# EX-OMNI-2D: EXPRESSIVE OMNI-MODAL DIALOGUE MODELS WITH NATIVE VISUAL PRESENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Omni-modal dialogue models can understand multimodal inputs and synthesize spoken replies, but a spoken answer still leaves the agent visually absent. We introduce **Ex-Omni-2D**, a framework that answers a multimodal query with coordinated text, personalized speech, and reference-conditioned video. The LLM first generates a structured *Visual Thought Plan* (VTP) followed by the response text. The speech generator then produces multi-codebook speech units from response-side representations. These speech units are decoded into audio and aligned with video frames, giving the speech and avatar modules a common timing signal while allowing them to learn from different data sources. The full-sequence video generator realizes the framework by conditioning on reference appearance, VTP semantics, and frame-aligned speech units. This is the configuration used for our main quality evaluation. For latency-sensitive settings, we additionally adapt it for few-step, block-causal streaming generation with a dedicated reference-image sink and speech-aligned chunks. This provides incremental output with lower startup latency than full-sequence decoding. The demo, code and data can be found in **uploaded supplementary materials** and **https://anonymous.4open.science/r/afojfm.**

## 1 INTRODUCTION

Omni-modal dialogue systems increasingly understand speech, text, and visual inputs and answer with natural spoken replies (Xu et al., 2025; Fang et al., 2025a;b; Xie & Wu, 2024a). For many interactive settings, however, a spoken answer is only part of the response: a user also expects a visible agent whose appearance, affect, and motion fit the current exchange. We study this setting with *dialogue-native video responses*: given a user query, reference image, and reference audio, the system generates response text, personalized speech, and a synchronized video of the referenced speaker.

Strong talking-avatar models already animate a portrait from speech and visual references (Peng et al., 2025; Gao et al., 2025; Gan et al., 2025; Tu et al., 2025; Chen et al., 2025b). They are less suited to dialogue by themselves because the visual behavior is usually controlled by a completed waveform and by an auxiliary prompt prepared outside the conversation in many systems. Joint audio-video generators provide tighter modality coupling (Ruan et al., 2023; Liu et al., 2026; Wang et al., 2025a), but they are typically prompt- or reference-driven content generators. The missing piece is a response interface that lets the LLM decide both what to say and how the speaking agent should look while saying it.

Ex-Omni-2D uses two lightweight interfaces for this purpose. A *Visual Thought Plan* (VTP) records the scene, emotion, movement style, and motion details needed by the avatar generator. Multi-codebook speech units provide the timing and acoustic content used by both the speech decoder and the video generator. This design has a practical training advantage: speech and dialogue modules can learn from speech-centric data, while the avatar module can learn from filtered video clips annotated with VTPs and teacher-forced speech units. At inference time, the VTP and generated unit stream join these pathways into one response.

We first train the full-sequence video generator with bidirectional self-attention over the entire latent sequence, conditioned on the reference appearance, VTP semantics, and frame-aligned speech units. For latency-sensitive deployment, we adapt this generator to block-causal, few-step streaming inference through distillation, drawing inspiration from prior work (Sun et al., 2026; Gu et al., 2026).



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.

The streaming version generates three-latent chunks, retaining bidirectional attention within each chunk and causal information flow across chunks through persistent historical KV caches. Our contributions are summarized as follows:

- We extend omni-modal dialogue models with native visual presence. Ex-Omni-2D generates coordinated response text, personalized speech, and reference-conditioned avatar video from a multimodal query. Its shared speech-unit interface connects the speech and video pathways, allowing them to learn from different data sources without requiring fully paired dialogue–speech–video responses.
- We introduce the Visual Thought Plan (VTP), a structured interface through which the LLM explicitly plans the scene, emotion, and motion of its visual response before generating the user-facing reply. The VTP gives the video generator dialogue-specific guidance for how the speaker should appear and behave.
- We explore streaming inference for the complete response pipeline. By adapting the video generator for few-step, block-causal generation and aligning its output chunks with generated speech units, Ex-Omni-2D can emit synchronized audio and video incrementally.

## 2 Related Work

**Omni-modal Dialogue.** Recent omni-modal large language models extend text-based interaction toward unified understanding and generation over speech, vision, and language. Representative systems explore speech-centered instruction following, real-time spoken response generation, cross-modal alignment, and unified multimodal generation within LLM-based architectures (Xie & Wu, 2024a;b; Luo et al., 2025; Xu et al., 2025; AI et al., 2025; Fu et al., 2026; Team, 2025). In parallel, speech language models further reduce the dependence on cascaded ASR–LLM–TTS pipelines by directly modeling spoken dialogue and speech generation (Ao et al., 2022; Zhang et al., 2024; Défossez et al., 2024; Zeng et al., 2024). However, most existing omni-modal dialogue systems still lack visual presence: they can understand speech and synthesize spoken responses, but do not directly generate a speaking visual agent.

**Unified Audio-Video Generation.** Aligned audio-video generation has been studied through both cascaded and joint paradigms. Audio-driven video synthesis first generates or receives a speech waveform and then animates a reference identity, which has led to strong progress in talking-head and human animation systems (Prajwal et al., 2020; Zhang et al., 2023b; Xu et al., 2024; Gan et al., 2025; Gao et al., 2025). The reverse direction, video-to-audio dubbing, generates sound or speech for a given silent video (Luo et al., 2023; Cheng et al., 2025; Sung-Bin et al., 2025). These conditional systems are effective visual or acoustic renderers, but their inputs usually do not include the conversational state from which response-specific visual behavior should be derived. Joint audio-video generation instead synthesizes both modalities within a unified or interacting generative process. Early diffusion-based work such as MM-Diffusion (Ruan et al., 2023) initiated this direction, followed by models that improve scalability and synchronization such as SyncFlow (Liu et al., 2024) and UniForm (Zhao et al., 2025). More recent systems, including JavisDiT (Liu et al., 2026), Ovi (Low et al., 2025), UniVerse-1 (Wang et al., 2025a), and UniAVGen (Zhang et al., 2026), further explore large-scale or human-centric joint audio-video generation (Liu et al., 2026; Low et al., 2025; Wang et al., 2025a; Zhang et al., 2026). UniVerse-1 stitches pre-trained audio and video experts to produce coordinated audio-video content, while UniAVGen introduces a dual-branch DiT with asymmetric cross-modal interactions and face-aware modulation for human-centric generation. These methods are the closest comparison group to our framework because they generate paired audio and video. Ex-Omni-2D connects this generation problem to dialogue: the model first plans visual behavior from the current multimodal query, then passes both semantic guidance and acoustic timing to the avatar generator.

## 3 Method

Ex-Omni-2D receives a multimodal user query $x = (x_t, x_s)$ containing text, speech, or both, together with a reference image $I^{\text{ref}}$ and reference audio $a^{\text{ref}}$. It generates response text $\mathbf{y}$, personalized speech, and a reference-conditioned video response $V$. Figure 1 shows the overall pipeline. The

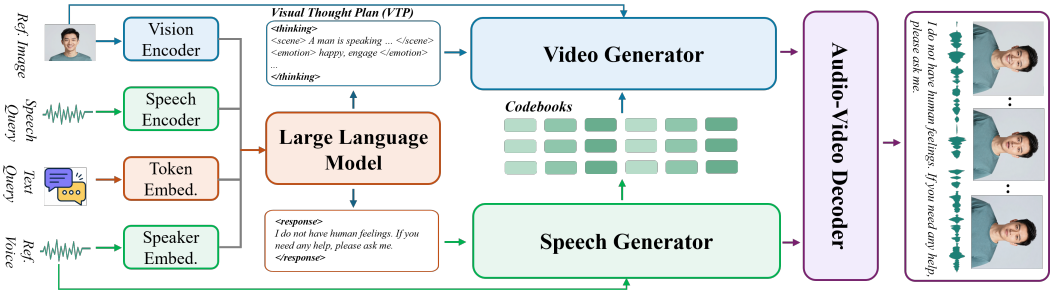


Figure 1: Overview of Ex-Omni-2D. Speech, text, and vision encoders provide multimodal context to the LLM, which produces a non-user-facing VTP and the user-facing response. The Speech Generator uses reference audio for voice conditioning and generates speech in a native multi-codebook space. The Video Generator encodes the reference image and VTP, and reuses the generated speech representation to synthesize the synchronized video response.

VTP $\mathbf{p}$ carries high-level visual intent, and the codebook speech units $\mathbf{U}$ carry the acoustic content and timing shared by speech and video. The LLM emits the VTP and response text; response-side hidden states drive the Speech Generator; the resulting units are decoded into waveform audio and mapped to frame-aligned video conditions. The full-sequence Video Generator uses bidirectional temporal context and provides the primary quality results. Its streaming adaptation renders chunks incrementally while retaining the reference and bounded historical KV state.

### 3.1 Multimodal Input

Let $x_t$ and $x_s$ denote the text and speech components of the current query, where either component may be empty but at least one is present. The text is embedded by the language model, while speech is encoded and projected into the same hidden space:

$$\mathbf{E}_t = \mathrm{Emb}_{\mathrm{text}}(x_t), \quad \mathbf{E}_s = P_{\mathrm{sp}}(E_{\mathrm{sp}}(x_s)) . \tag{1}$$

The two sequences are combined according to their order in the query to form the multimodal query representation $\mathbf{E}_x$.

The character conditions follow distinct pathways. For dialogue understanding and VTP generation, the reference image is processed by the vision tower of Qwen3-VL-2B (Bai et al., 2025) and projected into the dialogue-model space:

$$\mathbf{z}_I^{\mathrm{llm}} = P_{\mathrm{vis}}\big(E_{\mathrm{vis}}(I^{\mathrm{ref}})\big) . \tag{2}$$

Together, $\mathbf{E}_x$ and $\mathbf{z}_I^{\mathrm{llm}}$ form the LLM input. The same image is independently encoded inside the Video Generator by the Wan 3D VAE (Wan et al., 2025) to obtain an appearance-reference latent:

$$\mathbf{R}^{\mathrm{ref}} = E_{\mathrm{vae}}\big(I^{\mathrm{ref}}\big) . \tag{3}$$

Reference audio is handled inside the Speech Generator, where a speaker encoder extracts the voice embedding $\mathbf{s}^{\mathrm{ref}}$.

