STOP ON SIGHT: OMNIMODAL STANDALONE FULL-DUPLEX INTERACTION SUPPORTING REAL-TIME VISUAL INTERRUPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal full-duplex interaction is pivotal for enabling natural, human-like dialogue in interactive artificial intelligence systems. However, existing full-duplex models trigger interruptions almost exclusively via acoustic cues, halting generation only when speech barge-in is detected, while critically lacking vision-driven interruption. In practice, this capability is vital, as humans naturally cease speaking upon noticing urgent visual cues, such as an explicit stop gesture or an abrupt change in the scene. To address this fundamental limitation, we present **VIFI**, which, to our knowledge, is the first visual-interruptible standalone full-duplex interaction model. Operating on 400-ms streaming chunks, VIFI introduces a decoupled four-token interaction protocol that allows text emission and audio rendering to proceed at independent rates, ensuring a natural and stable speaking cadence. To support effective training under this paradigm, a robust construction pipeline is built with a dedicated focus on vision-driven interruption scenarios. Furthermore, we design a tailored reinforcement learning framework that further improves full-duplex turn-taking and interruption behavior. To facilitate comprehensive evaluation, we establish **VIBench**, to our knowledge the first benchmark specifically designed to assess visual interruption fidelity under continuous streaming settings. Experiments show that VIFI substantially outperforms existing full-duplex baselines, surpassing MiniCPM-o 4.5 by 40.4 percentage points in visual interruption accuracy under continuous streaming settings.

1 INTRODUCTION

Omnimodal full-duplex interaction represents a foundational milestone toward natural, human-like dialogue in interactive artificial intelligence (AI) systems. In practical scenarios such as embodied robotics, real-time virtual assistants, and interactive digital humans, conversational systems critically require an omnimodal full-duplex paradigm rather than isolated, turn-by-turn exchanges. To achieve authentic interaction of this kind, an intelligent system must perpetually process streaming visual and auditory inputs in real time. More importantly, much like humans in everyday communication, the system must retain the ability to see and listen concurrently even while speaking, thereby naturally engaging with and adapting to the dynamic physical environment. We use multimodal as a broad umbrella term, omnimodal for unified real-time interaction across perception and generation, and audio-visual specifically for paired audio and video signals, data, or benchmarks.

Toward realizing this interactive vision, the rapid evolution of multimodal large language models (LLMs) offers substantial promise for machines to perceive and interpret the physical world. Over the past few years, the community has achieved remarkable progress, progressing from visual-only video LLMs (Zhang et al., 2024; Liu et al., 2024; Zhang et al., 2025; Chai et al., 2025; Zhu et al., 2025; Li et al., 2025; Bai et al., 2025; An et al., 2026; Li et al., 2026a; Team et al., 2026) and audio-only LLMs (Tang et al., 2024; Gong et al., 2024; Ding et al., 2025; Wu et al., 2025; Ghosh et al., 2026; Yang et al., 2026b;a) towards unified omnimodal LLMs (Comanici et al., 2025; Cheng et al., 2024; Tang et al., 2025; Chen et al., 2025c; Ge et al., 2025; Ye et al., 2025; Tang et al., 2026; Li et al., 2026b; Team, 2026; Cui et al., 2026). Nevertheless, authentic conversational intelligence requires capabilities far beyond static, turn-isolated comprehension. It inherently demands continuous streaming perception, minimal response latency, and the full-duplex capacity to concurrently process

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.

visual and auditory signals. Driven by this imperative, several pioneering efforts have emerged to explore streaming and interactive paradigms, including streaming multimodal perception models (Chen et al., 2024; Qian et al., 2024; Chen et al., 2025a; Sun et al., 2025; Xu et al., 2026), audio-only full-duplex conversational systems (Défossez et al., 2024; Yu et al., 2024; Wang et al., 2024; Chen et al., 2025b; Fang et al., 2026), and nascent omnimodal full-duplex frameworks (Fu et al., 2024; Xie & Wu, 2024; Fu et al., 2026; Cui et al., 2026; Lu et al., 2026a; Orantqing et al., 2026; Zhao et al., 2026).

In full-duplex interaction, the ability to interrupt ongoing speech is a definitive hallmark of natural dialogue. In authentic human communication, interruption signals can arise across different sensory modalities. A prototypical source is auditory: when an interlocutor initiates a speech barge-in, the system must learn to cease generation and listen, a capability that existing models have actively investigated and optimized. Equally vital, interruption triggers can also originate from visual stimuli. Humans instinctively pause mid-sentence upon witnessing sudden, critical visual cues, such as a user raising a hand to signal "stop" or a vehicle rapidly approaching. Unfortunately, vision-driven interruption remains severely under-explored. Existing full-duplex models primarily rely on acoustic interruption cues and perform poorly on vision-driven interruption during ongoing generation. Moreover, dedicated benchmarks for evaluating visual interruption under real-time streaming settings remain scarce.

To bridge this fundamental gap, we propose **VIFI**, a visually-interruptible standalone full-duplex interaction model capable of continuous audio-visual streaming inputs. VIFI operates on 400-ms streaming chunks, generating streaming text responses that are concurrently rendered into speech via an external streaming text-to-speech (TTS) module. In existing open-source omnimodal duplex systems, maintaining temporal synchronization between text and audio typically relies on a rigid "lock-step" paradigm: systems either hardcode fixed text-to-audio ratios via hardware pre-computation (Huang et al., 2026) or dynamically throttle text generation rates to align with audio playback (Cui et al., 2026). Since text generation is typically faster than audio synthesis, this lock-step design often slows down text generation to keep them in sync, which makes the system speak unnaturally slowly. Moreover, this design heavily relies on stable computational power. In resource-limited environments, or when longer contexts increase inference latency, the system may suffer from audio stuttering and text starvation. In contrast, drawing inspiration from SALMONN-omni (Yu et al., 2024), VIFI adopts a redesigned four-token omnimodal duplex interaction protocol paired with a linear-attention-based LLM backbone. This architectural formulation ensures long-context, stable generation across both text and audio modalities by allowing text and audio to be generated continuously at their respective rates.

