EmbodiedHead: Real-Time Listening and Speaking Avatar for Conversational Agents

Yu Zhang[1] *, Kaiyuan Shen[1] *, and Yang Li[1,2] †

[1] School of Computer Science and Technology, East China Normal University, Shanghai, China
{51275900029,51285900014}@stu.ecnu.edu.cn, yli@cs.ecnu.edu.cn
[2] Garabido Intelligence, Shanghai, China

https://vpx-ecnu.github.io/EmbodiedHead-Website/

Abstract. We present EmbodiedHead, a speech-driven talking head framework that equips LLMs with real-time visual avatars for conversation. A practical embodied avatar must achieve real-time generation, unified listening-speaking behavior, and high rendered visual quality simultaneously. Our framework couples the first Rectified-Flow Diffusion Transformer (DiT) for this task with a differentiable renderer, enabling diverse, high-fidelity generation in as few as four sampling steps. Prior listening-speaking methods rely on dual-stream audio, introducing an interlocutor look-ahead dependency incompatible with causal user–LLM interaction. We instead adopt a single-stream interface with explicit per-frame listening-speaking state conditioning and a Streaming Audio Scheduler, suppressing spurious mouth motion during listening while enabling seamless turn-taking. A two-stage training scheme of coefficient-space pretraining and joint image-domain refinement further closes the gap between motion-level supervision and rendered quality. Extensive experiments demonstrate state-of-the-art visual quality and motion fidelity in both speaking and listening scenarios.

Keywords: Speech-driven 3D Talking Head Generation · Rectified Flow · Embodied Conversational Agent

1 Introduction

Large Language Models (LLMs) are now widely used for daily chat and assistance, but most systems still interact through plain text or voice. Without a visible face, users miss social cues such as eye contact, facial expression, and head motion that convey attention, emotion, and turn-taking. Embodied Social Presence (ESP) Theory argues that an embodied representation (*e.g.*, an avatar) becomes the focal point through which people experience co-presence and interpret social interaction; richer embodied cues strengthen perceived social presence and engagement in mediated communication [22].

* Equal contribution.
† Corresponding author.

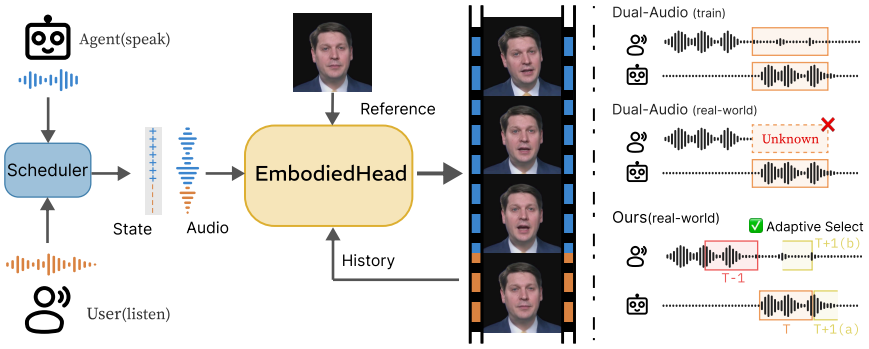

Fig. 1: We present EmbodiedHead, which generates a real-time head-embodied avatar for LLMs. Unlike dual-audio methods, it uses a single audio stream with explicit listening-speaking state conditioning to achieve unified conversational behavior.

In this paper we study *Head-Embodied LLMs*: equipping an LLM with a head-only visual avatar that listens and speaks in real time during conversation. A practical head-embodied LLM avatar must meet three goals at once: (i) real-time generation to support natural turn-taking; (ii) unified listening-speaking behavior to provide role-aware nonverbal signaling throughout the full interaction rather than only during speech; and (iii) high rendered visual quality to ensure perceptual plausibility and prevent visual artifacts from undermining credibility and social presence. Compared to 2D-based methods [9, 15], 3D-based methods [5, 10, 32, 37] generally enable lower inference cost, making them a promising route toward Head-Embodied LLMs.

Most current listening-speaking integrated methods [2, 7, 26, 42] employ a dual-stream audio architecture, whose datasets suffer from limited diversity in languages and scenarios. More importantly, dual-stream conditioning introduces an interlocutor look-ahead dependency: generating the current avatar response requires the interlocutor's reactive audio or visual feedback within the same window, which is not available at inference time in user–LLM interactions. Consequently, the model cannot reliably leverage cross-stream cues in a causal setting, and the dual-stream design may collapse into an effectively single-audio, alternately driven paradigm. Beyond the interaction architecture, most 3D talking-head methods treat mesh generation and image rendering as two separate steps. Motion is trained and evaluated in coefficient space, so the reported speed and quality numbers do not reflect the rendered visual quality that a head-embodied LLM avatar actually delivers to the user.

To address these issues, we propose **EmbodiedHead**, an end-to-end speech-driven talking-head framework designed for head-embodied LLMs (Fig. 1). Our pipeline couples the first Rectified-Flow [21] Diffusion Transformer (DiT) [3, 25] for this task with a differentiable renderer, retaining diffusion-model diversity while enabling high-fidelity generation in as few as four sampling steps. To remove the interlocutor look-ahead dependency, we adopt a single-stream audio

interface and inject an explicit per-frame *listening-speaking state* (LS-state) into both the model input and the audio cross-attention via FiLM modulation [27], giving the model a clear mode signal at every frame. A *Streaming Audio Scheduler* merges microphone and LLM audio into a unified window and emits aligned LS-states, enabling rapid turn-taking without future information. To close the gap between coefficient-space supervision and rendered quality, we adopt a two-stage training scheme: the DiT is first pretrained with flow matching in coefficient space for stable dynamics, then jointly fine-tuned with the renderer using image-domain losses. The straight-path property of Rectified Flow supports reliable one-step endpoint estimation, making this joint optimization both well-grounded and efficient.