### 3.2 Visual Response Planning

The LLM is instantiated with Qwen3-8B (Yang et al., 2025). It follows a structured assistant protocol that separates the internal visual plan from the user-facing response:

$$\mathbf{o} = [\texttt{<thinking>}, \mathbf{p}, \texttt{</thinking>}, \texttt{<response>}, \mathbf{y}, \texttt{</response>}]. \tag{4}$$

The `<thinking>` block is restricted to the structured VTP rather than an unconstrained chain-of-thought trace. Atomic boundary tokens and constrained decoding ensure that the VTP is completed before response generation and that the planning and user-facing spans can be routed independently. The VTP contains five fields:

$$\mathbf{p} = (p_{\mathrm{first}}, p_{\mathrm{scene}}, p_{\mathrm{emotion}}, p_{\mathrm{style}}, p_{\mathrm{motion}}), \tag{5}$$

which describe the first-frame scene, overall scene, emotion, movement style, and detailed motion. The plan is internal and is not displayed as the conversational answer. The Video Generator encodes it into a semantic condition

$$\mathbf{L}^{\mathrm{vtp}} = E_{\mathrm{text}}(\mathbf{p})\,. \tag{6}$$

Because the plan is text, it can be supervised with ordinary autoregressive training and inspected during evaluation. We retain the final-layer hidden states only over the response span, denoted $\mathbf{H}^{\ell}_{\mathbf{y}}$, for speech generation. The VTP provides high-level semantic guidance for how the response should look, while $\mathbf{H}^{\ell}_{\mathbf{y}}$ and $\mathbf{y}$ specify what should be spoken. The final video realization is jointly conditioned on the VTP and frame-aligned speech units.

### 3.3 Speech Generator and Shared Acoustic Interface

The Speech Generator is initialized from Qwen3-TTS-0.6B (Hu et al., 2026). Conditioned on the response states, response tokens, and reference voice, it predicts a sequence of multi-codebook acoustic units:

$$\mathbf{U} = G_{\mathrm{sp}}(\mathbf{H}^{\ell}_{\mathbf{y}}, \mathbf{y}, \mathbf{s}^{\mathrm{ref}}), \quad \mathbf{U} \in \mathbb{N}^{N\times C}, \tag{7}$$

where $C = 16$ is the number of Qwen3-TTS acoustic codebooks. The first codebook is generated autoregressively at each acoustic frame, and the residual codebooks refine its acoustic content. Text-token embeddings and $\mathbf{H}^{\ell}_{\mathbf{y}}$ are fused through gated cross-attention before acoustic decoding.

We treat $\mathbf{U}$ as the shared representation of the generated speech. A codec decoder renders it as waveform audio, while the avatar pathway maps the same units into video-conditioning features. For acoustic frame $n$, the codebook embeddings are aggregated and projected by a lightweight adapter:

$$\widetilde{\mathbf{A}}_n = f_{\mathrm{aud}}\left(\sum_{c=1}^{C} E_c(u_{n,c})\right). \tag{8}$$

The units are produced at 12.5 Hz and the video at 25 FPS, so each acoustic feature is repeated for two video frames:

$$\mathbf{A}_{2n-1} = \mathbf{A}_{2n} = \widetilde{\mathbf{A}}_n, \quad n = 1, \ldots, N, \tag{9}$$

where $\mathbf{A} \in \mathbb{R}^{2N\times d_a}$ is aligned with the video timeline. This interface uses all 16 codebooks and gives the video generator an explicit frame-level timing signal without re-encoding a completed waveform. It also makes the training data easier to use: speech modeling can use speech and spoken-dialogue data, while the video pathway can use avatar clips with teacher-forced speech units. At inference time, the video pathway receives the dialogue-derived VTP together with the same generated acoustic representation used to render the spoken response.

### 3.4 Video Generator

**Full-Sequence Video Generation.** Ex-Omni-2D first trains a full-sequence Video Generator based on Wan2.1-T2V-1.3B (Wan et al., 2025) and initialized with the corresponding OmniAvatar-1.3B LoRA weights (Gan et al., 2025). It contains a VTP text encoder, a 3D VAE, and a latent denoising backbone, and uses bidirectional temporal context over the target clip. This full-sequence realization is the primary Video Generator and later serves as the frozen target for the streaming adaptation in Section 3.4. It receives three complementary conditions: the reference latent $\mathbf{R}^{\mathrm{ref}}$ for appearance, the frame-aligned acoustic condition $\mathbf{A}$ for speech-driven motion, and the VTP condition $\mathbf{L}^{\mathrm{vtp}}$ for scene, affect, and movement semantics:

$$V = G_{\mathrm{vid}}(\mathbf{R}^{\mathrm{ref}}, \mathbf{A}, \mathbf{L}^{\mathrm{vtp}}). \tag{10}$$

For a ground-truth video $V^*$, the 3D VAE produces a clean latent $\mathbf{x}_0$. Following the flow-matching formulation of Gu et al. (2026), we sample Gaussian noise $\boldsymbol{\epsilon}$ and a noise level $\sigma$ and construct the flow-matching input

$$\mathbf{x}_\sigma = (1-\sigma)\mathbf{x}_0 + \sigma\boldsymbol{\epsilon}. \tag{11}$$

The denoising network is conditioned on $\mathbf{x}_\sigma$, $\sigma$, $\mathbf{R}^{\mathrm{ref}}$, $\mathbf{A}$, and $\mathbf{L}^{\mathrm{vtp}}$. After alignment to the 25-FPS timeline, the acoustic features are packed to the video-latent temporal resolution. Following OmniAvatar (Gan et al., 2025), the projected speech features are injected into DiT blocks 2–15. At an injected layer $\ell$, the aligned feature is tiled over the latent spatial grid, patchified, and added to the video tokens:

$$\mathbf{X}^{(\ell)} \leftarrow \mathbf{X}^{(\ell)} + \mathbf{G}^{(\ell)}. \tag{12}$$

**Streaming Adaptation.** To support latency-sensitive deployment, we distill the full-sequence generator into a block-causal streaming version that renders speech-aligned video chunks incrementally. The full-sequence generator provides the frozen teacher target for this stage, and the streaming model is trained as its student. Following StreamAvatar (Sun et al., 2026), streaming generation uses bidirectional attention within each three-latent chunk and causal attention across chunks. We further adapt the dual-time flow maps and on-policy distribution-matching objective of AnyFlow (Gu et al., 2026). For the $m$-th noisy window $\mathbf{x}_t^{(m)}$, destination time $r$, condition $\mathbf{c}^{(m)}$, and persistent state $\mathcal{S}_m$, the streaming model predicts

$$\mathbf{v}_\theta\left(\mathbf{x}_t^{(m)}, t, r; \mathbf{c}^{(m)}, \mathcal{S}_m\right), \qquad \mathbf{c}^{(m)} = \left(\mathbf{R}^{\mathrm{ref}}, \mathbf{A}^{(m)}, \mathbf{L}^{\mathrm{vtp}}\right). \tag{13}$$

Training proceeds in two phases: the first learns source-destination flow maps, while the second performs on-policy chunked rollouts with distribution matching and the flow-map objective.

Before streaming generation, the reference image is encoded as one clean latent and prefilled as a dedicated sink chunk. Every denoising window thereafter contains exactly three new latent positions; no latent from the preceding generated chunk is repeated:

$$\begin{aligned} \mathcal{S}_0 &= \mathrm{CLEANCache}_\theta(\mathbf{R}^{\mathrm{ref}}), \\ \mathbf{X}^{(m)} &= \mathbf{Z}_{1:3}^{(m)}, \quad m \geq 0. \end{aligned} \tag{14}$$

The reference latent also remains available as an image condition for each generated chunk, while the acoustic condition covers only its three new positions. Noisy intermediate states are never written to the persistent cache. After denoising, a separate forward pass at $t = r = 0$ extracts clean K/V for those three positions. The reference sink remains available throughout generation; the newest three generated chunks retain full-resolution K/V, while older positions use spatially compressed K/V with a bounded history. Temporal RoPE uses ordinary absolute positions for generated chunks and places the reference sink at a distance of 30 latent positions from the current chunk.

Once the VTP is generated, speech and video are streamed using the same acoustic units. Every six units at 12.5 Hz produce a 0.48-second waveform segment and condition 12 video frames at 25 FPS:

$$\mathbf{U}_{6m+1:6m+6} \longrightarrow V_{12m+1:12m+12}. \tag{15}$$

The reference prefill is internal and is not emitted as a repeated frame. A final incomplete block is padded for computation and cropped to its valid duration.

### 3.5 Audio-Video Decoder

The generated video latents are converted to frames by the 3D VAE decoder of Wan2.1-T2V-1.3B (Wan et al., 2025). During streaming inference, the decoder carries its feature state across chunks so that newly generated latents can be decoded incrementally. The generated multi-codebook speech units are converted to waveform audio by the Qwen3-TTS Tokenizer decoder (Hu et al., 2026). Both pretrained decoders are kept fixed during training.

### 3.6 Training Objectives and Stages

The core Ex-Omni-2D framework uses three pathway-specific supervision signals: an autoregressive language loss over the VTP and response protocol, a multi-codebook speech-generation loss conditioned on the response span, and a flow-matching video loss conditioned on the reference image, teacher-forced speech units, and VTP. For efficient incremental deployment, the frozen full-sequence generator then supervises the streaming adaptation through the flow-map and on-policy distribution-matching procedure described in Section 3.4 and Appendix C.

Training uses three core stages followed by one deployment-oriented distillation stage. **Stage 1: Speech Interface Alignment** uses approximately 800K ASR and 1M TTS examples to train the Speech Projector and Speech Generator while keeping the LLM frozen. **Stage 2: Omni-modal Response Adaptation** updates the LLM, Speech Projector, and Speech Generator with InstructS2S-200K (Fang et al., 2025a) and OmniCharacter (Zhang et al., 2025a), introducing the structured VTP–response protocol and adapting the shared speech interface to interactive responses. Because these dialogue datasets do not contain target response videos, Qwen3-VL-235B-A22B-Instruct

constructs the target VTP from the reference image, dialogue context, and target response using the five-field template in Appendix D. **Stage 3: Avatar Video Realization** operates on about 140K filtered SpeakerVid clips. Each clip is converted into an avatar-response record containing a reference frame, a five-field VTP annotation, aligned multi-codebook speech units, and the target video. These records train the full-sequence Video Generator to realize the semantic and acoustic-temporal interfaces produced by the response pathway. The resulting model is frozen as the distillation target. **Stage 4: Streaming Adaptation** then trains the deployment-oriented streaming version through Phase I flow-map learning and Phase II on-policy distribution matching. We report the streaming configuration with eight denoising steps in the main experiments. At inference time, VTP and speech units reconnect the pathways without requiring large-scale dialogues paired with response videos. Data construction, optimization, and inference configurations are provided in Appendix C.

