STEADYTALK: REMEMBERING EVERY SPEAKER FOR LONG-HORIZON AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

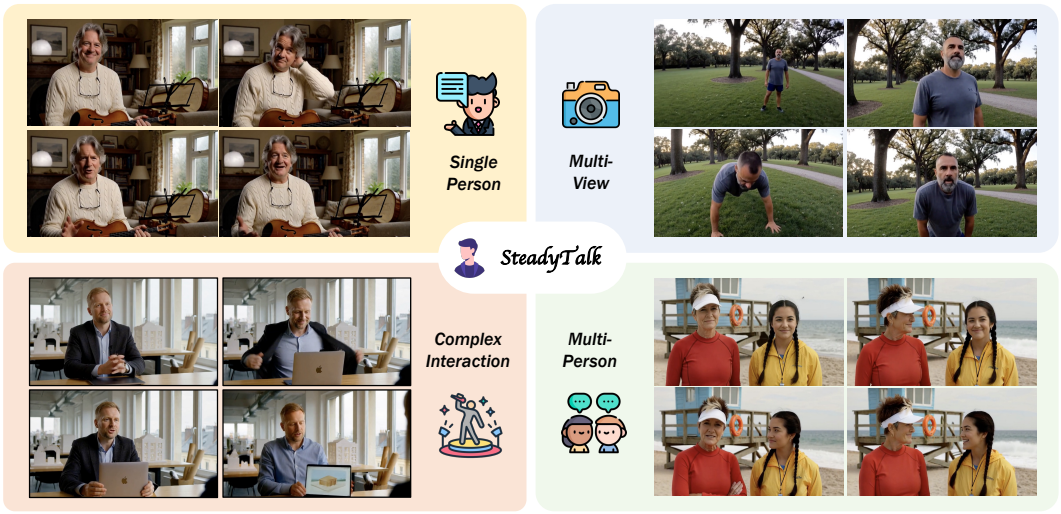

Figure 1: **Four settings supported by SteadyTalk:** single-person talking, multi-view generation with camera changes, complex interaction with actions and objects, and multi-person conversation.

ABSTRACT

Digital humans often appear in long videos with several speakers, and each speaker must keep the same voice and face throughout. Joint autoregressive generators drift in timbre and color over such long videos and forget a speaker's voice and face after a long silence. In audio-driven generators, the given audio dominates the output and limits control over the scene. To address these issues, we propose SteadyTalk, a long audio-video generation framework for single- and multi-person scenes, built upon three modules. First, Audio-First Asynchronous Generation drives the joint model with its own audio. A frozen audio branch generates the audio of each chunk first, and the video branch follows both this audio and the text prompt. Second, Multi-Person Audio-Visual Memory stores key audio segments and frames for each participant and recalls them with the current prompt, so a returning speaker keeps the same voice and face. Third, Reference-Anchored Drift Correction aligns the color statistics of each chunk to the opening frames and its audio latent statistics to the speaker's first utterance. On one-minute single- and two-person conversations, SteadyTalk outperforms existing methods.

1 INTRODUCTION

Joint audio-video diffusion models generate speech, sound, and video on a shared timeline from a text description (HaCohen et al., 2026; Low et al., 2025; SII-OpenMOSS Team, 2026; Liu et al., 2026b). They are a natural backbone for digital humans in live commerce and interactive dialogue. These applications need videos that last for minutes, often with several people who speak in turns. Each person must keep the same voice and appearance throughout. Such long videos are usually generated autoregressively, one chunk at a time, with each chunk conditioned on the previous ones (Yin et al., 2025; Huang et al., 2025b; Zhu 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.

Existing methods follow two routes. Joint autoregressive generators produce audio and video together, chunk by chunk (Su et al., 2026; Li et al., 2026; Zhou et al., 2026; Bai et al., 2026; Gan et al., 2026; Duan et al., 2026). Over long videos, they are prone to timbre drift and long-term semantic forgetting. Audio-driven generators animate a person from a given audio track (Kong et al., 2026; Yang et al., 2025). The audio dominates their output, and this limits the controllability and flexibility of the generated scenes. A better choice lies between the two routes: a joint audio-video model that is driven by its own generated audio. Such a model keeps the text control of joint generation and gives the video a stable audio signal to follow.

Long multi-person videos raise a second problem. To keep the cost of each chunk bounded, the generator attends only to a window of recent chunks. A participant who stays silent for a long time may have no utterance and no speaking frame left in this window. When this participant speaks again, the output no longer matches the voice and appearance generated earlier, and voices or faces can be swapped between participants. Memory mechanisms for long video generation store visual history only (Meng et al., 2026; Zhao et al., 2026; Lu et al., 2026; Lin et al., 2026), and audio-visual memories are designed for a single subject (Gan et al., 2026) or for cross-shot stories (Duan et al., 2026). Long multi-person videos need a memory that recalls the earlier voice and appearance of each participant and uses them in generation.

Errors also accumulate over long rollouts. An autoregressive generator is trained with clean context but continues from its own outputs at inference. Small errors therefore compound, and the video color and the voice timbre gradually drift. This drift has to be corrected on every chunk.

Based on these observations, we propose SteadyTalk, a long audio-video generation framework for single- and multi-person scenes, built on the open-source joint audio-video model MiniMax-H3[1]. Figure 1 shows the four settings it supports: single-person talking, multi-view generation, complex interaction, and multi-person conversation. SteadyTalk consists of three modules.

Audio-First Asynchronous Generation (Section 3.2) lets the base model drive itself with its own audio. For each chunk, the frozen audio branch generates the audio first. The video branch then generates the video from this finished audio and the text prompt. The audio gives lip and body motion a stable signal, and the prompt still controls actions, camera, and scene.

Multi-Person Audio-Visual Memory (Section 3.3) recalls the earlier voice and appearance of each participant. The memory keeps a reservoir of key audio segments and key frames for each participant. Audio segments are encoded with CAM++ (Wang et al., 2023), and frames with SigLIP 2 (Tschannen et al., 2025). A new entry is stored only when it differs enough from the last entry of the same participant, so the reservoir stays sparse. When the prompt changes, the memory returns the frames that best match the prompt and the earliest audio of the named speaker. These entries enter the generation context together with the recent chunks.

Reference-Anchored Drift Correction (Section 3.4) corrects the accumulated drift on every chunk. Following DynaMem (Lin et al., 2026), it aligns the color statistics of each chunk to those of the first frames in CIELAB space. It also aligns the channel-wise mean and variance of each audio latent to those of the current speaker's first utterance. Time-pooled statistics mainly carry timbre rather than spoken content (Chou et al., 2019), so this alignment keeps what is said. Both corrections are closed-form and add no denoising steps.

Our main contributions are summarized as follows:

- We propose Audio-First Asynchronous Generation, a generation scheme in which a joint audio-video model is driven by its own generated audio.
- We propose Multi-Person Audio-Visual Memory, a participant-partitioned memory that recalls the earlier voice and appearance of each participant.
- We propose Reference-Anchored Drift Correction, a closed-form correction that suppresses color and timbre drift in long generation.

[1] https://github.com/MiniMax-AI/MiniMax-H3

2 Related Work

Audio-video and talking-human generation. Video diffusion models build on denoising diffusion, latent-space generation, classifier-free guidance, and transformer denoisers (Ho et al., 2020; Song et al., 2021; Ho & Salimans, 2022; Rombach et al., 2022; Peebles & Xie, 2023; Blattmann et al., 2023b;a; HaCohen et al., 2024). Joint audio-video models generate both modalities from text within one network. Their designs range from a single transformer with separate audio and video streams to fused pretrained audio and video backbones (HaCohen et al., 2026; Low et al., 2025; SII-OpenMOSS Team, 2026; Liu et al., 2026a;b; Wang et al., 2025). These models generate each clip offline with many denoising steps. Audio-driven methods take the other route: the speech is given, and the model animates a face or a body to match it (Prajwal et al., 2020; Zhou et al., 2020; Wang et al., 2021; Zhang et al., 2023a; Ye et al., 2023; Tian et al., 2024; Xu et al., 2024). Recent diffusion-transformer methods extend audio-driven generation to full bodies, long videos, and conversations among several people (Cui et al., 2025; Gan et al., 2025; Tu et al., 2025; Gao et al., 2025; Yang et al., 2025; Kong et al., 2026). With the speech given in advance, these methods reach accurate lip synchronization, but they do not generate the speech themselves.

