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
[2] Garabido Shanghai Technology Co., Ltd., Shanghai, China
[*]Equal contribution. [†]Corresponding author.

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.

Project page: https://03skyboy.github.io/EmbodiedHead/

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 [24].

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:

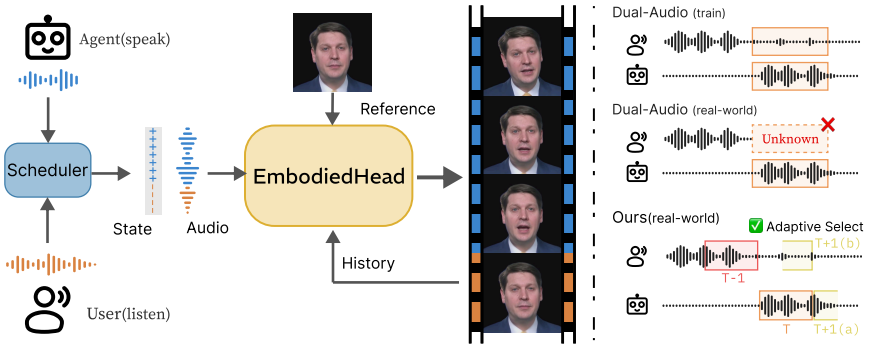

Fig. 1. We present EmbodiedHead, which generates a real-time head-embodied avatar for LLMs. Unlike dual-audio methods, it uses a single audio stream with explicit listening-speaking state conditioning to achieve unified conversational behavior.

(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 [16,10], 3D-based methods generally enable lower inference cost, making them a promising route toward Head-Embodied LLMs.

Most current listening-speaking integrated methods [28,3,8,45] 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 (figure 1). Our pipeline couples the first Rectified-Flow [22] Diffusion Transformer (DiT) [27,4] 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 [29], giving the model a clear mode signal at every frame. A *Streaming Audio Sched-*

uler 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 through 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 [9,11,40,35]. VOCA [9] established identity-decoupled regression from speech. FaceFormer [11] and CodeTalker [40] introduced Transformer modeling and discrete motion priors to improve lip sync and motion diversity. DiffPoseTalk [35] applied diffusion for style-conditioned generation, while ARTalk [6] and GLDiTalker [20] 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 [25,13,18] or 3D Gaussian Splatting [17,1]. 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 [16,10], 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 [26,28,3,8,45]. Learning to Listen [26] modeled stochastic 3D listener responses but did not address multi-turn dialogue. DualTalk [28] 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 [3] 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 [46] and UniLS [8] 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 [45] recognized the value of image-domain supervision and introduced 2D-lifted training alongside dual-stream audio. However, MANGO is built on DDPM [14], 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

figure 2 provides an overview of EmbodiedHead. After formalizing the task and introducing Rectified Flow (section 3.1), we detail the injection of multiple conditions into DiT blocks (section 3.2). To unify speaking and listening behaviors in conversational streams, we then propose a listening-speaking state (section 3.3), culminating in a two-stage training scheme for end-to-end image-domain refinement (section 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 [19] 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 $\tau \in \{1,\dots,T\}$, with a shared identity shape $\mathbf{s} \in \mathbb{R}^{300}$. The per-frame motion vector is defined as

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

We adopt Rectified Flow [22] 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

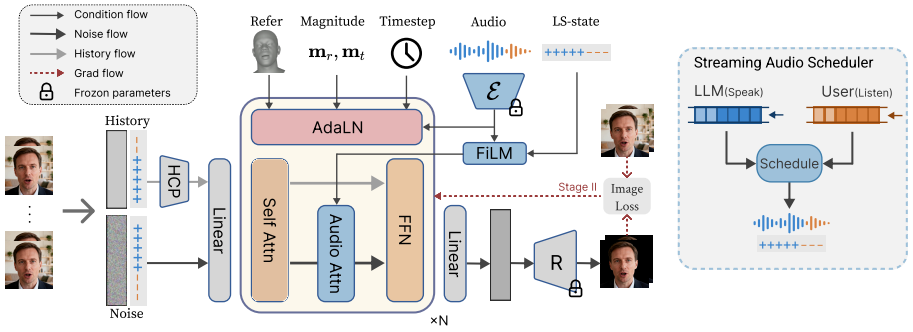

Fig. 2. EmbodiedHead employs a Rectified-Flow DiT to generate speech-driven talking-head animation in few steps. It conditions on reference, timestep, motion magnitude, and LS-state. A streaming scheduler merges user–LLM audio, enabling unified listening-speaking.