Overall, the contributions of this paper can be summarized as follows:

- We propose an end-to-end framework coupling a Rectified-Flow DiT with a differentiable renderer and multi-level conditioning, achieving real-time, diverse talking-head generation in few sampling steps.
- We enable unified listening-speaking behavior via explicit LS-state conditioning and a Streaming Audio Scheduler, suppressing listening-phase mouth hallucinations and supporting rapid turn-taking.
- A two-stage scheme jointly optimizes the DiT and renderer with image-domain supervision, closing the gap between coefficient-space training and rendered quality. Extensive experiments demonstrate state-of-the-art performance in both speaking and listening scenarios.

2 Related Work

2.1 Speech-driven 3D Talking Head Generation

Most work in speech-driven 3D talking-head generation follows a two-stage pipeline: a motion model generates 3D motion from audio, which is then rendered by a separate module [8, 10, 32, 37]. VOCA [8] established identity-decoupled regression from speech. FaceFormer [10] and CodeTalker [37] introduced Transformer modeling and discrete motion priors to improve lip sync and motion diversity. DiffPoseTalk [32] applied diffusion for style-conditioned generation, while ARTalk [5] and GLDiTalker [19] target real-time output via autoregressive codebooks and latent diffusion. Despite these advances, training and evaluation stay in coefficient space: the reported accuracy and speed reflect mesh-level performance, not the rendered visual quality that users actually perceive.

A parallel line of work bypasses parametric models and drives a per-identity neural scene directly with audio, using either NeRF [12, 17, 23] or 3D Gaussian Splatting [1, 16]. These methods can capture fine appearance details, but the reconstructed scene is largely static: motion is limited to the lip region while eye movement, head pose, and lighting variation are poorly handled, yielding a rigid and visually flat result. Per-identity scene reconstruction also adds a costly setup stage, reducing suitability for general head-embodied LLM deployment. Some 2D methods learn a direct mapping from audio to images and achieve

high rendered quality [9, 15], but their generation speed is too slow for real-time head-embodied LLM interaction.

2.2 Listening-Speaking Integrated 3D Avatar Generation

Unified listening-speaking generation has evolved from single-listener modeling toward multi-turn streaming frameworks [2, 7, 24, 26, 30, 34, 42]. Learning to Listen [24] modeled stochastic 3D listener responses but did not address multi-turn dialogue. DualTalk [26] extended this to multi-round conversations under a dual-stream setting, but processes complete audio sequences offline and generates outputs in a single pass, making streaming deployment infeasible. TIMAR [2] introduced turn-level causal modeling to address latency, yet relies on the interlocutor's visual stream as input, which adds extra acquisition cost in real-time scenarios. INFP [43] and UniLS [7] further improved naturalness and continuity under dual-stream audio settings. Despite this progress, all dual-stream methods share a common structural limitation: generating the current window requires the interlocutor's reactive audio for the same span, which is unavailable in real user–LLM chat. This interlocutor look-ahead dependency makes dual-stream conditioning incompatible with causal streaming inference, and in practice these methods tend to degrade into alternated single-audio driving with unreliable mode inference during listening phases. MANGO [42] recognized the value of image-domain supervision and introduced 2D-lifted training alongside dual-stream audio. However, MANGO is built on DDPM [13], which requires many sampling steps to produce reliable outputs. Single-step estimates from DDPM carry large approximation errors, so image-domain losses cannot be grounded in outputs that match actual inference behavior, limiting the practical effect of the image supervision.

3 Method

Fig. 2 provides an overview of EmbodiedHead. After formalizing the task and introducing Rectified Flow (Sec. 3.1), we detail the injection of multiple conditions into DiT blocks (Sec. 3.2). To unify speaking and listening behaviors in conversational streams, we then propose a listening-speaking state (Sec. 3.3), culminating in a two-stage training scheme for end-to-end image-domain refinement (Sec. 3.4).

3.1 Preliminaries

Given an input audio segment $\mathbf{a}$, we aim to generate a temporally coherent 3D motion sequence $\mathbf{x}_1 \in \mathbb{R}^{T\times D}$ for a streaming window of length T frames. We use $\mathbf{c}$ to denote the full set of conditioning variables.

We use FLAME [18] to represent facial geometry, parameterizing each frame by an expression vector $\mathbf{e}^\tau \in \mathbb{R}^{100}$ and a pose vector $\mathbf{p}^\tau$ for $\tau \in \{1, \ldots, T\}$, with a shared identity shape $\mathbf{s} \in \mathbb{R}^{300}$. The per-frame motion vector is defined as

$$\mathbf{x}_1^\tau = [\mathbf{e}^\tau,\ \mathbf{p}^\tau], \qquad D = 100 + \dim(\mathbf{p}^\tau). \tag{1}$$

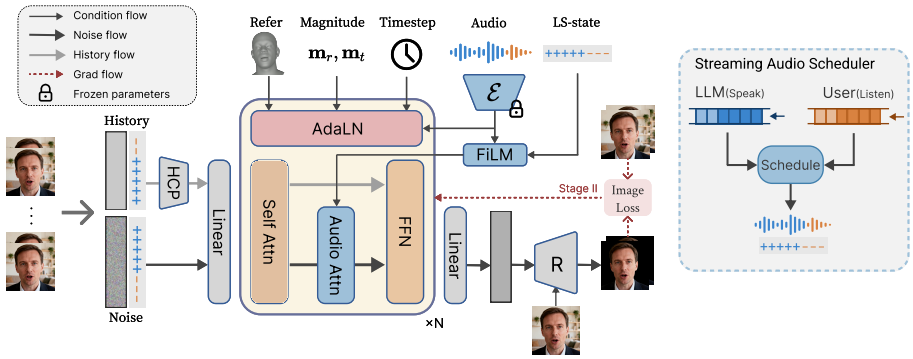

Fig. 2: EmbodiedHead employs a Rectified-Flow DiT to generate speech-driven talking-head animation in few steps. It conditions on reference, timestep, motion magnitude, and LS-state. A streaming scheduler merges user–LLM audio, enabling unified listening-speaking.