To facilitate effective duplex training, a comprehensive data construction pipeline for omnimodal full-duplex interaction is devised, with a dedicated focus on vision-driven interruption scenarios. Moreover, to provide standard evaluation, we introduce **VIBench**, an omnimodal interaction benchmark specifically designed to evaluate full-duplex visual interruption. We also design a tailored reinforcement learning (RL) framework that jointly optimizes response triggering and interruption dynamics to further refine real-time turn-taking behaviors. Experiments demonstrate that our RL strategy yields consistent performance gains across interaction benchmarks like OmniInteract (Lu et al., 2026b) and our VIBench. Finally, VIFI achieves 42.0% visual interruption accuracy on VIBench, significantly outperforming existing omnimodal duplex baselines such as MiniCPM-o 4.5, which achieves 1.6% accuracy.

Our primary contributions can be summarized as follows:

- We propose VIFI, an omnimodal standalone full-duplex interaction model supporting real-time visual interruption, underpinned by an asynchronous four-token interaction protocol with a 400-ms streaming chunk size.
- We establish VIBench, a standard benchmark tailored to evaluate full-duplex visual interruption under continuous omnimodal streaming conditions. We also construct a scalable, robust data construction pipeline for omnimodal full-duplex interaction, with special focus on vision-driven interruption scenarios.
- We present a dedicated RL framework tailored for full-duplex turn-taking and interruption dynamics, delivering consistent empirical gains across omnimodal interaction benchmarks.

2 BACKGROUND

2.1 NON-STANDALONE FULL-DUPLEX INTERACTION

Most multimodal LLMs (MLLMs) are inherently trained for turn-based conversations, passively generating responses only after receiving full user queries. To achieve real-time interaction without overhauling this paradigm, a prevalent strategy is to construct a non-standalone system that couples the multimodal LLM with external triggering and interruption modules, using techniques like Voice Activity Detection (VAD). In the speech interaction domain, Freeze-Omni (Wang et al., 2024) pairs a frozen speech LLM with an upstream VAD and an auxiliary state classifier to detect speech barge-in. Similarly, MinMo (Chen et al., 2025b) relies on external acoustic echo cancellation and VAD modules alongside an auxiliary predictor module to determine when the system should yield the floor. For omnimodal interaction, non-standalone architectures similarly depend on external coordination. VITA (Fu et al., 2024) employs an external VAD alongside a dual-model framework, running a dedicated monitoring model to halt generation whenever user speech is detected. AURA (Lu et al., 2026a) integrates external sensory tracking modules to arbitrate conversational turn-taking and system interruptions based on user presence. Furthermore, Mini-Omni 2 (Xie & Wu, 2024) relies on acoustic detection modules to predict frame-level interruption tokens upon hearing explicit user commands.

While these systems can engage in interactive conversations, non-standalone frameworks treat dialogue state transitions as an external engineering pipeline rather than an internal modeling task, and thus lack an intrinsic duplex strategy for modeling complex dialogue dynamics. Relying on decoupled controllers to switch between listening and speaking prevents the underlying model from developing real-time conversational reflexes, rendering it ill-equipped for speech barge-ins or visual interruptions that require concurrent perception, reasoning, and generation.

2.2 STANDALONE FULL-DUPLEX INTERACTION

Trained on massive corpora, modern foundation models naturally possess rich cross-modal reasoning capabilities, motivating researchers to empower the central model to regulate its own conversational states natively without external controllers. In spoken dialogue, Moshi (Défossez et al., 2024) enables simultaneous listening and speaking via multi-stream autoregression and inner monologues. SALMONN-omni (Yu et al., 2024) models streaming full-duplex conversations by introducing discrete state tokens into the token stream, discarding external voice activity detectors. Similarly, BayLing-Duplex (Fang et al., 2026) integrates turn-taking decisions directly into autoregressive decoding through dedicated interaction tokens to initiate or yield conversational turns dynamically.

Building on speech foundations, recent efforts have progressively extended standalone duplex interaction to the omnimodal domain. Compared to pure audio dialogues, incorporating visual streams introduces heavy computational overhead, requires continuous perceptual grounding under dynamically evolving scenes, and introduces complex, proactive interruption demands. To address these challenges, DuplexOmni (Huang et al., 2026) decouples interaction from deep reasoning via a real-time multimodal interaction layer and an asynchronous thinking layer. MiniCPM-o 4.5 (Cui et al., 2026) introduces an Omni-Flow design to coordinate audio-visual perception and streaming speech on a shared temporal timeline. Gander (Orantqing et al., 2026) adopts a Cerebellum-Brain architecture that flattens streaming audio-visual inputs to predict discrete turn-taking tokens, while Realtime-Venus (Zhao et al., 2026) introduces a dual-loop framework with unified stream serialization to align perception and interaction across a shared causal timeline.

3 METHOD

3.1 STANDALONE FULL-DUPLEX STRATEGY

Existing open-source standalone omnimodal full-duplex models often enforce a rigid rate-balancing constraint between text generation and audio synthesis. To eliminate the resulting issue of sluggish speaking cadence, we redesign the duplex interaction strategy by decoupling text emission from audio rendering, allowing the system to converse naturally at a stable pace while preserving full-duplex capabilities. As shown in Figure 1, VIFI is an omnimodal streaming LLM that continuously ingests audio-visual inputs and regulates its conversational states using four discrete control tokens. It is