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 [21]

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

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 (section 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 [43], we propose History Context Packing (HCP) to efficiently encode long-range context: $\tilde{\mathbf{x}}^{\mathrm{hist}}$ 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 [15,42] 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 [29] 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 $|i - j| \leq R$, and $M_{i,j} = -\infty$ otherwise, where R=2 frames.

Global conditioning via AdaLN Following mainstream DiT paradigms [27,4], we introduce global conditioning through AdaLN. 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 [35,6]. 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 section 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 and the LLM stream to the same conditioning window and outputs the aligned LS-state.

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 $(\mathbf{a}, \mathbf{q})$

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 [36] 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 section 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

FLAME pseudo-labels from monocular tracking are noisy and biased, especially in frames with occlusion, motion blur, extreme pose, or poor illumination, 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 [7] as the differentiable renderer. While it renders fewer facial details than high-fidelity avatar pipelines [30], its single-image avatar construction makes end-to-end finetuning 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 GAGAvatar 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 VGG-based perceptual feature extractor [44,32].

4 Experiments

4.1 Experiments 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

Table 1. Quantitative comparisons of 2D visual quality and 3D motion generation against baseline methods. The upper metrics are reported on EmbodiedHead speaking test set and lower metrics are reported on the public DualTalk test set. All methods evaluated with image-domain metrics are rendered using the GAGAvatar [7].

Method	LVE↓	FDD↓	MOD↓	BA↑	SID↑	PSNR↑	SSIM↑	LPIPS↓
DiffPoseTalk [35]	10.3	2.69	4.24	0.44	2.01	16.09	0.565	0.268
ARTalk [6]	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 [28]	7.94	1.95	2.84	**0.42**	2.47	–	–	–
Ours	**7.18**	**1.74**	**2.57**	0.39	**2.67**	–	–	–

of footage. The primary data source is filtered from the TFHP dataset [35], contributing 14.7 hours. To enhance linguistic and environmental diversity, we integrated 7.8 hours of data from the TalkVid [5] and VFHQ [39] datasets. We additionally curated 3.9 hours of segments from the RealTalk dataset [12] to specifically model listening behaviors. We extract 3D FLAME parameters from videos using the GAGAvatar [7] 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 use mHuBERT-147 [15,42] model for superior multilingual adaptation. The motion branch predicts 115-dimensional FLAME parameters (including expressions, jaw poses, eye poses, global rotation and translation) over a 100-frame window. Besides, 75 frames of historical motion are divided into 4 groups and compressed into 20 frames in total.

We train the model using a two-stage strategy on two NVIDIA RTX 3090 GPUs. In the first stage, we train for 80,000 iterations with a learning rate of 6e-4 and a total batch size of 64. In the second stage, we train for 150,000 iterations with a learning rate of 8e-5 and a total batch size of 4.

During inference, we use Euler integration with 4 sampling steps, and fix the motion-magnitude guidance to 0.3; unless otherwise stated, all experimental results are reported under this setting. For more implementation details, please refer to the supplementary materials.

Baselines To comprehensively evaluate EmbodiedHead, we select representative state-of-the-art baselines across two distinct tracks. For standard speech-driven motion generation, we compare our method with DiffPoseTalk [35] and ARTalk [6]. To explicitly assess interactive conversational dynamics (i.e., unified listening and speaking), we benchmark against DualTalk [28], a recent dual-audio-driven framework.

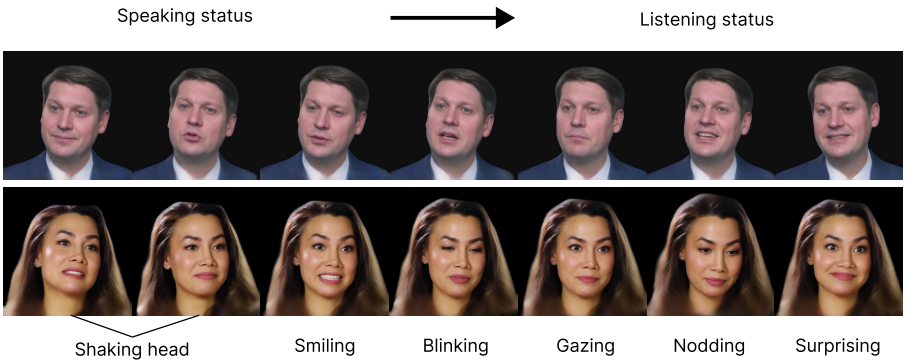

Fig. 3. Qualitative Examples of Natural Listening-Speaking Transitions and Conversational Behaviors

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 [38], and LPIPS [44], which directly reflect the rendered fidelity perceived by users. For *motion fidelity and listening-speaking behavior*, we use LVE [31] and MOD [40] for lip synchronization, FDD [40] for upper-face naturalness, BA [33] for head-motion rhythmic synchronization, and SID [28] for generation diversity. *Real-time efficiency* is evaluated via end-to-end throughput (FPS).