Streaming generation and distillation. A streaming generator produces a video chunk by chunk with causal attention and a KV cache. Such generation needs few-step sampling, which is usually obtained by distillation, for example progressive distillation, consistency training, or distribution matching (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Yin et al., 2024b;a). CausVid distills a bidirectional video model into a causal few-step student (Yin et al., 2025). A student trained on teacher trajectories sees clean history during training but must continue from its own outputs at inference. Self Forcing and Causal Forcing reduce this exposure bias by training on the student's own rollouts (Huang et al., 2025b; Zhu et al., 2026), and LongLive applies causal generation to long interactive videos (Yang et al., 2026a). The same recipe has been used for avatars and for joint audio-video generation. Live Avatar and AvatarForcing stream audio-driven avatars (Huang et al., 2025c; Cui et al., 2026), and other work distills joint audio-video models into causal few-step generators (Su et al., 2026; Li et al., 2026; Zhou et al., 2026; Bai et al., 2026; Gan et al., 2026; Duan et al., 2026). Within each chunk, these joint generators still denoise audio and video together. Hallo-Live lets each video block attend to a provisional future audio block, which reduces articulation lag (Li et al., 2026), but this audio is denoised in the same few-step passes as the video.

Long-context memory. Language models reach past a fixed context window with segment-level recurrence, compressed memories, retrieval over stored key-value pairs, attention sinks, retention of heavily attended tokens, and blockwise attention (Dai et al., 2019; Rae et al., 2020; Wu et al., 2022; Xiao et al., 2024; Zhang et al., 2023b; Liu et al., 2023). Video world models add memory through test-time weight adaptation (Hong et al., 2024), explicit spatial representations of the scene (Huang et al., 2025a; Li et al., 2025), or retrieval of past frames by camera pose or action (Xiao et al., 2025; Yu et al., 2025a), all of which rely on camera or action signals that a talking-human stream does not have. For text-driven long videos, recent methods manage the KV cache by recalling, selecting, sparsifying, consolidating, or re-anchoring past entries (Meng et al., 2026; Chen et al., 2026; Xu et al., 2026; Zhao et al., 2026; Lu et al., 2026; Yang et al., 2026b), or they carry a compact state from chunk to chunk (Henschel et al., 2025; Yu et al., 2025b; Dalva & Yanardag, 2026; Dou et al., 2026). DynaMem keeps a sparse set of key frames and retrieves them with the current prompt (Lin et al., 2026). All of these memories store visual content only. Audio-visual memory has been built for a single presenter (Gan et al., 2026) and for cross-shot stories, where memory slots taken from earlier shots are chosen for each new shot (Duan et al., 2026).

3 Method

3.1 Problem Formulation

Given a reference image and an ordered sequence of text prompts, we generate synchronized audio-video chunks. Let $\mathbf{z}_k = (\mathbf{z}_k^a, \mathbf{z}_k^v)$ denote the audio and video latents of the k-th chunk and e_k its prompt. Consecutive chunks that share a prompt form a segment. Each prompt names the participant who speaks in its segment, and we denote this participant by the tag $s_k \in \{1, \dots, P\}$. This work

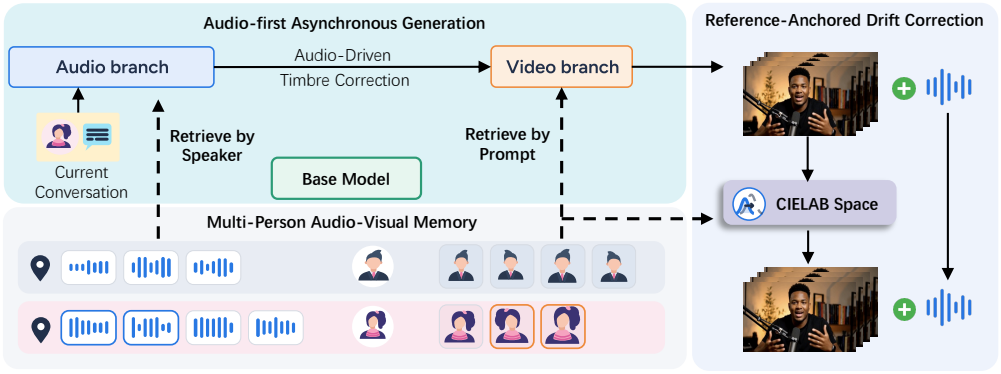

Figure 2: **Overview of SteadyTalk.** *Audio-first asynchronous generation* (top left): the frozen audio branch generates the audio of each chunk, and the video branch follows the timbre-corrected audio. *Multi-person audio-visual memory* (bottom): a reservoir partitioned by participant stores key audio segments and crops; the audio branch recalls the speaker's earliest audio, and the video branch recalls the crops that best match the prompt. *Reference-anchored drift correction* (right): the colors of each chunk are aligned to reference statistics in CIELAB space.

considers scenes with one or two participants, $P \leq 2$. Each prompt also describes the camera, so the viewpoint changes of the multi-view setting in Figure 1 need no extra control signal.

Two kinds of history matter when chunk k is generated. Lip motion, gestures, and phonetic continuity depend on the most recent chunks. The generator reads them through a bounded context $\mathcal{C}_{k-1} = (\mathcal{C}^a_{k-1}, \mathcal{C}^v_{k-1})$, which holds the first chunk of the video as reference frames and a few recent clean chunks at explicit temporal positions. The voice and appearance of a participant, in contrast, may have been observed long before chunk k and must stay available after those chunks have left $\mathcal{C}_{k-1}$. We keep this evidence in a reservoir $\mathcal{R}_{k-1}$ and read from it a small working memory $\mathcal{W}_k = (\mathcal{W}^a_k, \mathcal{W}^v_k)$ for each segment. SteadyTalk factorizes the generation of a chunk as:

$$\begin{aligned} \mathbf{z}^a_k &\sim p_\phi\big(\mathbf{z}^a_k \mid e_k, \mathcal{C}^a_{k-1}, \mathcal{W}^a_k\big), \qquad \hat{\mathbf{z}}^a_k = \Gamma_a\big(\mathbf{z}^a_k\big), \\ \mathbf{z}^v_k &\sim p_\theta\big(\mathbf{z}^v_k \mid e_k, \hat{\mathbf{z}}^a_k, \mathcal{C}^v_{k-1}, \mathcal{W}^v_k\big), \\ \mathcal{R}_k &= U\big(\mathcal{R}_{k-1}, \hat{\mathbf{z}}^a_k, \Gamma_v(\mathbf{z}^v_k), e_k, s_k\big), \end{aligned} \tag{1}$$

where ϕ denotes the frozen audio branch, θ the trainable parameters of the video branch, Γ_a and Γ_v the audio and video drift corrections of Section 3.4, and U the reservoir update of Section 3.3. The audio of a chunk depends only on the prompt and on audio history. The video of a chunk depends on the finished and corrected audio of the same chunk, so the model is driven by its own audio. The same ordering also lets the video follow audio from an external source, such as the lines of an off-screen interviewer (Appendix A.3). Because $\mathcal{C}_{k-1}$ and $\mathcal{W}_k$ have bounded size, the cost of conditioning a chunk stays constant as the video gets longer. Figure 2 gives an overview.

3.2 Audio-First Asynchronous Generation

In the base model, audio and video are denoised together at a shared timestep. A causal generator distilled from it samples the audio with the few steps of the video, and each video step attends to still-noisy audio. SteadyTalk instead generates the audio of a chunk first and lets the video follow it. The model is thus driven by its own audio, and the prompt still controls the scene.