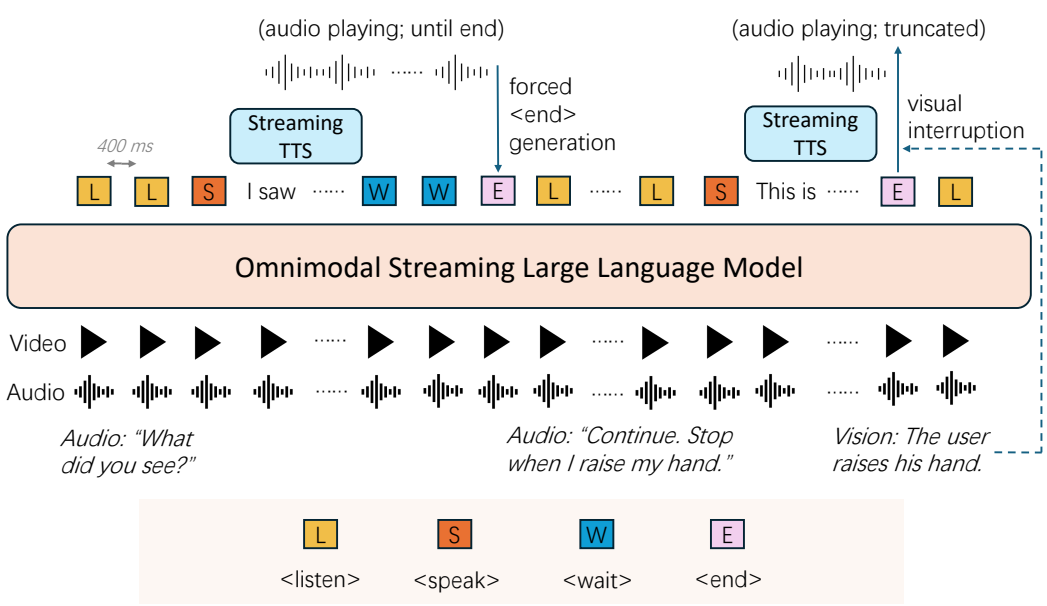

Figure 1: Overview of the standalone full-duplex interaction strategy in VIFI. The model operates on continuous 400 ms streaming chunks, using four discrete control tokens (<listen>, <speak>, <wait>, and <end>) to govern conversational state transitions.

coupled with an external streaming TTS module for speech output. Here, standalone refers specifically to the interaction policy: VIFI does not rely on external VAD or semantic turn-taking controllers. The streaming TTS module is used only for speech rendering and playback synchronization, while interruption decisions are predicted by the model itself.

Specifically, VIFI processes continuous audio and video streams in streaming chunks with a duration of $t_{\text{chunk}} = 400\,\text{ms}$, generating outputs in real time. During interaction, the model perceives environmental context and produces responses concurrently. The interactive workflow is orchestrated by four specialized control tokens: <listen>, <speak>, <wait>, and <end>.

The tokens <listen> and <speak> represent the two primary conversational states commonly adopted in duplex systems. Under default conditions where no user request is present, the model remains in the listening state. Within each chunk, the model simply outputs a single <listen> token to indicate active monitoring. When the user finishes speaking or contextual cues require a proactive response, the model initiates speech by emitting a <speak> token within that chunk. This token marks the state transition from listening to speaking, functioning as an autonomous internal trigger. In subsequent chunks throughout the active speaking turn, the model directly generates textual response tokens without repeating <speak>.

The <wait> token is specifically designed to handle the temporal discrepancy between fast text emission and slower acoustic playback. Because autoregressive text generation is typically faster than real-time speech rendering, the model often concludes semantic text generation while audio playback is still underway. Under such circumstances, the system is physically still speaking to the user. Rather than terminating the turn prematurely, the model emits <wait> tokens as non-semantic placeholders, indicating semantic text generation has finished while the system is still speaking.

The <end> token signifies the definitive termination of the speaking state, which can occur under two distinct scenarios:

1. **Normal Completion:** In the absence of interruptions, once the final piece of audio completes playback, the runtime inserts an <end> token to synchronize the model state with the completion of audio playback. This synchronization step transitions the model from the speaking state back to the listening state. Only after observing the <end> token does the model resume emitting <listen> tokens in subsequent chunks.

2. **Autonomous Interruption:** When an interruption intent is detected, whether triggered by user speech barge-in or visual cues, the model itself proactively predicts an <end> token during generation. The moment this <end> token is emitted, the audio synthesis and playback thread immediately truncates ongoing playback.

3.2 Data Construction Pipeline

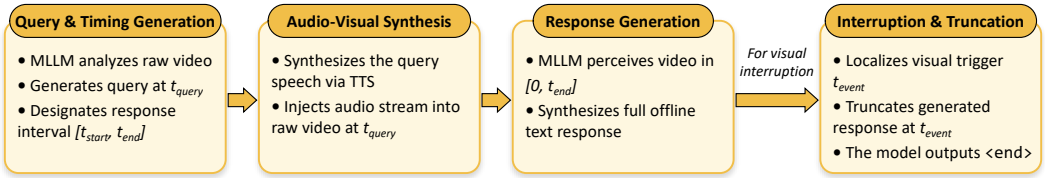

Figure 2: Overview of our data construction pipeline for omnimodal full-duplex interaction, comprising query and timing generation, audio-visual synthesis, response generation, and specialized visual interruption formatting.

Due to the extreme scarcity of real-time omnimodal conversational data, we build an automated pipeline to synthesize interactive duplex dialogues from offline video repositories, as illustrated in Fig. 2. In the initial query and timing generation stage, an offline MLLM, such as Gemini-3.5-flash or the base model of VIFI, performs analysis on the raw video. Operating in the persona of a user, the MLLM formulates a contextually grounded query at an arbitrary timestamp t_{query} and designates the expected response interval $[t_{\text{start}}, t_{\text{end}}]$.

Next, a TTS model, such as Qwen3-TTS (Hu et al., 2026), converts the query text into speech, which is subsequently injected into the video at t_{query}. To preserve the natural timbre and cadence of the speech without causing visual dissonance, we slow down the video frame rate at the insertion point to match the duration of the audio. In addition, to simulate realistic auditory environments, we blend background ambient noise or mix the original audio track of the video, preventing the model from training solely on clean acoustic signals.

In the subsequent response generation stage, we transfer the perceptual understanding of the MLLM into the full-duplex conversational paradigm. Specifically, we use the base model of VIFI as the MLLM to perceive the video segment spanning $[0, t_{\text{end}}]$ and generate a textual response answering the user query. This synthesis naturally supports multi-turn conversations with dialogue history, where the model incorporates all preceding questions and answers when addressing subsequent queries. By deriving responses directly from the base model, the subsequent training primarily encourages the model to master duplex turn-taking protocols rather than learning fundamental video comprehension from scratch. This formulation effectively decouples duplex interaction dynamics from general omnimodal reasoning, which may facilitate smoother optimization.