2D Visual Quality Evaluation A practical head-embodied LLM avatar must deliver high rendered visual quality to users, not merely accurate 3D coefficients. We therefore evaluate the end-to-end rendered output by routing all methods through the identical GAGAvatar [7] pipeline and comparing PSNR, SSIM, and LPIPS on our speaking test set (table 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 our speaking test set, and an interactive listening-speaking mode on the DualTalk test set. Quantitative comparisons of 3D motion generation performance are summarized in table 1.

In the standard speaking mode, EmbodiedHead significantly outperforms previous baselines, achieving state-of-the-art results across all evaluated metrics.

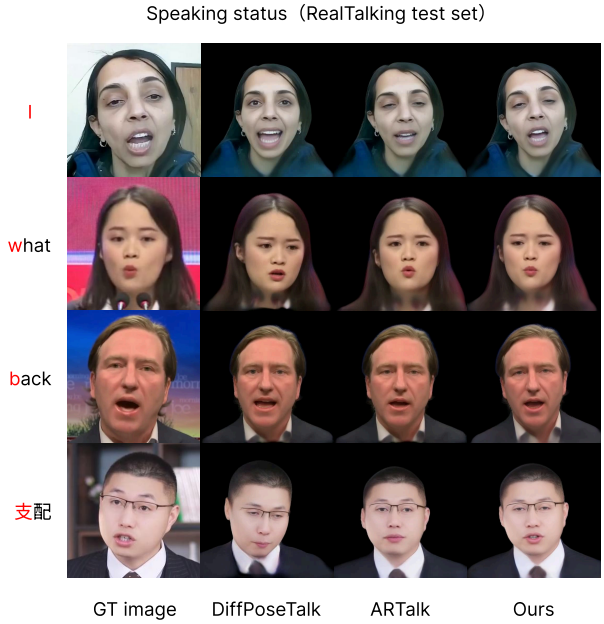

Fig. 4. Qualitative comparisons of 2D visual quality against other baseline methods on our EmbodiedHead test set.

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 (section 3.3), EmbodiedHead surpasses the dual-audio-driven DualTalk in LVE, FDD, and MOD. Although DualTalk attains a slightly higher BA score (0.42 vs. 0.39), this likely stems from its use of inter-locutor 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. If we measure 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 ARTalk and DualTalk can reach real-time speeds at the motion-coefficient level, neither ex-

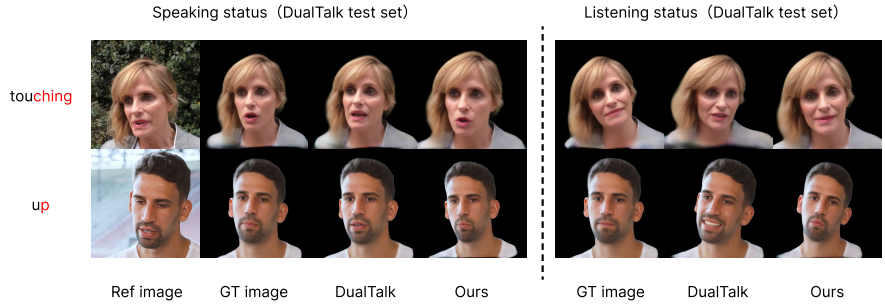

Fig. 5. Qualitative comparisons of 2D visual quality against DualTalk in the listening-speaking scenario on the public DualTalk test set.

plores end-to-end image generation, so their reported throughput does not reflect the actual speed of producing the final rendered output.

4.3 Qualitative Evaluation

To demonstrate our interactive capabilities, figure 3 visualizes the natural listening-speaking transitions and conversational behaviors generated by EmbodiedHead. 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 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 (figure 4), we found that 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 (figure 5), despite producing rich dynamics, DualTalk occasionally hallucinates unreasonable mouth openings during the "listening" phase. By contrast, equipped with an explicit state modulation mechanism, EmbodiedHead completely suppresses unnatural mouth hallucinations when listening, while ensuring highly accurate and expressive lip synchronization when speaking.

