OMNIPERSONA: PERSONA-EMBODIED AGENTS FOR MULTIMODAL CONVERSATIONAL INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current interactive talking avatar models remain fundamentally unaware of dialogue context, requiring explicit textual or audio driving signals for each response, thereby functioning more like video renderers than autonomous avatar agents. Inspired by recent advances in unified understanding and generation frameworks, we present OmniPersona, an agentic model that integrates a pretrained multimodal large language model (MLLM) with an adapted joint audio-video generator. Our framework unifies dialogue understanding with synchronized audio-video avatar generation within a single architecture. Central to this design are learnable queries that compress historical multimodal context into a fixed-size non-verbal memory. Together with verbal content, this memory enables our generator to produce more responsive avatar behaviors, particularly in the listening state, which remains underexplored in prior work centered on talking scenarios. Additionally, we incorporate a suite of anti-drifting techniques, including a masked fusion attention design, to support persona consistency (audio-visual identity and synchronization) across multi-turn conversations. This attention design enables customized voice conditioning without compromising audio-video consistency. Extensive experiments on three datasets and human evaluation demonstrate that OmniPersona achieves superior performance.

1 INTRODUCTION

The emergence of talking avatars has opened new possibilities for human-agent interaction, spanning applications from virtual assistants to digital companions. Despite recent progress, current avatar systems (Huang et al., 2025b; Gan et al., 2025; Sun et al., 2026; Low et al., 2025) remain unaware of dialogue context: they treat conversation as a series of isolated lip-sync video generations, requiring explicit textual descriptions or pre-generated audio for each response rather than a continuous, context-aware dialogue. This limitation stems from an architectural disconnect: existing systems separate conversational understanding from visual generation, preventing avatars from naturally adapting their behavior based on accumulated interaction history and non-verbal cues.

Recent breakthroughs in unified understanding and generation, exemplified by Omni-Video (Tan et al., 2025) and UniVid (Luo et al., 2025b), demonstrate the promise of integrating multimodal comprehension with visual generative capabilities within a unified architecture. However, none of these approaches is designed for conversational interaction: which requires interactive (face-to-face dyadic communication), synchronized multimodal (text, audio, and video), and multi-turn identity-preserving (audio-visual persona) dialogue generation.

In this direction, OmniResponse (Luo et al., 2025a) takes an initial step by integrating visual and acoustic generation into an LLM pipeline via next-token prediction, enabling multimodal responses spanning text, audio, and video. However, its reliance on low-dimensional, animation-based visual representations and unsynchronized audio-video prediction substantially limits generation quality. SpeakerVid-5M (Zhang et al., 2026) adopts an MLLM as a backbone and injects textual embeddings as conditions for joint audio-video generation; yet it focuses exclusively on speaking portrait generation, without modeling listener behavioral dynamics or supporting multi-turn dialogue.

To this end, we introduce **OmniPersona**, a unified model that seamlessly combines a pretrained multimodal large language model (MLLM) with an adapted audio-video generative model to create identity-customizable avatars capable of context-aware, audio-synchronized, multi-turn interactions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

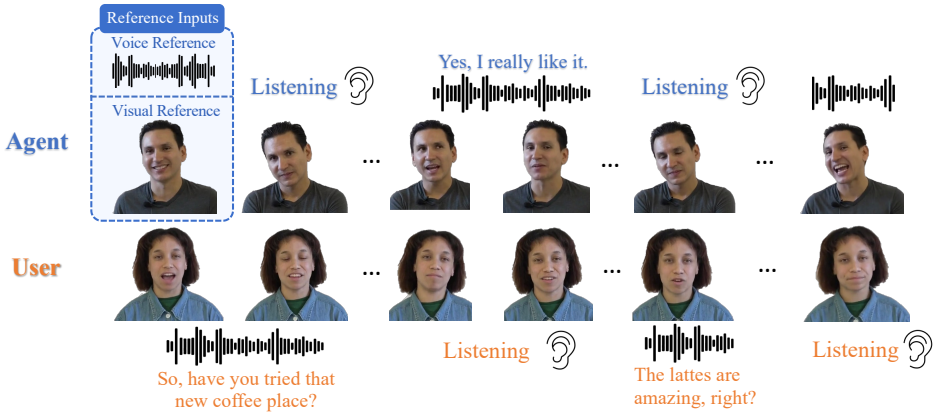


Figure 1: OmniPersona supports context-aware conversational interaction. Given accumulated dialogue history, it generates synchronized text, audio and video responses while preserving audio-visual identity across turns, enabling natural long-form human-agent interaction.

(Fig. 1; Tab. 1). OmniPersona is built around two core design challenges that prior work has left unresolved: *how* to model the speaking and listening behavioral dynamics throughout natural conversation, and *how* to sustain identity consistency while mitigating temporal drift across multi-turn auto-regressive generations.

Modeling such behavioral dynamics requires recognizing that speaking and listening are not mutually exclusive: a natural agent produces backchannels (e.g., "uh-huh", "mm") with corresponding facial cues while listening, and exhibits natural pauses and rhythmic modulation while speaking. OmniPersona addresses this via two parallel streams: a *verbal* stream that autoregressively generates transcriptions from prior conversational context, and a *non-verbal* stream that distills affective cues (e.g., tone, prosody, behavioral tendencies) into a compact memory condition via learnable queries. Both streams jointly condition a downstream audio-video diffusion model, synchronizing *what* is said and *how* it is naturally expressed.

A further challenge in auto-regressive diffusion generation is the drifting issue from teacher forcing (Chen et al., 2024; Huang et al., 2025a; Chen et al., 2025a; Yang et al., 2026b; Xiao et al., 2025; He et al., 2025). We address this through a suite of complementary anti-drifting techniques. First, *Diffusion Forcing* (Chen et al., 2024) is adopted to mitigate temporal drifting by imposing varying noise levels across sequential frames, reducing error propagation during auto-regressive synthesis. Second, to prevent mutual interference between the audio and video streams, we introduce a carefully designed *Masked Fusion Cross Attention* module within the cross-attention layers that connect audio and video modalities, enabling customized voice conditioning without compromising audio-video consistency. Third, *Strided RoPE* augments rotary position embeddings with a strided encoding scheme, sustaining long-range visual coherence. Together, these techniques enable stable, identity-preserving avatar generation over long conversational interactions.

Overall, our contributions can be summarized as follows:

- We introduce **OmniPersona**, a unified framework that integrates multimodal understanding with synchronized text, audio, and visual generation, enabling context-aware conversational interactions.

Table 1: Comparisons with recent methods. *Joint A&V*: joint audio and video generation. *Persona Control*: reference face and voice conditioning. LiveAvatar (speech-to-video) and OVI (text-to-audio-video) rely on an external MLLM for driving conditions.

Method	Multi-Turn	Context-aware	Joint A&V	Persona Control
OmniResponse (Luo et al., 2025a)	✓	✓	✗	Face only
SpeakerVid-5M (Zhang et al., 2026)	✗	✓	✓	Face only
LiveAvatar (Huang et al., 2025b)	✓	✗	✗	Face only
OVI (Low et al., 2025)	✗	✗	✓	Face only
OmniPersona	✓	✓	✓	Face + Voice

- We establish a bridge between an MLLM and an adapted audio-video diffusion generator via learnable queries, which compress historical context into a non-verbal memory condition. Combined with verbal response, this design significantly enhances the naturalness and coherence of generated content.

- We develop a suite of anti-drifting techniques to ensure long-term generation stability, including diffusion forcing and strided RoPE to alleviate visual drifting, and a masked fusion cross-attention design that enables customized voice conditioning without compromising audio-video consistency.