We adopt Rectified Flow [21] to retain the diversity of diffusion models while reducing sampling steps. Let $\mathbf{x}_1$ be a data sample and $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be Gaussian noise. Rectified Flow defines a straight interpolation path

$$\mathbf{x}_t = \mathbf{x}_0 + t\,(\mathbf{x}_1 - \mathbf{x}_0), \qquad t \sim \mathcal{U}[0,1]. \tag{2}$$

Along this path, the target velocity is a constant vector $\mathbf{u}_t = \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_1 - \mathbf{x}_0$. We train a conditional velocity field $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$ by minimizing the flow-matching objective [20]

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[\big\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\big\|_2^2\Big]. \tag{3}$$

3.2 Multi-Condition Rectified-Flow DiT

We parameterize $\mathbf{v}_\theta$ with a DiT to model the conditional velocity field. The condition is instantiated as $\mathbf{c} = (\mathbf{a}, \mathbf{s}, \mathbf{x}^{\mathrm{ref}}, \mathbf{x}^{\mathrm{hist}}, \mathbf{q}, \mathbf{m})$, where $\mathbf{x}^{\mathrm{ref}}$ and $\mathbf{s}$ represent the reference frame, $\mathbf{x}^{\mathrm{hist}}$ denotes history FLAME parameters, $\mathbf{q}$ is the per-frame listening-speaking state (Sec. 3.3) and $\mathbf{m}$ controls normalized rotation/translation magnitudes. To inject these conditions in a stable and controllable manner, we adopt three complementary pathways.

Model input. We build a single sequence of motion tokens by concatenating a history FLAME $\mathbf{x}^{\mathrm{hist}} \in \mathbb{R}^{L\times D}$ and the current noisy window $\mathbf{x}_t \in \mathbb{R}^{T\times D}$, and explicitly append the listening-speaking state $\mathbf{q} = [\mathbf{q}^{\mathrm{hist}}, \mathbf{q}^{\mathrm{cur}}] \in \{-1, +1\}^{L+T}$ as an extra channel to every frame feature. Given the flow time t, we form augmented frame features by channel-wise concatenation

$$\tilde{\mathbf{x}}^{\mathrm{hist}} = [\mathbf{x}^{\mathrm{hist}}, \mathbf{q}^{\mathrm{hist}}], \qquad \tilde{\mathbf{x}}_t = [\mathbf{x}_t, \mathbf{q}^{\mathrm{cur}}]. \tag{4}$$

Inspired by [40], we propose History Context Packing (HCP) to efficiently encode long-range context: $\tilde{\mathbf{x}}^{\text{hist}}$ is partitioned into G groups of exponentially growing temporal spans, and each group is compressed into a fixed number of tokens via a learnable linear projection, yielding $\tilde{\mathbf{h}} \in \mathbb{R}^{H\times(D+1)}$. Finally, we concatenate these packed history tokens with the current tokens and project them to the transformer dimension: $\mathbf{Z}_0 = \text{Linear}([\tilde{\mathbf{h}}, \tilde{\mathbf{x}}_t]) \in \mathbb{R}^{(H+T)\times d}$.

Frame-level audio attention. We use mHuBERT [14, 39] as our speech encoder, fusing all hidden layers with learnable softmax weights, and inject frame-level audio features into the noise via cross-attention. Notably, in each DiT block, the self-attention layer runs over all $(H + T)$ motion tokens, while the audio cross-attention layer is applied only to the T current-window tokens.

To help the model interpret the audio source, we apply a state-conditioned FiLM modulation [27] to audio features before cross-attention:

$$\hat{\mathbf{A}}^{\tau} = \mathbf{A}^{\tau} \odot \big(1 + \boldsymbol{\alpha}(q^{\tau})\big) + \boldsymbol{\beta}(q^{\tau}), \tag{5}$$

where $[\boldsymbol{\alpha}(q^{\tau}), \boldsymbol{\beta}(q^{\tau})] \in \mathbb{R}^{2D}$ is produced by a lightweight zero-initialized MLP. Each transformer block has its own modulation parameters, while parameters are shared across time steps within a block. For temporal alignment, we also use a diagonal locality mask in cross-attention: $M_{i,j} = 0$ if $|i - j| \le R$, and $M_{i,j} = -\infty$ otherwise, where R=2 frames.

Global conditioning via AdaLN. Following mainstream DiT paradigms [3, 25], we introduce global conditioning through AdaLN [38]. Specifically, a unified global conditioning vector is constructed by concatenating: (i) the flow timestep embedding $\mathbf{t}$, (ii) a reference embedding from shape and reference motion ($[\mathbf{s}, \mathbf{x}_{\text{ref}}]$), (iii) a mean-pooled global audio embedding computed over current-window audio tokens only, using a separate set of learnable layer-fusion weights independent of those used in cross-attention, and (iv) the motion magnitude guidance $\mathbf{m} = [\mathbf{m}_{\text{r}}, \mathbf{m}_{\text{t}}] \in [0, 1]^2$.

To mitigate over-averaging caused by weak audio-motion correlation, we explicitly inject motion magnitude, which offers superior interpretability and practical utility over implicit style features [5, 32]. During training, $\mathbf{m}_{\text{r}}$, $\mathbf{m}_{\text{t}}$ are computed from the ground-truth target window as the mean successive-frame rotational displacement (SO(3) geodesic angle) and translational displacement magnitude, and then normalized to $[0, 1]$.

3.3 Listening-Speaking State Conditioning

As discussed in Sec. 1, dual-stream audio conditioning is ill-suited for causal user–LLM interaction because future user audio is unavailable. We therefore adopt a *single-stream* conditioning interface and explicitly represent the behavioral mode with a per-frame listening-speaking state (LS-state). Training uses a unified single-stream waveform with aligned LS-state supervision. At inference time, a Streaming Audio Scheduler causally maps the microphone stream

Algorithm 1 Streaming audio scheduler

Require: Listening queue $\mathcal{Q}_{\mathrm{L}}$, Speaking queue $\mathcal{Q}_{\mathrm{S}}$, cursors $(c_{\mathrm{L}}, c_{\mathrm{S}})$, previous mode m_{prev}, window length T
Ensure: Waveform $\mathbf{a}$ and LS-state $\mathbf{q} \in \{-1, +1\}^T$