4.4 Ablation Study

Effect of the Global Condition We validate the global condition module in the top half of table 2. Baseline only has timestep condition. Progressively adding reference and audio embeddings to AdaLN steadily improves all metrics. And motion magnitude condition further offers intuitive amplitude control: as shown in figure 6(c), low values (e.g., 0.1) yield restrained head motions while high values (e.g., 0.9) produce markedly more expressive movements, all from the same

Table 2. Ablations on global and listening modules. Upper: Listening-Speaking test set; Lower: listening test set.

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
+input	6.21	**1.46**	2.10	0.36
+input+FiLM	**6.13**	**1.46**	**1.99**	**0.38**

Table 3. 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

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 table 2, and is the 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 figure 6(b), it tends to interpret weak or distant speech as a listening state, and under uncertain conditions may open its mouth without producing clear speaking motions. Adding the state to 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 table 1 and figure 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 figure 6(a), our optimized geometry successfully captures subtle, natural mouth openings, visually surpassing the pseudo-GT mesh. This corroborates our hypothesis that 2D supervision can effectively rectify 3D tracker biases.

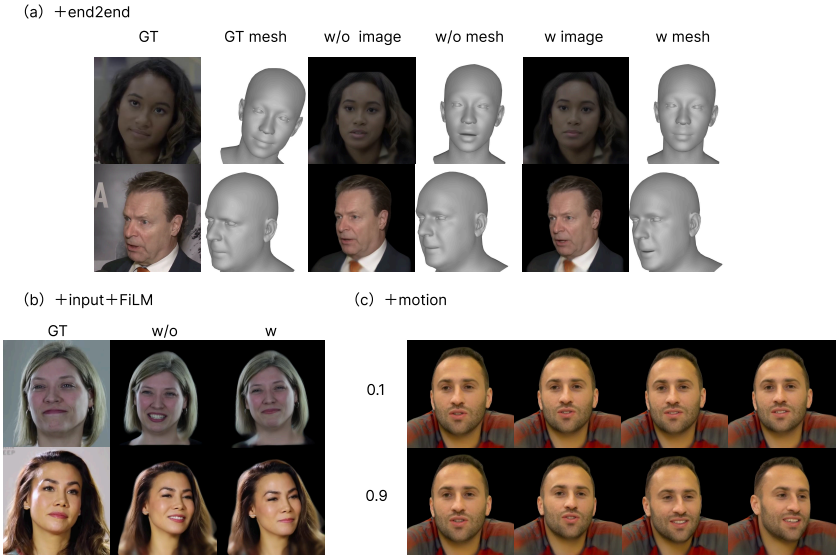

Fig. 6. Visual results of ablation study.

Effect of the Inference Step Our inference step analysis (table 3) 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 ablation results and complete table, 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.

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. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. 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
4. 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
5. 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)
6. 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
7. Chu, X., Harada, T.: Generalizable and animatable gaussian head avatar. Advances in Neural Information Processing Systems **37**, 57642–57670 (2024)
8. Chu, X., Liu, R., Huang, Y., Liu, Y., Peng, Y., Zheng, B.: UniLS: End-to-End Audio-Driven Avatars for Unified Listening and Speaking (2025). https://doi.org/10.48550/arXiv.2512.09327
9. 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)
10. 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 Diffusion Transformer Networks (2025). https://doi.org/10.48550/arXiv.2412.00733
11. 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)
12. Geng, S., Teotia, R., Tendulkar, P., Menon, S., Vondrick, C.: Affective faces for goal-driven dyadic communication (2023), https://arxiv.org/abs/2301.10939
13. 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)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
15. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021). https://doi.org/10.1109/TASLP.2021.3122291
16. 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 (2024)

17. 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/
18. 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)
19. 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
20. Lin, Y., Fan, Z., Wu, X., Xiong, L., Peng, L., Li, X., Kang, W., Lei, S., Xu, H.: Glditalker: Speech-driven 3d facial animation with graph latent diffusion transformer (2025), https://arxiv.org/abs/2408.01826
21. 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
22. 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
23. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (2017). https://doi.org/10.48550/arXiv.1711.05101
24. 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
25. 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
26. 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)
27. 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)
28. 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)
29. 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)
30. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians (2024). https://doi.org/10.48550/arXiv.2312.02069
31. 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)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015), https://arxiv.org/abs/1409.1556
33. 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)
34. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing (2024)

35. Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y.H., Yu, M., Liu, Y.J.: Diff-PoseTalk: Speech-Driven Stylistic 3D Facial Animation and Head Pose Generation via Diffusion Models (2024). https://doi.org/10.48550/arXiv.2310.00434
36. 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
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
38. 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)
39. 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)
40. 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)
41. 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)
42. 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
43. 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)
44. 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)
45. 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
46. 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)

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

Supplementary Material

This supplementary material provides additional details to support the main manuscript. section A details the model architecture. section B formalizes the evaluation metrics used in our experiments. section C presents additional experimental results, including evaluations on the VOCASET [9] dataset and additional ablation tables. Finally, section D discusses the limitations of our current framework and outlines directions for future research.

A Model Architecture

In the configuration used in our experiments, the Rectified-Flow [22] DiT [27,4] predicts FLAME [19] motion coefficients directly. Each target frame contains 100 expression coefficients together with jaw pose, eyes pose, global rotation, and global translation, giving a 115-dimensional motion vector in total. The model predicts a 100-frame target window at 25 FPS. We append a one-dimensional listening-speaking state to every motion frame before tokenization, so the per-token input dimension becomes 116. A linear projection maps each token to a 448-dimensional transformer [37] space. The DiT backbone uses 8 blocks, 8 attention heads, and an MLP ratio of 4, yielding a head dimension of 56 and a feed-forward hidden width of 1792.

For temporal context, the model consumes 75 historical frames in addition to the current noisy window. Following our History Context Packing design, these 75 frames are split into four groups with receptive spans of 5, 10, 20, and 40 frames, respectively. Each group is compressed into 5 tokens by an independent linear projector, producing 20 packed history tokens in total. These 20 tokens are concatenated with the 100 current-window tokens, so self-attention is performed over a sequence of 120 motion tokens. The reference condition is formed by concatenating the first 100 FLAME shape coefficients with one reference motion frame, resulting in a 215-dimensional vector.

We use frozen mHuBERT-147 [15,42] as the audio encoder. All hidden layers are fused by learnable softmax weights, with separate weight sets for the local cross-attention branch and the global conditioning branch. The extracted audio sequence is aligned to the motion rate by a strided 1D convolution with kernel size 5 and stride 2, followed by LayerNorm [2], producing 448-dimensional audio tokens. In cross-attention, we further apply a diagonal locality mask with radius 2 frames, so each motion token attends only to a narrow temporal neighborhood in the aligned audio sequence.

Global conditions are injected through AdaLN [41]. The conditioning vector has dimension 896 and is formed by concatenating four parts: a 448-dimensional

sinusoidal timestep embedding, a 112-dimensional reference embedding, a 224-dimensional mean-pooled global audio embedding, and a 112-dimensional motion-magnitude embedding derived from the normalized rotation and translation guidance scalars. Each DiT block consists of RoPE self-attention, RoPE audio cross-attention, and a feed-forward layer, with rotary base $\theta = 10000$ [34]. Before audio cross-attention, the aligned audio features are modulated by a block-specific FiLM [29] network conditioned on the listening-speaking state; this modulator uses a hidden size of 112 and zero-initialized output weights so that training starts from an identity mapping. The AdaLN modulation layers are also zero-initialized for stable optimization. Cross-attention is applied only to the 100 current-window tokens, while the packed history tokens contribute through self-attention only. Finally, a last AdaLN-modulated projection maps the final 100 tokens back to 115-dimensional velocity predictions.

B Experimental Setup and Metric Formalization

B.1 Training Hyperparameters and Loss Weights

Both training stages utilize the AdamW [23] optimizer. For the flow-matching objective (Equation (6) of the main manuscript), the parameter group weights are empirically set to: $\lambda_{\text{expr}} = 0.6$, $\lambda_{\text{jaw}} = 0.1$, $\lambda_{\text{eye}} = 0.1$, $\lambda_{\text{rot}} = 0.1$, and $\lambda_{\text{trans}} = 0.1$. The pose smoothing weight is set to $\lambda_{\text{smooth}} = 0.1$.