- Extensive experiments on two existing benchmarks, complemented by dedicated multi-turn evaluations on listening behavior accuracy and long-term temporal stability, demonstrate that OmniPersona achieves superior performance in generating contextually responsive, audio-video consistent avatars ranging from single-turn to multi-turn conversations.

2 Related Work

2.1 Unified Models for Understanding and Generation

Recent advances have converged on architectures that unify visual understanding and generation within single models. Show-o (Xie et al., 2025a), Show-o2 (Xie et al., 2025b) and BAGEL (Deng et al., 2025) demonstrate that autoregressive and diffusion objectives can coexist in a single model, scaling to both image and video modalities. BLIP3-o (Chen et al., 2025b) takes a complementary route by generating semantically rich CLIP features via flow-matching diffusion transformers rather than discrete token prediction. On the video side, UniVideo (Wei et al., 2026), Omni-Video (Tan et al., 2025), and Omni-Video 2 (Yang et al., 2026a) establish that MLLMs can serve as semantic controllers for diffusion-based video generation, while MetaQueries (Pan et al., 2025) offers a lightweight bridging mechanism that preserves a frozen MLLM's understanding capabilities. Our work extends this unified paradigm to multi-turn audio-video generation, where the model must condition on an evolving multimodal dialogue history.

2.2 From Talking to Agentic Avatar

Audio-driven talking avatar methods (Yang et al., 2025; Gao et al., 2025; Gan et al., 2025; Zhu et al., 2025; Huang et al., 2026) take a complete audio sequence as input and generate the full video non-causally, making them unsuitable for interactive scenarios. Streaming systems (Guo et al., 2025; Ki et al., 2026; Huang et al., 2025b; Sun et al., 2026) also rely on complete audio input but generate video frames causally; since speech synthesis runs at low latency, these systems can be composed with an external MLLM to support low-latency interaction. In both cases, however, the avatar module remains a passive renderer that depends entirely on upstream models for language and speech decisions, with no access to dialogue history.

This coupling bottleneck motivates agentic designs that integrate an MLLM into the generation pipeline, enabling contextual understanding and generation within a unified framework. Body of Her (Ao, 2024) extends an LLM to jointly model speech and upper-body motion via next-token prediction. OmniResponse (Luo et al., 2025a) introduces temporal alignment tokens to synchronize text and visual streams. SpeakerVid-5M (Zhang et al., 2026) pairs Qwen2.5-Omni (Xu et al., 2025) with a hybrid autoregressive-diffusion head for person-generic generation. X-Streamer (Xie et al., 2025c) separates conversational reasoning from video synthesis through a thinker-actor dual-transformer, sustaining consistent interactions over long sessions. MIDAS (Chen et al., 2025c) routes multimodal conditioning signals from a minimally modified LLM to a diffusion decoding head. Following this paradigm, our method conditions generation on the accumulated dialogue history to improve response accuracy while maintaining cross-turn audio-visual identity consistency and synchronization.

3 METHOD

3.1 TASK FORMULATION AND ARCHITECTURE OVERVIEW

Our goal is to automatically generate multi-round responses that are contextually coherent, cross-modally aligned, and consistent with the specified agent identity including voice and image reference. The conversational context is defined as

$$\mathcal{C} = \{(T_n^r, A_n^r, V_n^r)\}_{n=1}^{N}, \quad r \in \{\text{user}, \text{agent}\}, \tag{1}$$

where T_n^r, A_n^r, and V_n^r denote the text, audio, and video signals of the n-th turn for role r, respectively. The agent's identity is specified by visual and vocal references:

$$\mathcal{I} = (V^{\mathrm{R}}, A^{\mathrm{R}}), \tag{2}$$

where V^{R} comprises one reference image and A^{R} contains several seconds of reference speech. The reference audio is user-specified at inference time and can come from any speaker. The task is to learn a conditional generation function:

$$(\hat{T^S}, \hat{A}, \hat{V}) \sim p(T^S, A, V \mid \mathcal{C}, \mathcal{I}), \tag{3}$$

where $\hat{T^S}, \hat{A}, \hat{V}$ are the generated textual response, synthesized speech, and avatar video, respectively.

Fig. 2 illustrates the overall architecture of our model, which accepts multimodal inputs and produces joint audio-video responses alongside textual output. In the first turn, the MLLM generates a textual response T^S conditioned on the user's input alone; from the second turn onward, prior dialogue is encoded and compressed into a compact non-verbal memory M^n via learnable queries $\mathcal{Q}$ (Pan et al., 2025) and a Transformer-based Connector. Concatenating M^n with T5 embeddings E^{T5} (Raffel et al., 2020) of T^S forms the conditioning signal for the Diffusion Transformer (DiT (Peebles & Xie, 2023)), which further takes prefix frames V^P, noised video V^N, reference image V^R, prefix audio A^P, noised audio A^N, and reference voice A^R to synthesize synchronized audio-video responses that are appended to the dialogue history for coherent multi-turn interaction.

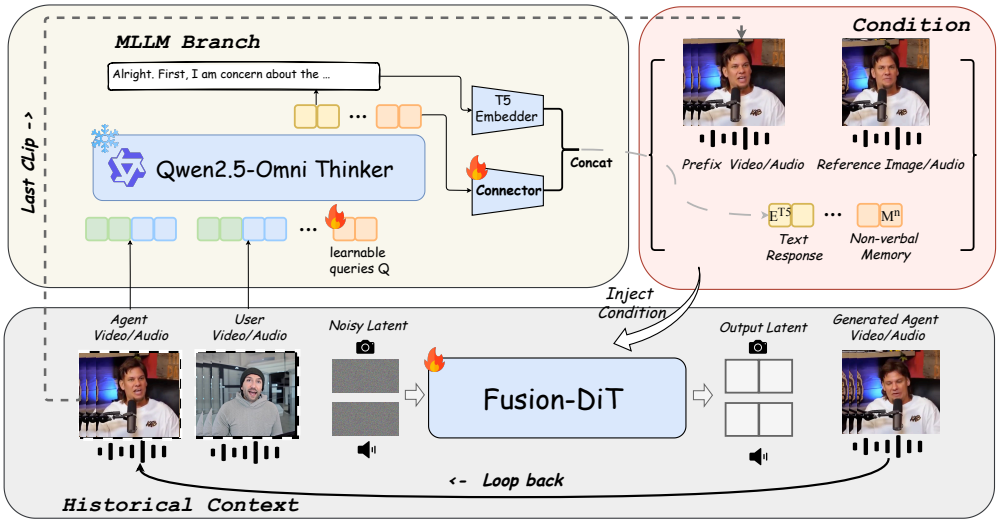


Figure 2: Overview of our OmniPersona framework for multi-turn multimodal response generation. OmniPersona consists of three main modules: an MLLM Branch to understand the historical multi-modal context and generate the text response T^S, a DiT-based diffusion generator for synchronized audio-video generation, and a connector to connect the MLLM and the diffusion model by compressing the long historical context into fixed-size non-verbal memory tokens M^n and mapping them into the video diffusion latent space. The generated video, audio, and text are autoregressively fed back into the historical context for subsequent turns.

Specifically, we instantiate the MLLM with Qwen2.5-Omni (Xu et al., 2025) for multimodal dialogue understanding and response T^S generation, and adopt OVI (Low et al., 2025) as the DiT-based backbone for joint audio–video synthesis. Sections 3.2–3.5 detail the connector, joint generation architecture, drifting mitigation and training objective, respectively.

3.2 Connecting MLLM with Diffusion Model

We build a connector that bridges the MLLM with the diffusion model by augmenting text conditioning with a non-verbal memory compressed from the multimodal dialogue context $\mathcal{C}$ (historical text, video, and audio from both user and agent) via learnable queries.

