VORCH-STREAMER: EXTENDING HUMAN AUDIO-VISUAL GENERATION TO REAL-TIME LONG-FORM STREAMING

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-time long-form avatar audio–video generation requires causal, continuous synthesis while maintaining audiovisual synchronization and visual consistency. Finetuning a pretrained bidirectional model for this setting presents two key dilemmas. First, autoregressively reusing generated blocks as context introduces exposure bias, causing errors and drift to accumulate over long rollouts. Second, conditioning on a global speech utterance cannot tell a causal generator which portion should be spoken next when only limited local context is available. We present **Vorch-Streamer**, a post-training framework that addresses these dilemmas and enables real-time long-form Text-to-Audio-Video (T2AV) streaming generation. We address the first dilemma through long-horizon Self Forcing, which exposes the model to its own long-term rollout distribution under the supervision of a bidirectional teacher. For the second dilemma, we introduce explicit speech progression control, where an external language model predicts discrete speech-planning tokens that condition the audio diffusion branch and align each causal block with the content it should speak. With bounded causal context and four-step denoising, Vorch-Streamer jointly generates audio and video from text at 27.12 FPS, exceeding the real-time efficiency while maintaining competitive audio–lip synchronization and strong identity preservation over long-form generation.

1 INTRODUCTION

Real-time avatar audio-video generation is essential for applications such as education, entertainment, virtual communication, and interactive agents. Recent audio-video diffusion models support diverse conditioning modes, including text, image, and audio, and can generate expressive faces with natural speech (Liu et al., 2024; Wang et al., 2024; Ma et al., 2025; Seedance et al., 2026). However, most of them are designed for short offline clips, as their bidirectional attention assumes access to the complete sequence during denoising. In contrast, streaming generation must produce each audio-video segment before future content is available, requiring not only low latency but also causal modeling, long-term synchronization, and stability under self-generated context (Huang et al., 2026c; Yin et al., 2025; Zhu et al., 2026).

This setting exposes two dilemmas. **First**, long-form streaming relies on autoregressive rollout, where self-generated histories create exposure bias that compounds errors under causal attention, leading to temporal drift, artifacts, and unstable long-horizon generation. **Second**, Text-to-Audio-Video (T2AV) faces a fundamental mismatch between bidirectional and causal generation. With full sequence context, bidirectional DiT can jointly denoise all tokens and implicitly plan the content of each audio segment. In streaming, however, limited historical context and fixed past predictions prevent the DiT from reliably deciding what to say next, making speech planning becoming an ill-posed problem and causing segment-level disorder or repetition. This becomes a major bottleneck in long-form streaming, especially when the complete transcript is unavailable in advance or speech is frequently switched in short segments.

We present **Vorch-Streamer**, a post-training framework that extends the pretrained LTX2.3 audio–video foundation model to real-time, long-form streaming. We first use LTX2.3 to synthesize 80K high-quality avatar audio–video clips, and train a causal block-autoregressive model with a sample-

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.

level mixture of Teacher Forcing and Diffusion Forcing, together with causal attention masks. The causal model is further optimized with long-horizon Self Forcing, performing self-rollouts over 12–21 seconds with four-step denoising distillation. This aligns the model with its inference-time rollout distribution and transfers the pretrained model's quality to long causal trajectories. Together, causal training and long-horizon distillation yield a stable four-step causal long-form streaming generator, addressing the first dilemma of exposure bias.

To address the second dilemma, we introduce explicit speech planning with the Fun-CosyVoice LLM (Lyu et al., 2025). The planner predicts discrete 40 ms speech-planning units, which are converted into continuous features using an extended lookup table (LUT) with a learnable silence token. These features are injected into the audio branch via cross-attention and gated fusion, decoupling speech planning from audio generation: the planner determines speech content and timing, while the generative model focuses on synchronized audio-visual realization, timbre, prosody, and expressiveness. This design also supports arbitrary speech interruption or switching, with the learnable silence token enabling explicit silent-listening for interactive applications. Experiments show that Vorch-Streamer is the only evaluated native T2AV method to exceed real-time playback while maintaining competitive long-form audio–lip synchronization and speech accuracy, demonstrating that post-training, explicit speech planning, and causal inference can effectively transform a powerful bidirectional foundation model into a real-time text-driven avatar generator.