For Stage I, we employ a linear learning rate warmup for the first 2,000 iterations and train with a total batch size of 64. During the Stage II end-to-end refinement, we employ 1,000 warmup steps and render the images at a resolution of 256×256. The loss weights (Equation (8) of the main manuscript) are set as follows: the coefficient-space flow-matching weight is set to $\lambda_{\text{coef}} = 0.2$, while the image-domain L1 and perceptual loss weights are set to $\lambda_1 = 0.2$ and $\lambda_p = 0.2$, respectively. We also scale the learning rate of the GAGAvatar [7] renderer by a factor of 0.5 relative to the base learning rate. Due to the memory footprint of the differentiable renderer, the total batch size in Stage II is reduced to 4.

B.2 Metric Calculation Details

To comprehensively evaluate both the 3D motion fidelity and the 2D visual quality, our metrics (LVE [31], FDD [40], MOD [40], BA [33], SID [28], PSNR, SSIM [38], and LPIPS [44]) span geometric, temporal, and image domains. In alignment with established evaluation protocols, we formalize the metrics as follows.

3D Motion Metrics For 3D geometry-based metrics, the predicted FLAME [19] coefficients are mapped to mesh vertices $\mathbf{v}$ and 3D landmarks $\mathbf{p}$. To strictly evaluate local facial articulations, we isolate facial expressions by eliminating the influence of global head rotation across all localized geometric metrics (LVE [31], FDD [40], MOD [40]).

Lip Vertex Error (LVE ↓) LVE [31] evaluates temporal lip synchronization by computing the maximum L_2 error across all vertices in the lip region (V_{lip}) for each frame t, and averaging over the sequence of length T:

$$\text{LVE} = \frac{1}{T}\sum_{t=1}^{T}\max_{i\in V_{\text{lip}}}\left\|\mathbf{v}_{t,i}^{\text{pred}} - \mathbf{v}_{t,i}^{\text{gt}}\right\|_2 \tag{9}$$

where $\mathbf{v}_{t,i} \in \mathbb{R}^3$ denotes the 3D coordinate of vertex i at frame t.

Face Dynamics Deviation (FDD ↓) FDD [40] assesses the naturalness of upper-face motions. Let V_{upper} denote the subset of FLAME [19] mesh vertices corresponding to the eye and forehead regions. We first compute the relative vertex motion to the initial frame $\mathbf{m}_{t,i} = \mathbf{v}_{t,i} - \mathbf{v}_{1,i} \in \mathbb{R}^3$ to remove static offsets. We then measure the difference in the temporal standard deviation (σ_i) between the prediction and ground truth:

$$\sigma_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{m}_{t,i} - \bar{\mathbf{m}}_i\|_2^2} \tag{10}$$

where $\bar{\mathbf{m}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{m}_{t,i}$ is the temporal mean. The FDD is formulated as:

$$\text{FDD} = \frac{1}{|V_{\text{upper}}|}\sum_{i\in V_{\text{upper}}}\left|\sigma_i^{\text{pred}} - \sigma_i^{\text{gt}}\right| \tag{11}$$

Mouth Opening Difference (MOD ↓) MOD [40] focuses on the similarity of mouth opening amplitudes. Let $d_t = \|\mathbf{p}_{t,66} - \mathbf{p}_{t,62}\|_2$ be the distance between the upper (idx: 62) and lower (idx: 66) lip landmarks. MOD [40] is the mean absolute error of this distance:

$$\text{MOD} = \frac{1}{T}\sum_{t=1}^{T}\left|d_t^{\text{pred}} - d_t^{\text{gt}}\right| \tag{12}$$

Beat Alignment (BA ↑) BA [33] quantifies the rhythmic synchronization of head movements. We extract the angular velocity of head rotation and detect motion peaks (beats) B using a dynamic threshold. The alignment score is calculated symmetrically using a Gaussian kernel with $\sigma = 0.1$s:

$$\begin{aligned}\text{BA} = \frac{1}{2}\Bigg(&\frac{1}{|B^{\text{pred}}|}\sum_{b\in B^{\text{pred}}}\exp\Big(-\frac{\min_{b'\in B^{\text{gt}}}(b-b')^2}{2\sigma^2}\Big)\\ &+\frac{1}{|B^{\text{gt}}|}\sum_{b'\in B^{\text{gt}}}\exp\Big(-\frac{\min_{b\in B^{\text{pred}}}(b'-b)^2}{2\sigma^2}\Big)\Bigg)\end{aligned} \tag{13}$$

where $b, b' \in \mathbb{R}$ denote the timestamps of the detected beats in seconds.