Specifically, learnable queries $\mathcal{Q} \in \mathbb{R}^{N_q \times d_m}$ serve as the compression bottleneck, where N_q controls the memory capacity and d_m is the hidden dimension of Qwen2.5-Omni. We concatenate $\mathcal{C}$ and $\mathcal{Q}$ and feed them into Qwen2.5-Omni thinker backbone to obtain contextualized query representations:

$$(\hat{\mathcal{C}}, \hat{\mathcal{Q}}) = \text{Qwen}(\mathcal{C}, \mathcal{Q}), \quad \hat{\mathcal{Q}} \in \mathbb{R}^{N_q \times d_m}. \tag{4}$$

Next, we discard $\hat{\mathcal{C}}$ and map $\hat{\mathcal{Q}}$ to the T5 hidden dimension d_{T5} with a zero-initialized projection layer and refine it with a K-layer Transformer encoder operating on the query sequence via self-attention:

$$M^n = \text{TransEnc}\Big(\text{Proj}_{\text{T5}}(\hat{\mathcal{Q}})\Big) \in \mathbb{R}^{N_q \times d_{\text{T5}}}. \tag{5}$$

The resulting M^n is concatenated with the T5 embeddings E^{T5} of the speech transcription T^S to form the augmented conditioning input for the diffusion generator:

$$\tilde{e} = [E^{\text{T5}};\ M^n], \tag{6}$$

where $[\cdot\,;\cdot]$ denotes sequence-dimension concatenation. This would be cross-attended by the noised latent with the prefix and reference latent in the later stage for denoising purposes.

3.3 Audio-Video Joint Generation

We build on OVI (Low et al., 2025), a pretrained DiT with a dual-branch architecture with bidirectional cross-modal fusion for synchronized audio-video generation (Figure 3). Each modality branch has N transformer blocks with self-attention, text cross-attention, fusion cross-attention, and FFN.

Specifically, each branch concatenates its latents along the sequence dimension: $[V^P, V^N, V^R]$ for video and $[A^P, A^N, A^R]$ for audio. Self-attention is then applied within each branch, where A^N attends to A^R for timbre conditioning and A^P for temporal continuity, while V^N attends to V^R for identity and V^P for frame consistency. Next, the latents cross-attend to the verbal and non-verbal condition embeddings $\tilde{e}$, after which the two branches interact via fusion cross-attention.

The core design is asymmetric fusion cross-attention (Figure 3, right), where each branch cross-attends to the other:

$$\text{Fusion}_{\text{vid}} = \text{CrossAttn}\big(Q = [V^P; V^N; V^R],\ KV = [A^P; A^N]\big), \tag{7}$$

$$\text{Fusion}_{\text{aud}} = \text{CrossAttn}\big(Q = [A^P; A^N],\ KV = [V^P; V^N; V^R]\big). \tag{8}$$

Crucially, A^R is excluded from all visual features in both fusion modules. This isolation serves a dual purpose: (1) preventing irrelevant acoustic content of A^R from leaking into the visual stream

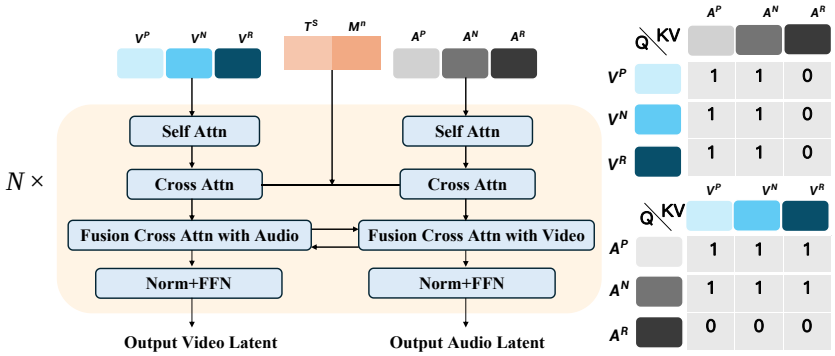


Figure 3: Proposed network architecture for joint audio-video generation with context modeling (left) and cross-modal attention maps (right). Superscripts P, N, R denote prefix, noised, and reference entities; A and V denote audio and video (see Sec. 3.1 for the detailed denotations).

and disrupting audio-video synchronization; (2) keeping the timbre representation encoded in A^R free from visual semantics, enabling controllable voice customization where the target voice and face originate from distinct sources (See Appendix A and **supplementary video** for more details).

3.4 Mitigating Video and Audio Drifting

3.4.1 Video Anti-drifting

To reduce the training–inference gap caused by autoregressive prefix conditioning in the video, we adopt Diffusion Forcing (Chen et al., 2024) and Strided RoPE in the video branch.

Diffusion Forcing. Specifically, the noised video latent V^N is partitioned into chunks of size c, resulting in C chunks $V^N = [V_1^N, \ldots, V_C^N]$, together with the prefix V^P, finally forms $\tilde{V} = [V^P, V_1^N, \ldots, V_C^N]$. Note that the reference latent V^R is kept clean throughout training to provide stable identity guidance. Each chunk V_i in $\tilde{V}$ is assigned an independently sampled diffusion timestep:

$$V_i = \alpha_{t_i} V_i + \sigma_{t_i} \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, I). \tag{9}$$

This chunk-wise noising prevents the model from relying on clean historical latents, encouraging it to extract only temporal continuity from the prefix while relying on the clean reference V^R for identity information, thereby reducing identity and color drift in multi-turn autoregressive generation.

Strided RoPE for Reference Anchoring. Since the reference image V^R is discontinuous from $\tilde{V}$, we offset its RoPE positional index (Su et al., 2024). If the sequence length of $\tilde{V}$ is L, the reference position is set to $p_R = L + \delta$, where $\delta \sim \mathcal{U}(20, 50)$. This offset prevents V^R from being interpreted as a neighboring frame, allowing it to act as a global identity anchor that future frames converge to while preserving temporal consistency. Appendix D.3 explains the choice of the training offset range and analyzes how different inference offsets affect identity consistency and pose variation across multiple turns.

3.4.2 Audio Anti-drifting

We empirically observe that audio drifting is less severe than video drifting and can be effectively mitigated with a simplified diffusion forcing strategy. During training, the noisy audio latent A^N is treated as a single chunk and noised with a diffusion timestep, while the prefix audio A^P is independently noised and the reference audio A^R remains clean. This training strategy improves robustness against noisy prefix audio and stabilizes multi-turn autoregressive audio generation.

In addition, we isolate the reference audio A^R from the visual stream in the fusion cross-attention as Sec. 3.3. Introducing A^R into audio-video fusion would couple speaker timbre with visual generation and easily lead to audio-video misalignment. In practice, even short reference audio A^R in $Fusion_{vid}$ can cause abnormal visual endings, which then propagate back through $Fusion_{aud}$ and corrupt the timbre representation. By restricting fusion to $[A^P, A^N]$, while keeping A^R only for audio self-attention, we prevent this feedback loop and stabilize both audio and video generation.

3.5 Training Objectives

We train the learnable queries and connector module with full parameters. To adapt the DiT backbone without disrupting its pre-trained priors, we apply LoRA (Hu et al., 2022) to all linear layers in its transformer blocks. The model is optimized with the standard diffusion objective defined as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0,\, \boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),\, t} \left[\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t,\, \mathcal{P},\, T^S,\, M^n,\, \mathcal{I}) \|_2^2 \right], \tag{10}$$

and the final loss is a weighted combination of the audio and video losses:

$$\mathcal{L}_{\text{diff}} = \lambda_A \mathcal{L}_{\text{diff}}^A + \lambda_V \mathcal{L}_{\text{diff}}^V, \tag{11}$$

where λ_A and λ_V are the audio and video loss weights. Here $\mathcal{P}$ is the audio-video prefix from the previous turn, T^S the speech transcript, M^n the non-verbal memory defined in Sec. 3.2, and $\mathcal{I}$ the reference image and voice; $\boldsymbol{\epsilon}_\theta$ is the LoRA-adapted diffusion backbone. Additional details are provided in Appendix B.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION

Training Dataset. We use Seamless Interaction dataset (Agrawal et al., 2025) for training, which is a large-scale dyadic face-to-face interaction video dataset collected in laboratory setting. It contains 4,000+ hours in-person human face-to-face interaction videos with 4,000+ participants, featuring naturalistic conversations between familiar pairs and professional actors. During training, we use all samples from the training set and crop them to participants' facial regions, resulting in videos with 448×448 resolution at 25 fps.