Audio stage. The audio branch of the base model generates $\mathbf{z}^a_k$ from the prompt, the audio context, and the audio working memory with S_a sampling steps, without video tokens in its input. We keep ϕ frozen and do not distill it. The audio therefore comes from the weights and sampler of the base model and is not bound to the few-step schedule of the video.

Video stage. The video branch denoises $\mathbf{z}^v_k$ in S_v steps while attending to the finished audio. The corrected audio latent $\hat{\mathbf{z}}^a_k$ is clean and does not change during video denoising. We therefore run the frozen audio branch once on it at timestep zero and cache the audio keys and values of every cross-modal attention layer, which all video denoising steps of the chunk reuse. The clean audio gives lip and body motion a fixed target to follow.

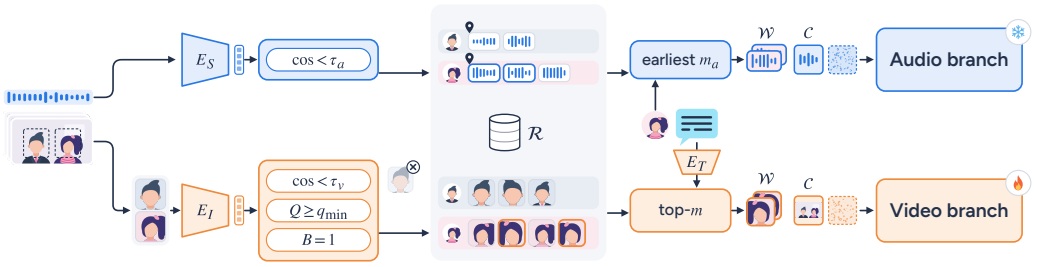

Figure 3: **Multi-person audio-visual memory.** Key audio segments and participant crops that pass the gates of Eq. (2) are filed under participant tags in the reservoir $\mathcal{R}$. The earliest m_a audio entries of the current speaker and the top-m crops for the prompt form the working memory $\mathcal{W}$, which enters the frozen audio branch and the trainable video branch together with the context $\mathcal{C}$.

Training the video branch. The base model always denoises both modalities at a shared noise level, so its video branch has never been conditioned on clean audio while its own latent is noisy. We therefore adapt the video branch with LoRA (Hu et al., 2022) and keep the audio branch and all base weights frozen. During training, the video branch receives the ground-truth audio latent at timestep zero without added noise, which matches the condition it receives from the audio stage at inference. We distill the video branch into an S_v-step causal generator with distribution matching distillation (DMD) (Yin et al., 2024b;a), using the multi-step base model as the teacher. A critic tracks the distribution of the student's outputs, and the difference between the teacher and critic scores on noised student samples gives the gradient for the student. The teacher sees the audio noised to the same timestep as the video, as in its own joint generation. It therefore stays within its training distribution, and only the student leaves the shared-timestep regime. Following causal self-rollout training (Huang et al., 2025b; Gan et al., 2026), we distill the student on its own rollouts. Appendix A.2 gives the details.

3.3 Multi-Person Audio-Visual Memory

A bounded context cannot hold the voice and appearance of a participant who has long been silent. Keeping every past chunk, however, makes storage and retrieval grow with duration, although most chunks repeat their neighbors. The reservoir is partitioned by participant so that each entry stays attached to the right person. Figure 3 shows how the memory is written and read.

Participant-partitioned reservoir. The reservoir $\mathcal{R} = \{(\mathcal{F}^{(p)}, \mathcal{A}^{(p)})\}_{p=1}^{P}$ holds a set of key frames $\mathcal{F}^{(p)}$ and a set of key audio segments $\mathcal{A}^{(p)}$ for each participant p. Key frames are stored as crops of individual participants. When a frame shows two people, we crop each of them and file each crop under the tag of the participant whose appearance description in the prompt it matches. A crop x is represented by its SigLIP 2 image embedding $E_I(x)$ (Tschannen et al., 2025). An audio segment a is filed under the tag of the speaker named in the prompt and represented by its CAM++ speaker embedding $E_S(a)$ (Wang et al., 2023).

Gated insertion. Consecutive chunks are largely redundant, so an observation enters the reservoir only when it adds new content. After chunk k has been generated and corrected, a crop x of participant p is inserted into $\mathcal{F}^{(p)}$ if

$$\cos\left(E_I(x), E_I(x_{\text{last}}^{(p)})\right) < \tau_v, \qquad Q(x) \geq q_{\min}, \qquad B(x) = 1, \tag{2}$$

where $x_{\text{last}}^{(p)}$ is the most recently stored crop of participant p, Q is an aesthetic score, and B is a blur check that returns one for sharp crops. The audio segment a_k of the chunk is inserted into $\mathcal{A}^{(s_k)}$ if $\cos\left(E_S(a_k), E_S(a_{\text{last}}^{(s_k)})\right) < \tau_a$, where $a_{\text{last}}^{(s_k)}$ is the most recently stored segment of the same speaker. The first observation of each participant is always inserted. The audio set $\mathcal{A}^{(p)}$ therefore begins with the audio of the first chunk in which participant p speaks, which we call the first utterance of p and which Section 3.4 uses as the timbre reference. The reservoir grows with new poses, expressions, and speaking styles rather than with the number of generated chunks.

Prompt-guided retrieval. Whenever the prompt changes, SteadyTalk rebuilds the working memory. The visual part $\mathcal{W}_k^v$ consists of the m crops, taken from all participants, whose image embeddings are most similar to the SigLIP 2 text embedding $E_T(e_k)$ of the current prompt. The audio part $\mathcal{W}_k^a$ consists of the earliest m_a entries of the current speaker in $\mathcal{A}^{(s_k)}$. These entries are selected by tag rather than by similarity; they include the first utterance and precede most of the drift that accumulates later in the video. The retrieved crops and audio segments are encoded by the video and audio VAEs, and the resulting latents enter the corresponding branches as clean context tokens next to $\mathcal{C}_{k-1}$. They are prefilled into the KV cache at the start of the segment and are never denoised, so they condition every denoising step of the segment at the cost of a fixed number of additional tokens. Insertion costs one image-encoder pass per candidate crop and one speaker-encoder pass per chunk, and retrieval is a single similarity search over a sparse reservoir; both are small compared with one forward pass of the denoiser. The video branch learns to read the working memory during distillation, as detailed in Appendix A.2.

3.4 Reference-Anchored Drift Correction

Over long rollouts, errors accumulate as slow shifts in global statistics: the video changes in brightness and saturation, and each voice drifts from its initial timbre. We align the first- and second-order statistics of every chunk to references fixed at the start of the video. The alignment is closed-form and adds no denoising steps.

Video color alignment. Let $\mathbf{x}_f$ be a decoded frame of chunk k, ψ the RGB-to-CIELAB conversion, and $\boldsymbol{\mu}_f, \boldsymbol{\sigma}_f \in \mathbb{R}^3$ the channel-wise spatial mean and standard deviation of $\psi(\mathbf{x}_f)$. The reference moments $\boldsymbol{\mu}_{\mathrm{ref}}$ and $\boldsymbol{\sigma}_{\mathrm{ref}}$ are averaged over the first N_r frames of the video. The corrected frame is

$$\hat{\mathbf{x}}_f = (1-\alpha)\,\mathbf{x}_f + \alpha\,\psi^{-1}\!\left(\frac{\boldsymbol{\sigma}_{\mathrm{ref}}}{\boldsymbol{\sigma}_f + \epsilon} \odot \big(\psi(\mathbf{x}_f) - \boldsymbol{\mu}_f\big) + \boldsymbol{\mu}_{\mathrm{ref}}\right), \tag{3}$$

where $\odot$ is channel-wise multiplication, ϵ is a small constant, and $\alpha \in [0,1]$ sets the correction strength. Blending with the original frame avoids over-correction when a legitimate change in expression, pose, or framing alters the color statistics.