Diversity Metric (SID ↑) To measure the generation diversity from the same audio input, we use the Speaker Identity Diversity (SID) [28]. We fit a K-Means clustering model on the ground truth feature distribution and predict the cluster assignments for the generated sequences. The diversity for a specific component k is defined as the Shannon entropy of its cluster assignment histogram:

$$\mathrm{SID}_k = -\sum_{c=1}^{K} p_c \log_2(p_c + \epsilon) \tag{14}$$

where $p_c = n_c/N$ is the probability of a sample falling into cluster c, and $\epsilon = 10^{-8}$ is a small constant for numerical stability. For a comprehensive assessment, we compute the mean of the SID across three semantic groups: jaw ($\mathrm{SID}_{\mathrm{jaw}}$), the first 50 dominant dimensions of expression ($\mathrm{SID}_{\mathrm{exp50}}$), and rotation ($\mathrm{SID}_{\mathrm{rot}}$):

$$\mathrm{SID} = \frac{1}{3}\left(\mathrm{SID}_{\mathrm{jaw}} + \mathrm{SID}_{\mathrm{exp50}} + \mathrm{SID}_{\mathrm{rot}}\right) \tag{15}$$

2D Metrics (PSNR ↑, SSIM ↑, LPIPS ↓) To ensure a fair evaluation of the final visual quality, we synthesize 2D image frames by rendering the FLAME-driven 3D mesh via the same GAGAvatar [7] pipeline. The rendered images are centrally cropped to a 256×256 region of interest (ROI) surrounding the face.

- **PSNR & SSIM:** We compute the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [38] frame-by-frame between the rendered outputs and the preprocessed ground-truth video, then average across the temporal dimension. SSIM [38] is evaluated using an 11×11 Gaussian window with $\sigma = 1.5$.
- **LPIPS (VGG):** We adopt the VGG-based LPIPS [44] implementation to measure perceptual similarity, reporting the average distance across all frames.

C Additional Experimental Results

C.1 Results on Public Dataset——VOCASET

To validate the generalization capability of our model, we evaluate our method on the public VOCASET [9] dataset. As shown in table S1, our method significantly outperforms existing baselines in both LVE [31] and FDD [40], demonstrating that our approach achieves highly accurate lip synchronization and natural upper-face dynamics even on unseen data domains. To accommodate its input requirements for dual audio streams and the interlocutor's mesh data, DualTalk [28] is evaluated under a single-speaker setting with a muted secondary speaker.

We observe a slight performance drop in the MOD [40]—which measures the Euclidean distance between two specific landmarks on the upper and lower lips—compared to DiffPoseTalk [35] and ARTalk [6]. This primarily stems from

the inherent trade-off between expressive generation capability and absolute distance-based metrics. Specifically, DiffPoseTalk [35] restricts jaw articulation to a single degree of freedom (1-DOF). Meanwhile, constrained by its autoregressive architecture, ARTalk [6] tends to produce over-smoothed motion patterns. In contrast, our model predicts fully expressive three-degree-of-freedom (3-DOF) jaw articulations ($\text{jaw}_{x,y,z}$). Although generating such rich and dynamic variations introduces a slight mathematical penalty in MOD [40], it significantly enhances the overall topological accuracy (as evidenced by our leading performance in LVE [31]) and the visual realism of the generated animations.

Furthermore, we do not report the BA [33] and SID [28] metrics in this evaluation. This is because the VOCASET [9] dataset provides vertex-only annotations, lacking the ground-truth global head rotation parameters strictly required for BA [33] computation. Additionally, since VOCASET [9] consists of tightly controlled read-aloud speech with inherently limited motion variance, diversity metrics like SID [28] are largely uninformative and hold little evaluation value on this dataset. Consequently, the generation diversity of our model is exclusively evaluated on the rich and highly dynamic EmbodiedHead dataset presented in the main manuscript.

Table S1. Quantitative comparison on the VOCASET benchmark. Our method achieves state-of-the-art performance in lip synchronization (LVE) and facial dynamics (FDD).

Method	**LVE ↓**	**FDD ↓**	**MOD ↓**
DualTalk [28]†	13.24	2.37	6.86
ARTalk [6]	9.78	2.29	5.20
DiffPoseTalk [35]	10.82	2.65	**4.99**
Ours	**7.97**	**2.18**	6.66

† *Evaluated under a single-speaker setting with a muted secondary speaker.*

C.2 Additional Quantitative Ablation Results

In this section, we provide more comprehensive quantitative ablation results to complement the main manuscript, specifically detailing the impact of inference-step scheduling and motion magnitude guidance.