Table 1: Audio-video comparison across three generation paradigms. DD indicates how often substantial motion occurs. A higher score does not establish whether that motion is appropriate.

| Method | Dialogue | Ref. Voice | Parameters | | | Audio Quality | | | Video Quality | | | A-V Sync |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LLM | Speech Gen. | Video Gen. | PQ ↑ | CU ↑ | SIM ↑ | SC ↑ | IQ ↑ | DD | Sync-C ↑ |
| Video Generation | | | | | | | | | | | | |
| EchoMimic | ✗ | ✗ | - | - | 1.3B | - | - | - | **97.87** | 54.37 | **74.00** | 4.82 |
| StableAvatar-1.3B | ✗ | ✗ | - | - | 1.3B | - | - | - | 97.40 | 67.51 | 42.00 | 3.38 |
| OmniAvatar-1.3B | ✗ | ✗ | - | - | 1.3B | - | - | - | 97.74 | 66.65 | 15.50 | **5.64** |
| FantasyTalking | ✗ | ✗ | - | - | 1.3B | - | - | - | 96.78 | 64.21 | 6.50 | 3.43 |
| Video-Audio Generation | | | | | | | | | | | | |
| UniVerse-1 | ✗ | ✗ | - | 3.5B | 1.3B | 4.16 | 3.98 | - | 97.35 | 67.72 | 34.00 | 1.75 |
| UniAVGen | ✗ | ✓ | - | 1.4B | 5B | - | - | 0.291 | 96.07 | **68.02** | 55.00 | 4.15 |
| Dialogue-Based Omni-Modal Response Generation | | | | | | | | | | | | |
| Ex-Omni-2D (full-sequence) | ✓ | ✓ | 8B | 0.6B | 1.3B | **7.53** | **7.06** | **0.417** | 94.62 | 67.31 | 72.00 | 4.95 |
| Ex-Omni-2D (streaming) | | | | | | | | | 93.91 | 61.15 | 48.00 | 4.00 |

## 4 EXPERIMENTS

**Benchmarks.** We evaluate audio-video generation on the 200-query CommonEval speech-QA split from VoiceBench (Chen et al., 2024a), pairing each query with fixed reference image and speech conditions sampled from the held-out SpeakerVid test split (Zhang et al., 2025b). Spoken question answering uses AlpacaEval, CommonEval, and BBH from VoiceBench; multi-turn dialogue quality uses the 400-dialogue OmniCharacter benchmark (Zhang et al., 2025a). For cascaded audio-video baselines, Qwen2.5-Omni-7B (Xu et al., 2025) provides the shared upstream response, while each method is evaluated through the modalities supported by its public interface. Full baseline lists and modality-specific comparison details are provided in Appendix E.

**Metrics.** We report audio quality (PQ, CU, SIM), video quality (SC, IQ, DD), audio–video synchronization (Sync-C), VoiceBench speech-QA scores, and OmniCharacter dialogue scores. Detailed metric definitions and scoring protocols are provided in Appendix E.

**Audio-Video Comparison Results.** Table 1 compares video generation, joint video-audio generation, and dialogue-based omni-modal response generation under the protocol above. Among the downstream baselines, EchoMimic obtains the highest subject consistency (SC 97.87) and largest motion incidence by DD (74.00), UniAVGen obtains the highest imaging quality (IQ 68.02), and OmniAvatar-1.3B obtains the strongest synchronization (Sync-C 5.64). Our primary full-sequence Ex-Omni-2D result obtains SIM 0.417, DD 72.00, and Sync-C 4.95 while additionally generating the dialogue response, personalized speech, and query-derived visual plan. This configuration does not dominate every standalone rendering metric, but demonstrates the complete dialogue-to-video capability. The streaming configuration has lower video quality and synchronization scores than the full-sequence configuration while enabling incremental output. Appendix Table 13 studies the effect of replacing the cascaded baselines' upstream controller with Ex-Omni-2D.

Table 2: Full multi-turn dialogue evaluation on OmniCharacter. KE: Knowledge-Exposure, KA: Knowledge-Accuracy, KH: Knowledge-Hallucination, PB: Persona-Behavior, PU: Persona Utterance, Flu.: Fluency, Coh.: Coherency, Cons.: Consistency, HL: Human-Likeness, CS: Communication Skill, ED: Expression Diversity, and Emp.: Empathy.

| Model | Character Consistency | | | | | Conversational Ability | | | Role-playing Attractiveness | | | | Avg. ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | KE ↑ | KA ↑ | KH ↑ | PB ↑ | PU ↑ | Flu. ↑ | Coh. ↑ | Cons. ↑ | HL ↑ | CS ↑ | ED ↑ | Emp. ↑ | |
| ChatGLM3-6B | 2.016 | 2.792 | 2.704 | 2.455 | 2.812 | 3.269 | 3.647 | 3.283 | 3.064 | 2.932 | 1.969 | 2.993 | 2.828 |
| Baichuan2-7B | 1.813 | 2.849 | 2.929 | 2.830 | 3.081 | 3.551 | 3.894 | 3.827 | 3.670 | 2.728 | 2.115 | 2.984 | 3.023 |
| InternLM-7B | 1.782 | 2.800 | 2.781 | 2.719 | 3.016 | 3.527 | 3.823 | 3.744 | 3.546 | 2.622 | 2.070 | 2.897 | 2.944 |
| CharacterGLM-6B | 1.640 | 2.819 | 2.738 | 2.301 | 2.969 | 3.414 | 3.717 | 3.737 | 3.738 | 2.265 | 1.966 | 2.937 | 2.853 |
| Llama-3.1-8B | 2.197 | 2.701 | 2.615 | 3.130 | 2.704 | 3.059 | 3.477 | 3.071 | 2.922 | 2.934 | 2.634 | 2.759 | 2.850 |
| Qwen-7B | 1.956 | 2.728 | 2.633 | 2.605 | 2.780 | 3.187 | 3.564 | 3.229 | 3.036 | 2.791 | 2.052 | 2.838 | 2.783 |
| Qwen2-7B-Instruct | 1.966 | 2.537 | 2.412 | 2.313 | 2.436 | 2.864 | 3.171 | 2.743 | 2.655 | 2.612 | 1.867 | 2.654 | 2.519 |
| Qwen2.5-7B-Instruct | 2.172 | 3.012 | 2.887 | 3.526 | 2.879 | 3.321 | 3.657 | 3.326 | 3.278 | 3.127 | 2.728 | 3.217 | 3.094 |
| Qwen3-8B | **2.722** | 3.199 | 2.950 | **3.565** | 3.044 | 3.497 | 3.795 | 3.318 | 3.148 | **3.566** | **3.030** | 3.338 | 3.264 |
| OmniCharacter | 2.230 | 3.040 | 2.918 | 3.531 | 2.988 | 3.369 | 3.768 | 3.410 | 3.374 | 3.261 | 3.002 | 3.187 | 3.173 |
| Qwen2.5-Omni-7B | 2.462 | 3.266 | 2.943 | 3.148 | 3.070 | 3.692 | 3.951 | 3.617 | 3.486 | 3.481 | 2.502 | **3.474** | 3.258 |
| **Ex-Omni-2D** | 2.076 | **3.410** | **2.994** | 2.911 | **3.354** | **3.812** | **4.100** | **3.902** | **3.893** | 3.362 | 2.239 | 3.341 | **3.283** |

**Multi-turn Dialogue Quality.** As shown in Table 2, we use OmniCharacter (Zhang et al., 2025a) to evaluate multi-turn dialogue quality: whether the model remains fluent, coherent, and consistent when responding with access to the preceding conversation. Ex-Omni-2D obtains the highest reported Fluency, Coherency, and Consistency values among the compared methods, averaging 3.938 across the three metrics compared with 3.537 for its Qwen3-8B backbone. Its twelve-dimension average is 3.283, close to Qwen3-8B at 3.264; the metric-wise results include both increases and decreases relative to Qwen3-8B, showing that visual-response training changes the capability profile. Appendix Table 12 directly evaluates the effect of providing dialogue history.

Table 3: Speech QA comparison. * indicates results reproduced using open-source code.

| Model | AlpacaEval ↑ | CommonEval ↑ | BBH ↑ |
|---|---|---|---|
| Qwen2.5-Omni | **4.49** | **3.93** | **60.80*** |
| Moshi | 2.01 | 1.60 | 47.40 |
| VITA-1.5 | 4.21 | 3.66 | 55.30 |
| Mini-Omni2 | 2.32 | 2.18 | 46.40 |
| SLAM-Omni | 1.90 | 1.79 | 48.80 |
| Ex-Omni-2D | 4.28 | 3.71 | 58.70 |

Table 4: Inference Latency Analysis. FAC/FVC: time to first audio/video chunk output.

| Metric | Full-seq. | Streaming |
|---|---|---|
| Text TTFT (s) ↓ | 0.057 | 0.066 |
| Speech TTFC (s) ↓ | 0.206 | 0.242 |
| Video TTFC (s) ↓ | 30.692 | 0.714 |
| FAC (s) ↓ | 5.644 | 5.986 |
| FVC (s) ↓ | 39.189 | 6.641 |
| Text TPS (token/s) ↑ | 37.725 | 36.632 |
| Speech TPS (unit/s) ↑ | 30.958 | 30.079 |
| Video FPS (frame/s) ↑ | 3.556 | 46.217 |

**Speech QA Comparison Results.** As shown in Table 3, Ex-Omni-2D ranks second on three VoiceBench tasks, scoring 4.28 on AlpacaEval, 3.71 on CommonEval, and 58.70 on BBH. The gaps to Qwen2.5-Omni are 0.21, 0.22, and 2.10 points, respectively, while Ex-Omni-2D outperforms every other listed speech-capable model. These results show that the framework retains competitive spoken-question-answering ability while also producing explicit visual plans, personalized speech, and video responses. The VTP ablation in Table 6 shows that visual-planning supervision incurs a measurable cost on CommonEval and BBH. The strong overall ranking therefore reflects a trade-off across response capabilities.

**Inference Latency Analysis.** After adapting Ex-Omni-2D to the vLLM-Omni inference framework, we compared its full-sequence and streaming configurations under warmed conditions on 12 matched CommonEval requests. Table 4 shows that streaming reduces the mean time to first video output (FVC) from 39.189 to 6.641 seconds, a 32.548-second reduction, while time to first audio output (FAC) remains close to six seconds in both settings (5.644 versus 5.986 seconds). Video generation throughput rises from 3.556 to 46.217 frames per second. The main latency gain lies in making the visual response available soon after speech begins: the gap between mean FAC and mean FVC falls from 33.545 to 0.655 seconds. Streaming begins video generation after the first six speech units and emits 12-frame chunks; full-sequence generation waits for the complete speech-unit sequence and emits a preview of up to 125 frames. The faster visual response is accompanied by lower video quality

and synchronization scores in Table 1. Text TTFT may include VTP tokens before the user-facing reply, and Video FPS measures generation throughput rather than uninterrupted playback. Appendix Table 17 reports results by resolution, while Appendix F examines the reference sink.