$U \leftarrow \textsc{GetUnconsumedLength}(\mathcal{Q}_{\mathrm{S}}, c_{\mathrm{S}})$ ▷ unconsumed frames in speaking queue
if $U \geq T$ **then** ▷ speaking queue alone fills the window
  $\mathbf{a} \leftarrow \textsc{SliceHead}(\mathcal{Q}_{\mathrm{S}}, c_{\mathrm{S}}, T)$; $\mathbf{q} \leftarrow +\mathbf{1}_T$
  $c_{\mathrm{S}} \leftarrow c_{\mathrm{S}} + T$; $m_{\mathrm{prev}} \leftarrow$ speak
else if $U > 0$ **then** ▷ partial Speak: fill remainder with listening tail
  $x_{\mathrm{S}} \leftarrow \textsc{SliceHead}(\mathcal{Q}_{\mathrm{S}}, c_{\mathrm{S}}, U)$; $x_{\mathrm{L}} \leftarrow \textsc{SliceTail}(\mathcal{Q}_{\mathrm{L}}, T - U)$
  $\mathbf{a} \leftarrow (m_{\mathrm{prev}} = \text{speak})\ ?\ x_{\mathrm{S}} \| x_{\mathrm{L}} : x_{\mathrm{L}} \| x_{\mathrm{S}}$ ▷ preserve temporal continuity
  $\mathbf{q} \leftarrow \textsc{LabelBySource}(\mathbf{a})$; $c_{\mathrm{S}} \leftarrow c_{\mathrm{S}} + U$
  $m_{\mathrm{prev}} \leftarrow (m_{\mathrm{prev}} = \text{speak})\ ?\ \text{listen} : \text{speak}$ ▷ track window tail state
else ▷ no pending Speak: pure listening mode
  $\mathbf{a} \leftarrow \textsc{SliceTail}(\mathcal{Q}_{\mathrm{L}}, T)$; $\mathbf{q} \leftarrow -\mathbf{1}_T$; $m_{\mathrm{prev}} \leftarrow$ listen
end if
return $(\mathbf{a}, \mathbf{q})$

and the LLM stream to the same conditioning window and outputs the aligned LS-state.

LS-state definition and acquisition. We define $q^\tau \in \{-1, +1\}$ with $+1$ for speaking and -1 for listening, and denote the window state as $\mathbf{q} \in \{-1, +1\}^T$. During training, we run TalkNet-ASD [33] on each video to obtain a frame-level speaking confidence, apply a short temporal smoothing to suppress jitter, and then binarize it into $\mathbf{q}$. During inference, $\mathbf{q}$ is constructed from *audio provenance*: frames sourced from LLM are labeled speaking $(+1)$, while frames sourced from the environment are labeled listening (-1).

LS-state injection. We inject $\mathbf{q}$ through two mechanisms described in Sec. 3.2. We concatenate q^τ to each motion token in the history and current window. This provides an explicit mode indicator at the input representation and yields a mode-consistent temporal context. We apply a state-conditioned FiLM to frame-level audio features before cross-attention in each DiT block. This maintains the conditioning signal throughout the network depth.

Streaming Audio Scheduler. We maintain two audio queues: $\mathcal{Q}_{\mathrm{L}}$ as a rolling buffer (microphone) and $\mathcal{Q}_{\mathrm{S}}$ as a monotonically consumed queue. At each tick, we output a fixed-length window of T frames by prioritizing unconsumed samples in $\mathcal{Q}_{\mathrm{S}}$ and filling any deficit with the most recent context from $\mathcal{Q}_{\mathrm{L}}$. When both sources appear within one window, we order the two segments using the previous window's tail mode m_{prev} to reduce boundary discontinuities; $\mathbf{q}$ is emitted alongside $\mathbf{a}$ by labeling each frame by its source (Alg. 1).

3.4 Two-Stage Training

The FLAME parameters estimated from monocular videos are inevitably noisy and biased, especially in frames with occlusion, motion blur, or extreme pose, where tracker confidence is low. Pure coefficient-space supervision therefore forces the model to fit unreliable targets that are not always consistent with image evidence. We thus adopt a two-stage strategy: coefficient pretraining for stable dynamics, followed by end-to-end image refinement for bias correction.

Stage I: Coefficient-space pretraining. We partition the per-frame motion vector $\mathbf{x}_1^\tau = [\mathbf{e}^\tau,\ \mathbf{p}^\tau]$ into five parameter groups $\mathcal{K}$ (expression, jaw, eye, rotation, and translation) and optimize flow matching with per-group reweighting. The Stage-I objective is

$$\mathcal{L}_{\mathrm{I}} = \sum_{k\in\mathcal{K}} \lambda_k \left\| \mathbf{v}_\theta^{(k)} - (\mathbf{x}_1^{(k)} - \mathbf{x}_0^{(k)}) \right\|_2^2 + \lambda_{\mathrm{smooth}}\, \mathcal{L}_{\mathrm{smooth}}^{\mathrm{pose}}. \tag{6}$$

$\mathcal{L}_{\mathrm{smooth}}^{\mathrm{pose}} = \sum_\tau \|\mathbf{p}^{\tau+1} - \mathbf{p}^\tau\|_2^2$ is applied only to rotation and translation dimensions to suppress jitter from video cuts and tracking instability; expression and jaw are excluded to avoid over-smoothing lip articulation.

Stage II: End-to-end image refinement. For image-domain refinement, we use GAGAvatar [6] as the differentiable renderer. While it renders fewer facial details than high-fidelity avatar pipelines [28], its single-image avatar construction makes end-to-end fine-tuning practical on large-scale 2D video datasets. Moreover, GAGAvatar explicitly models intra-oral regions (*e.g.*, teeth and tongue), which is crucial for realistic speech appearance. In Stage II, we unfreeze the renderer and jointly optimize both the DiT and differentiable renderer parameters with image-domain losses, allowing the renderer to co-adapt to the generated motion distribution.

Using Rectified Flow, we can estimate the endpoint with a single Euler update,

$$\hat{\mathbf{x}}_1 \approx \mathbf{x}_t + (1-t)\, \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}), \tag{7}$$

where $\mathbf{x}_t$ is sampled by Eq. (2). The straight path has a constant target velocity, so flow matching learns a displacement field; consequently, the inference step (setting t=0) coincides with the rendered endpoint, and image losses supervise exactly what will be sampled.