Our contributions are summarized as follows:

- We introduce Vorch-Streamer, a post-training framework that adapts a pretrained bidirectional audio-video diffusion model to causal, real-time, long-form streaming generation.
- We combine mixed Teacher Forcing/Diffusion Forcing with long-horizon Self Forcing to align causal training with streaming inference.
- We propose an LLM-based speech-planning approach for T2AV that supplies continuous planning features to the audio branch, supports interruption and speech switching through the decoupling of planning and generation.
- Experiments demonstrate real-time, long-form T2AV streaming at 27.12 FPS, together with competitive synchronization, speech accuracy, and long-horizon identity preservation.

2 Related Work

Real-Time and Long-Form Video Generation. Real-time streaming generation requires causal temporal modeling, efficient few-step sampling, and robustness to self-generated history. CausVid and Self Forcing (Yin et al., 2025; Huang et al., 2026c) distill bidirectional models into causal generators, with Self Forcing reducing train–test mismatch through autoregressive rollouts. Causal Forcing (Zhu et al., 2026) uses an autoregressive teacher for ODE initialization but evaluates only short silent videos. LongLive (Yang et al., 2025) supports long-form T2V and online prompt changes through long tuning, windowed causal attention, and KV recaching, but does not generate audio. OmniForcing (Su et al., 2026) is the closest precursor to causal joint audio-video generation, distilling LTX-2 with asymmetric block alignment, audio sink tokens, joint Self Forcing, and rolling KV caches. However, it focuses on short-form generation and does not address long-form speech progression or interruption. Vorch-Streamer instead targets long-horizon joint audio-video generation and explicitly resolves the mismatch between global speech specifications and the limited local context available to causal generators.

Human-Centric Audio-Video Generation. Human-centric generation requires identity preservation, lip synchronization, expressive motion, and long-term stability. Existing approaches can be broadly divided into audio-driven and text-driven joint generation. Live Avatar (Huang et al., 2025) and StreamAvatar (Sun et al., 2026) convert large video diffusion models into causal streaming generators with persistent reference mechanisms, but rely on external audio input. Thus, neither jointly generates speech and audiovisual content. Among text-driven joint avatar methods, HalloLive (Li et al., 2026a) adopts asynchronous dual-stream diffusion and preference-guided distillation but focuses on short-form evaluation. StreamChar (Tian et al., 2026) combines an LLM orchestrator with a joint audio-video DiT for long-form streaming, but assumes a complete transcript and does not support online speech replacement; it also requires task-specific fine-tuning and substantial adaptation of a Wan-derived backbone. JoyAI-Echo (Li et al., 2026b) preserves appearance and

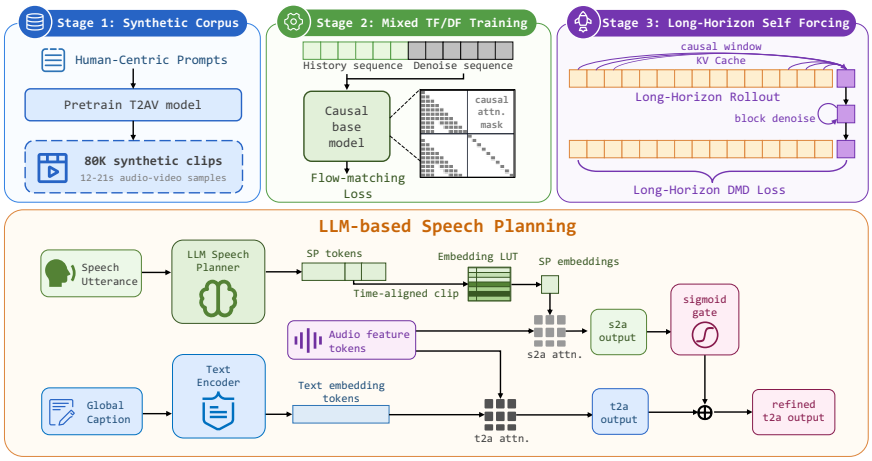

Figure 1: Overview of Vorch-Streamer.