During query generation, the pipeline creates vision-driven interruption requests, such as "Describe the scene and stop when I raise my hand". While the initial stage estimates a coarse response window, this interval usually cannot pinpoint the exact moment of the visual event. Therefore, we introduce a specialized step for this task, utilizing an MLLM to carefully inspect the candidate video clip and identify the precise timestamp of the visual trigger, denoted as t_{event}. We then truncate the generated response at t_{event} to simulate speech halted mid-utterance, supervising the model to emit an <end> token at this exact moment to reinforce visual interruption capabilities.

3.3 Training Strategy

Continuous duplex interaction rapidly inflates the omnimodal context, incurring heavy key-value (KV) cache overhead under standard attention. Therefore, unlike existing omnimodal LLMs such as Qwen3-Omni (Xu et al., 2025) and MiniCPM-o 4.5 (Cui et al., 2026), VIFI is built upon an internal base model that adapts the video-SALMONN 2 (Tang et al., 2025) framework by replacing its LLM backbone with Qwen3.5-9B (Qwen Team, 2026). Note that Qwen3.5 primarily relies on linear attention without traditional KV caching. We first performed supervised fine-tuning (SFT) on the base model. During SFT, training instances are structured into an interleaved streaming format.

Continuous audio-visual streams are segmented into discrete chunks with a fixed temporal duration of $t_{\text{chunk}} = 400$ ms. After each chunk, the model is supervised to output the corresponding tokens conditioned on the historical context. Appendix A provides more details of the data format.

Although SFT establishes basic cross-modal alignment, it remains insufficient for reliable turn-taking. We empirically observe that SFT models exhibit extreme overconfidence at turn-taking boundaries, assigning near-unity probability to maintaining their current state even when a transition should occur. Consequently, standard online RL exploration is highly inefficient, as stochastically sampling valid transition trajectories is exceedingly rare. To address this challenge, we introduce an offline trajectory-guided RL algorithm. Specifically, we collect offline rollout trajectories on training samples with reference annotation y_{ref}, identify erroneous decision steps where state transitions fail based on y_{ref}, and directly optimize the model at these discrete decision points using Direct Preference Optimization (DPO) (Rafailov et al., 2023).

Formally, at each identified step with context x, we construct a preference pair (y_w, y_l), where y_w is the desirable target token and y_l is the undesirable rollout prediction. Letting π_θ denote the policy model and π_{ref} represent the reference policy initialized from the base model, the token-level DPO loss at this decision point is defined as:

$$\mathcal{L}_{\text{DPO}}(y_w, y_l \mid x) = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right), \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function and β is a temperature hyperparameter controlling the divergence from π_{ref}. To stabilize training and preserve general conversational ability, we jointly optimize the preference loss alongside a supervised cross-entropy anchor:

$$\mathcal{L} = \frac{1}{|G|} \sum_{(x, y_w, y_l) \in G} \mathcal{L}_{\text{DPO}}(y_w, y_l \mid x) + \lambda \mathcal{L}_{\text{SFT}}(y_{\text{ref}} \mid X), \tag{2}$$

where G denotes the collection of all identified decision points in the trajectory, $\mathcal{L}_{\text{SFT}}$ represents the standard next-token prediction loss on y_{ref} conditioned on the video input X, and λ is the coefficient.

To construct the preference pairs $(y_w, y_l) \in G$, we categorize full-duplex transition errors into five distinct failure modes:

- **Missed Initiation:** When y_{ref} indicates that the model should begin speaking with <speak> but the rollout continues emitting <listen> tokens, we assign $y_w =$ <speak> and $y_l =$ <listen> to encourage proactive conversational triggering.
- **Missed Interruption:** When y_{ref} signals the cessation of speech via <end> due to auditory or visual interruption, yet the rollout fails to halt and continues speaking, we designate the ongoing speech token as y_l and supervise the transition with $y_w =$ <end>.
- **False Initiation:** When the rollout inappropriately generates a <speak> token while the reference remains in the listening state, we assign $y_w =$ <listen> and $y_l =$ <speak> to penalize unprompted barge-ins.
- **False Interruption:** When the rollout prematurely outputs an <end> token while the reference is actively in a speaking phase without any interruption trigger, we set $y_l =$ <end> to suppress unwarranted halting, and sample y_w from valid text tokens.
- **Premature Waiting:** When the rollout emits a <wait> placeholder in the middle of an incomplete sentence during speech production, we set $y_l =$ <wait> and sample y_w from valid text tokens to discourage premature stalling before natural sentence completion.

4 Experimental Setup

4.1 Data Specifications

For SFT, we collect an internal corpus of approximately one million first-person egocentric videos. Following the pipeline in Sec. 3.2, we use Gemini-3.5-flash and our base model to formulate interactive queries, while our base model synthesizes the corresponding responses. In total, we generate roughly 760k real-time omnimodal duplex dialogue samples, including 260k samples

dedicated to vision-driven interruption. To adapt the model to streaming interaction while preserving general omnimodal capabilities, we blend roughly equal proportions of streaming speech recognition, streaming video captioning, and offline audio-visual comprehension data into the training mixture. For RL, we sample 100k instances from the duplex training corpus to collect rollout trajectories, filtering out approximately 70k trajectories with state transition errors to train the policy.

For evaluation, benchmarks for real-time visual interruption remain scarce. As constructing complex evaluation scenarios via video generation is a common and effective approach (He et al., 2026; Chen et al., 2026), we synthesize video clips using MiniMax-H3 (MiniMax Team, 2026) to build VIBench, a benchmark dedicated to duplex visual interruption. After fine-grained MLLM annotation and meticulous manual verification by the authors, we curate 116 accurately labeled test samples. Each video pairs an embedded spoken instruction with a precisely annotated visual trigger timestamp, requiring the model to respond and immediately halt speech once the visual event appears. Performance is measured by the visual interruption success rate within a temporal window around the trigger onset. Appendix B provides additional details. Note that we do not include any similar AI-generated video data originating from VIBench in the training set; videos from VIBench are out-of-domain for our models.