We then render the reconstructed endpoint with GAGAvatar $\hat{\mathbf{I}} = \mathcal{R}(\hat{\mathbf{x}}_1, \mathbf{s})$, and apply image-domain losses to jointly refine both the motion model and the renderer,

$$\mathcal{L}_{\mathrm{II}} = \lambda_{\mathrm{coef}}\, \mathcal{L}_{\mathrm{I}} + \lambda_{\mathrm{img}} \Big(\lambda_1 \|\hat{\mathbf{I}} - \mathbf{I}\|_1 + \lambda_p \|\psi(\hat{\mathbf{I}}) - \psi(\mathbf{I})\|_2^2 \Big), \tag{8}$$

where $\psi(\cdot)$ is a frozen AlexNet-based LPIPS perceptual feature extractor [41].

4 Experiments

4.1 Experimental Setup

Datasets. To ensure the robustness and generalization of our proposed model, we curated a diverse dataset comprising 11,372 samples, totaling 27.2 hours of footage. The primary data source is filtered from the TFHP dataset [32], contributing 14.7 hours. To enhance linguistic and environmental diversity, we integrated 7.8 hours of data from the TalkVid [4] and VFHQ [36] datasets. We additionally curated 3.9 hours of segments from the RealTalk dataset [11] to specifically model listening behaviors. We extract 3D FLAME parameters from videos using the GAGAvatar [6] tracking pipeline. For evaluation, we randomly sample 100 speaking clips and 50 listening clips as the test set.

Implementation Details. For the audio condition, we adopt mHuBERT-147 [14, 39] for multilingual adaptation. The motion branch predicts FLAME parameters over a 100-frame window, with 75 historical motion frames compressed into 20 frames using HCP. We train the model in two stages on two NVIDIA RTX 3090 GPUs: 80K iterations with lr=6e-4 and batch size 64, followed by 150K iterations with lr=8e-5 and batch size 4. During inference, we use 4-step Euler integration and set the motion-magnitude guidance to 0.3. More details are in the supplementary materials.

Baselines. To comprehensively evaluate EmbodiedHead, we select state-of-the-art baselines across two distinct tracks. For standard speech-driven motion generation, we compare our method with FaceFormer [10], CodeTalker [37], DiffPoseTalk [32], and ARTalk [5]. To explicitly assess interactive conversational dynamics, we benchmark against DualTalk [26], UniLS [7], and TIMAR [2], which are recent frameworks for unified listening-and-speaking motion generation. For visualization and image-domain evaluation, all baselines are rendered using the GAGAvatar [6].

4.2 Quantitative Evaluation

Metrics. Our evaluation metrics are organized around the three core objectives of a head-embodied LLM avatar. For *high visual quality*, we adopt standard image-domain metrics: PSNR, SSIM [35], and LPIPS [41], which directly reflect the rendered fidelity perceived by users. For *motion fidelity and listening-speaking behavior*, we use LVE [29] and MOD [37] for lip synchronization, FDD [37] for upper-face naturalness, BA [31] for head-motion rhythmic synchronization, and SID [26] for generation diversity. *Real-time efficiency* is evaluated by the generation speed measured in frames per second (FPS).

Table 1: Quantitative comparisons of 2D visual quality and 3D motion generation against baselines. The upper part reports on the EmbodiedHead speaking test set to evaluate standard speaking ability; BA is inapplicable to FaceFormer and CodeTalker because they do not predict head motion. The lower part reports on the public DualTalk test set to evaluate interactive listening-speaking ability; image-domain metrics are unavailable because the DualTalk test set does not provide ground-truth RGB frames.

Method	LVE↓	FDD↓	MOD↓	BA↑	SID↑	PSNR↑	SSIM↑	LPIPS↓
FaceFormer [10]	11.02	3.46	3.84	–	1.21	17.64	0.605	0.203
CodeTalker [37]	10.99	3.09	3.89	–	1.32	17.68	0.607	0.202
DiffPoseTalk [32]	10.3	2.69	4.24	0.44	2.01	16.09	0.565	0.268
ARTalk [5]	6.86	2.42	3.21	0.46	2.39	16.81	0.579	0.221
Ours (Stage1)	**5.70**	1.66	**2.98**	**0.47**	**2.64**	17.53	0.600	0.199
Ours (Stage2)	5.71	**1.58**	3.04	0.46	**2.64**	**18.28**	**0.617**	**0.184**
DualTalk [26]	7.94	1.95	2.84	**0.42**	2.47	–	–	–
UniLS [7]	12.73	3.04	4.83	0.36	1.57	–	–	–
TIMAR [2]	8.02	1.94	2.98	0.41	2.57	–	–	–
Ours	**7.18**	**1.74**	**2.57**	0.39	**2.67**	–	–	–

2D Visual Quality Evaluation. A practical head-embodied LLM avatar must deliver high rendered visual quality to users. We evaluate the end-to-end rendered output by routing all methods through the identical render pipeline and comparing PSNR, SSIM, and LPIPS on the EmbodiedHead speaking test set (Tab. 1, upper). As shown, EmbodiedHead achieves the best performance across all metrics, validating the efficacy of our DiT and two-stage training paradigm. Unlike previous methods reliant on noise-prone coefficient-space supervision, our Rectified Flow formulation enables direct one-step endpoint generation ($\hat{\mathbf{x}}_1$). This property seamlessly facilitates joint fine-tuning of the DiT and renderer in the second stage, effectively mitigating tracking biases, leading to superior 2D visual outcomes.