Evaluation Setup. We evaluate our method following the protocols of OmniResponse (Luo et al., 2025a), SpeakerVid-5M (Zhang et al., 2026), and our introduced benchmark based on the SeamlessInteraction (Agrawal et al., 2025). OmniResponse assesses text reply accuracy and serves as a multi-turn benchmark, yet lacks metrics for audio-video long-term consistency (e.g., identity preservation and voice similarity) as well as evaluation of listening-phase behavior. SpeakerVid-5M offers more comprehensive audio-video metrics, but is limited to short responder talking videos, making it insufficient for capturing the full dynamics of multi-turn interaction.

To address these gaps, we introduce an additional evaluation on the SeamlessInteraction dataset (Agrawal et al., 2025), comprising 60 randomly selected video clips from the test set spanning three interaction states: mainly speaking, mainly listening, and rapid speaking–listening alternation, each around 50 seconds long. We adopt five metrics: FVD (Unterthiner et al., 2018) to measure long-term temporal video quality, ArcFace (Deng et al., 2019a) to evaluate identity consistency across frames, SIM-o (Chen et al., 2025d) to assess voice similarity between generated and reference speech, FID_{emo} (Deng et al., 2019b) and FD (Zhou et al., 2022; Zhu et al., 2025) to quantify the naturalness and accuracy of facial expressions during the listening phase.

Inference Efficiency. On four H200 GPUs, OmniPersona generates a 5-second, 448×448 clip in 4.6 seconds using 10 diffusion steps, including MLLM encoding and VAE decoding. Timing comparisons and configurations are detailed in Appendix E.

4.2 QUANTITATIVE COMPARISONS

SpeakerVid-5M Benchmark. We follow the original evaluation protocol and compare against the SpeakerVid-5M baseline (Zhang et al., 2026) and a cascade pipeline *Qwen2.5-Omni* (Xu et al., 2025) $\rightarrow$ Text Prompt $\rightarrow$ OVI (Low et al., 2025). All three approaches pair the same MLLM with a text-to-audio-video generator, differing only in how the MLLM is bridged: non-verbal memory with text (ours), text hidden states, or text-only conditioning. Tab. 2 shows strong overall performance, with FVD reduced from 28.82/18.50 to 5.18 and improvements in expression quality, synchronization, and identity preservation. When all methods use the same SpeakerVid-5M response text, OmniPersona retains the best scores on all four evaluated generation metrics (Appendix D.2).

Table 2: Comparison of different methods on the SpeakerVid-5M dataset. The best results are highlighted in **bold**.

Method	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	ArcFace ↑	CLIP_{diag} ↑	Sync_{conf} ↑	FID_{emo} ↓	SIM-o ↑
SpeakerVid-5M	32.35	28.82	17.55	0.66	0.772	0.643	2.698	3.22	**0.65**
Qwen2.5-Omni + OVI	44.13	18.50	14.25	0.53	0.533	0.590	3.203	0.69	0.42
Ours	**20.66**	**5.18**	**17.65**	**0.68**	**0.774**	**0.764**	**3.935**	**0.20**	**0.65**

OmniResponse Benchmark. Following the OmniResponse (Luo et al., 2025a) protocol, we evaluate text with METEOR (Banerjee & Lavie, 2005), BERTScore_{F1} (Zhang et al., 2020), ROUGE-L (Lin, 2004), and Distinct-2 (Li et al., 2016); audio with UTMOSv2 (Baba et al., 2024) and LSE-D (Prajwal et al., 2020); video with FVD (Unterthiner et al., 2018). Qwen2.5-Omni + OVI is excluded, as OVI is limited to 10-second single-turn outputs while OmniResponse clips average $\sim$70 seconds, making fair comparison infeasible.

Tab. 3 shows improvements in text quality, acoustic naturalness, and video fidelity. LSE-D decreases to 5.59, below even Ground-Truth (8.96), whose audio-visual misalignment is illustrated in the supplementary video.

Table 3: Quantitative Results on OmniResponse (Luo et al., 2025a) benchmark. The best results are highlighted in **bold** and the second-best results are underlined.

Model	Text				Audio		Video
	METEOR ↑	BERTScore$_{F1}$ ↑	ROUGE-L ↑	Distinct-2 ↑	LSE-D ↓	UTMOSv2 ↑	FVD ↓
Ground-Truth	–	–	–	0.835	8.96	1.56	–
LSTM (Hochreiter & Schmidhuber, 1997)	0.042	0.716	0.000	0.000	9.72	1.21	320.92
Audio-Visual LLM (Abdin et al., 2024)	0.030	0.662	0.020	0.155	10.03	1.32	681.55
OmniResponse (Luo et al., 2025a)	0.141	0.806	**0.081**	**0.882**	9.56	1.41	314.94
OmniPersona (Ours)	**0.151**	**0.820**	**0.081**	0.833	**5.59**	**1.50**	**284.87**

Table 4: Comparison and ablation study on the SeamlessInteraction dataset for multi-turn interaction quality. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	FVD ↓	ArcFace ↑	SIM-o ↑	FID$_{emo}$ ↓	FD ↓	
					Exp	Pose
F5-TTS (Chen et al., 2025d)	–	–	0.53	–	–	–
LiveAvatar (Huang et al., 2025b)	12.55	**0.81**	–	0.55	18.46	0.06
OmniAvatar (Gan et al., 2025)	59.43	0.68	–	1.43	22.17	0.09
Ours	**11.93**	**0.81**	**0.66**	**0.31**	**14.28**	**0.04**
w/o DF	21.72	0.74	0.65	0.34	16.92	0.05
w/o FM	78.74	0.45	0.62	1.36	21.25	0.08
w/o MQ	13.44	0.77	**0.66**	0.62	19.04	0.07
w/o SR	12.04	0.73	0.65	0.36	16.39	0.05

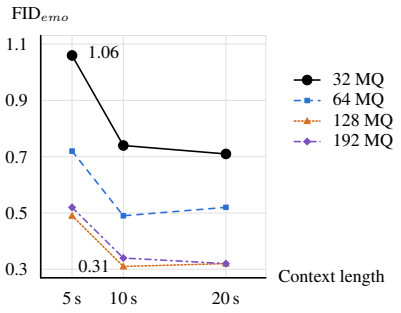


Figure 4: FID$_{emo}$ vs. MQ capacity and context length.

Evaluation on SeamlessInteraction Dataset. We evaluate multi-turn interaction quality on the SeamlessInteraction dataset. Since no existing model supports joint audio-video generation across turns, we assess video and audio separately. For video, we compare against LiveAvatar (Huang et al., 2025b) and OmniAvatar (Gan et al., 2025) on FVD, ArcFace, FID$_{emo}$ and FD; for audio, we compare against F5-TTS (Chen et al., 2025d) on SIM-o.

Table 4 shows the best overall performance in video fidelity and expression modeling, alongside strong visual identity and voice consistency. OmniPersona also outperforms LiveAvatar and w/o MQ on a 20-sample anger-dominant subset (Appendix D.4).