voice across minute-scale multi-shot stories, but relies on pre-authored shot sequences rather than causal continuous streaming. Wan-Streamer (Huang et al., 2026b;a) unifies language, audio, and video in a Transformer for full-duplex interaction, but requires end-to-end multimodal pretraining and lacks standardized minute-scale evaluation. In contrast, Vorch-Streamer post-trains an existing bidirectional audio-video foundation model for causal long-form streaming, retaining its pretrained generative capabilities while using a pretrained speech language model for explicit speech planning. This enables speech interruption and switching without requiring the complete future utterance in advance, providing a complementary solution for real-time, long-form text-driven avatar generation.

3 Method

Vorch-Streamer extends the pretrained LTX2.3 audio–video diffusion model to real-time, long-form avatar audio–video generation. As illustrated in Fig. 1, it consists of three post-training stages. We first construct an 80K synthetic avatar corpus with the pretrained bidirectional LTX2.3 model for Stage 2 and Stage 3 training. Stage 2 adopts a mixed Diffusion Forcing and Teacher Forcing strategy, with the Fun-CosyVoice LLM (Lyu et al., 2025) providing explicit speech planning. Stage 3 further applies long-horizon Self Forcing to align the model with its inference-time rollout distribution. The speech-planning tokens are integrated into the audio branch via cross-attention and gated injection, resolving the mismatch between global textual conditioning and causal audio generation.

3.1 Causal Audio–Video Streaming

Synchronized block generation. Let $\mathbf{x}_0^v$ and $\mathbf{x}_0^a$ denote the clean video and audio latents. We partition them into synchronized audio–video blocks, with each regular block containing three video latent frames and 25 audio latent tokens, corresponding to approximately one second since the LTX2.3 VAE operates at 3 video latent frames and 25 audio latent tokens per second. Following OmniForcing (Su et al., 2026), the first block contains one additional video frame and one additional audio token due to causal VAE alignment.

Within each block, we retain bidirectional attention to capture fine-grained audio–video interactions, while enforcing causal dependencies across blocks. During training, each block attends to the first three blocks as persistent global context and the most recent three preceding blocks as local context. This bounded global-local window preserves initial appearance and recent motion while maintaining constant memory. The same dependency is implemented during causal training with attention masks and during long-form training and inference with a windowed KV cache and cache eviction.

Mixed Teacher Forcing and Diffusion Forcing. Directly applying Teacher Forcing exposes the model only to clean ground-truth histories, whereas streaming inference conditions on previously generated and potentially imperfect blocks. To reduce this train–test discrepancy, we train the causal

model with a sample-level mixture of Teacher Forcing and Diffusion Forcing. For modality $m \in \{v, a\}$, a clean latent is corrupted through flow-matching interpolation $\mathbf{x}_\sigma^m = (1-\sigma)\mathbf{x}_0^m + \sigma\boldsymbol{\epsilon}^m$, where $\boldsymbol{\epsilon}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Each sample is organized as a dual-sequence input containing a history sequence and a denoising sequence with same content. For 10% of the samples, the history sequence remains clean, corresponding to Teacher Forcing. For the remaining 90%, the history sequence is corrupted, corresponding to Diffusion Forcing. The causal mask allows the current denoising block to access preceding history blocks but prevents it from reading the clean target of the current block, as shown in Fig. 1.