**Effect of VTP and Reference Speech.** We ablate response-specific VTP and personalized reference speech on the same 200 CommonEval samples used in Table 1. Replacing the generated VTP with a fixed neutral plan lowers SC from 94.62 to 93.58 and Sync-C from 4.95 to 4.65 when the reference speech is held fixed; the same trend appears with the fixed public reference utterance. These results suggest that VTP conditioning contributes to subject consistency and audio–video synchronization, although these metrics do not establish whether the motion is semantically appropriate to the dialogue. DD increases when the response-specific VTP is removed, but DD measures the prevalence of substantial motion rather than its quality. Replacing the matched reference speech with a fixed public utterance raises PQ and CU from 7.53 and 7.06 to 8.11 and 7.45, while SIM against the original speaker falls from 0.417 to 0.015. Together, these ablations reveal distinct trade-offs: the response-specific VTP improves subject consistency and synchronization despite lower DD, while the matched reference speech preserves speaker similarity despite lower PQ and CU in this comparison.

Table 5: Complete ablation of response-specific VTP and personalized reference speech. SIM is evaluated against each sample's original matched reference speech.

| Setting | Specific VTP | Personalized Speech | PQ ↑ | CU ↑ | SIM ↑ | SC ↑ | IQ ↑ | DD | Sync-C ↑ |
|---|---|---|---|---|---|---|---|---|---|
| Full | ✓ | ✓ | 7.53 | 7.06 | **0.417** | 94.62 | **67.31** | 72.00 | **4.95** |
| w/o response-specific VTP | ✗ | ✓ | 7.53 | 7.06 | **0.417** | 93.58 | 67.26 | **81.50** | 4.65 |
| w/o personalized ref. speech | ✓ | ✗ | **8.11** | **7.45** | 0.015 | **94.76** | 67.22 | 70.50 | 4.87 |
| w/o both | ✗ | ✗ | **8.11** | **7.45** | 0.015 | 93.64 | 67.16 | 72.50 | 4.65 |

Table 6: Effect of VTP supervision on speech QA. Both settings use the same response-generation evaluation protocol.

| LLM Training Target | CommonEval ↑ | BBH ↑ |
|---|---|---|
| Response only (w/o VTP) | **3.82** | **61.10** |
| VTP + response (ours) | 3.71 | 58.70 |

Table 7: Controlled upstream-source comparison.

| Downstream | Upstream | SC ↑ | IQ ↑ | Sync-C ↑ |
|---|---|---|---|---|
| OmniAvatar-1.3B | Qwen2.5-Omni | 97.74 | 66.65 | **5.64** |
| | Ex-Omni-2D | **98.19** | **66.89** | 5.48 |
| UniAVGen | Qwen2.5-Omni | **96.07** | **68.02** | 4.15 |
| | Ex-Omni-2D | 95.92 | 65.75 | **5.07** |

**Effect of VTP Supervision.** We isolate the effect of supervising the LLM to emit a structured VTP before its user-facing response. As shown in Table 6, response-only training obtains 3.82 on CommonEval and 61.10 on BBH, whereas VTP-plus-response training obtains 3.71 and 58.70, decreases of 0.11 and 2.40 points, respectively. Thus, the explicit visual-planning interface incurs a cost on standalone spoken question answering and reasoning in this comparison. This ablation quantifies the language-side trade-off of VTP supervision; the effect of using the resulting VTP to condition video generation is shown separately in Table 5.

**Effect of the Upstream LLM.** As shown in Table 7, replacing Qwen2.5-Omni with Ex-Omni-2D as the upstream model, while keeping the downstream renderer and reference conditions fixed, produces mixed and renderer-dependent changes. For OmniAvatar, Ex-Omni-2D slightly improves SC and IQ but does not improve Sync-C. For UniAVGen, it substantially improves Sync-C, while SC and IQ decrease. Moreover, OmniAvatar retains the highest Sync-C among the tested renderers under either upstream source. These results show that the upstream LLM and its generated speech can affect video generation, but changing the LLM alone does not yield consistent improvements across renderers and metrics. Our advantage instead stems from the native coupling of response-aware VTP semantics and frame-aligned speech units with a video generator specifically trained to consume them.

**Visualization Analysis.** Figure 2 presents generated examples. Panel (a) demonstrates that Ex-Omni-2D can generate visually coherent and appealing videos across different dialogue scenarios. In Panel (b), we manually modify the `emotion` field of the VTP to examine emotion controllability.




Figure 2: Qualitative visualization of Ex-Omni-2D. **(a)** General VTP-guided omni-modal responses with head motion, facial expressions, and gestures. **(b)** Affect realization in dialogue-generated video responses with explicit emotion descriptions in the model-generated VTP.

These edits produce visible expression changes in some cases, but the requested emotion may also appear weak or fail to emerge. This shows that the VTP acts as high-level semantic guidance rather than a strict control signal. The final motion may be jointly determined by the VTP, frame-aligned speech condition, and text/audio CFG scales.

## 5 Conclusion

Ex-Omni-2D extends omni-modal dialogue from spoken replies to visible, synchronized responses. The LLM generates a structured visual plan and response text, while the Speech Generator produces multi-codebook speech units from response-side representations. The plan gives the avatar generator dialogue-specific scene and motion cues, while the units provide the timing shared by speech and video. This interface lets the speech, dialogue, and avatar pathways use the supervision available to each one, then reconnects them at inference. The full-sequence video generator provides the primary quality results; its reference-sink streaming adaptation generates speech-aligned chunks incrementally using bounded multi-resolution KV history for latency-sensitive settings.

## AI USE STATEMENT

Generative AI tools were used to assist with source-code review and language polishing. All AI-assisted content was checked against the implementation and experimental results and was reviewed by the authors.

## ETHICS STATEMENT

Our system generates personalized talking-avatar videos from reference images and speech, and may be misused for impersonation or deceptive content. Responsible deployment should require authorization for reference media, clearly disclose synthetic outputs, and adopt appropriate provenance and access controls. Potential biases across identities, languages, and accents should also be evaluated before deployment.

## REPRODUCIBILITY STATEMENT

The paper and appendix provide the model architecture, training objectives, implementation details, evaluation protocols, and hyperparameters. Code is provided in the supplementary material to facilitate reproduction.

## REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, Guangming Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, Yifei Wu, Yongjie Lyu, Ziping Ma, Zipeng Feng, Zhijiang Fang, Zhihao Qiu, Ziyuan Huang, and Zhengyu He. Ming-omni: A unified multimodal model for perception and generation. *CoRR*, abs/2506.09344, 2025.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5723–5738, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS 2020*, 2020.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, abs/2309.16609, 2023. doi: 10.48550/ARXIV.2309.16609.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

Gary Bradski. *The Opencv Library*, volume 25. 11 2000.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

Brandon Castellano. Pyscenedetect. URL https://www.scenedetect.com. Video Cut Detection and Analysis Tool.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, et al. Slam-omni: Timbre-controllable voice interaction system with single-stage training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2262–2282, 2025a.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *CoRR*, abs/2505.20156, 2025b.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024a.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditioning, 2024b.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911, June 2025.

J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *CoRR*, abs/2410.00037, 2024.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, 2025a.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni 2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 18617–18629. Association for Computational Linguistics, 2025b.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Advances in Neural Information Processing Systems*, 38:75300–75320, 2026.

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

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. *CoRR*, abs/2508.18621, 2025.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuchao Gu, Guian Fang, Yuxin Jiang, Weijia Mao, Song Han, Han Cai, and Mike Zheng Shou. Anyflow: Any-step video diffusion model with on-policy flow map distillation. *arXiv preprint arXiv:2605.13724*, 2026.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890, 2024.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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, June 2026.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Lindevs. Yolov8-face. https://github.com/lindevs/yolov8-face. Accessed: 2026-08-11.

Haohe Liu, Gael Le Lan, Xinhao Mei, Zhaoheng Ni, Anurag Kumar, Varun Nagaraja, Wenwu Wang, Mark D. Plumbley, Yangyang Shi, and Vikas Chandra. Syncflow: Toward temporally aligned joint audio-video generation from text, 2024.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. 2026.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Xiaobo Xia, Hamid Alinejad-Rokny, Fei Huang, and Min Yang. Openomni: Advancing open-source omnimodal large language models with progressive multimodal alignment and real-time emotional speech synthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 48855–48876. Curran Associates, Inc., 2023.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, July 2024. Accessed: 2026-09-01.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

Ziqiao Peng, Yi Chen, Yifeng Ma, Guozhen Zhang, Zhiyao Sun, Zixiang Zhou, Youliang Zhang, Zhengguang Zhou, Zhaoxin Fan, Hongyan Liu, Yuan Zhou, Qinglin Lu, and Jun He. Actavatar: Temporally-aware precise action control for talking avatars. *CoRR*, abs/2512.19546, 2025.

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*, pp. 484–492, 2020. ISBN 9781450379885.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10219–10228. IEEE, 2023.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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, June 2026.

Kim Sung-Bin, Jeongsoo Choi, Puyuan Peng, Joon Son Chung, Tae-Hyun Oh, and David Harwath. Voicecraft-dub: Automated video dubbing with neural codec language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14623–14632, October 2025.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Meituan LongCat Team. Longcat-flash-omni technical report. *CoRR*, abs/2511.00279, 2025.

Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347, pp. 402–419. Springer, 2020.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *CoRR*, abs/2508.08248, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *CoRR*, abs/2509.06155, 2025a.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9891–9900, 2025b.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *International Conference on Computer Vision (ICCV)*, 2023.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming. *CoRR*, abs/2408.16725, 2024a.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *CoRR*, abs/2410.11190, 2024b.

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. *CoRR*, abs/2503.20215, 2025.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024a. doi: 10.48550/ARXIV.2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024b. doi: 10.48550/ARXIV.2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4210–4220, 2023.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *CoRR*, abs/2412.02612, 2024.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechgpt-gen: Scaling chain-of-information speech generation. *CoRR*, abs/2401.13527, 2024. doi: 10.48550/ARXIV.2401.13527.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions, 2026.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, et al. Omnicharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26318–26331, 2025a.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. $S^3$fd: Single shot scale-invariant face detector, 2017.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8652–8661, June 2023b.

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. *arXiv preprint arXiv:2507.09862*, 2025b.

Lei Zhao, Linfeng Feng, Dongxu Ge, Rujin Chen, Fangqiu Yi, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Uniform: A unified multi-task diffusion transformer for audio-video generation, 2025.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*, 2023.