4.2 Training and Evaluation Specifications

Across both training and inference, VIFI processes continuous inputs and produces textual outputs with a uniform streaming chunk window of 0.4 s. Built upon the video-SALMONN 2 framework, VIFI's base model utilizes an audio encoder based on the Whisper non-streaming architecture (Radford et al., 2023). To support real-time streaming audio perception, we convert its attention masks into a causal formulation. During SFT, we also insert a LoRA adapter with rank $r = 128$ and scaling factor $\alpha = 256$ into the LLM backbone, optimizing both the LoRA parameters and the audio encoder. SFT is executed across 128 NVIDIA H20 GPUs for approximately 24k steps. During the RL stage, we freeze the audio encoder and exclusively optimize the LoRA adapter, training on 32 NVIDIA H20 GPUs for about 600 steps.

During real-time duplex evaluation, to balance inference speed with textual fluency, we cap each input video frame at 76,800 pixels. For speech synthesis, the generated text tokens are fed into the dots-TTS model (dots.tts Team, 2026) to produce continuous audio output. In our experimental setup, VIFI generates up to 5 tokens per 400 ms chunk. In practical deployment scenarios, this generation budget can be dynamically adjusted based on the computational speed. For offline evaluation, we sampled video frames at 2 frames per second (FPS) and capped at a maximum of 512 frames.

5 Experimental Results

5.1 Main Results

We evaluate the interactive capabilities of different models in Table 1. Specifically, we assess full-duplex performance on OmniInteract (Lu et al., 2026b) and our newly established VIBench. In addition, we evaluate models on half-duplex interaction benchmarks, including OmniVChat (He et al., 2026) and ODU-Bench (Chen et al., 2026). We compare VIFI against other competitive open-source standalone omnimodal full-duplex LLMs, namely MiniCPM-o 4.5 (Cui et al., 2026), Gander (Orantqing et al., 2026), and Realtime-Venus (Zhao et al., 2026).

As shown in Table 1, existing full-duplex models show limited performance in vision-driven interruption. Notably, MiniCPM-o 4.5 and Realtime-Venus attain very low visual interruption success rates of only 1.6% and 4.2%, respectively, while Gander achieves 17.9%. This indicates that prior standalone duplex models exhibit limited sensitivity to visual interruption cues during ongoing generation, failing to react promptly to occurring visual events in the surrounding environment. In sharp contrast, VIFI achieves a visual interruption success rate of 42.0% on VIBench, substantially outperforming all existing baselines. This large improvement supports the effectiveness of our data construction strategy, especially considering that the synthesized videos in VIBench serve as out-of-domain evaluation data for our model as well.

On general omnimodal duplex benchmarks such as OmniInteract, VIFI shows mixed but competitive performance across different interaction settings. Furthermore, on turn-based half-duplex benchmarks

such as OmniVChat and ODU-Bench, VIFI maintains strong conversational reasoning, performing better than MiniCPM-o 4.5.

Table 1: Evaluation results on full-duplex and half-duplex interaction interaction benchmarks. "-" denotes evaluation runtime failures left unaddressed, as this is not the primary focus of this work.

Model	VIBench	OmniInteract		OmniVChat	ODU-Bench
		1q1a	1qna		
MiniCPM-o 4.5	1.6	45.6	1.5	30.4	39.1
Gander	17.9	**45.7**	0.5	-	-
Realtime-Venus	4.2	26.7	**3.9**	-	-
VIFI	**42.0**	34.7	2.1	**33.5**	**67.8**

Beyond interactive evaluations, VIFI also exhibits robust capabilities across general video and omnimodal comprehension tasks. As shown in Table 2, we evaluate models on streaming video understanding benchmarks, including StreamingBench (Lin et al., 2026) and OVO-Bench (Niu et al., 2025), alongside audio-visual benchmarks such as Daily-Omni (Zhou et al., 2025), WorldSense (Hong et al., 2026), and Video-MME (Fu et al., 2025). On streaming video understanding, VIFI remains highly competitive, achieving 68.3% on StreamingBench and 62.2% on OVO-Bench. Meanwhile, across offline audio-visual benchmarks, VIFI achieves top-tier performance among open-source standalone full-duplex LLMs, attaining 82.6% on Daily-Omni, 57.3% on WorldSense, and 81.6% on Video-MME. This indicates that our four-token interaction protocol and two-stage training pipeline are able to adapt an offline foundation MLLM into an omnimodal full-duplex agent while preserving its core omnimodal capabilities, avoiding catastrophic forgetting to some degree.

Table 2: Evaluation results on streaming video understanding and offline audio-visual comprehension benchmarks. "-" denotes evaluation runtime failures left unaddressed.

Model	StreamingBench	OVO-Bench	Daily-Omni	WorldSense	Video-MME
MiniCPM-o 4.5	67.9	60.7	80.2	55.7	70.4
Gander	56.7	56.3	78.5	49.6	-
Realtime-Venus	**70.2**	**64.7**	81.3	54.0	-
VIFI	68.3	62.2	**82.6**	**57.3**	**81.6**

5.2 Ablation on Supervised Loss Weight λ in RL

During full-duplex RL, the initial SFT model frequently exhibits severe overconfidence in its existing state, assigning near-zero likelihood to the correct state transition token. Directly optimizing the policy via token-level DPO under such substantial probability discrepancies might induce extreme gradient instability, often causing the policy to collapse into degenerate repetition loops. To prevent this distributional shift and stabilize the optimization, we incorporate the supervised loss $\mathcal{L}_{\text{SFT}}(y_{\text{ref}} \mid X)$ as an anchor in Eq. 2. We conducted ablation experiments on the loss weight λ, training the policy for 600 steps across varying values of $\lambda \in \{0, 1, 10, 100\}$ and evaluating performance on VIBench and OmniInteract.

Table 3: Ablation study on the supervised loss weight λ in Eq. 2 over 600 training steps.