Audio timbre alignment. Let $\mathbf{z}_k^a \in \mathbb{R}^{C_a \times F_a}$ be the audio latent of chunk k, and let $\boldsymbol{\mu}_k^a$ and $\boldsymbol{\sigma}_k^a$ be its channel-wise mean and standard deviation over the F_a temporal positions. The reference is the first utterance of the current speaker, i.e., the first entry of $\mathcal{A}^{(s_k)}$, whose latent moments we denote by $\boldsymbol{\mu}_{\mathrm{ref}}^{(s_k)}$ and $\boldsymbol{\sigma}_{\mathrm{ref}}^{(s_k)}$. The first chunk in which a participant speaks has no reference yet, so it is left uncorrected and becomes the reference. Since the tag s_k comes from the prompt, the reference belongs to the participant who is supposed to speak, regardless of how the generated voice currently sounds. The corrected latent is

$$\hat{\mathbf{z}}_k^a = \frac{\boldsymbol{\sigma}_{\mathrm{ref}}^{(s_k)}}{\boldsymbol{\sigma}_k^a + \epsilon} \odot \big(\mathbf{z}_k^a - \boldsymbol{\mu}_k^a\big) + \boldsymbol{\mu}_{\mathrm{ref}}^{(s_k)}. \tag{4}$$

Instance-normalization voice conversion shows that channel statistics pooled over time carry speaker characteristics, whereas the normalized temporal pattern carries content (Chou et al., 2019). Equation (4) changes only per-channel statistics and keeps the normalized pattern $(\mathbf{z}_k^a - \boldsymbol{\mu}_k^a)/\boldsymbol{\sigma}_k^a$ unchanged, so it moves the timbre toward the reference without altering what is said or when. The corrected outputs of both modalities are the ones written to the reservoir (Appendix A.3).

4 Experiments

4.1 Implementation Details

Benchmark. Existing audio-video benchmarks use short clips and cannot test whether a participant is recalled correctly after a long silence. We therefore evaluate on SteadyBench, our set of interview-style conversations in Mandarin and English. In each scenario, an off-screen interviewer asks four questions, and the on-screen subjects listen and answer. The interviewer's lines are given as synthesized speech. SteadyBench contains 150 scenarios: 70 single-presenter conversations, 20

complex-action conversations, and 60 two-person conversations. In the complex-action conversations, the presenter performs a body action in the middle of an answer. Each scenario is described by twelve 5-second prompts, and every output contains 60 seconds of video at 864×480 and 24 fps with 32 kHz stereo audio. Appendix A.1 describes the prompts in detail.

Training and inference. The causal video generator is initialized from a distilled MiniMax-H3 checkpoint and optimized with DMD against the frozen bidirectional MiniMax-H3 teacher. Only the LoRA parameters of the video branch are trained, using filtered synthetic trajectories from the bidirectional MiniMax-H3. At inference, the frozen audio branch samples the audio with more steps than the few-step video branch. Appendices A.2 and A.3 give the full settings.

4.2 Comparison with State-of-the-Art Methods

Metrics and Baselines. For synchronization, we report LipSync and Desync from VABench (Hua et al., 2026) and Listener, the share of frames in which a silent on-screen subject keeps the mouth closed. For speech, we report WER, computed with Whisper (Radford et al., 2023) and Qwen3-ASR (Shi et al., 2026), and the WavLM speaker similarity Spk-Sim (Chen et al., 2022). For identity, we report the ArcFace similarity ID-Sim (Deng et al., 2018) and, in two-person scenarios, the participant swap rate Swap. Color Δ and Long Consistency measure stability. We compare with bidirectional joint audio-video models (LTX-2.3 (HaCohen et al., 2026), Ovi (Low et al., 2025), and MOVA (SII-OpenMOSS Team, 2026)), audio-driven methods fed with our generated audio (LiveAvatar (Huang et al., 2025c), InfiniteTalk (Yang et al., 2025), and MultiTalk (Kong et al., 2026)), and autoregressive joint audio-video models (OmniForcing (Su et al., 2026), Hallo-Live (Li et al., 2026), and JoyAI-Echo-1.5 (Duan et al., 2026)). All methods receive the same prompts and reference images, and Appendix A.4 gives the metric definitions and the setup of each baseline.

Table 1: Comparison on SteadyBench. Best and second-best results are in bold and underlined, respectively. †Audio-driven methods receive the audio generated by SteadyTalk, so Desync, speech metrics, and Swap are not reported for them. ‡Evaluated on the first 30 seconds, their maximum length; Swap is undefined because each participant answers once in this span.

		Synchronization			Speech		Identity		Stability	
Method	Params	LipSync↑	Desync↓	Listener (%)↑	WER (%)↓	Spk-Sim↑	ID-Sim↑	Swap↓	Color Δ ↓	Long Cons.↑
Bidirectional joint audio-video generation										
LTX-2.3 (HaCohen et al., 2026)	22B	2.349	<u>0.050</u>	<u>73.2</u>	3.64	<u>0.569</u>	<u>0.671</u>	0.0667	0.04387	0.3745
Ovi (Low et al., 2025)	11B	3.493	0.135	55.8	3.98	0.482	0.492	0.0667	0.01797	0.8252
MOVA (SII-OpenMOSS Team, 2026)	32B/18B	3.860	0.060	63.2	<u>2.85</u>	0.513	0.568	<u>0.0334</u>	0.02825	0.7129
Audio-driven video generation†										
LiveAvatar (Huang et al., 2025c)	14B	<u>5.396</u>	–	70.1	–	–	0.571	–	<u>0.00541</u>	<u>0.9619</u>
InfiniteTalk (Yang et al., 2025)	14B	5.124	–	66.3	–	–	0.609	–	0.00578	0.9523
MultiTalk (Kong et al., 2026)	14B	4.967	–	64.9	–	–	0.591	–	0.02132	0.7034
Autoregressive joint audio-video generation										
OmniForcing‡ (Su et al., 2026)	22B	1.590	**0.045**	51.6	43.9	0.290	0.412	–	0.02060	0.3404
Hallo-Live‡ (Li et al., 2026)	11B	1.782	0.068	44.9	39.7	0.287	0.378	–	0.02741	0.2815
JoyAI-Echo-1.5 (Duan et al., 2026)	22B	5.266	0.061	71.4	3.87	0.551	0.654	0.3334	0.00892	0.8821
SteadyTalk (ours)	32B	**6.96**	0.057	**78.5**	**0.85**	**0.596**	**0.696**	**0.0000**	**0.00150**	**0.971**

Quantitative Comparison. Table 1 shows that SteadyTalk obtains the best result on eight of the nine metrics. OmniForcing and Hallo-Live denoise audio and video together in a few steps, and they lose most of the speech. Their WER exceeds 39%, and their Spk-Sim and LipSync are the lowest in the table. SteadyTalk also denoises the video in a few steps, but its WER stays below 1%. Its audio comes from the undistilled audio branch. JoyAI-Echo-1.5 is the strongest autoregressive baseline, and its LipSync is close to that of the audio-driven methods. It still gives a third of the answers the wrong voice or face. This suggests that memory slots chosen per shot cannot keep two alternating speakers apart in one continuous scene. The bidirectional models produce clear speech but swap participants in some answers. Their appearance is also less stable, most visibly for LTX-2.3. LiveAvatar and InfiniteTalk receive finished speech and reach high LipSync and stability. Their lower Listener scores show that they also move the listener's lips during the interviewer's lines. On Desync, OmniForcing and LTX-2.3 are slightly better than SteadyTalk. In a user study with 50 expert annotators, SteadyTalk is preferred over every autoregressive baseline (Appendix A.6).