# A METHOD DETAILS

**Reference Sink and Cache Implementation.** The clean reference latent is prefilled as a dedicated one-latent chunk before generation and retained as a persistent sink. Each generated chunk samples noise for three new latent positions, with no repeated latent or acoustic overlap from the preceding chunk. Denoising reads the persistent state without writing noisy intermediate K/V. After the final transition, clean-cache forwards at $t = r = 0$ update the full-resolution and compressed-resolution history:

$$\begin{aligned} \widehat{\mathbf{x}}_0^{(m)} &= \mathrm{FlowMap}_\theta\left(\mathbf{x}_{t_0}^{(m)}; t_0 \to \cdots \to 0, \mathcal{S}_m\right), \\ \mathcal{S}_{m+1} &= \mathrm{CLEANCache}_\theta\left(\widehat{\mathbf{x}}_0^{(m)}, \mathcal{S}_m\right). \end{aligned} \tag{16}$$

Only the three newly generated positions are appended. The reference sink is stored separately at full resolution and is not duplicated in either rolling cache. Attention combines this sink with full-resolution K/V from the newest three generated chunks and spatially compressed K/V from older history, bounded to 96 latent positions. The full and compressed patch sizes are $1 \times 2 \times 2$ and $1 \times 4 \times 4$, respectively. K/V are stored before RoPE and rotated when read: generated positions retain their ordinary absolute temporal coordinates, whereas the sink's temporal RoPE coordinate is set to $\max(0, f_{\mathrm{cur}} - 30)$ for current latent position $f_{\mathrm{cur}}$. Compressed tokens use spatial RoPE frequencies pooled to their patch resolution.

**Rate-Aligned Streaming Chunks.** The speech interface produces multi-codebook units at 12.5 Hz, while video is decoded at 25 FPS; hence, each speech unit aligns with two video frames. We group six units into a 0.48-second block, corresponding to 12 output frames or three new VAE latent positions: $\frac{6}{12.5} = 0.48$ s. The reference prefill is an internal condition and is excluded from the emitted video chunks. The final partial block is padded for computation, after which both waveform and video outputs are cropped to the valid duration. The same chunk definition is used during streaming adaptation and inference.

**Optimization Objectives.** We train the framework with different objectives according to the available supervision in each stage. Let $\mathbf{o}^* = (o_1^*, \ldots, o_K^*)$ denote the complete structured assistant sequence containing the target VTP $\mathbf{p}^*$ and user-facing response $\mathbf{y}^*$. The language-modeling loss supervises both spans autoregressively:

$$\mathcal{L}_{\mathrm{lm}} = -\sum_{k=1}^{K} \log p\left(o_k^* \mid o_{<k}^*, \mathbf{E}_{\mathrm{in}}\right). \tag{17}$$

The speech generator is conditioned only on hidden states and text tokens from the response span. Let $\mathbf{U}^* \in \mathbb{N}^{N \times C}$ denote the target multi-codebook sequence. Its loss includes the autoregressive first codebook and the residual codebooks predicted within each acoustic frame:

$$\mathcal{L}_{\mathrm{sp}} = -\sum_{n=1}^{N} \sum_{c=1}^{C} \log p\left(u_{n,c}^* \,\middle|\, \begin{matrix} \mathbf{U}_{<n}^*, \mathbf{u}_{n,<c}^*, \\ \mathbf{H}_{\mathbf{y}}^{\ell}, \mathbf{y}^*, \mathbf{s}^{\mathrm{ref}} \end{matrix}\right). \tag{18}$$

The VTP requires no separate regression loss because it is directly supervised as part of $\mathcal{L}_{\mathrm{lm}}$. During video training, the target VTP is serialized and encoded by the frozen video text encoder to obtain $\mathbf{L}^{\mathrm{vtp}}$. Following the flow-matching parameterization, the denoising model predicts the flow target $\boldsymbol{\epsilon} - \mathbf{x}_0$ from the noised latent and the three video conditions:

$$\mathcal{L}_{\mathrm{vid}} = \left\| D_\theta(\mathbf{x}_\sigma, \sigma; \mathbf{R}^{\mathrm{ref}}, \mathbf{A}, \mathbf{L}^{\mathrm{vtp}}) - (\boldsymbol{\epsilon} - \mathbf{x}_0) \right\|_2^2. \tag{19}$$

Across the heterogeneous training stages, the framework-level objective can be summarized as

$$\mathcal{L} = \mathcal{L}_{\mathrm{lm}} + \mathcal{L}_{\mathrm{sp}} + \mathcal{L}_{\mathrm{vid}}, \tag{20}$$

where each term is enabled for the pathway and supervision available in the current batch, allowing the stages to use speech, dialogue, and video data without requiring all three annotations in every example.

## B Data Construction

**ASR and TTS Data.** Our Stage-1 speech data contain about 800K ASR examples and 1M TTS examples. The ASR portion consists of roughly 300K examples from LibriSpeech (Panayotov et al., 2015) and 500K randomly sampled examples from Emilia (He et al., 2024). For TTS supervision, we randomly sample 1M text examples from Emilia and synthesize speech with Qwen3-TTS (Hu et al., 2026), matching the tokenizer and generator family used by the speech pathway.

**Dialogue Data.** For Stage-2 omni-modal response adaptation, we use examples from InstructS2S-200K (Fang et al., 2025a) and OmniCharacter (Zhang et al., 2025a). InstructS2S-200K provides speech-to-speech conversational supervision, while OmniCharacter contributes character-grounded, multi-turn dialogue examples. We construct target VTPs offline with Qwen3-VL-235B-A22B-Instruct: the model receives the reference image together with the dialogue context and target response, then fills the same five VTP fields used by the video pathway. These annotations supervise the structured `<thinking>`–`<response>` protocol without using target response videos. Together, the dialogue examples update the LLM, Speech Projector, and Speech Generator, adapting the shared speech interface toward interactive and character-consistent responses.

**Video Generation Data.** For the avatar-realization pathway in Stage 3, we use SpeakerVid (Zhang et al., 2025b) as the source video dataset. We apply the automated filtering pipeline described below, obtaining about 140K training clips. For every retained clip, Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025) constructs a five-field VTP describing the first frame, scene, emotion, movement style, and motion details using the prompt in Appendix D. We then tokenize its aligned speech with the Qwen3-TTS Tokenizer (Hu et al., 2026) to obtain the multi-codebook video condition. Each retained clip becomes an avatar-realization record for training the diffusion model to map a reference image, VTP annotation, and aligned speech units to target video.

Table 8: Filtering pipeline for avatar-video training data.

| Step | Filter | Model or Algorithm | Main Acceptance Rule |
|---|---|---|---|
| 2a | Scene continuity | PySceneDetect (Castellano) AdaptiveDetector | No scene cut |
| 2b | Face visibility | YOLOv8x-face (Lindevs) + IoU tracking | One visible face with $\geq 98\%$ track coverage |
| 2c | Pose and motion | YOLOX-L (Ge et al., 2021) + DWPose (Yang et al., 2023) | Complete facial landmarks and stable non-static motion |
| 2d | Audio–visual sync | S3FD (Zhang et al., 2017) + SyncNet (Chung & Zisserman, 2016; Prajwal et al., 2020) | Sync-C $\geq 7.0$ and Sync-D $\leq 10.0$ |
| 2e | Visual quality | DOVER (Wu et al., 2023) + codec metadata | Height $\geq 480$ and fused DOVER $\geq 0.5$ |
| 2f | Exposure and flicker | OpenCV (Bradski, 2000) | No prolonged under/overexposure or strong flicker |
| 2g | Sharpness | HyperIQA (Su et al., 2020) | HyperIQA $\geq 55$ |

**SpeakerVid Filtering and Record Construction.** Audio segmentation first produces an aligned 16-kHz waveform and a 25-FPS video clip. We then apply the several filters in Table 8, retaining clips that pass every check. For each retained clip, the first frame serves as the appearance reference; Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025) annotates the five-field VTP, and the Qwen3-TTS Tokenizer (Hu et al., 2026) converts the aligned speech into 16 codebook streams at 12.5 Hz. The final diffusion record contains the target video, reference frame, VTP, and speech tokens. At 25 FPS, each speech-token frame aligns with exactly two video frames, supplying the conditions required by the avatar generator.

## C Implementation Details

**Framework Training Strategy.** We use three data-oriented stages followed by one deployment-oriented distillation stage.

**Stage 1: Speech Interface Alignment.** We first use approximately 800K ASR and 1M TTS examples to establish the bidirectional speech interface while freezing the LLM, Speech Encoder,

and Vision Encoder. ASR batches update the Speech Projector, grounding speech inputs in the LLM representation space. TTS batches update the Qwen3-TTS-initialized (Hu et al., 2026) Speech Generator, teaching it to predict native multi-codebook units from response-side representations. Thus, Stage 1 updates the Speech Projector and Speech Generator, with each batch activating only the branch for which it provides supervision.

**Stage 2: Omni-modal Response Adaptation.** We then jointly update the LLM, Speech Projector, and Speech Generator using examples from InstructS2S-200K (Fang et al., 2025a) and OmniCharacter (Zhang et al., 2025a). Qwen3-VL-235B-A22B-Instruct provides the target VTP for each dialogue example from the reference image, dialogue context, and target response. The LLM learns the structured `<thinking>`–`<response>` protocol, where the thinking block contains the VTP, while speech supervision remains restricted to the response span. This stage adapts the aligned speech interface toward interactive and character-consistent spoken replies without requiring a target response video.

**Stage 3: Avatar Video Realization.** Finally, we train the full-sequence avatar diffusion model on the 140K SpeakerVid-derived records described above. Each training example provides a reference image, an automatically annotated VTP, teacher-forced multi-codebook speech units, and a target video. The diffusion model is therefore optimized to realize the same semantic and acoustic-temporal interfaces that the response pathway produces at inference time. The interface-based decomposition lets the response and video pathways use their respective data sources: dialogue supervision teaches what to say and how to plan a response, whereas SpeakerVid supervision teaches how to render a supplied plan and speech-unit sequence as video.

**Stage 4: Streaming Adaptation.** After Stage 3, the full-sequence Video Generator is frozen as the distillation teacher. Stage 4 trains the streaming version as its student in two successive phases. Phase I learns source–destination flow maps from the full-sequence target, and Phase II performs on-policy chunked streaming rollouts with distribution matching while retaining the flow-map objective. Table 9 summarizes the optimization settings for all stages. For the video pathway, Stage 3 uses 125-frame SpeakerVid clips assigned to the $400 \times 720$, $720 \times 400$, and $720 \times 720$ buckets, while both Stage 4 phases use 121-frame windows with the same buckets.

**Hyperparameters.** We implement and train the framework with PyTorch 2.7.0 and CUDA 12.6. Each GPU has 16,896 CUDA cores, 132 streaming multiprocessors, and 141 GB of memory. Table 9 summarizes the stage-wise optimization settings and the two-phase streaming adaptation protocol; the corresponding data and trainable modules are specified in the stage descriptions above. Table 10 reports the inference configuration used for the main quality evaluation. Both configurations use empty negative prompts and deterministic dialogue-response decoding (temperature 0 and top-$p$ 1), while the main streaming results use one validation-selected reference-sink checkpoint.