Ablation Study. Table 4 ablates diffusion forcing (DF), masked fusion (FM), memory queries (MQ), and strided RoPE (SR). Removing FM causes the largest degradation, raising FVD from 11.93 to 78.74. Removing MQ doubles FID$_{emo}$ from 0.31 to 0.62; complementary probes show that combining MQ with text improves emotion and arousal prediction over text alone (Appendix D.1). Removing SR weakens identity preservation, while removing DF degrades temporal quality. Figure 4 shows diminishing returns from increasing query capacity and context length; we use 128 queries and a 10s context for their balance of quality and computation.

Human Evaluation. We invite 20 participants to evaluate 20 pairs randomly sampled from the 60 pairs used in our SeamlessInteraction evaluation, comparing our method against OmniAvatar and LiveAvatar. Participants rate each video on a 1–5 scale (higher is better) across three dimensions following the Mean Opinion Scores (MOS) rating protocol: listening expression appropriateness (EA), talking audio-visual synchronization (AVS), and overall quality (OQ). As shown in Tab. 5, our method achieves superior scores across all evaluated dimensions (evaluation protocol in Appendix C).

Table 5: Mean Opinion Score (MOS) user study on SeamlessInteraction. Best in **bold**.

Method	EA ↑	AVS ↑	OQ ↑
OmniAvatar	2.76	3.86	2.55
LiveAvatar	3.62	4.31	3.74
Ours	**4.21**	**4.65**	**3.98**

4.3 Qualitative Comparison

Fig. 5 presents qualitative comparisons on the SpeakerVid-5M benchmark. The SpeakerVid-5M baseline produces blurry frames with artifacts around the mouth and eyes. Qwen2.5-Omni + OVI pipeline may introduce unintended scene transitions due to scene-level prompt generation. In contrast, our method maintains stable identity and preserves fine-grained facial shading across frames. Fig. 6 presents qualitative comparisons against baseline methods and ablation variants on SeamlessInteraction dataset, consistent with the quantitative results in Table 4. Comparisons with



Figure 5: Qualitative comparison on the SpeakerVid-5M benchmark. See the supplementary video for clearer visual and acoustic comparisons.

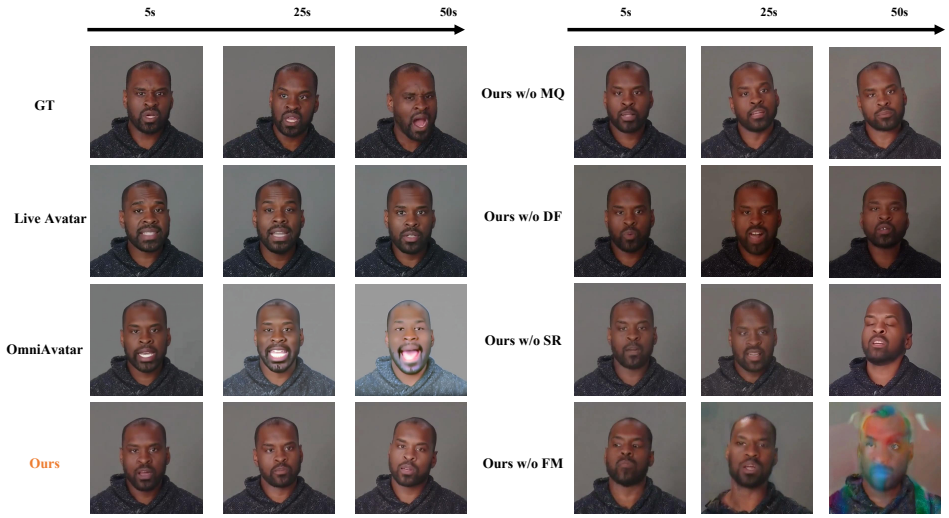


Figure 6: Qualitative comparison with other methods and ablation study. MQ, DF, SR, and FM denote non-verbal memory query, diffusion forcing, strided RoPE offset, and the fusion attention mask, respectively.

OmniResponse and concurrent methods (MIDAS (Chen et al., 2025c) and X-Streamer (Xie et al., 2025c)) are included in the **supplementary video**.

5 Limitation and Conclusion

In this paper, we present OmniPersona, a unified framework that bridges pre-trained MLLMs with joint audio–visual generation for interactive avatar agents. By compressing dialogue history into compact non-verbal memory queries and combining them with the textual response, our system enables coherent multi-turn conversations with natural listening and speaking behaviors. Together with masked fusion attention and anti-drifting mechanisms, OmniPersona achieves stable identity preservation, synchronized audio–visual generation, and temporally consistent long-horizon interactions.

Despite these advantages, several limitations remain. First, although our model achieves faster-than-real-time clip generation on four H200 GPUs (Appendix E), limited computational resources prevented us from training a distilled block-causal variant, which could further reduce latency and better support streaming. Second, the current system focuses primarily on facial regions. We leave streaming-oriented interaction and full-body digital human generation for future work. Overall, we hope OmniPersona can provide useful insights toward building more interactive and responsive digital agents.

AI USE STATEMENT

Generative AI tools were used to assist with implementing the proposed method, as well as translation and language polishing of the manuscript. The research ideas originated from the authors. AI tools were not used to generate research data or analyze experimental results. The authors take responsibility for the final content of the paper and the implementation.

ETHICS STATEMENT

OmniPersona uses publicly available human video datasets in accordance with their respective licenses and usage conditions. Human participants were involved only in rating generated dialogue videos, and no private information about the participants is disclosed. The evaluation procedure is described in Appendix C. Realistic identity-conditioned speech and facial animation can be misused for impersonation, deceptive synthetic media, harassment, or non-consensual identity replication. We advocate authorized use of real-person identities and clear disclosure of synthetic content. Appendix F discusses these risks and planned release safeguards, including usage restrictions, documentation of intended uses and limitations, and consideration of staged release or access controls for high-risk assets.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

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

Audiovisual Motion Modeling and Large-Scale Dataset. *arXiv preprint arXiv:2506.22554*, 2025. URL https://arxiv.org/abs/2506.22554.

Tenglong Ao. Body of Her: A Preliminary Study on End-to-End Humanoid Agent. *arXiv preprint arXiv:2408.02879*, 2024. URL https://arxiv.org/abs/2408.02879.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2aee1c4159e48407d68fe16ae8e6e49e-Abstract-Conference.html.

Jinnan Chen, Lingting Zhu, Zeyu Hu, Shengju Qian, Yugang Chen, Xin Wang, and Gim Hee Lee. MAR-3D: Progressive Masked Auto-regressor for High-Resolution 3D Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11083–11092, 2025a. doi: 10.1109/CVPR52734.2025.01035. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_MAR-3D_Progressive_Masked_Auto-regressor_for_High-Resolution_3D_Generation_CVPR_2025_paper.html.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A family of fully open unified multimodal models—architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025b.

Ming Chen, Liyuan Cui, Wenyuan Zhang, Haoxian Zhang, Yan Zhou, Xiaohan Li, Songlin Tang, Jiwen Liu, Borui Liao, Hejia Chen, Xiaoqiang Liu, and Pengfei Wan. MIDAS: Multimodal Interactive Digital-humAn Synthesis via Real-time Autoregressive Video Generation. *arXiv preprint arXiv:2508.19320*, 2025c. URL https://arxiv.org/abs/2508.19320.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271. Association for Computational Linguistics, 2025d. doi: 10.18653/v1/2025.acl-long.313. URL https://aclanthology.org/2025.acl-long.313/.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019a. doi: 10.1109/CVPR.2019.00482.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019b. URL https://openaccess.thecvf.com/content_CVPRW_2019/html/AMFG/Deng_Accurate_3D_Face_Reconstruction_With_Weakly-Supervised_Learning_From_Single_Image_CVPRW_2019_paper.html.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation, 2025. URL https://arxiv.org/abs/2506.18866.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-s2v: Audio-driven cinematic video generation, 2025. URL https://arxiv.org/abs/2508.18621.