Figure 4: **Qualitative comparison.** Frames at 10, 25, 60, and 120 seconds generated from the same product-presentation prompt. Hallo-Live and OmniForcing cannot generate beyond 30 seconds, so their panels at 60 and 120 seconds repeat their frame at 30 seconds. Because of the 100 MB size limit, the supplementary material includes only the first 20 seconds of these videos and a few demos.

Qualitative Comparison. Figure 4 compares SteadyTalk with Hallo-Live (Li et al., 2026), OmniForcing (Su et al., 2026), and JoyAI-Echo-1.5 (Duan et al., 2026), shown as Joy-Echo, on the same product-presentation prompt at 10, 25, 60, and 120 seconds. Hallo-Live and OmniForcing cannot generate beyond 30 seconds, so their panels at 60 and 120 seconds repeat their frame at 30 seconds. Within 30 seconds, their frames have low quality and oversaturated colors. Errors build up after about five seconds, and the subject becomes severely deformed. Both methods follow the instructions poorly, their lip motion does not match the speech, and the speech itself is unclear. They also support only a few languages. JoyAI-Echo-1.5 is designed for multi-shot stories. It generates longer videos but often skips frames, so it cannot produce a continuous single-take presentation. Its lip synchronization is weaker, and it sometimes renders garbled subtitles or hallucinated content. By the end of its video, the presenter has changed, and the requested product has disappeared. SteadyTalk keeps the presenter, the scene, and the product consistent over the full 120 seconds, and the presenter keeps showing the product as the prompt requests.

4.3 Ablation Study

We ablate each component of SteadyTalk, changing one factor at a time and keeping the rest fixed.

Table 2: **Audio-first ablation** on SteadyBench. All variants share the backbone, data, and training.

Variant	LipSync↑	Desync↓	WER (%)↓	Spk-Sim↑
Joint few-step	5.23	0.083	5.93	0.406
Audio-first, $S_a = 4$	6.12	0.069	3.91	0.493
Audio-first, $S_a = 10$ (ours)	**6.96**	**0.057**	**0.85**	**0.596**

Audio-First Generation. We train a joint few-step baseline on the same backbone with the same data, DMD objective, and causal rollouts. It denoises audio and video together at a shared timestep in $S_v = 4$ steps, as existing autoregressive joint generators do. Its memory and drift correction are the same as in SteadyTalk. We also run SteadyTalk with $S_a = 4$ audio steps. This variant separates the effect of the audio sampling steps from the effect of letting the video follow finished audio. Table 2 reports the results on all scenarios of SteadyBench. Both parts of the design help. With the same four audio steps as the joint baseline, audio-first generation already improves all four metrics. Desync and LipSync improve as well, and these two metrics depend only on how the video follows

the audio. The video now attends to a fixed audio track instead of a still-noisy estimate, and the frozen audio branch produces clearer speech than the jointly distilled audio. Giving the audio its own sampler brings the larger part of the speech gain. Much of the loss in intelligibility and timbre under joint few-step generation therefore comes from forcing the audio onto the few-step schedule of the video. The cleaner audio also improves LipSync.

Multi-Person Audio-Visual Memory. We evaluate memory variants on the 60 two-person scenarios. In these scenarios, each subject speaks again after the other has answered. All variants use the same trained model and differ only at inference. The first variant removes the memory tokens and relies on the recent context alone. Two variants keep only the visual or only the audio part of the working memory. A shared-reservoir variant drops the participant partition and injects the stored audio of both participants. Table 3 reports the results. Without memory, a quarter of the later answers go to the wrong participant, and voice and face similarity fall well below the full model. The recent context alone cannot keep two alternating speakers apart. Either part of the memory removes these swaps when its entries are separated by participant. With the audio of both participants injected, swaps return in about one of six answers, and ID-Sim falls almost to the level without memory. This suggests that conflicting evidence about who is speaking also disturbs the faces. With the full memory, the listener also keeps the mouth closed more often while the other participant speaks.

Table 3: **Ablation of the multi-person audio-visual memory** on the 60 two-person scenarios of SteadyBench. All variants use the same trained model.

Visual	Audio	Participant tags	Spk-Sim↑	ID-Sim↑	Swap↓	Listener (%)↑
–	–	–	0.459	0.507	0.2500	72.1
✓	–	✓	0.485	0.635	0.0000	82.3
–	✓	✓	0.538	0.569	0.0000	78.5
✓	✓	–	0.516	0.532	0.1667	80.2
✓	✓	✓	**0.544**	**0.663**	**0.0000**	**85.3**

Drift Correction. We enable the video and audio corrections separately and together on all scenarios of SteadyBench. Besides Color Δ, Long Consistency, and Spk-Sim, we report WER and LipSync to check the spoken content and its timing. Table 4 compares the variants, and Table 5 tracks Spk-Sim and color shift over the minute. Without correction, drift builds up steadily. The color shift grows at every ten-second mark, and Spk-Sim falls as the voice moves away from the first answer. With both corrections, both curves stay nearly flat. The video correction removes most of the color drift, and the audio correction removes most of the timbre drift. Each correction changes the metrics of the other modality much less, so the two target different sources of drift. Applying both gives the best result on every metric, and WER and LipSync also improve.

Table 4: **Ablation of drift correction** on SteadyBench.

Video	Audio	Color Δ ↓	Long Cons.↑	Spk-Sim↑	WER (%)↓	LipSync↑
–	–	0.0431	0.729	0.504	1.81	6.02
✓	–	0.0024	0.957	0.532	1.82	6.74
–	✓	0.0381	0.741	0.589	1.03	6.78
✓	✓	**0.0015**	**0.971**	**0.596**	**0.85**	**6.96**

Table 5: **Spk-Sim and color shift over time** with and without correction.

Metric	Correction	10s	20s	30s	40s	50s	60s
Spk-Sim↑	w/o	0.611	0.591	0.582	0.569	0.532	0.507
	w/	0.608	0.599	0.601	0.594	0.588	0.591
Color shift↓	w/o	0.0017	0.0034	0.0097	0.0162	0.0213	0.0298
	w/	0.0012	0.0019	0.0014	0.0017	0.0029	0.0016

5 CONCLUSION

We presented SteadyTalk for long single- and multi-person audio-video generation. SteadyTalk drives a joint audio-video model with its own audio, recalls each participant's voice and appearance with a participant-partitioned memory, and suppresses color and timbre drift with a closed-form alignment. On SteadyBench, it obtains the best result on eight of the nine metrics. The current system handles at most two on-screen participants, takes speaker identities from the prompt, and generates audio without seeing the video. Extending the memory to larger groups and letting the audio branch read the video are left for future work.

AI USE STATEMENT

We used large language models (LLMs) as writing and editing assistants. They helped draft and revise text based on content and decisions provided by the authors, polish wording, correct grammar, format tables, and carry out routine computations on results collected by the authors, such as converting vote counts to percentages. The research ideas, method design, experiments, and data are the authors' own, and all citations were chosen and checked by the authors. The authors reviewed and edited all text and take full responsibility for the content of this paper.

ETHICS STATEMENT

SteadyTalk generates realistic speech and video of people, which could be misused to impersonate real individuals or to produce misleading content. In our experiments, the voices of the on-screen participants are generated from text descriptions in the prompts rather than cloned from recordings. The user study asked annotators, recruited through a professional annotation vendor, only for preference votes between pairs of generated videos. We encourage clear labeling of generated content and the use of watermarking and detection tools when such models are deployed. The authors have read and adhere to the ICLR Code of Ethics.

REFERENCES

Lichen Bai, Tianhao Zhang, Shitong Shao, Dingwei Tan, Qiyu Zhong, et al. Mainecoon: Pursuing a real-time audio-visual social world model. *arXiv preprint arXiv:2606.17800*, 2026. URL https://arxiv.org/abs/2606.17800.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a. URL https://arxiv.org/abs/2311.15127.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Blattmann_Align_Your_Latents_High-Resolution_Video_Synthesis_With_Latent_Diffusion_Models_CVPR_2023_paper.html.