3D Motion Evaluation. To assess motion fidelity, we evaluate the models under two distinct datasets: a standard speaking mode on the EmbodiedHead speaking test set, and an interactive listening-speaking mode on the DualTalk test set. Quantitative comparisons of 3D motion generation performance are summarized in Tab. 1. In the standard speaking mode, EmbodiedHead significantly outperforms previous baselines, achieving state-of-the-art results across all evaluated metrics. The substantial improvements in LVE and FDD indicate that our model captures highly accurate lip articulations and natural upper-face dynamics to support realistic rendering. In the interactive listening-speaking mode, we train on the DualTalk train split and evaluate on its test set. Despite relying on a single audio stream with explicit LS-state conditioning (Sec. 3.3), EmbodiedHead surpasses the dual-audio-driven baseline methods in LVE, FDD, MOD, and SID. Although DualTalk and TIMAR attain slightly higher BA scores

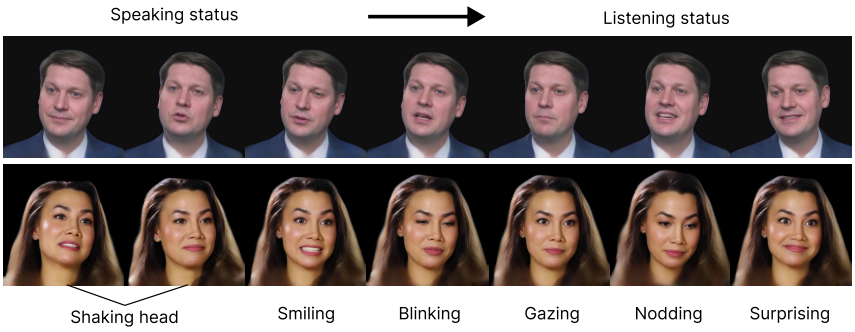

Fig. 3: Qualitative examples of natural listening-speaking transitions and conversational behaviors.

(0.42 and 0.41 vs. 0.39), this likely stems from their use of interlocutor visual cues to better capture conversational rhythm.

Real-Time Performance. Real-time generation is a prerequisite for deploying a head-embodied LLM avatar in live conversation. Our full pipeline runs in real time with GAGAvatar as the renderer, achieving 59 FPS on a single NVIDIA RTX 3090 GPU. When measuring only the motion-coefficient generation module, our method reaches over 900 FPS, largely enabled by the efficiency of Rectified Flow. Compared with other methods, DiffPoseTalk is constrained by its DDPM architecture and cannot achieve real-time performance. Although some baselines can reach real-time speeds at the motion-coefficient level, none of them explores end-to-end image generation, so their reported efficiency does not reflect the actual speed of producing the final rendered output.

Table 2: User study results. Each value reports the percentage of all pairwise comparisons in which participants preferred our method over the corresponding baseline, with ties included.

Comparison	Overall Realism	Lip-Sync Accuracy	Listening Naturalness	Turn-Taking Smoothness
Ours vs. DualTalk [26]	89.1%	81.8%	85.5%	85.0%
Ours vs. TIMAR [2]	91.8%	90.9%	86.8%	87.7%
Ours vs. UniLS [7]	76.8%	72.7%	69.5%	68.6%

4.3 Qualitative Evaluation

To demonstrate our interactive capabilities, Fig. 3 visualizes the natural listening-speaking transitions and conversational behaviors generated by EmbodiedHead.

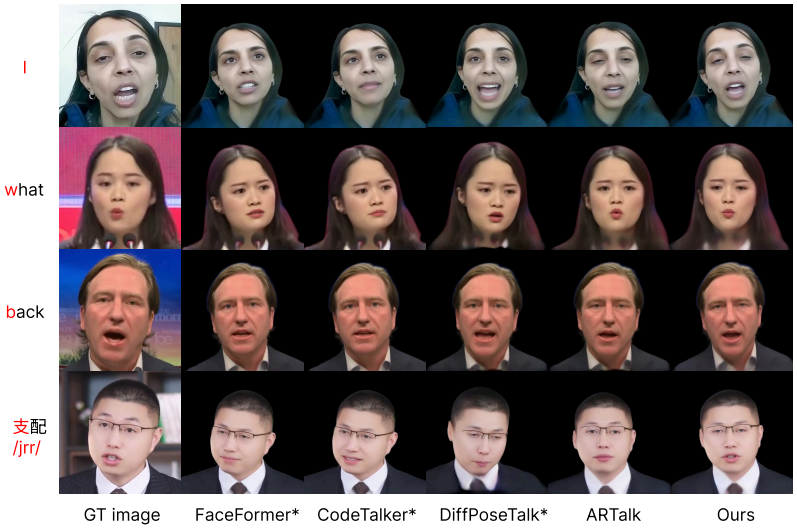

Fig. 4: Qualitative comparisons of 2D visual quality against other baseline methods on the EmbodiedHead speaking test set. * indicates baselines are additionally rendered.

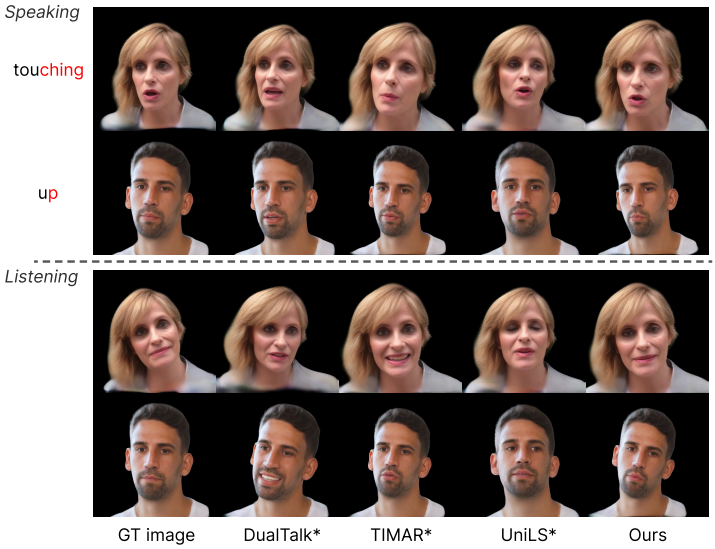

Fig. 5: Qualitative comparisons of 2D visual quality against interactive baseline methods on the DualTalk test set, covering both speaking (upper) and listening (lower) statuses. * indicates baselines are additionally rendered.

Modulated by the explicit LS-state, our model achieves seamless state switching without abrupt visual artifacts. Beyond accurate lip synchronization, the rendered 2D avatars exhibit rich, context-aware non-verbal dynamics, such as

Table 3: Ablations on global and listening modules. Upper: global condition ablation; lower: listening-speaking module ablation. *mag*, *ref*, and *audio* denote motion magnitude, reference embedding, and audio embedding, respectively.