Configuration	VIBench	OmniInteract 1q1a			
		Real-time	Proactive	Nested	Global
SFT Baseline	40.2	31.7	16.5	39.1	30.8
RL ($\lambda = 0$)	*Model collapsed (repetitive generation)*				
RL ($\lambda = 1$)	*Model collapsed (repetitive generation)*				
RL ($\lambda = 10$)	**42.0**	**37.7**	**16.8**	**41.4**	**34.7**
RL ($\lambda = 100$)	39.5	31.8	14.8	41.2	31.2

As shown in Table 3, omitting the supervised anchor ($\lambda = 0$) or setting an insufficiently small weight ($\lambda = 1$) leads to catastrophic model collapse, where the policy model starts to generate repetitive tokens during interaction. In contrast, setting a moderate weight of $\lambda = 10$ provides sufficient regularization to maintain linguistic fluency while allowing preference optimization to correct state transition boundaries. This configuration yields consistent performance gains over the SFT baseline, improving visual interruption accuracy on VIBench from 40.2% to 42.0% and boosting the OmniInteract 1q1a global score from 30.8% to 34.7%. When λ becomes too large, the supervised loss dominates the optimization landscape, marginalizing the gradient signals from preference pairs. Under this condition, the model exhibits minor fluctuations around the SFT baseline without realizing meaningful interactive gains. Therefore, we adopt $\lambda = 10$ as the setting in the final model.

5.3 Case Study on Speaking Cadence and Fluency

Unlike prior systems that enforce rigid lock-step synchronization between textual emission and acoustic playback, VIFI, to a certain extent, decouples text generation from speech synthesis. This formulation allows speech rendering to proceed at a natural tempo without suffering from stuttering or pauses caused by delayed textual supply. To evaluate conversational delivery in practice, we conduct four Mandarin case studies to compare VIFI against MiniCPM-o 4.5 under identical conditions. For each interaction, both systems process the same continuous video stream and spoken prompt to provide detailed scene descriptions, averaging approximately 25 seconds per session. We manually evaluate these sessions by calculating the average speaking rate (characters per second) and tracking the occurrences of unnatural cadence decelerations.

Table 4: Results of the case studies comparing conversational cadence and fluency between MiniCPM-o 4.5 and the VIFI system.

System	Chunk Window	Speaking Rate (chars/s)	Unnatural Decelerations (Avg.)
MiniCPM-o 4.5	1.0 s	4.04	1.5
VIFI	400 ms	5.86	0

As presented in Table 4, the VIFI-based system delivers a faster and more fluent speaking cadence, achieving a speaking rate of 5.86 chars/s compared to 4.04 chars/s for MiniCPM-o 4.5. Crucially, MiniCPM-o 4.5 sometimes exhibits unnatural deceleration during generation to throttle its pace for audio-text alignment, averaging 1.5 occurrences per session, whereas VIFI maintains a consistent speaking tempo during the interaction with zero observed decelerations. These results suggest that our duplex strategy can reduce text throughput starvation, supporting a more stable speaking pace.

6 Conclusion

In this work, we propose VIFI, an omnimodal standalone full-duplex LLM equipped with real-time visual interruption capabilities. Operating on 400-ms streaming chunks, VIFI is governed by a decoupled four-token duplex interaction protocol: it emits a `<listen>` token while actively monitoring ambient inputs, precedes speech generation with a `<speak>` token, outputs `<wait>` placeholder tokens once textual response concludes to await the completion of acoustic playback, and utilizes the `<end>` token to execute immediate interruptions or conclude turns naturally. This design enables VIFI to continuously monitor incoming visual streams throughout speech production, enabling speech interruptions triggered by visual events during ongoing generation. To support full-duplex training, we introduce an automated pipeline to synthesize interactive omnimodal conversations from offline videos. We also establish VIBench, a novel benchmark dedicated to evaluating duplex visual interruption. On top of SFT, we develop a tailored RL framework that pinpoints erroneous transition tokens along model trajectories and optimizes turn-taking dynamics via token-level DPO. Experimental results demonstrate that our RL stage further improves turn-taking behavior by correcting erroneous state transitions, enabling VIFI to achieve a 42.0% visual interruption accuracy on VIBench, outperforming the open-source MiniCPM-o 4.5 by 40.4 percentage points. In a practical case study, VIFI delivers a more stable speaking cadence with fewer unnatural pauses. We believe that VIFI takes a step toward perceptually grounded full-duplex agents that can adapt their speaking behavior to ongoing visual events.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, implementing methods (specifically, scripting data requests for model serving, implementing the loss calculation for the reinforcement learning algorithm, and testing select models), and assisting with translation. We have not used generative AI tools for developing theoretical models or conceptual frameworks, designing or providing feedback on research methodology or experiments, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code and editing the paper to improve readability. We have reviewed all AI-assisted work. Specifically, all code generated by large language models was thoroughly reviewed and verified through correctness testing by the authors. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-Onevision-2: Towards Next-generation Perceptual Intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark. In *Proc. ICLR*, Singapore, 2025.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online Video Large Language Model for Streaming Video. In *Proc. CVPR*, 2024.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Live: Learning Video LLM with Streaming Speech Transcription at Scale. In *Proc. CVPR*, 2025a.

Qi Chen, Yunfei Chu, Haolin He, Yifan Yang, Zihan Liu, Yuxuan Wang, Ziyang Ma, Ruiyang Xu, Meng Gao, Yinsong Yan, et al. Omni Demand Understanding: A Benchmark for Contextual User-Intent Inference in Multimodal Interaction. *arXiv preprint arXiv:2609.21392*, 2026.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, et al. MinMo: A Multimodal Large Language Model for Seamless Voice Interaction. *arXiv preprint arXiv:2501.06282*, 2025b.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Yuanxing Zhang, Qiang Liu, Pengfei Wan, et al. AVoCaDO: An Audiovisual Video Captioner Driven by Temporal Orchestration. *arXiv preprint arXiv:2510.10395*, 2025c.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing Spatial-temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. MiniCPM-o 4.5: Towards Real-time Full-duplex Omni-modal Interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A Speech-text Foundation Model for Real-time Dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-Audio Technical Report. *arXiv preprint arXiv:2504.18425*, 2025.

dots.tts Team. dots.tts technical report. *arXiv preprint arXiv:2606.07080*, 2026.

Qingkai Fang, Shoutao Guo, and Yang Feng. BayLing-Duplex: Native Full-Duplex Speech Dialogue with a Single Autoregressive LLM. *arXiv preprint arXiv:2606.14528*, 2026.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. VITA: Towards Open-source Interactive Omni Multimodal LLM. *arXiv preprint arXiv:2408.05211*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The First-ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proc. CVPR*, 2025.