Jiayu Chen, Junbei Tang, Wenbiao Zhao, Maoliang Li, Jiayi Luo, Zihao Zheng, et al. Pyramid forcing: Head-aware pyramid kv cache policy for high-quality long video generation. *arXiv preprint arXiv:2605.13111*, 2026. URL https://arxiv.org/abs/2605.13111.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*, 2019.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21086–21095, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cui_Hallo3_Highly_Dynamic_and_Realistic_Portrait_Image_Animation_with_Video_CVPR_2025_paper.html.

Liyuan Cui, Wentao Hu, Wenyuan Zhang, Zesong Yang, Fan Shi, and Xiaoqiang Liu. Avatarforcing: One-step streaming talking avatars via local-future sliding-window denoising. *arXiv preprint arXiv:2603.14331*, 2026. URL https://arxiv.org/abs/2603.14331.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285/.

Yusuf Dalva and Pinar Yanardag. Adastate: Self-evolving anchors for streaming video generation. *arXiv preprint arXiv:2605.30349*, 2026. URL https://arxiv.org/abs/2605.30349.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. URL https://arxiv.org/abs/1801.07698.

Weijia Dou, Hui Li, Jiahao Cui, Lei Zhou, Jingdong Wang, and Siyu Zhu. Slotmemory: Object-centric kv memory for streaming long-video generation. *arXiv preprint arXiv:2605.31033*, 2026. URL https://arxiv.org/abs/2605.31033.

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, et al. Long-horizon audio-visual generation for persistent stories and interactive worlds. *arXiv preprint arXiv:2608.23383*, 2026.

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

Qijun Gan, Chenwei Zhang, Meiguang Jin, Junfeng Ma, and Qiu Shen. Taomate: Anchor-guided memory bridging evolving and reference states for real-time audio-video digital human generation. *arXiv preprint arXiv:2607.24359*, 2026.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025. URL https://arxiv.org/abs/2508.18621.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL https://arxiv.org/abs/2501.00103.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. URL https://arxiv.org/abs/2601.03233.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2568–2577, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Yining Hong, Beide Liu, Maxine Wu, Yuanhao Zhai, Kai-Wei Chang, Linjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, et al. Slowfast-vgen: Slow-fast learning for action-driven long video generation. *arXiv preprint arXiv:2410.23277*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. VABench: A comprehensive benchmark for audio-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025c. URL https://arxiv.org/abs/2512.04677.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019. URL https://arxiv.org/abs/1907.06724.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *Advances in Neural Information Processing Systems*, 38:70990–71013, 2026.

Chunyu Li, Jiaye Li, Ruiqiao Mei, Haoyuan Xia, Hao Zhu, Jingdong Wang, and Siyu Zhu. Hallo-live: Real-time streaming joint audio-video avatar generation with asynchronous dual-stream and human-centric preference distillation. *arXiv preprint arXiv:2604.23632*, 2026.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025.

Jingyu Lin, Xinyi Shang, Peng Sun, Cunjian Chen, and Zhiqiang Shen. Dynamem: Consistent long video generation via hierarchical memory and motion priors. In *Forty-third International Conference on Machine Learning*, 2026.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023. URL https://arxiv.org/abs/2310.01889.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Scofield Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, volume 2026, pp. 139160–139194, 2026a.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026b. URL https://arxiv.org/abs/2602.19163.

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

Yu Lu, Junjie Yang, Piotr Koniusz, Yuxin Song, and Yi Yang. Fademem: Distance-aware memory consolidation for autoregressive video diffusion. *arXiv preprint arXiv:2606.10671*, 2026. URL https://arxiv.org/abs/2606.10671.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. URL https://arxiv.org/abs/2310.04378.

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Jiang, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. Tethercache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *arXiv preprint arXiv:2606.13035*, 2026. URL https://arxiv.org/abs/2606.13035.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

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. doi: 10.1145/3394171.3413532. URL https://doi.org/10.1145/3394171.3413532.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.05507.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://iclr.cc/virtual/2022/spotlight/6538.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

SII-OpenMOSS Team. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026. URL https://arxiv.org/abs/2602.08794.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation. In *European Conference on Computer Vision*, pp. 566–584. Springer, 2026.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive – generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260, 2024. doi: 10.1007/978-3-031-73010-8_15. URL https://link.springer.com/chapter/10.1007/978-3-031-73010-8_15.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 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. *arXiv preprint arXiv:2508.08248*, 2025. URL https://arxiv.org/abs/2508.08248.

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. *arXiv preprint arXiv:2509.06155*, 2025.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*, 2023.

Suzhen Wang, Lincheng Li, Yu Ding, Chen Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1098–1105, 2021. doi: 10.24963/ijcai.2021/152. URL https://www.ijcai.org/proceedings/2021/152.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TrjbxzRcnf-.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e5fd18f863cbe6d8ae392a93fd271c9-Abstract-Conference.html.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

Boxun Xu, Yuming Du, Zichang Liu, Siyu Yang, Ziyang Jiang, et al. Sparse forcing: Native trainable sparse attention for real-time autoregressive diffusion video generation. *arXiv preprint arXiv:2604.21221*, 2026. URL https://arxiv.org/abs/2604.21221.

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. *arXiv preprint arXiv:2406.08801*, 2024. URL https://arxiv.org/abs/2406.08801.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=nCAODkpsPJ.

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region RoPE for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026b.

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=YfwMIDhPccD.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-1505. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54dcf25318f9de5a7a01f0a4125c541e-Abstract-Conference.html.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025a.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, et al. Videossm: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025b. URL https://arxiv.org/abs/2512.04519.

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*, pp. 8652–8661, 2023a. doi: 10.1109/CVPR52729.2023.00836. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_SadTalker_Learning_Realistic_3D_Motion_Coefficients_for_Stylized_Audio-Driven_Single_CVPR_2023_paper.html.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

Lin Zhao, Yushu Wu, Yifan Gong, Yanzhi Wang, and Pu Zhao. Omnimem: Scalable and adaptive memory retrieval for long video generation. *arXiv preprint arXiv:2605.30519*, 2026. URL https://arxiv.org/abs/2605.30519.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39 (6), 2020. doi: 10.1145/3414685.3417774. URL https://doi.org/10.1145/3414685.3417774.

Yupeng Zhou, Lianghua Huang, Zhifan Wu, Jiabao Wang, Yupeng Shi, et al. Mutual forcing: Dual-mode self-evolution for fast autoregressive audio-video character generation. *arXiv preprint arXiv:2604.25819*, 2026. URL https://arxiv.org/abs/2604.25819.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026. URL https://arxiv.org/abs/2602.02214.

A Appendix

A.1 Benchmark Details

Each scenario of SteadyBench comprises four question-answer turns. It is represented by a temporally ordered sequence of twelve 5-second prompts, one segment per question and two per answer, and each prompt carries 4–13 words of transcript. Every prompt specifies the subject appearance and pose, scene, camera, speaker identity with a stable speaker ID, voice, and transcript, together with either the listener's reaction or the speaker's delivery. The subject, scene, camera, and voice descriptions remain consistent across a sequence, whereas the transcript, reactions, and actions evolve. The interviewer is never shown. The interviewer's lines are synthesized by a text-to-speech model and inserted as the audio condition of the corresponding segments, and the prompts of these segments render them as off-screen voiceover with the lips of the on-screen subjects explicitly closed. In the complex-action category, the presenter performs a body action mid-answer, such as standing up, removing a jacket or glasses, or pouring a drink, and the subject description switches to the post-action state for the remaining segments. In the two-person category, the interviewer addresses each on-screen subject by name, and each subject answers two of the questions.

A.2 Training Details

Causal rollouts. The teacher's trajectories are conditioned on clean, bidirectional context, whereas at inference the student continues from its own outputs. Following causal self-rollout training (Huang et al., 2025b; Gan et al., 2026), we distill on rollouts whose horizon is sampled from several lengths and whose prefixes come from clean trajectories or detached student generations. Within a rollout, the frozen audio branch generates the audio of each chunk first, as at inference, and the video branch follows it. Before a generated chunk is reused as context, forward flow matching moves its latent to a random noise level, while the reference frames stay clean. The student thus learns to follow the audio from the imperfect contexts it meets at inference, and the reference frames anchor appearance.