Method	LVE↓	FDD↓	MOD↓	BA↑
Baseline	5.97	1.70	2.89	0.41
+mag	6.20	1.58	2.96	0.40
+mag+ref	5.84	1.63	2.83	0.42
+mag+ref+audio	**5.76**	**1.55**	**2.63**	**0.43**
Baseline	6.58	1.53	2.43	0.37
+LS-state	6.21	**1.46**	2.10	0.36
+LS-state+FiLM	**6.13**	**1.46**	**1.99**	**0.38**

attentive nodding and blinking during listening, alongside vivid facial expressions during speaking. These high-fidelity interactive behaviors significantly enhance the realism and engagement of the conversational agent.

In standard speaking scenarios (Fig. 4), FaceFormer, CodeTalker, and ARTalk often yield restricted mouth amplitudes, failing to fully articulate phonetic details. Although DiffPoseTalk exhibits larger motion amplitudes, it still suffers from noticeable lip-audio desynchronization. In conversational settings (Fig. 5), despite producing rich dynamics, interactive baselines occasionally hallucinate unreasonable mouth openings during the "listening" phase. By contrast, equipped with an explicit state modulation mechanism, EmbodiedHead effectively suppresses unnatural mouth hallucinations when listening, while ensuring highly accurate and expressive lip synchronization when speaking.

We conducted a randomized pairwise preference study to evaluate the perceptual quality of the generated conversational avatars. Our user study focuses on interactive listening-speaking baselines, including DualTalk [26], TIMAR [2], and UniLS [7], since these methods generate interactive behaviors whose naturalness is best assessed by human judgment. The study involved 22 voluntary participants, each evaluating 48 pairwise comparisons across speaking and listening scenarios using four criteria: Overall Realism, Lip-Sync Accuracy, Listening Naturalness, and Turn-Taking Smoothness. For each comparison, participants selected the better result or indicated a tie. As shown in Tab. 2, our method is consistently preferred over all baselines across the four criteria, with preference rates of at least 68.6%, demonstrating its perceptual advantages in conversational listening-speaking interaction. Further details are provided in the supplementary material.

4.4 Ablation Study

Effect of the Global Condition. We validate the global condition module in the top half of Tab. 3. The baseline only uses the timestep condition. Progressively adding reference and audio embeddings to AdaLN steadily improves

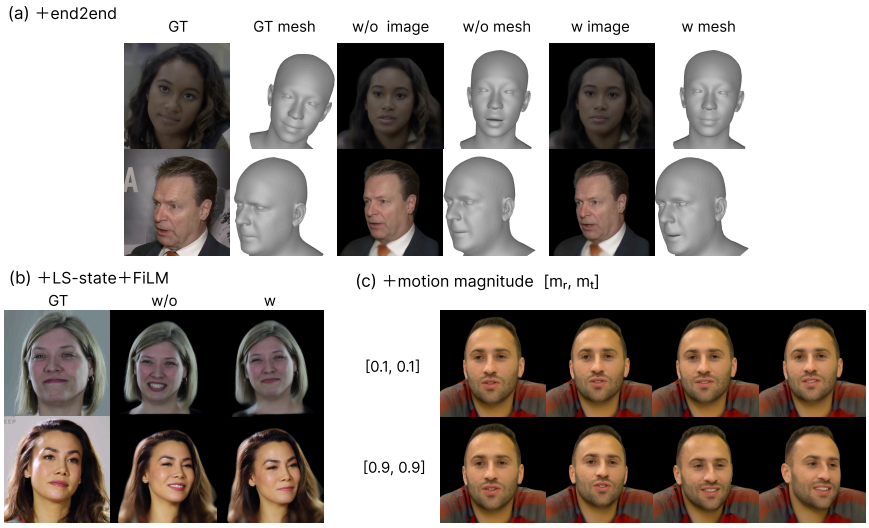

Fig. 6: Visual results of ablation studies. (a) The second-stage image-domain supervision improves both rendered quality and mesh geometry. (b) LS-state with FiLM modulation suppresses unnatural mouth openings during listening. (c) The motion magnitude condition controls head-motion amplitude under the same audio input.

Table 4: Ablation on inference-step scheduling.

Step	LVE↓	FDD↓	MOD↓	BA↑	PSNR↑	SSIM↑	LPIPS↓
1	5.80	1.76	2.73	**0.48**	17.08	0.571	0.204
4	**5.76**	1.55	**2.63**	0.43	**17.29**	**0.578**	**0.202**
10	5.99	1.46	2.67	0.43	17.13	0.573	0.206
25	6.18	**1.44**	2.73	0.41	17.10	0.572	0.208

all metrics. The motion magnitude condition is defined as $\mathbf{m} = [m_r, m_t]$, where m_r and m_t denote the normalized rotation and translation amplitudes, respectively. As shown in Fig. 6(c), small motion magnitude values (*e.g.*, 0.1) yield restrained head motions while large values (*e.g.*, 0.9) produce markedly more expressive movements, all from the same audio input. This continuous controllability demonstrates effective disentanglement of motion amplitude from the driving signal, enhancing both diversity and user-level customizability.

Effect of the Listening-Speaking Module. We assess explicit LS-state injection in the bottom half of Tab. 3, which is the component most directly tied to our unified listening-speaking objective. When LS-state conditioning is removed, the model attempts to infer conversational states directly from the audio signal. As illustrated in Fig. 6(b), it tends to interpret weak or distant speech as a listen-

ing state, and under uncertain conditions may open its mouth without producing clear speaking motions. Adding the state to the model input reduces MOD from 2.43 to 2.10, and further applying FiLM modulation to cross-attention brings it to 1.99, effectively eliminating spurious articulatory artifacts during listening and enabling reliable state-dependent behavior.

Effect of the Two-Stage Training. We evaluate the impact of the second-stage image-domain supervision. As shown in Tab. 1(upper) and Fig. 6(a), this strategy significantly enhances 2D visual fidelity. Furthermore, it effectively mitigates inherent tracking noise. For instance, as depicted in the second row of Fig. 6(a), our optimized geometry successfully captures subtle, natural mouth openings, visually surpassing the tracked pseudo-GT FLAME mesh. This corroborates our hypothesis that 2D supervision can effectively rectify 3D tracker biases.