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. *Proc. NeurIPS*, 2026.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *arXiv preprint arXiv:2406.20094*, 2024.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. ARC-Hunyuan-Video-7B: Structured Video Comprehension of Real-world Shorts. *arXiv preprint arXiv:2507.20939*, 2025.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Proc. NeurIPS*, 2026.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, Think, and Understand. In *Proc. ICLR*, Vienna, 2024.

Haolin He, Yunfei Chu, Qi Chen, Wen Huang, Yuan Feng, Muzhi Zhu, Zheqi Dai, Haoning Xu, Dongchao Yang, Chunyat Wu, et al. OmniVChat: Synthesizing, Benchmarking, and Training for Native Audio-Visual Dialogue. *arXiv preprint arXiv:2609.21465*, 2026.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs. In *Proc. ICLR*, 2026.

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. *arXiv preprint arXiv:2601.15621*, 2026.

Muye Huang, Lingling Zhang, Xingyu Yu, Lei Shi, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, and Jun Liu. Duplexomni: Real-time listening, seeing, thinking, and speaking for full-duplex interaction. *arXiv preprint arXiv:2606.09186*, 2026.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. VideoChat3: Fully Open Video MLLM for Efficient and Generalist Video Understanding. *arXiv preprint arXiv:2607.14935*, 2026a.

Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Improving LLM Video Understanding with 16 Frames Per Second. In *Proc. ICML*, Vancouver, 2025.

Yixuan Li, Guangzhi Sun, Yudong Yang, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN-R3: Learning to ReWatch, ReAsk, and ReAnswer for Efficient Video Understanding. *arXiv preprint arXiv:2606.24477*, 2026b.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the Gap for MLLMs to Achieve Streaming Video Understanding. In *Proc. ICASSP*, 2026.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. NVILA: Efficient Frontier Visual Language Models. *arXiv preprint arXiv:2412.04468*, 2024.

Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, et al. AURA: Always-on Understanding and Real-time Assistance via Video Streams. *arXiv preprint arXiv:2604.04184*, 2026a.

Xudong Lu, Xueying Li, Annan Wang, Yang Bo, Jinpeng Chen, Zengliang Li, Nianzu Yang, Rui Liu, Xue Yang, Jingwen Hou, et al. OmniInteract: Benchmarking Real-World Streaming Interaction for Real-Time Omnimodal Assistants. *arXiv preprint arXiv:2605.26485*, 2026b.

MiniMax Team. MiniMax H3: An Open Model Breaking the Boundaries Between Tasks and Modalities, July 2026. URL `https://www.minimax.io/blog/minimax-h3`.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. OVO-Bench: How Far is Your Video-LLMs from Real-world Online Video Understanding? In *Proc. CVPR*, 2025.

Orantqing, Shengpeng Ji, Junlong Tong, Jialong Zuo, Dongjie Fu, Di Cao, Yangzhuo Li, Shangda Wu, Franz, Evan, Theron Veyra, Changhao Pan, Jingyu Lu, Dongchao Yang, Zhifei Xie, Yang Tan, Xiaoyu Shen, Xiaoda Yang, Wenfu Wang, Teddy Sun, Steve Yves, and Zhou Zhao. Multimodal Duplex Interaction Agent. *arXiv preprint arXiv:2609.08977*, 2026.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming Long Video Understanding with Large Language Models. *Proc. NeurIPS*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-scale Weak Supervision. In *Proc. ICML*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Proc. NeurIPS*, 2023.

Guangzhi Sun, Yixuan Li, Xiaodong Wu, Yudong Yang, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN S: Memory-Enhanced Streaming Audio-Visual LLM. *arXiv preprint arXiv:2510.11129*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *Proc. ICLR*, Vienna, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Captioning-Enhanced Audio-Visual Large Language Models. *arXiv preprint arXiv:2506.15220*, 2025.

Changli Tang, Tianyi Wang, Fengyun Rao, Jing Lyu, and Chao Zhang. D-ORCA: Dialogue-centric Optimization for Robust Audio-visual Captioning. *arXiv preprint arXiv:2602.07960*, 2026.

Kwai Keye Team, Bin Wen, Changyi Liu, Chengru Song, Chongling Rao, Guowang Zhang, Han Li, Haonan Fan, Hengrui Ju, Jiankang Chen, et al. Kwai Keye-VL-2.0 Technical Report. *arXiv preprint arXiv:2606.10651*, 2026.

Qwen Team. Qwen3.5-Omni Technical Report. *arXiv preprint arXiv:2604.15804*, 2026.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM. *arXiv preprint arXiv:2411.00774*, 2024.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-Audio 2 Technical Report. *arXiv preprint arXiv:2507.16632*, 2025.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. StreamingVLM: Real-time Understanding for Infinite Video Streams. In *Proc. ICLR*, 2026.

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.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio Technical Report. *arXiv preprint arXiv:2606.01802*, 2026a.

Xiaoyu Yang, Xuenan Xu, Wenyi Yu, Siyin Wang, Changli Tang, Terumi Chiba, Siyuan Hou, Ziyang Zhang, Wen Wu, Baoxiang Li, et al. SALMONN-2: Advancing General-Purpose Hearing Abilities with Self-Supervised Representations. *arXiv preprint arXiv:2607.17079*, 2026b.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. OmniVinci: Enhancing Architecture and Data for Omni-Modal Understanding LLM. *arXiv preprint arXiv:2510.15870*, 2025.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. SALMONN-Omni: A Codec-free LLM for Full-duplex Speech Understanding and Generation. *arXiv preprint arXiv:2411.18138*, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video Instruction Tuning with Synthetic Data. *arXiv preprint arXiv:2410.02713*, 2024.