Training with memory. The video branch learns to read recalled evidence during distillation. Before training, we run the insertion rule of Eq. (2) and its audio counterpart over each long training video, which yields a reservoir per video. Each training sample is a clip cut from a random position of the video, and its working memory is retrieved only from entries that precede the clip. The model therefore learns to use memory that shows the same participants earlier in the same video, as it does at inference. The frozen audio branch receives the retrieved audio in the same form as its audio history and is used without training.

Hyperparameters. The generator and critic use learning rates of 3×10^{-5} and 5×10^{-6}, respectively, and the generator EMA decay is 0.999. We train on 15,000 filtered synthetic trajectories generated by the bidirectional MiniMax-H3 at 864×480. Most of them last 15 to 60 seconds. Rollout horizons of five and ten chunks are sampled with probabilities 0.8 and 0.2, and detached student-generated prefixes replace clean prefixes with probability 0.1. Generated chunks reused as context are re-noised with $\sigma \sim \mathcal{U}(0, 0.15)$, and the reference frames stay clean. We train for 3,000 steps with a batch size of 8 on 32 NVIDIA H100 GPUs.

A.3 Inference Details

Segment windows. To keep the position-bearing context bounded, each prompt segment is generated as a window. At a window boundary, the KV cache is discarded and prefilled again from the reference frames, the latest clean chunks, and the newly retrieved working memory, with positions restarting at the start of the window.

Feedback through memory. The audio correction is applied as soon as the audio stage finishes a chunk, so the video branch follows the corrected audio, which is also the output audio. The corrected outputs of both modalities are the ones written to the reservoir. Since the working memory of later segments is drawn from the reservoir, corrected evidence re-enters the context of later chunks through retrieval, and the correction limits drift in the conditioning as well as in the output.

Off-screen speech. The interviewer's lines in SteadyBench are given as synthesized speech (Appendix A.1). For these segments, SteadyTalk skips the audio stage, and the video branch follows the given audio in the same way as generated audio. No on-screen participant speaks here, so the given audio is neither inserted into the reservoir nor corrected by the timbre alignment of Eq. (4).

Hyperparameters. The audio stage samples each chunk with $S_a = 10$ steps, and the video stage uses $S_v = 4$ steps. The working memory holds $m = 7$ participant crops and the earliest $m_a = 2$ audio entries of the current speaker. Each 5-second prompt segment is generated as a window of four chunks of 39, 34, 34, and 17 frames. The context keeps the first chunk of the video and the two most recent chunks. The insertion thresholds of the memory are $\tau_v = 0.8$ and $\tau_a = 0.6$. The video color correction uses the first $N_r = 5$ frames of the video as its reference.

A.4 Evaluation Protocol

All metrics are computed per scenario and then averaged over scenarios. Speech metrics are computed on answer turns, the 10-second spans in which an on-screen subject answers a question, and exclude the interviewer lines. In two-person scenarios, detected faces are assigned to participants by their horizontal position, which the scenario metadata records for each participant.

Synchronization. LipSync and Desync follow the VABench protocol (Hua et al., 2026) and are computed on the face of the current speaker in each answer turn. Listener is computed over the question segments and, in two-person scenarios, over the answers of the other subject, excluding 0.5 seconds at each segment boundary. For every frame, we compute the mouth aspect ratio from MediaPipe face landmarks (Kartynnik et al., 2019),

$$\mathrm{MAR} = \frac{\|p_{13} - p_{14}\|}{\|p_{78} - p_{308}\|}, \tag{5}$$

where p_{13} and p_{14} are the midpoints of the upper and lower inner lips and p_{78} and p_{308} are the inner mouth corners. After a five-frame median filter, a frame counts as closed when $\mathrm{MAR} < \tau_{\mathrm{close}}$, with $\tau_{\mathrm{close}} = 0.1$. Listener is the percentage of closed frames among the frames with a detected face.

Speech. We transcribe English answer turns with Whisper large-v3 (Radford et al., 2023) and Mandarin answer turns with Qwen3-ASR (Shi et al., 2026). English transcripts are normalized with the Whisper English text normalizer, and Mandarin transcripts are converted to simplified characters with punctuation removed. Within a scenario, we accumulate the edit operations over all answer turns and divide them by the number of reference words for English or reference characters for Mandarin; WER averages this rate over scenarios. Spk-Sim embeds each answer turn with a WavLM-Large speaker verification model (Chen et al., 2022) and, for each participant, averages the cosine similarity between the embeddings of the later turns and that of the participant's first turn. We use WavLM rather than CAM++ for this metric because CAM++ is part of the timbre correction in SteadyTalk.

Identity. ID-Sim samples frames at 2 fps and embeds the detected faces with ArcFace (Deng et al., 2018). The reference of a participant is the mean embedding over the first second in which they appear. For every segment after the first, we average the face embeddings of the participant within the segment and compute the cosine similarity to the reference; ID-Sim averages these similarities over segments and participants. Swap considers each answer turn of a participant except the first. A turn counts as swapped if its WavLM embedding, or the ArcFace embedding of the participant in that turn, is closer to the reference of the other participant than to the participant's own.

Stability. Color Δ and Long Consistency follow the evaluation protocol of Gan et al. (2026) and are computed on fixed-length video intervals. Color Δ compares neighboring intervals, so it responds to abrupt color changes at their boundaries. Long Consistency compares color and brightness across the whole video, so it responds to slow global drift. Lower Color Δ and higher Long Consistency indicate a more stable appearance. For Table 5, color shift at time t is the distance between the normalized color statistics, with RGB values scaled to $[0, 1]$, of the one-second window ending at t and those of the first second, computed with the same script as Color Δ. Spk-Sim at time t is computed on the speech of each participant in the ten seconds before t.

Baselines. The bidirectional joint models generate each scenario segment by segment. They use chunked continuation where the model supports it and otherwise join clips of native length. The audio-driven methods animate the reference image with the audio generated by SteadyTalk. This setting tests how well they follow a given audio track. MultiTalk also handles the two-person scenarios, and the single-subject methods are evaluated on the single-person categories only. Hallo-Live and OmniForcing generate at most 30 seconds, so we evaluate them on the first 30 seconds of each scenario. Methods without an audio input generate the interviewer lines from the transcripts, together with the scene. All methods keep their native resolution and frame rate.

A.5 Results by Category

Table 6 breaks the results down by category. Two-person conversations are the hardest case for speech: Desync and WER nearly double relative to single presenters and Spk-Sim is lowest, since the model must switch voices and move the right person's lips at every turn. Even so, no answer turn is attributed to the wrong participant. Complex actions mainly affect appearance statistics: Color Δ is largest and Long Consistency lowest, which is expected when the prompts ask the presenter to remove a jacket or change pose, altering the colors in the frame. Single presenters obtain the best lip-sync and identity scores but the lowest Listener, as they listen only during interviewer lines.

Table 6: **Results of SteadyTalk on each category** of SteadyBench.

Category	Synchronization			Speech		Identity		Stability	
	LipSync↑	Desync↓	Listener (%)↑	WER (%)↓	Spk-Sim↑	ID-Sim↑	Swap↓	Color Δ ↓	Long Cons.↑
Single-presenter	7.20	0.055	75.6	0.78	0.608	0.711	–	0.0012	0.975
Complex-action	6.38	0.048	85.3	0.80	0.582	0.658	–	0.0028	0.960
Two-person	6.41	0.093	85.3	1.44	0.544	0.663	0.0000	0.0014	0.970

A.6 User Study

Protocol. We run a side-by-side user study with 50 expert annotators recruited through a professional annotation vendor. We randomly sample 50 scenarios from SteadyBench and pair SteadyTalk with each autoregressive baseline: OmniForcing, Hallo-Live, and JoyAI-Echo-1.5. For the two 30-second baselines, both videos of a pair are cut to their first 30 seconds. The two videos are played side by side with audio, in random left-right order and without method names. Each annotator watches pairs from several scenarios. For each baseline, the annotator casts one vote on overall preference and one on each of three criteria: audio-visual synchronization, consistency of identity and voice, and visual and audio quality. Every vote allows a tie, and every comparison receives 50 votes. Table 7 lists the vote rates, and Figure 5 shows the full vote distribution.