Video clips are decoded at 25 FPS. We assign each clip to its nearest aspect-ratio bucket among $400 \times 720$, $720 \times 400$, and $720 \times 720$, resize isotropically, and center-crop to the exact bucket without padding. A random contiguous window of 125 frames is used for full-sequence training and 121 frames for streaming adaptation; short clips repeat their final frame. Window starts are aligned to the 12.5-Hz speech-token grid. For CommonEval, each sample uses a fixed reference image taken from the first frame and its speaker-matched reference speech. Reference speech is converted to mono, resampled to 24 kHz for the Speech Generator, and deterministically cropped to 3–10 seconds using seed 42; query speech and video-conditioning audio are mono at 16 kHz.

## D Prompt Templates

**VTP Annotation Prompt.** The following prompt is provided to Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025) together with each retained training video.

```
You are a meticulous video annotator for speech-driven human avatar generation data.

Watch the video carefully and describe visible appearance, speaking behavior, emotional expression, and human
motion in a generation-friendly manner.

Global annotation rules:
1. Describe only clearly visible and visually verifiable content.
```

Table 9: Unified optimization configuration. The upper panel compares the stage-wise device, batch, budget, learning-rate, and schedule settings; the lower panel reports the detailed hyperparameters of the two Stage 4 distillation phases.

| Setting | Stage 1 | Stage 2 | Stage 3 | Stage 4, Phase I | Stage 4, Phase II |
|---|---|---|---|---|---|
| GPUs | 8 | 8 | 24 | 40 | 40 |
| Batch | 16 | 1 | 1 | 1 | 1 |
| Budget | 1 epoch | 2 epochs | 10K steps | 6K steps | 3K steps |
| LR | Proj.: $1 \times 10^{-3}$; Gen.: $1 \times 10^{-4}$ | LLM: $1 \times 10^{-6}$; Proj./Gen.: $2 \times 10^{-5}$ | $2 \times 10^{-5}$ | $5 \times 10^{-5}$ | $2 \times 10^{-6}$ |
| Schedule | Cosine; 30% warmup | Cosine; 10% warmup | Constant | 1K-step warmup | No warmup |

**Stage 4 distillation details.** Streaming cache lengths below are measured in latent positions unless stated otherwise.

| Hyperparameter | Phase I | Phase II |
|---|---|---|
| Diffusion horizon / sigma shift | 1,000 / 5.0 | 1,000 / 5.0 |
| Streaming inference steps | 8 | 8 |
| Gate / destination-time type | 0.25 / $r$ | 0.25 / $r$ |
| Diffusion / consistency weights | 0.5 / 0.25 | 0.5 / 0.25 |
| Flow-map guidance / $\epsilon$ | 3.0 / 5.0 | 3.0 / 5.0 |
| DMD / co-training weights | – | 0.5 / 1.0 |
| Real-score guidance | – | 2.5 (CFG 3.5) |
| Text drop / bidirectional prob. | 0.1 / 0.1 | 0.1 / 0.0 |
| Supervised chunks / long-context ratio | 3 / 0.5 | 3 / 1.0 |
| New / overlapping latent positions | 3 / 0 | 3 / 0 |
| Reference sink / sink RoPE distance | 1 / 30 | 1 / 30 |
| Recent full-resolution chunks / compressed history | 3 / 96 | 3 / 96 |

Table 10: Video inference protocol used for the main quality experiments. Text/audio CFG lists the two guidance scales in that order. Streaming output length follows the generated speech duration.

| Setting | Full-sequence | Streaming |
|---|---|---|
| Temporal attention | Bidirectional | Block causal |
| Output frames | 125 | Speech aligned |
| Denoising steps | 50 | 8 |
| Scheduler | FlowMatch | FlowMap discrete |
| Training horizon / sigma shift | 1,000 / 5.0 | 1,000 / 5.0 |
| Text/audio CFG | 3.5 / 8.5 | 1.0 / 1.0 |
| Generation/prompt seed | 42 / 42 | 42 / 42 |
| Frame rate | 25 FPS | 25 FPS |

```
2. Use concise, concrete, generation-oriented language.
3. Prefer observable motion and appearance over abstract interpretation.

Avoid: inferred intentions or beliefs, invisible actions, story-like narration, cinematic language, or unsupported
semantic speculation.

Note:
1. Output ONLY XML-style tags in the exact order below.
2. Each tag must appear exactly once.
3. Do not add explanations or extra text outside the XML tags.

Template:
<first_frame_scene>
One sentence describing only the first frame: appearance, framing (e.g., close-up, medium shot, full body), pose,
facial expression, background, and lighting.
</first_frame_scene>

<scene>
One sentence describing the overall situation, atmosphere, and visible background activity when present.
</scene>

<emotion>
One or two emotion labels chosen from: neutral, happy, sad, angry, surprised, fearful, disgusted, calm, concerned,
confident, excited, nervous, thoughtful, engaged, amused.
Compound emotions are allowed using comma separation. If the emotion shifts during the video, use an arrow (e.g.,
neutral -> excited).
</emotion>

<movement_style>
```

```
A short phrase describing the high-level motion style, overall energy level, expressiveness, and vibe (e.g.,
"formal and restrained", "energetic and theatrical"). Do not describe specific body parts here.
</movement_style>

<motion_description>
One or two sentences detailing the specific physical kinematics: visible head, facial, hand, shoulder, torso, and
body movements, including their directions and approximate amplitudes.
</motion_description>
```

**Dialogue-data VTP Construction.** For Stage 2 dialogue data, Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025) receives the reference image, the available role or dialogue context, and the target response. The instruction asks it to fill the same five XML fields as above, but to describe a plausible response-specific visual plan rather than an observed target video. The first-frame field describes the supplied reference image, while the scene, emotion, movement style, and motion-description fields describe the intended speaking behavior for the target response. No target response video is provided for these dialogue examples.

**VTP Visual-Grounding Judge Prompt.** For the main protocol in Table 14, Gemini 3.1 Pro receives the following system prompt. The first supplied image is the reference image; the remaining inputs are chronological generated-video samples excluding frame 0.

```
You are a strict reference-image and video-to-plan alignment evaluator.
Judge only visible evidence in the supplied images. Do not assume that a plan is correct merely because it is
plausible. The first supplied image is the ORIGINAL REFERENCE IMAGE. Use it only to judge first_frame_scene. All
remaining images are chronological GENERATED-video samples; video frame 0, which repeats the reference image, has
been excluded. Use only these generated-video samples to judge scene, emotion, movement_style, and
motion_description.

For each of the five VTP fields, assign exactly one binary point:
- 1: the field's central visual meaning is visibly supported.
- 0: the central meaning is contradicted, absent, or too weak to verify.

Judge at field level, not by exact word matching. Do not fail an otherwise aligned field because of a minor
omitted prop, crop, paraphrase, or left/right ambiguity caused by camera viewpoint or mirroring. For a field with
several details, award the point when its principal subject, setting, affect, style, or action is preserved and
there is no major semantic contradiction.

For movement fields, require visible temporal evidence across generated-video samples. For first_frame_scene,
judge only the original reference image and focus on its principal subject, setting, and pose. Return JSON only.
```

## E Evaluation Protocol and Metric Details

**Baselines and Comparison Protocol.** Table 1 groups the audio-video comparison by supported interface: video-only renderers (EchoMimic (Chen et al., 2024b), StableAvatar-1.3B (Tu et al., 2025), OmniAvatar-1.3B (Gan et al., 2025), and FantasyTalking (Wang et al., 2025b)), video-audio generators (UniVerse-1 (Wang et al., 2025a) and UniAVGen (Zhang et al., 2026)), and the full-sequence and streaming configurations of Ex-Omni-2D. Table 2 compares against representative role-playing and dialogue models from the OmniCharacter protocol, including ChatGLM3-6B (GLM et al., 2024), Baichuan2-7B (Baichuan, 2023), InternLM-7B (Cai et al., 2024), CharacterGLM-6B (Zhou et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), OmniCharacter (Zhang et al., 2025a) and Qwen-family baselines (Bai et al., 2023; Yang et al., 2024b; 2025; 2024a; Xu et al., 2025). Table 3 compares with speech-enabled omni-modal systems on VoiceBench (Chen et al., 2024a): Qwen2.5-Omni (Xu et al., 2025), Moshi (Défossez et al., 2024), VITA-1.5 (Fu et al., 2026), Mini-Omni2 (Xie & Wu, 2024b), and SLAM-Omni (Chen et al., 2025a).

For audio-video baselines that require an upstream response, Qwen2.5-Omni-7B is used as the common response controller, whose parameter scale is close to our 8B LLM backbone. A2V renderers receive speech from this controller, so PQ and CU are not attributed to the downstream renderer; SIM is reported only when the evaluated interface supports reference speech. UniAVGen's main video metrics use A2V mode because its joint A-V mode supports only 5-second outputs, while its SIM score is obtained separately in RAVG reference-speech mode. All Ex-Omni-2D rows share the same generated speech and therefore have identical audio metrics. Full-sequence and streaming quality comparisons reuse the same text response, VTP, speech units, waveform, reference image, and reference speech, so their reported differences isolate the video generator. Table 1 evaluates practical capability under each supported interface, and Appendix Table 13 diagnoses the effect of changing the upstream response source while keeping each downstream renderer fixed.

**Audio Metrics.** We use Audiobox-Aesthetics (Tjandra et al., 2025) to report Production Quality (PQ) and Content Usefulness (CU). PQ assesses technical audio quality, including clarity, fidelity, dynamics, and audible artifacts. CU estimates the usefulness of an audio clip as source material for content creation; it does not measure the correctness or semantic completeness of the spoken answer. We additionally report speaker similarity (SIM) following the Seed-TTS evaluation protocol (Anastassiou et al., 2024): a WavLM-Large speaker-verification model extracts embeddings from generated and matched reference speech, and SIM is their cosine similarity.

**Video Metrics.** We use VBench (Huang et al., 2024) to report Subject Consistency (SC), Imaging Quality (IQ), and Dynamic Degree (DD). SC measures the consistency of the subject's appearance across frames using DINO features. IQ estimates frame-level imaging quality, including distortions such as blur, noise, and overexposure. DD uses optical flow to identify videos with substantial motion and reports their proportion. A higher DD indicates that more videos contain noticeable motion, but does not imply more natural or dialogue-appropriate motion. None of these metrics requires a paired ground-truth video.

**A-V Sync.** We use SyncNet confidence (Sync-C) (Chung & Zisserman, 2016; Prajwal et al., 2020) to evaluate whether mouth movements are temporally aligned with the generated speech. Sync-C measures the separation between the best audio-video alignment and competing temporal offsets, so higher values indicate more confident synchronization.