Ablation on Inference-Step Scheduling. We first present the complete ablation study on the inference steps in table S2. The results demonstrate that our Rectified-Flow DiT architecture enables high-quality generation with extremely few steps. Specifically, 1-step inference yields performance highly comparable to the 25-step setting across both 3D motion and 2D image metrics. Weighing these findings, we ultimately select 4 steps as our default inference configuration to

achieve an optimal balance between real-time speed, visual quality, and motion fidelity.

To further validate the mathematical grounding of our Stage-II image-domain fine-tuning, we report a special "$1^{\dagger}$" setting. This setting denotes one-step generation using randomly sampled timesteps that are strictly consistent with the training phase distribution. The strong performance of this practical one-step output strongly supports the validity of our Stage-II optimization, rendering the end-to-end image-domain refinement both computationally feasible and highly efficient.

Table S2. Detailed ablation on inference-step scheduling. Best results are highlighted in **bold** except for $1^{\dagger}$. $1^{\dagger}$ denotes one-step generation with randomly sampled timesteps (training-consistent); results support the validity of Stage-II image-domain fine-tuning from practical one-step outputs.

Step	LVE↓	FDD↓	MOD↓	BA↑	PSNR↑	SSIM↑	LPIPS↓
1	5.80	1.76	2.73	**4.78**	17.08	0.571	0.204
$1^{\dagger}$	4.45	0.85	2.44	4.67	17.09	0.576	0.193
2	5.79	1.74	2.67	4.50	17.23	**0.578**	0.206
3	5.91	1.60	2.68	4.44	17.27	**0.578**	**0.202**
4	**5.76**	1.55	**2.63**	4.31	**17.29**	**0.578**	**0.202**
5	5.95	1.53	2.69	4.15	17.22	0.576	**0.202**
10	5.99	1.46	2.67	4.32	17.13	0.573	0.206
15	6.12	1.47	2.73	4.23	17.09	0.571	0.208
25	6.18	**1.44**	2.73	4.13	17.10	0.572	0.208

† *Not directly comparable to standard fixed-step inference settings.*

Ablation on Motion Magnitude. We quantitatively evaluate the explicit controllability introduced by our global condition module. As visualized in the main manuscript, modulating the motion magnitude guidance vector ($\mathbf{m} = [\mathbf{m}_{\mathrm{r}}, \mathbf{m}_{\mathrm{t}}]$) affords intuitive control over the kinematic intensity of the generated head dynamics. table S3 details how varying these conditional scalars influences the actual generated rotation and translation magnitudes. Notably, the results exhibit a degree of disentanglement: altering the rotation condition ($\mathbf{m}_{\mathrm{r}}$) primarily scales the rotational amplitude while leaving the translation magnitude largely stable, and vice versa. For reference, the original Ground Truth (GT) sequences exhibit a mean rotation magnitude of 0.61 and a translation magnitude of 0.81. This baseline serves as a standard anchor, demonstrating that our model can effectively control the motion intensity, capable of generating both subtle, restrained motions (when setting small condition values, e.g., 0.1) and highly expressive, exaggerated dynamics (when setting large condition values, e.g., 0.9).

Table S3. Quantitative ablation on motion magnitude guidance. We report the actual generated *Rotation* (Rot.) and *Translation* (Trans.) magnitudes under different combinations of input conditions $[\mathbf{m}_r, \mathbf{m}_t]$. The results demonstrate that our model not only effectively controls the magnitude but also partially disentangles rotation and translation.

Input Trans. ($\mathbf{m}_t$)	**Input Rot.** ($\mathbf{m}_r$)					
	0.1		**0.5**		**0.9**	
	Rot.	*Trans.*	*Rot.*	*Trans.*	*Rot.*	*Trans.*
0.1	0.50	0.88	1.05	0.95	1.77	1.07
0.5	0.55	1.41	1.03	1.56	1.64	1.72
0.9	0.76	2.57	1.10	3.05	1.31	2.63
GT	*Rot.*: 0.61			*Trans.*: 0.81		

D Limitations and Future Work

While EmbodiedHead achieves high-fidelity conversational dynamics, it presents two main limitations. First, the avatar's current listening behaviors are primarily driven by acoustic cues rather than semantic-level dialogue understanding, limiting its intent-driven expressiveness. Second, unlike causal autoregressive (AR) architectures, our diffusion-based framework inherently relies on window-based processing, which introduces a theoretical latency floor. Future work will focus on integrating textual semantics from LLMs to enable context-aware listening responses, as well as exploring hybrid AR-diffusion structures to further minimize streaming latency.