Ying Guo, Xi Liu, Cheng Zhen, Pengfei Yan, and Xiaoming Wei. ARIG: Autoregressive interactive head generation for real-time conversations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12956–12965, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Guo_ARIG_Autoregressive_Interactive_Head_Generation_for_Real-time_Conversations_ICCV_2025_paper.html.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025. URL https://arxiv.org/abs/2508.13009.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025a. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live Avatar: Streaming Real-time Audio-Driven Avatar Generation with Infinite Length, 2025b. URL https://arxiv.org/abs/2512.04677.

Yubo Huang, Weiqiang Wang, Sirui Zhao, Tong Xu, Lin Liu, and Enhong Chen. Bind-Your-Avatar: Multi-Character-Talking Video Generation with Dynamic 3D-mask-based Embedding Router. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4440–4449, 2026. URL https://yubo-shankui.github.io/bind-your-avatar/.

Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, and Sung Ju Hwang. Avatar forcing: Real-time interactive head avatar generation for natural conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18074–18084, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ki_Avatar_Forcing_Real-Time_Interactive_Head_Avatar_Generation_for_Natural_Conversation_CVPR_2026_paper.html.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, 2016. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025. URL https://arxiv.org/abs/2510.01284.

Cheng Luo, Jianghui Wang, Bing Li, Siyang Song, and Bernard Ghanem. OmniResponse: Online multimodal conversational response generation in dyadic interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 125869–125891, 2025a. doi: 10.52202/085713-4196. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b734c30b9c955c535e333f0301f5e45c-Abstract-Conference.html.

Jiabin Luo, Junhui Lin, Zeyu Zhang, Biao Wu, Meng Fang, Ling Chen, and Hao Tang. Univid: The open-source unified video model. *arXiv preprint arXiv:2509.24200*, 2025b.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 484–492. ACM, October 2020. doi: 10.1145/3394171.3413532. URL http://dx.doi.org/10.1145/3394171.3413532.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Yong-Jin Liu. StreamAvatar: Streaming Diffusion Models for Real-Time Interactive Human Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10887–10897, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sun_StreamAvatar_Streaming_Diffusion_Models_for_Real-Time_Interactive_Human_Avatars_CVPR_2026_paper.html.

Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. Omni-video: Democratizing unified video understanding and generation. *arXiv preprint arXiv:2507.06119*, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. UniVideo: Unified understanding, generation, and editing for videos. In *International Conference on Learning Representations*, pp. 113905–113933, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b940430f7acd23302bbaf0680f458e61-Abstract-Conference.html.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10086–10096, October 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. In *International Conference on Learning Representations*, pp. 28240–28264, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45f0d179ef7e10eb7366550cd4e574ae-Abstract-Conference.html.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 47490–47518, 2025b. doi: 10.52202/085713-1584. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/43e8fd8b9581faa71a6a61602bc28435-Abstract-Conference.html.

You Xie, Tianpei Gu, Zenan Li, Chenxu Zhang, Guoxian Song, Xiaochen Zhao, Chao Liang, Jianwen Jiang, Hongyi Xu, and Linjie Luo. X-Streamer: Unified Human World Modeling with Audiovisual Interaction. *arXiv preprint arXiv:2509.21574*, 2025c. URL https://arxiv.org/abs/2509.21574.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Hao Yang, Zhiyu Tan, Jia Gong, Luozheng Qin, Hesen Chen, Xiaomeng Yang, Yuqing Sun, Yuetan Lin, Mengping Yang, and Hao Li. Omni-Video 2: Scaling MLLM-Conditioned Diffusion for Unified Video Generation and Editing. *arXiv preprint arXiv:2602.08820*, 2026a. URL https://arxiv.org/abs/2602.08820.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, and Xiaoming Wei. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing, 2025. URL https://arxiv.org/abs/2508.14033.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time Interactive Long Video Generation. In *International Conference on Learning Representations*, pp. 90037–90059, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/91a1610c6ed9e02d33f826b46f472b92-Abstract-Conference.html.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_SkeHuCVFDr.html.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. SpeakerVid-5M: A Large-Scale High-Quality Dataset for Audio-Visual Dyadic Interactive Human Generation. In *International Conference on Learning Representations*, pp. 117896–117926, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/bf7dbac50ed7f6e12ad529c5b9396bc4-Abstract-Conference.html.

Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: A benchmark dataset and baseline. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 124–142, Cham, 2022. Springer Nature Switzerland.

Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. INFP: Audio-driven interactive head generation in dyadic conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10667–10677, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhu_INFP_Audio-Driven_Interactive_Head_Generation_in_Dyadic_Conversations_CVPR_2025_paper.html.

Appendix

A Qualitative Videos

We provide supplementary videos to better illustrate the qualitative performance of OmniPersona. The videos include three categories of results:

(1) Additional qualitative results. We present more examples demonstrating multi-turn conversational interactions, where the avatar naturally switches between *speaking* and *listening* states. The examples cover both identities from the *test set* and *in-the-wild* identities, showing that our model can independently control the avatar's appearance and voice characteristics from *distinct sources.*

(2) Videos corresponding to qualitative comparisons in the main paper. We include the video results for all examples shown in the *Qualitative Comparison* section of the main paper, as well as selected examples from the OmniResponse benchmark to show obvious audio-visual misalignment in the ground-truth recordings.

(3) Visual comparisons with concurrent work. We further provide side-by-side comparisons with concurrent methods to highlight the advantages of our approach in identity consistency, audio–visual synchronization, and conversational expressiveness.

B Implementation Details

B.1 Model Architecture

Our system combines a multimodal large language model (MLLM) with a diffusion-based joint audio–video generation backbone. Specifically, we adopt Qwen2.5-Omni-7B as the MLLM backbone, which provides high-level conversational reasoning. To bridge the MLLM and the diffusion generator, we introduce a learnable connector that first maps the MLLM query representations to the T5 hidden dimension through a zero-initialized projection layer and then refines them with Transformer encoder layers. This connector transforms the MLLM outputs into conditioning tokens suitable for the diffusion backbone.

The diffusion backbone is initialized from the pretrained OVI-v1.1 model, a 960×960 resolution video diffusion model trained on 10-second clips. The generated video and audio latents are decoded using Wan2.2 VAE and MMAudio VAE, respectively. In addition, a separate T5-XXL text encoder (UMT5) provides textual conditioning signals for the diffusion process.

Table 6: Architecture configuration of the proposed model.

Component	Configuration
MLLM backbone	Qwen2.5-Omni-7B
Connector layers	8 Transformer encoder layers
Projection layers	MLLM-to-T5 projection layers
Learnable queries	$\mathcal{Q} = 128$
Diffusion backbone	OVI-v1.1 (10s, 960×960)
Video decoder	Wan2.2 VAE
Audio decoder	MMAudio VAE
Text encoder	UMT5 (T5-XXL)

B.2 Training Setup

Training is conducted using bf16 mixed precision on $8\times$ H200 GPUs for 48k steps, with a per-device batch size of 5. The model is optimized using AdamW with a cosine learning rate schedule, where the learning rate decays to a minimum of 1×10^{-6} after a warmup of 100 steps. A weight decay of 0.1 is applied during optimization, and gradient clipping with a maximum norm of 1.0 is used to stabilize training. The loss weights for audio and video generation are set to $\lambda_A = 0.5$ and $\lambda_V = 1.0$, respectively.

LoRA is applied to both the video and audio branches of the diffusion model, with the LoRA rank and alpha both set to 128. The LoRA parameters are trained with a learning rate of 1×10^{-5}, while the MLLM backbone remains frozen and only the connector module together with the token embedding layer are updated using a learning rate of 5×10^{-5}.