Ruixiang Zhao, Hualei Wang, Renhe Sun, Enzhi Zhou, Jincenzi Wu, Xujie Song, Kexin Shi, Zihang Liu, Pengcheng Zhu, Jiayi Zhou, et al. Realtime-Venus: A Full-duplex Interaction System with Asynchronous Delegation. *arXiv preprint arXiv:2609.13814*, 2026.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-Omni: Towards Audio-visual Reasoning with Temporal Alignment Across Modalities. *arXiv preprint arXiv:2505.17862*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

A Details of Data Format

To support real-time full-duplex interaction, all conversational training instances are structured into a temporally synchronized, chunk-interleaved token sequence. Let a continuous omnimodal interaction stream of temporal duration T be partitioned into $K = \lceil T/\Delta t_{\text{chunk}} \rceil$ discrete chunks with a uniform window size of $\Delta t_{\text{chunk}} = 400\,\text{ms}$. The full training sequence $\mathcal{S}$ is formulated as a concatenation of K sequential chunk units:

$$\mathcal{S} = [\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_K], \tag{3}$$

where each chunk unit $\mathbf{C}_k$ ($k \in \{1, 2, \ldots, K\}$) is systematically constructed by interleaving time anchors, audio-visual observation tokens, and model generation tokens:

$$\mathbf{C}_k = [T_k, \mathbf{V}_k, \mathbf{A}_k, \mathbf{Y}_k]. \tag{4}$$

The individual components within each chunk unit $\mathbf{C}_k$ are defined as follows:

- **Timestamp Anchor (T_k):** A discrete temporal token indicating the elapsed physical time since the inception of the interaction stream, structured as <τ_k seconds>, where $\tau_k = (k-1) \cdot \Delta t_{\text{chunk}}$.
- **Visual Observation Tokens ($\mathbf{V}_k$):** A sequence of visual tokens $\mathbf{V}_k = [v_{k,1}, v_{k,2}, \ldots, v_{k,N_v}]$ extracted from the video frame sampled within the k-th chunk interval.
- **Acoustic Observation Tokens ($\mathbf{A}_k$):** A sequence of audio feature tokens $\mathbf{A}_k = [a_{k,1}, a_{k,2}, \ldots, a_{k,N_a}]$ encoded from the streaming audio signal recorded over the temporal window $[(k-1)\Delta t_{\text{chunk}}, k\Delta t_{\text{chunk}}]$.
- **Supervision Target Tokens ($\mathbf{Y}_k$):** The sequence of textual or control tokens $\mathbf{Y}_k = [y_{k,1}, y_{k,2}, \ldots, y_{k,M_k}]$ assigned to the k-th chunk. When the model is actively monitoring, $\mathbf{Y}_k$ consists of a single control token <listen>. During speech generation, $\mathbf{Y}_k$ contains the corresponding response text tokens, or appropriate transition tokens such as <speak>, <wait>, and <end>.

Under this formulation, the audio-visual conditioning input stream introduced in Sec. 3.3 is represented as $X = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K]$, where $\mathbf{X}_k = [T_k, \mathbf{V}_k, \mathbf{A}_k]$ denotes the observation prefix for the k-th chunk. Correspondingly, the complete reference sequence is formed by the concatenation of all supervisory tokens $y_{\text{ref}} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K]$. During SFT, the objective $\mathcal{L}_{\text{SFT}}(y_{\text{ref}} \mid X)$ in Eq. 2 is calculated strictly via autoregressive next-token prediction over the target tokens $\mathbf{Y}_k$, while all observation tokens in X are excluded through loss masking:

$$\mathcal{L}_{\text{SFT}}(y_{\text{ref}} \mid X) = -\sum_{k=1}^{K} \sum_{m=1}^{M_k} \log P_\theta \left(y_{k,m} \mid \mathcal{S}_{<(k,m)} \right), \tag{5}$$

where $\mathcal{S}_{<(k,m)}$ denotes the entire historical context prefix preceding the target token $y_{k,m}$ within the interleaved sequence $\mathcal{S}$.

This interleaved chunk design provides direct compatibility with streaming inference runtimes. Under online deployment, the model processes each incoming chunk unit $\mathbf{C}_k$ sequentially. Upon ingesting the observation prefix $\mathbf{X}_k$, the model updates its internal representation and generates the output tokens $\mathbf{Y}_k$ autoregressively. Specifically, in linear attention architectures, the recurrent hidden state is updated incrementally chunk by chunk, maintaining constant computational complexity and memory footprint across extended interactions. In standard attention architectures, this design similarly facilitates seamless streaming inference, as newly arrived chunk representations can simply be appended to the existing KV cache without recomputing historical states.

B Details of VIBench

To construct realistic visual interruption instances for VIBench, we prompt Qwen3-235B-A22B-Instruct (Yang et al., 2025) using diverse personas from Persona Hub (Ge et al., 2024) as seeds to draft interactive scripts. These scripts specify spoken instructions and visual triggers that demand immediate interruption, and are fed into MiniMax-H3 for audio-visual video generation. While capturing the target scenarios, the raw generated clips exhibit three primary limitations: visual triggers occurring prematurely before user speech ends, trigger onset drifting from scripted timestamps, and inconsistent visual quality.

To resolve these issues, we re-synthesize user instructions using Qwen3-TTS and anchor them at the start of each video. If the audio exceeds two seconds, we temporally stretch the initial two seconds of the video to match the audio duration, ensuring that visual triggers transpire only after the instruction completes. To localize the exact trigger onset timestamp t_{event}, we probe candidate videos incrementally every 0.5 s using the base model of VIFI across prefixes $[0, t]$, identifying videos that exhibit a clear transition from negative to sustained positive detections. Finally, the authors perform meticulous manual verification to filter out problematic clips, retaining a final set of 116 high-quality, verified test samples. Note that some existing full-duplex video benchmarks share a similarly compact scale due to data scarcity (e.g., OmniInteract contains only 170 and 40 samples for its 1q1a and 1qna tasks).

During evaluation, the model must enter the speaking state to answer the user query and emit an <end> token once the visual trigger appears. Let t_{event} denote the annotated trigger onset and $t_{\text{interrupt}}$ denote the timestamp where the model emits <end>. We evaluate visual interruption within an

asymmetric tolerance window:

$$t_{\text{interrupt}} \in [t_{\text{event}} - 0.5\,\text{s},\ t_{\text{event}} + 1.0\,\text{s}]. \tag{6}$$

The left tolerance of 0.5 s accounts for the 0.5 s probing interval where visual cues may become perceptible slightly earlier, while the right tolerance of 1.0 s accommodates chunking delay and system response delay. Performance is evaluated via the visual interruption success rate measuring the proportion of samples satisfying this criterion.