered meta-face: Personalized speaking style adaptation for audio-driven 3d talking face animation. *arXiv preprint arXiv:2408.09357*, 2024. 1

[45] Yixiang Zhuang, Baoping Cheng, Yao Cheng, Yuntao Jin, Renshuai Liu, Chengyang Li, Xuan Cheng, Jing Liao, and Juncong Lin. Learn2talk: 3d talking face learns from 2d talking face. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3