Results. Annotators prefer SteadyTalk over every baseline, overall and on each criterion. In all twelve comparisons, win votes exceed lose votes with $p < 0.01$ under a two-sided sign test on the non-tie votes. The margins are largest against the two 30-second baselines, in line with their low visual quality and oversaturated colors in Figure 4. Synchronization gives the most uniform advantage: at least three of four annotators prefer SteadyTalk against every baseline. JoyAI-Echo-1.5 is the closest competitor. Annotators give it the most ties on quality and consistency, but consistency still favors SteadyTalk by a wide margin, in line with the changed presenter and the missing product in Figure 4.

Table 7: **User study.** Vote rates (%) of SteadyTalk against each baseline, from one vote per annotator and criterion. The attribute columns give the win rate of SteadyTalk on each criterion.

Compared method	Overall preference (%)			Attribute win (%)		
	Win	Tie	Lose	Sync.	Consistency	Quality
OmniForcing (Su et al., 2026)	84.0	12.0	4.0	80.0	68.0	90.0
Hallo-Live (Li et al., 2026)	90.0	8.0	2.0	80.0	86.0	82.0
JoyAI-Echo-1.5 (Duan et al., 2026)	64.0	18.0	18.0	76.0	60.0	56.0

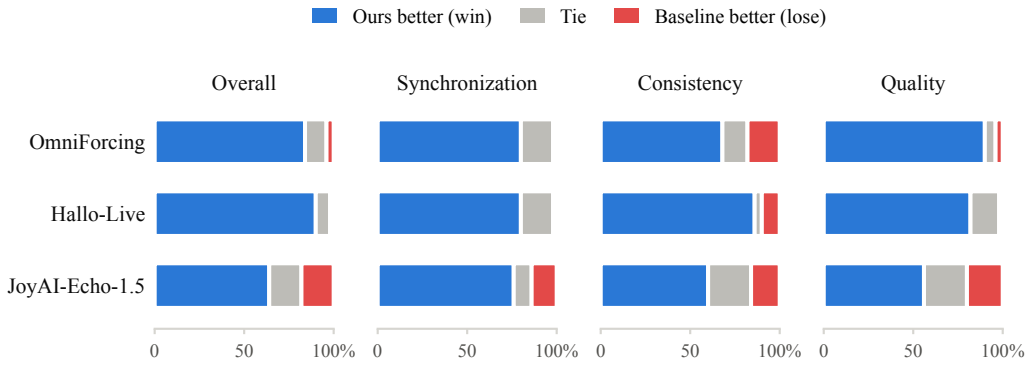

Figure 5: **User study.** Side-by-side votes of 50 expert annotators on SteadyTalk versus each baseline.

A.7 Two-Person Examples

Figure 6 shows frames from four two-person scenarios of SteadyBench, and Table 8 lists their participants. Each prompt names the participant who answers, describes the voice of every participant, and asks the listener to keep the lips closed, while the interviewer stays off screen. The scenarios differ in the order in which the two subjects answer. In the kitchen and hiking scenarios, the subjects alternate, so each of them is silent for 20 seconds before answering again. In the sailing and classroom scenarios, one subject answers the first two questions, and the other speaks for the first time after 35 seconds. Throughout the minute, both subjects keep their face, hairstyle, clothing, and position, and the background and lighting stay unchanged; this also holds for Grace and Kwame, who only listen during the first 35 seconds. In the frames taken during an answer, the listener keeps the mouth closed and looks toward the speaker, as the prompts request.



Figure 6: **Two-person examples from SteadyBench.** Each column is one scenario and each row a frame from the one-minute stream. Labels give the time and the participant who speaks, with (L) and (R) marking the left and right subject.

Table 8: **Participants of the scenarios in Figure 6**, with the voice descriptions from the prompts and the order in which the left (L) and right (R) subjects answer the four questions.

Scenario	Left subject	Right subject	Answer order
Grandmother and grandson	Rosa: warm, slightly raspy, low female voice with a gentle Italian accent	Marco: bright, youthful tenor male voice with a quick, cheerful pace	R, L, R, L
Couple on a sailboat	Grace: cheerful, mid-pitched female voice with a light New Zealand accent	Samuel: deep, resonant, unhurried male voice with a warm Caribbean accent	R, R, L, L
Teacher and former student	Margaret: clear, gentle, slightly high female voice with a precise, teacherly pace	Kwame: rich, warm baritone male voice with a light British accent	L, L, R, R
Hiking channel friends	Diego: loud, upbeat, mid-pitched male voice with a light Spanish accent	Ethan: soft, dry, slightly nasal male voice with a laid-back Canadian accent	R, L, R, L

A.8 Reservoir Growth

We bound the number of entries that the gated insertion of Section 3.3 can store. Consider one participant and one modality, let $e_1, \dots, e_C$ be the ℓ_2-normalized embeddings of the candidates in the order in which they are tested, and write $\angle(u, v) = \arccos(u^\top v)$ for the angle between two unit vectors. The insertion condition $\cos(e_i, e_{\text{last}}) < \tau$ is equivalent to $\angle(e_i, e_{\text{last}}) > \theta_\tau$ with $\theta_\tau = \arccos \tau$. Let $L = \sum_{i=1}^{C-1} \angle(e_i, e_{i+1})$ be the angular path length traced by the candidates.

Proposition. The number of stored entries satisfies $N \le \min\{C,\, 1 + L/\theta_\tau\}$.

Proof. The first candidate is always stored. Let $i_1 < \cdots < i_N$ be the indices of the stored candidates. Candidate i_{j+1} is compared with e_{i_j}, the most recently stored entry, so $\angle(e_{i_j}, e_{i_{j+1}}) > \theta_\tau$; quality checks on crops only reject more candidates and leave this inequality intact. The angle is a metric on the unit sphere, so the triangle inequality gives $\angle(e_{i_j}, e_{i_{j+1}}) \le \sum_{i=i_j}^{i_{j+1}-1} \angle(e_i, e_{i+1})$. Summing over j yields $(N-1)\,\theta_\tau < \sum_{i=i_1}^{i_N-1} \angle(e_i, e_{i+1}) \le L$, and $N \le C$ holds trivially. □

Interpretation. The bound depends on how far the embeddings travel rather than on how many candidates are tested. A participant whose appearance or voice changes little adds few entries however long the stream runs, and each new entry requires an accumulated change of more than $\theta_{\tau_v} = \arccos 0.8 \approx 0.64$ rad for crops and $\theta_{\tau_a} = \arccos 0.6 \approx 0.93$ rad for audio. Because a candidate is compared only with the most recent entry, a participant who alternates between two states more than θ_τ apart adds an entry at every switch, so N then grows with the number of switches; comparing each candidate with all stored entries would remove these duplicates at a cost linear in N.

Cost. Insertion costs one encoder pass and one inner product per candidate. Retrieval computes N_v inner products between the prompt embedding and the stored crops and selects the top m, which takes $\mathcal{O}(N_v d + N_v \log m)$ time for embedding dimension d, and it runs only when the prompt changes. The working memory always contains m crops and m_a audio entries, so the number of memory tokens seen by the generator does not depend on N. In a one-minute scenario, generated as twelve segments of four chunks, audio candidates arise only in chunks where a participant answers, which limits a participant's audio entries to 32 in single-presenter and 16 in two-person scenarios.