To better model conversational dynamics, the training data is sampled with a 4:1 ratio between mainly speaking segments and mainly listening segments. Furthermore, the loss for listening samples is doubled to encourage the model to learn richer non-verbal reactions.

Table 7: Training hyperparameters.

Parameter	Value
GPU configuration	$8\times$ H200
Precision	bf16
Batch size (per GPU)	5
Training steps	48k
Warmup steps	100
Scheduler	cosine
Minimum LR	1×10^{-6}
Learning rate (LoRA)	1×10^{-5}
Learning rate (others)	5×10^{-5}
Weight decay	0.1
Gradient clipping	1.0
Audio loss weight λ_A	0.5
Video loss weight λ_V	1.0

B.3 Conditioning for the Diffusion Model

The diffusion model is conditioned on seven inputs: the prefix video latent V_P, the noised video latent V_N, a reference video latent V_R, the prefix audio latent A_P, the noised audio latent A_N, a reference audio latent A_R, and a response embedding $\tilde{e}$.

Video latents are produced by a 3D Causal VAE. The first latent frame corresponds to 1 video frame, while each subsequent latent frame corresponds to 4 frames. For reference image V_R, it is randomly selected from the whole video clip during training while provided by user during inference. During training, we either select the prefix latent length as 1 or 5 (1 or 17 frames, also $0.04\,\mathrm{s}$ or $0.68\,\mathrm{s}$ at 25 FPS), so that the model can handle both scenarios at inference time: the first turn where only a single reference image is available, and later turns where a longer prefix segment can be extracted from already generated video to maintain temporal continuity. The target video latent V_N is set as 31 latent frames (about $4.96\,\mathrm{s}$) for training.

Audio and video latents are temporally aligned according to the downsampling of video and audio VAEs. Given a video frame rate of 25 FPS, audio sampling rate of $16\,\mathrm{kHz}$, and audio VAE downsampling factor of 512, the temporal ratio becomes $\frac{25/4}{16000/512} = 0.2$, meaning that 1 video latent frame corresponds to 5 audio latent frames. Since the first latent frame of the 3D Causal VAE represents only one frame rather than 4, we append 3 silent audio latent frames in addition to the true aligned audio to maintain the fixed audio–video temporal ratio.

For prefix audio latent A_P and target audio latent A_N at training, they are extracted from the whole audio clip to temporally align with V_P, V_N respectively. As for the reference audio, a 4-6 s speaking audio clip is randomly selected from the whole audio during training while provided by the user during inference.

The response embedding $\tilde{e}$ concatenates non-verbal memory tokens and a T5-based verbal response representation. The memory tokens are produced by the MLLM encoding the audio–video context of the previous conversational turn from both the user and the agent (5–20s during training and inference), while the verbal response is generated by the MLLM conditioned on the full dialogue history.

During training, classifier-free guidance is applied with a dropout ratio of 0.1 to both the response embedding $\tilde{e}$ and the reference audio latent A_R.

Table 8: Conditioning signals used by the diffusion models.

Condition	Description	Latent Length	CFG Drop
V_P	prefix video context	1 or 5	–
V_N	noised target video segment	31	–
V_R	reference video identity frame	1	–
A_P	prefix audio context	5 or 25	–
A_N	noised target audio segment	155	–
A_R	reference audio clip	4–6s * 16000 // 512	0.1
$\tilde{e}$	response embedding (memory + T5 text)	128 + T5 embedding sequence	0.1

B.4 INFERENCE

During inference, sampling is performed using UniPC with 10 steps. The response embedding $\tilde{e}$ guidance scales are set to 3.0 for video and 4.0 for audio, while the reference-audio guidance scale is set to 4.0. Negative prompts are applied to both video (*"jitter, bad hands, blur, distortion"*) and audio (*"robotic, muffled, echo, distorted"*) to improve generation quality.

C HUMAN EVALUATION PROTOCOL

We invited 20 participants to evaluate 20 dialogue pairs randomly sampled from the 60 pairs in our SeamlessInteraction evaluation, comparing OmniPersona, LiveAvatar, and OmniAvatar. Participants viewed the full dialogue, with speaking and listening intervals explicitly marked, and rated each criterion on a 1–5 Mean Opinion Score (MOS) scale.

General instructions and evaluation scope. Before rating, participants were instructed to judge each criterion independently, avoid carrying impressions from one criterion into another, and disregard personal preferences for the person's appearance. Listening expression appropriateness (EA) assessed contextual appropriateness within the marked listening intervals. Audio-visual synchronization (AVS) assessed synchronization during speaking intervals. Overall quality (OQ) assessed audio-visual artifacts and identity consistency over the full video. This separation was intended to reduce interference between criteria; the three scores should be interpreted jointly, since high EA does not compensate for poor synchronization or identity consistency.

The AVS and OQ criteria are similar to those commonly used in prior talking-avatar evaluations, so we do not elaborate on their scoring procedures here. We focus instead on EA, which evaluates listening behavior in its conversational context and requires explicit guidance on how to judge the appropriateness of a listener's reactions. We therefore detail the EA instructions and rating anchors below.

Listening expression appropriateness. EA was a composite contextual-appropriateness rating, rather than separate subscores for emotion, timing, and intensity. It considered semantic and emotional congruence with the conversation, reaction timing, and reaction intensity and attentiveness. Image quality, identity consistency, and speaking lip synchronization were assessed through the other criteria rather than the EA score. After the general instructions, participants received the following EA-specific prompt:

> Please focus only on the person while they are listening. Considering what the other person is saying, their tone of voice, and the surrounding dialogue context, how appropriate is the listener's facial reaction at that moment? Consider whether the reaction matches the meaning and emotion of the conversation and whether its timing and intensity are appropriate rather than emotionally unrelated or unresponsive.

This prompt was specific to EA and was not the complete instruction set. Before the formal evaluation, participants viewed examples corresponding to different rating levels to calibrate their interpretation of the criterion and scale.

Table 9: Rating anchors for listening expression appropriateness.

Score	Criterion
1	Completely inappropriate: the expression clearly conflicts with the context, such as an opposite emotion or an evidently incorrect reaction time.
2	Mostly inappropriate: the expression is largely inconsistent in emotion, timing, or intensity.
3	Acceptable: the expression is generally plausible but may be neutral, ambiguous, weakly responsive, or occasionally mismatched.
4	Appropriate: the expression aligns well with the conversational meaning, emotion, and timing, with only minor imperfections.
5	Highly appropriate: the expression is highly consistent with the context and convincingly matches its emotion, timing, and intensity.

D MORE EXPERIMENTS

D.1 NON-VERBAL MEMORY ANALYSIS

Affective information in the memory. We analyze the information that can be decoded from the verbal and non-verbal conditions over ten 5-second conversational turns. We separately average the Text and memory query (MQ) tokens and train linear probes using Text, MQ, or their combined features (Text+MQ). Probes are trained on the validation set and evaluated on the test set to predict eight emotion classes and arousal scores. The targets are derived by averaging the official Imitator outputs provided with SeamlessInteraction over the video frames of each turn; they are automatically estimated affective annotations, not human emotion labels. Tables 10 and 11 report selected turns and the average over all ten turns.

Table 10: Emotion linear probes: Macro-F1 (higher is better).

Feature	Turn 1	Turn 2	Turn 5	Turn 10	Avg. 1–10
Text	0.432	0.428	0.447	0.434	0.435
MQ	0.578	0.559	0.553	0.541	0.554
Text+MQ	**0.651**	**0.648**	**0.644**	**0.642**	**0.649**

Table 11: Arousal linear probes: MAE (lower is better).