**Speech QA Metrics.** For spoken question answering, we use AlpacaEval (Li et al., 2023), CommonEval (Chen et al., 2024a), and BBH (Suzgun et al., 2022) from VoiceBench (Chen et al., 2024a) and follow the open-source VoiceBench evaluation code. AlpacaEval and CommonEval use GPT (OpenAI, 2024) to judge open-ended response quality on a five-point scale, where 5 is the best score; AlpacaEval emphasizes general instruction following, whereas CommonEval focuses on common spoken QA. BBH evaluates complex QA requiring multi-step reasoning, logical composition, and abstract thinking, and is reported as accuracy. Each BBH subtask then applies task-specific string matching or regular expressions to convert free-form responses into binary labels, such as yes/no answers, option choices, navigation decisions, truthfulness judgments, or sports-plausibility judgments. The final score is computed as the percentage of correct predictions.

## F Additional Analysis

**Backbone Capability Comparison.** To analyze how multimodal adaptation affects the language capability inherited from initialization, we compare Ex-Omni-2D with its Qwen3-8B (Yang et al., 2025) backbone under the same open-ended VoiceBench protocol. As shown in Table 11, Ex-Omni-2D scores 4.48 on AlpacaEval and 4.35 on CommonEval with text input, compared with 4.82 and 4.51 for Qwen3-8B. The relatively small reductions indicate that Ex-Omni-2D largely retains the backbone's text capability, while incurring a modest trade-off after multimodal adaptation. This degradation may result from interference when the shared language backbone is optimized across heterogeneous modalities and tasks.

Table 11: Backbone capability comparison

| Method | AlpacaEval ↑ | CommonEval ↑ |
|---|---|---|
| Qwen3-8B (text input) | **4.82** | **4.51** |
| Qwen3-8B (speech input) | N/A | N/A |
| Ours (text input) | 4.48 | 4.35 |
| Ours (speech input) | 4.28 | 3.71 |

**Multi-turn Context Ablation.** We isolate dialogue history on the 400-dialogue OmniCharacter test set. Both settings receive the benchmark Role Card and are scored on the final turn with the same CharacterEval reward model. The context-free setting receives only the final user query, whereas the full setting additionally receives all preceding question–answer turns. Table 12 shows higher scores in every evaluated dimension when dialogue history is provided, with especially large differences in persona behavior, persona utterance, expression diversity, and response consistency. Across the

twelve dimensions, the dialogue-level macro average increases from 2.769 (95% CI: 2.725–2.812) to 3.283 (3.242–3.322); the paired difference is 0.514 with a 95% bootstrap CI of 0.464–0.562. Under this evaluation protocol, preceding dialogue history is therefore associated with a consistent improvement over using only the final query.

Table 12: Multi-turn context ablation on OmniCharacter. KE: Knowledge-Exposure, KA: Knowledge-Accuracy, KH: Knowledge-Hallucination, PB: Persona-Behavior, PU: Persona Utterance, Flu.: Fluency, Coh.: Coherency, Cons.: Consistency, HL: Human-Likeness, CS: Communication Skill, ED: Expression Diversity, and Emp.: Empathy. Both settings retain the same Role Card and differ only in whether preceding dialogue turns are provided.

| Model | KE ↑ | KA ↑ | KH ↑ | PB ↑ | PU ↑ | Flu. ↑ | Coh. ↑ | Cons. ↑ | HL ↑ | CS ↑ | ED ↑ | Emp. ↑ | Avg. ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/o dialogue context | 1.798 | 2.864 | 2.461 | 2.035 | 2.583 | 3.461 | 3.743 | 3.410 | 3.198 | 2.987 | 1.535 | 3.145 | 2.769 |
| w/ dialogue context | **2.076** | **3.410** | **2.994** | **2.911** | **3.354** | **3.812** | **4.100** | **3.902** | **3.893** | **3.362** | **2.239** | **3.341** | **3.283** |

**Upstream Response Source Ablation.** Table 13 extends the controlled upstream comparison to all downstream generators. Within each pair, we keep the requests, reference conditions, downstream checkpoint, and evaluation protocol fixed, and change only the upstream source. Consistent with the main-text analysis, the effect is mixed and renderer-dependent. EchoMimic and OmniAvatar change only slightly; StableAvatar and FantasyTalking obtain lower Sync-C, whereas UniVerse-1 and UniAVGen obtain higher Sync-C at the cost of decreases in other metrics. FantasyTalking shows the largest SC drop, while UniAVGen shows the largest Sync-C increase. Nevertheless, OmniAvatar retains the highest Sync-C under both upstream sources. These results show that changing the upstream model affects individual scores but provides no consistent advantage across renderers. We therefore use Qwen2.5-Omni as the common upstream source for external generators in the main comparison and report these paired results as a sensitivity analysis.

Table 13: Ablation of the upstream response source for fixed downstream generators. A2V methods do not report upstream audio metrics. UniAVGen video metrics use its A2V mode because joint A-V generation is limited to 5-second outputs; its SIM is separately measured in RAVG reference-speech mode.

| Downstream Generator | Upstream Source | PQ ↑ | CU ↑ | SIM ↑ | SC ↑ | IQ ↑ | DD | Sync-C ↑ |
|---|---|---|---|---|---|---|---|---|
| EchoMimic | Qwen2.5-Omni | – | – | – | 97.87 | 54.37 | 74.00 | 4.82 |
| | Ex-Omni-2D | – | – | – | 97.90 | 54.63 | 72.00 | 4.88 |
| StableAvatar-1.3B | Qwen2.5-Omni | – | – | – | 97.40 | 67.51 | 42.00 | 3.38 |
| | Ex-Omni-2D | – | – | – | 97.43 | 67.56 | 38.50 | 2.87 |
| OmniAvatar-1.3B | Qwen2.5-Omni | – | – | – | 97.74 | 66.65 | 15.50 | 5.64 |
| | Ex-Omni-2D | – | – | – | 98.19 | 66.89 | 22.50 | 5.48 |
| FantasyTalking | Qwen2.5-Omni | – | – | – | 96.78 | 64.21 | 6.50 | 3.43 |
| | Ex-Omni-2D | – | – | – | 93.19 | 63.60 | 8.00 | 3.20 |
| UniVerse-1 | Qwen2.5-Omni | 4.16 | 3.98 | – | 97.35 | 67.72 | 34.00 | 1.75 |
| | Ex-Omni-2D | 4.11 | 3.74 | – | 97.21 | 67.60 | 29.00 | 1.94 |
| UniAVGen | Qwen2.5-Omni | – | – | 0.291 | 96.07 | 68.02 | 55.00 | 4.15 |
| | Ex-Omni-2D | – | – | 0.291 | 95.92 | 65.75 | 54.50 | 5.07 |

**VTP Grounding and Observational Alignment.** We additionally evaluate whether the generated VTP is visually grounded rather than merely structurally valid. On 200 CommonEval examples, a third-party multimodal judge assigns one binary score to each VTP field based on visible evidence. As shown in Table 14, the `first_frame_scene` field obtains 43.00%, lower than the other fields. We believe this is because that the field is evaluated against a single initial image, whereas the VTP describes an evolving visual response whose scene and motion may change over time. The remaining fields are evaluated using nine chronological samples from the generated video and achieve an average alignment of 85.38%, including 89.50% for emotion and 94.00% for movement style. Overall, the results show that Ex-Omni-2D generally translates its generated VTP into observable visual behavior. The first frame scene field obtains 43.00%, substantially lower than the remaining

fields. This indicates that the generated VTP does not always faithfully preserve the visual details of the supplied reference image in its first-frame description. In contrast, the remaining fields achieve an average alignment of 85.38% when evaluated against generated-video observations. We note that the realized video is jointly conditioned on the VTP and frame-aligned speech units, so the VTP serves as high-level semantic guidance rather than an exclusive control signal over the generated motion and appearance. Judge prompts are provided in Appendix D.

Table 14: VTP grounding and observational alignment on 200 CommonEval examples. First-frame is evaluated against the reference image; the remaining fields are evaluated on chronological samples from the generated video.

| First-frame | Scene | Emotion | Movement | Motion | Video Avg. |
|---|---|---|---|---|---|
| 43.00 | 84.50 | 89.50 | 94.00 | 73.50 | 85.38 |

**VTP Motion-Field Intervention.** To complement the observational alignment analysis above, we run a paired intervention that changes only the VTP motion fields. On 16 CommonEval examples, we fix the reference image, generated waveform, speech units, sampling configuration, random seed, and the first-frame, scene, and emotion VTP fields. We then generate two full-sequence videos per example by replacing only `movement_style` and `motion_description`: the restrained condition asks for subtle lip movement, small natural head motion, and minimal upper-body or hand movement, while the expressive condition asks for larger facial-expression changes, frequent head motion, and visible upper-body or hand gestures when the framing allows. We score the paired outputs with two local motion diagnostics: mean adjacent-frame gray-scale difference and mean optical-flow magnitude (Teed & Deng, 2020) over decoded frames. These metrics measure visible frame-to-frame motion and can also respond to artifacts or drift, so they are used as auxiliary diagnostics rather than semantic motion-quality scores. Table 15 shows that the expressive motion fields increase both diagnostics on average, and the expressive video has the larger value in 14 of 16 paired examples for both metrics. The paired bootstrap confidence intervals for the mean deltas are $[-0.0010, 0.0057]$ for frame difference and $[-0.0754, 0.3685]$ for optical-flow magnitude. These results provide evidence that VTP motion fields tend to bias the full-sequence generator toward stronger visible motion, although the bootstrap confidence intervals of the mean differences include zero.