Effect of the Inference Step. Our inference step analysis (Tab. 4) demonstrates that the proposed architecture enables high-quality one-step generation. Specifically, 1-step inference yields performance highly comparable to the 25-step setting. For additional details and ablation results, please refer to the supplementary material.

5 Conclusion

We presented EmbodiedHead, an end-to-end speech-driven talking-head framework for head-embodied LLM avatars. By coupling a Rectified-Flow DiT with a differentiable renderer, our pipeline achieves 59 FPS on a single GPU with only four sampling steps, satisfying the real-time requirement for live conversation. Explicit per-frame listening-speaking state conditioning and a Streaming Audio Scheduler replace the dual-stream paradigm, enabling unified listening and speaking behavior without interlocutor look-ahead. A two-stage training strategy that augments coefficient-space flow matching with image-domain refinement further bridges the gap between mesh-level accuracy and rendered visual fidelity. Experiments on both our curated dataset and public benchmarks validate state-of-the-art performance across motion, rendering, and interaction metrics.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62572194). This article was also supported by the Key Technology Research and Development Program Project of the Shanghai Science and Technology Commission under Grants (25511107200).

References

1. Aneja, S., Sevastopolsky, A., Kirschstein, T., Thies, J., Dai, A., Nießner, M.: Gaussianspeech: Audio-driven gaussian avatars (2024), https://arxiv.org/abs/2411.18675
2. Chen, J., Wang, F., Huang, Z., Zhou, Q., Li, K., Guo, D., Zhang, L., Yang, X.: Towards Seamless Interaction: Causal Turn-Level Modeling of Interactive 3D Conversational Head Dynamics (2025). https://doi.org/10.48550/arXiv.2512.15340
3. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023), https://arxiv.org/abs/2310.00426
4. Chen, S., Huang, H., Liu, Y., Ye, Z., Chen, P., Zhu, C., Guan, M., Wang, R., Chen, J., Li, G., et al.: Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis. arXiv preprint arXiv:2508.13618 (2025)
5. Chu, X., Goswami, N., Cui, Z., Wang, H., Harada, T.: ARTalk: Speech-Driven 3D Head Animation via Autoregressive Model (2025). https://doi.org/10.48550/arXiv.2502.20323
6. Chu, X., Harada, T.: Generalizable and animatable gaussian head avatar. Advances in Neural Information Processing Systems **37**, 57642–57670 (2024)
7. Chu, X., Liu, R., Huang, Y., Liu, Y., Peng, Y., Zheng, B.: UniLS: End-to-End Audio-Driven Avatars for Unified Listening and Speaking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25142–25152 (Jun 2026)
8. 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 (CVPR) (June 2019)
9. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21086–21095 (June 2025)
10. 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 (CVPR). pp. 18770–18780 (June 2022)
11. Geng, S., Teotia, R., Tendulkar, P., Menon, S., Vondrick, C.: Affective faces for goal-driven dyadic communication (2023), https://arxiv.org/abs/2301.10939
12. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5784–5794 (October 2021)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
14. 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). https://doi.org/10.1109/TASLP.2021.3122291
15. Jiang, J., Liang, C., Yang, J., Lin, G., Zhong, T., Zheng, Y.: Loopy: Taming Audio-Driven Portrait Avatar with Long-Term Motion Dependency. In: The Thirteenth International Conference on Learning Representations (2025)

16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
17. Li, J., Zhang, J., Bai, X., Zhou, J., Gu, L.: Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7568–7578 (2023)
18. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813
19. Lin, Y., Fan, Z., Wu, X., Xiong, L., Li, X., Kang, W., Peng, L., Lei, S., Xu, H.: Glditalker: Speech-driven 3d facial animation with graph latent diffusion transformer. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 1548–1556. International Joint Conferences on Artificial Intelligence Organization (8 2025). https://doi.org/10.24963/ijcai.2025/173, main Track
20. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023), https://arxiv.org/abs/2210.02747
21. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow (2022), https://arxiv.org/abs/2209.03003
22. Mennecke, B.E., Triplett, J.L., Hassall, L.M., Conde, Z.J.: Embodied social presence theory. In: 2010 43rd Hawaii International Conference on System Sciences. pp. 1–10 (2010). https://doi.org/10.1109/HICSS.2010.179
23. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis (2020), https://arxiv.org/abs/2003.08934
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 (CVPR). pp. 20395–20405 (June 2022)
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (October 2023)
26. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21055–21064 (2025)
27. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2018)
28. 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)
29. Richard, A., Zollhöfer, M., Wen, Y., et al.: 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)
30. Siniukov, M., Chang, D., Tran, M., Gong, H., Chaubey, A., Soleymani, M.: Ditailistener: Controllable high fidelity listener video generation with diffusion (2025), https://arxiv.org/abs/2504.04010

31. Siyao, L., Yu, W., Gu, T., et al.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11050–11059 (2022)
32. 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) (2024). https://doi.org/10.1145/3658221
33. 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. p. 3927–3935. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475587, https://doi.org/10.1145/3474085.3475587
34. Tran, M., Chang, D., Siniukov, M., Soleymani, M.: Dyadic interaction modeling for social behavior generation (2024), https://arxiv.org/abs/2403.09069
35. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
36. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: 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 (CVPR) Workshops. pp. 657–666 (June 2022)
37. 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 (CVPR). pp. 12780–12790 (June 2023)
38. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM computing surveys **56**(4), 1–39 (2023)
39. Zanon Boito, M., Iyer, V., Lagos, N., Besacier, L., Calapodescu, I.: mHuBERT-147: A Compact Multilingual HuBERT Model. In: Interspeech 2024. pp. 3939–3943 (2024). https://doi.org/10.21437/Interspeech.2024-938
40. Zhang, L., Cai, S., Li, M., Wetzstein, G., Agrawala, M.: Frame context packing and drift prevention in next-frame-prediction video diffusion models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
41. Zhang, R., Isola, P., Efros, A.A., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
42. Zhu, L., Lin, L., Zhu, Y., Wu, J., Hou, X., Li, Y., Liu, Y., Chen, J.: MANGO:Natural Multi-speaker 3D Talking Head Generation via 2D-Lifted Enhancement (2026). https://doi.org/10.48550/arXiv.2601.01749
43. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10667–10677 (June 2025)