Feature	Turn 1	Turn 2	Turn 5	Turn 10	Avg. 1–10
Text	0.118	0.126	0.109	0.139	0.123
MQ	0.061	0.056	0.064	0.072	0.066
Text+MQ	**0.035**	**0.039**	**0.040**	**0.043**	**0.040**

MQ outperforms Text on both tasks, while Text+MQ performs best, indicating complementary affective information in the two conditions. Text+MQ Macro-F1 changes from 0.651 at Turn 1 to 0.642 at Turn 10, and arousal MAE from 0.035 to 0.043. These probes characterize decodable affective information across the evaluated turns; they do not assign a semantic interpretation to individual memory tokens or establish a causal effect on generated expressions.

Memory capacity and context length. Table 12 provides the exact values plotted in Fig. 4. Increasing the query count from 32 or 64 to 128 improves expression modeling, whereas 192 queries yield no further improvement in these settings. Increasing the context from 5 to 10 seconds also helps, with little benefit from extending it to 20 seconds. We therefore use 128 queries and a 10-second context by default.

Table 12: FID_{emo} for different query counts and context lengths (lower is better).

Queries	5 s	10 s	20 s
32	1.06	0.74	0.71
64	0.72	0.49	0.52
128	**0.49**	**0.31**	**0.32**
192	0.52	0.34	**0.32**

D.2 SAME-RESPONSE-TEXT CONTROLLED COMPARISON

The main SpeakerVid-5M comparison evaluates end-to-end systems, whose response texts can differ. To separate response-content differences from downstream generation performance, we additionally condition all three methods on the response text generated by SpeakerVid-5M, keeping the evaluation set and reference inputs unchanged. Table 13 shows that OmniPersona retains the best scores on all four evaluated generation metrics. This control fixes response text without assuming that the methods have identical generators or training procedures.

D.3 MULTI-TURN ROPE-OFFSET ANALYSIS

Placing the reference image at $p_R = L + \delta$ separates the identity reference from adjacent prefix and target frames. During training, sampling $\delta \sim \mathcal{U}(20, 50)$ exposes the model to different positional

Table 13: Controlled comparison using the same SpeakerVid-5M response text.

Method	FVD ↓	ArcFace ↑	Sync$_{conf}$ ↑	FID$_{emo}$ ↓
SpeakerVid-5M	28.82	0.772	2.698	3.22
Qwen2.5-Omni + OVI	18.39	0.541	3.256	0.67
OmniPersona	**5.25**	**0.773**	**3.944**	**0.21**

gaps rather than a single fixed location. This empirical range balances identity anchoring and pose flexibility. At inference, smaller offsets strengthen anchoring but can constrain the generated pose toward the static reference. We progressively decrease the offset over turns to retain pose flexibility early and strengthen identity correction as autoregressive errors accumulate.

To examine the offset trade-off over longer generation, we evaluate fixed offsets using the same checkpoint, conversations, and random seeds for ten autoregressive turns. Table 14 reports ArcFace and Pose$_{ref}$, where Pose$_{ref}$ is the average frame-wise pose distance to the reference image. Higher Pose$_{ref}$ indicates greater deviation from the reference pose, not necessarily better generation quality; it is a rigidity diagnostic and differs from the pose FD in Table 4.

Table 14: Fixed-offset results over ten turns. Each cell reports ArcFace / Pose$_{ref}$.

Offset	Turn 1	Turn 2	Turn 5	Turn 10	Avg. 1–10
$\delta = 20$	.854/.023	.842/.024	.815/.026	.782/.027	.819/.024
$\delta = 30$	.852/.026	.838/.026	.804/.029	.776/.031	.812/.029
$\delta = 40$	.852/.027	.829/.028	.796/.032	.769/.034	.803/.031
$\delta = 50$	.847/.031	.811/.033	.778/.036	.762/.042	.795/.037

ArcFace decreases from Turn 1 to Turn 10 for all offsets, while Pose$_{ref}$ generally increases. Smaller offsets preserve higher identity similarity with less deviation from the reference pose. These observations motivate the decreasing-offset strategy, but this fixed-offset comparison does not establish that a particular dynamic schedule is optimal.

D.4 Emotion-intensive Scenarios

We select 20 anger-dominant samples from the official SeamlessInteraction test set using its per-frame Anger scores and arousal estimates. These annotations are automatically extracted by the dataset's Imitator model rather than manually assigned. We compare OmniPersona with LiveAvatar and the same w/o-MQ checkpoint used in our ablation study. Table 15 shows that OmniPersona achieves lower FID$_{emo}$ and expression FD on this subset, while removing MQ degrades expression modeling. The finding is limited to this selected subset and does not establish robustness across all high-intensity emotions.

Table 15: Expression modeling on 20 anger-dominant samples.

Method	FID$_{emo}$ ↓	Expression FD ↓
LiveAvatar	0.537	17.95
Ours w/o MQ	0.689	19.81
Ours	**0.323**	**14.46**

E Inference Efficiency

We distinguish complete-pipeline latency and post-MLLM generation latency for 5-second (125-frame), 448×448 outputs. These measurements concern clip generation, rather than streaming time to the first frame or interruption handling.

Complete inference pipeline. Table 16 compares the methods on one H200, including their method-specific context conditioning and prompt encoding, as well as final VAE decoding. OmniPersona additionally includes MLLM-based contextual memory encoding. The row excluding MLLM encoding retains the remaining generation pipeline and separates the cost of that stage; it is not the w/o-MQ quality ablation. The methods accept different types of conditioning, so these timings compare the listed generation pipelines under their respective input requirements.

Table 16: Complete-pipeline efficiency on one H200. The w/o MLLM encoding row excludes only that stage.

Method	Total time (s) ↓	Peak VRAM (GB) ↓
OVI	31.5	41.8
OmniAvatar	110.4	43.3
LiveAvatar	14.2	77.3
OmniPersona w/o MLLM encoding	8.6	45.7
OmniPersona	12.3	59.7

Hardware configuration and timing scope. Table 17 separates complete inference from the post-MLLM pipeline, which includes the connector, DiT, and VAE. On two H200s, the conditional and unconditional classifier-free guidance (CFG) branches run on separate GPUs and the VAE layers are sharded. Including sharded MLLM encoding gives 7.2 seconds for the complete pipeline. On four H200s, the MLLM and VAE are sharded across the GPUs, and two GPUs are assigned to each CFG branch, reducing complete-pipeline latency to 4.6 seconds. Thus, faster-than-real-time generation of a 5-second clip is achieved with four H200s; the 4.8-second result on two H200s excludes MLLM encoding. All OmniPersona settings use 10 diffusion steps.

Table 17: OmniPersona latency by timing scope and hardware configuration. Dashes denote unreported measurements.

Timing scope	1 H200	2 H200s	4 H200s
Post-MLLM generation (s)	8.6	4.8	–
Complete pipeline (s)	12.3	7.2	4.6

F Impact and Responsible Release

Broader impact. OmniPersona aims to make multimodal human-agent interaction more natural and accessible by combining dialogue understanding, speech generation, and expressive avatar animation in a unified system. Potential positive uses include educational assistants, remote communication, accessibility support, and interactive digital services where visual and spoken feedback can improve user engagement. At the same time, realistic talking-avatar generation may also create risks, including impersonation, misleading synthetic media, harassment, and privacy or consent violations. These risks are especially important when a generated avatar resembles a real person or is used in sensitive contexts. We therefore view consent, transparency, and careful deployment boundaries as necessary conditions for practical use.

Responsible release safeguards. Upon acceptance, we plan to release code, models, and examples with safeguards that reduce misuse while supporting reproducible research. The release will include clear usage terms prohibiting impersonation, deception, harassment, and non-consensual identity replication; documentation describing intended uses, limitations, and known risks; and examples based on consent-respecting or synthetic identities. For components that could increase misuse risk, we will consider staged release, access control, or withholding high-risk assets. We will also recommend visible disclosure and provenance practices for generated avatar content so that downstream users can distinguish synthetic outputs from real recordings.