Table 15: Paired VTP motion-field intervention on 16 short CommonEval examples. Only `movement_style` and `motion_description` are changed between conditions. "Expressive higher" reports the fraction of paired examples where the expressive condition has the larger metric value.

| Metric | Restrained | Expressive | Δ | Expressive higher |
|---|---|---|---|---|
| Frame difference | 0.0135 | 0.0162 | +0.0026 | 14/16 |
| Optical-flow magnitude | 0.7400 | 0.9071 | +0.1670 | 14/16 |

**Reference-Sink Ablation.** We isolate the effect of retaining the reference-image attention sink after the first generated chunk. The first generated chunk receives the reference-image latent and attends to the prefilled sink in both variants. For this ablation, from the second chunk onward, both use the previous chunk's final generated latent as the image condition. Figure 3 compares sink-off and sink-on frames at the same late time point within each example. The sink-on frames better retain the reference subject's facial and clothing details and the scene structure, whereas sink-off shows more pronounced appearance degradation after 13–15 streaming chunks. We quantify this with DINO (Zhang et al., 2023a) features: mean consistency averages similarity to the reference image and to the previous chunk endpoint. At chunk 1, the scores are nearly identical (0.9695 without the sink and 0.9696 with it). At chunk 16, among the 99 matched examples that reach that chunk, sink-on increases mean consistency from 0.8359 to 0.8743, a paired gain of 0.0385 (95% bootstrap CI: 0.0304–0.0453). Reference-image similarity rises from 0.6862 to 0.7625, a paired gain of 0.0763 (95% CI: 0.0625–0.0879), while adjacent-chunk similarity remains similar (0.9856 versus 0.9862). Thus the sink preserves reference appearance without sacrificing local continuity. It does not eliminate long-term drift: gradual background changes may be less noticeable during continuous playback than in these side-by-side frames. Our reference sink mitigates accumulated drift with a 1.3B video backbone, but fully controlling drift over long streaming sequences at this model scale

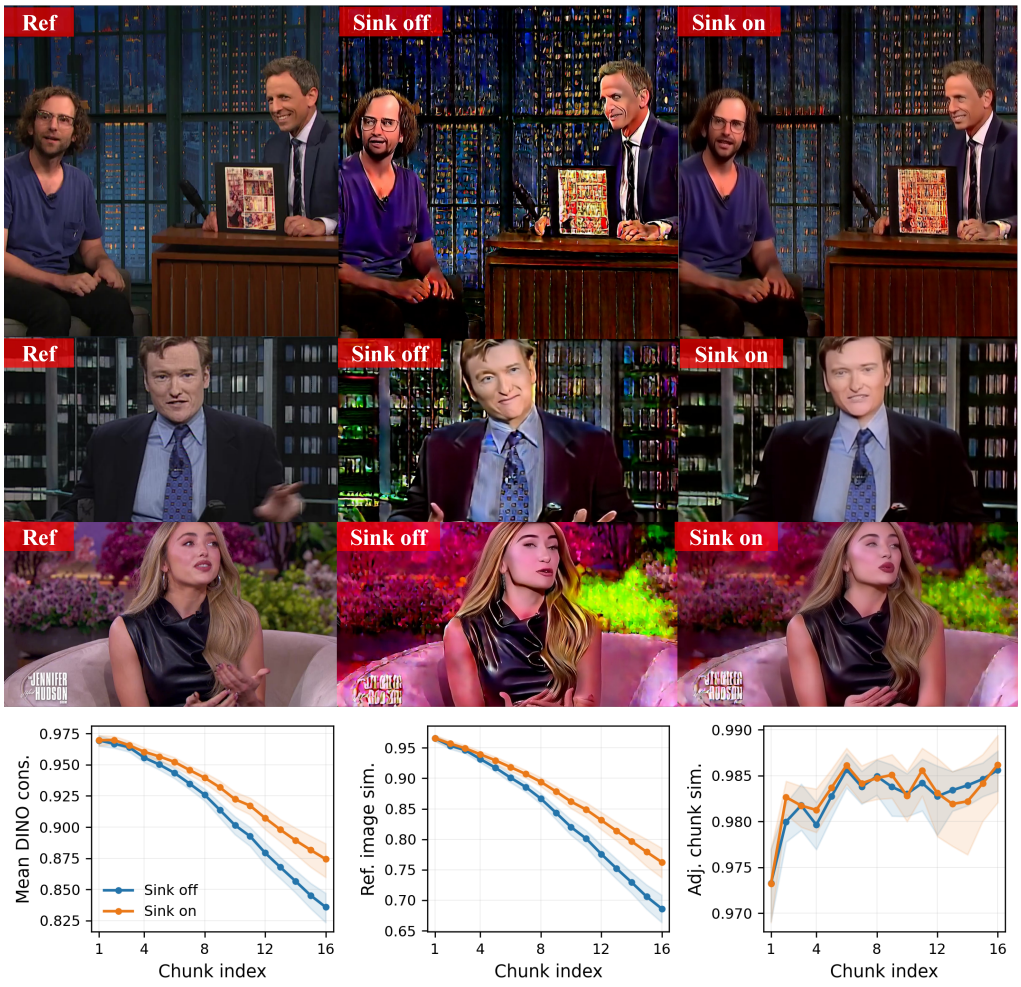


Figure 3: Reference-sink ablation under shared previous-chunk-latent conditioning from chunk 2 onward. **Top:** CommonEval examples. **Bottom:** three plots show mean DINO consistency, reference-image similarity, and adjacent-chunk similarity, from left to right. Shading indicates 95% bootstrap confidence intervals.

requires further research. Ki et al. (2026) also identifies long-horizon drift as a challenge for 1.3B streaming avatar generation.

**Ablation on Speech Conditioning Pathway.** As shown in Table 16, we compare waveform–wav2vec (Baevski et al., 2020) conditioning with native single- and 16-codebook interfaces on the same 200 CommonEval utterances. Waveform–wav2vec conditioning obtains Sync-C 5.83 but requires waveform decoding, resampling, and re-encoding. Single-codebook units reduce conditioning latency to 0.003 seconds but obtain Sync-C 2.07. The 16-codebook interface obtains Sync-C 4.95 with 0.011-second conditioning latency, making the measured trade-off explicit without requiring waveform reconstruction. The waveform baseline remains stronger for synchronization, but it conditions the video path on an already decoded and re-encoded audio signal; the unit interface is the lower-latency representation naturally emitted by the response model and is therefore used in the streaming pipeline.

**Full Inference Scaling Results.** Table 17 compares one-copy full-sequence and streaming inference on eight GPUs. Both modes use the same four matched CommonEval speech requests at each supported resolution, after one warmup per resolution. Full-sequence generation uses 50 denoising steps with CFG 4 and emits previews after each 125-frame window; streaming uses eight steps and

Table 16: Synchronization and conditioning latency of three video-conditioning interfaces. Latency starts from available speech units and ends before the first video denoising step. The waveform interface includes code2wav, in-memory resampling, and wav2vec encoding. Results are averaged over 200 CommonEval utterances on one GPU with the same hardware configuration.

| Method | Sync-C ↑ | Conditioning Latency (s) ↓ |
|---|---|---|
| Waveform + wav2vec | **5.83** | 0.051 |
| Single-codebook units | 2.07 | **0.003** |
| 16-codebook units | 4.95 | 0.011 |

emits 12-frame chunks. Full-sequence assigns six GPUs to sequence-parallel video generation, while streaming uses five DiT GPUs and a dedicated VAE GPU; both reserve one GPU each for the Thinker and speech pathway. Rectangular and overall rows pool the individual requests.

Table 17: Full inference results. All pools all three shapes (12 requests). The first five metrics are seconds. FAC/FVC measure first PCM/RGB availability from request arrival; full-sequence FVC uses the first preview window. Video TTFC and Video FPS start when video input is ready: all speech units for full-sequence, the first six for streaming. Text TPS, Speech TPS, and Video FPS are generated tokens/s, speech units/s, and frames/s.

| Resolution | Text TTFT | Speech TTFC | Video TTFC | FAC | FVC | Text TPS | Speech TPS | Video FPS |
|---|---|---|---|---|---|---|---|---|
| Full-sequence (50 steps; 8 GPUs) | | | | | | | | |
| 400 × 720 | 0.067 | 0.218 | 24.940 | 5.677 | 33.493 | 37.704 | 30.845 | 4.030 |
| 720 × 400 | 0.051 | 0.200 | 23.035 | 5.628 | 31.506 | 37.706 | 30.988 | 4.361 |
| 720 × 720 | 0.054 | 0.199 | 44.102 | 5.626 | 52.567 | 37.765 | 31.041 | 2.276 |
| **All** | 0.057 | 0.206 | 30.692 | 5.644 | 39.189 | 37.725 | 30.958 | 3.556 |
| Streaming (8 steps; 8 GPUs) | | | | | | | | |
| 400 × 720 | 0.067 | 0.241 | 0.580 | 5.963 | 6.479 | 36.821 | 29.431 | 51.900 |
| 720 × 400 | 0.066 | 0.243 | 0.572 | 6.040 | 6.555 | 36.218 | 30.399 | 53.149 |
| 720 × 720 | 0.065 | 0.241 | 0.991 | 5.955 | 6.889 | 36.856 | 30.408 | 33.602 |
| **All** | 0.066 | 0.242 | 0.714 | 5.986 | 6.641 | 36.632 | 30.079 | 46.217 |

**Human Preference Evaluation against Recent Methods.** We conduct a blind human preference study on 200 CommonEval requests, comparing Ex-Omni-2D with OmniAvatar and UniAVGen. We recruit five professional annotators with extensive experience in video-caption annotation and speech/ASR annotation. For each request, annotators rank the anonymized outputs in terms of lip synchronization and facial-motion naturalness, with ties allowed. Table 18 reports the mean rank and first-tier rate. OmniAvatar achieves the strongest preference on both criteria. Ex-Omni-2D and UniAVGen show complementary strengths: UniAVGen performs better on lip synchronization, while Ex-Omni-2D obtains a slightly better mean rank for facial naturalness. The two methods therefore show broadly comparable overall performance, but both remain behind OmniAvatar. This gap also reflects the advantage of OmniAvatar as a model specialized for audio-driven avatar rendering, whereas Ex-Omni-2D addresses unified dialogue, speech, and video response generation.

Table 18: Blind human comparison with recent methods. Each cell reports mean rank / first-tier rate (%). Lower mean rank is better; first-tier rates include ties.

| Method | Lip Sync | Facial Naturalness |
|---|---|---|
| OmniAvatar-1.3B | **1.290 / 75.6** | **1.396 / 64.9** |
| UniAVGen | 1.683 / 48.4 | 1.932 / 29.9 |
| Ex-Omni-2D | 1.755 / 42.6 | 1.866 / 32.5 |

**Human Preference across Guidance Scales.**

## G LIMITATIONS

Multi-codebook speech units provide a shared speech–video interface, but their video-conditioning synchronization remains below that of waveform-derived features in our experiments. We have not

Table 19: Human evaluation of full-sequence guidance scales. Each cell reports mean rank / first-tier rate (%). Lower mean rank is better; first-tier rates include ties.

| CFG / Audio CFG | Lip Sync | Facial Naturalness | Gesture |
|---|---|---|---|
| 1 / 1 | 3.226 / 14.7 | 2.565 / 27.5 | 2.013 / 47.7 |
| 2 / 2 | 2.564 / 23.2 | 2.187 / 35.8 | 1.763 / 56.5 |
| 3.5 / 8.5 | 2.275 / 27.3 | 2.510 / 20.3 | 1.700 / 57.0 |
| 4 / 4 | **1.780 / 50.0** | **1.863 / 45.7** | 1.580 / 63.1 |
| 8 / 8 | 2.530 / 21.4 | 2.985 / 13.8 | **1.571 / 63.9** |

established whether additional training data, longer optimization, or greater model capacity can close this gap. Streaming avatar generation with a 1.3B-parameter video backbone is particularly challenging: the model must preserve identity, scene structure, and temporal coherence across incrementally generated chunks while using bounded historical context. In our late-chunk ablation under previous-latent conditioning, the persistent reference sink improves reference-image consistency but does not eliminate gradual drift. The streaming model also trails the full-sequence model in image quality and audio–video synchronization, reflecting the current quality cost of faster incremental output.