The model is optimized with the standard audio–video flow-matching objective,

$$\mathcal{L}_{\mathrm{CF}} = \sum_{m\in\{v,a\}} \|\widehat{\mathbf{v}}^m - (\boldsymbol{\epsilon}^m - \mathbf{x}_0^m)\|_2^2 . \tag{1}$$

This stage transfers the pretrained model's joint audio–video generation ability to a causal architecture and produces a short-form streaming model that serves as the initialization for long-horizon training, while adapts the injection of speech planning tokens.

3.2 Long-Horizon Self Forcing

Although mixed forcing improves robustness to locally corrupted context, it does not fully capture the accumulated errors of autoregressive inference. Over long sequences, small errors in appearance, motion, speech, or synchronization can propagate through the causal history and cause severe degradation. We therefore further apply long-horizon Self Forcing to address this problem.

Starting from the causal model obtained above, the student generates a 12–21 second sequence block by block from its own predictions, with each block conditioned on previous outputs through the same bounded attention window used in causal training and implemented via KV-cache eviction. This directly exposes the model to its long-horizon rollout distribution, closing the exposure gap between training and streaming inference while keeping memory usage constant.

To preserve the quality and semantic fidelity of pretrained LTX2.3, we retain the original bidirectional model as a frozen teacher and perform distribution matching distillation (DMD) (Yin et al., 2024). Specifically, the causal *student* generates long-horizon rollouts, the frozen bidirectional *real-score* model represents the pretrained target distribution, and a trainable bidirectional *fake-score* model tracks the student's evolving distribution through flow-matching training.

Given a student rollout $\widetilde{\mathbf{x}}_0^m$, we first perturb it to a random noise level, *i.e.*, $\widetilde{\mathbf{x}}_\sigma^m = (1-\sigma)\widetilde{\mathbf{x}}_0^m + \sigma\boldsymbol{\epsilon}^m$. The same noisy audio–video sample is then evaluated by the real-score and fake-score models,

$$\begin{aligned} \mathbf{p}_{\mathrm{real}}^m &= \mathrm{real_score}^m\left(\widetilde{\mathbf{x}}_\sigma^v, \widetilde{\mathbf{x}}_\sigma^a, \mathbf{c}, \sigma\right), \\ \mathbf{p}_{\mathrm{fake}}^m &= \mathrm{fake_score}^m\left(\widetilde{\mathbf{x}}_\sigma^v, \widetilde{\mathbf{x}}_\sigma^a, \mathbf{c}, \sigma\right). \end{aligned} \tag{2}$$

For the real-score network, conditional and unconditional predictions are combined with CFG to form the guided teacher target, while the fake-score network is evaluated on the conditional student distribution. Thus, $\mathbf{p}_{\mathrm{real}}^m$ indicates how the frozen bidirectional teacher would denoise the sample toward the pretrained data distribution, whereas $\mathbf{p}_{\mathrm{fake}}^m$ describes the current student distribution. Their difference provides the DMD update direction,

$$\overline{\mathbf{g}}^m \propto \mathbf{p}_{\mathrm{fake}}^m - \mathbf{p}_{\mathrm{real}}^m. \tag{3}$$

The causal student is optimized using the corresponding stop-gradient surrogate

$$\mathcal{L}_{\mathrm{DMD}} = \sum_{m\in\{v,a\}} \frac{1}{2}\left\|\widetilde{\mathbf{x}}_0^m - \mathrm{sg}\left(\widetilde{\mathbf{x}}_0^m - \overline{\mathbf{g}}^m\right)\right\|_2^2, \tag{4}$$

which moves the student samples toward the real-score prediction and away from the distribution currently captured by the fake-score model.

The fake-score model is updated separately on detached student rollouts. It is trained with the standard audio–video flow-matching objective,

$$\mathcal{L}_{\mathrm{fake}} = \sum_{m\in\{v,a\}} \|\widehat{\mathbf{v}}_{\mathrm{fake}}^m - (\boldsymbol{\epsilon}^m - \widetilde{\mathbf{x}}_0^m)\|_2^2 . \tag{5}$$

Alternating the student and fake-score updates continually refreshes the estimate of the student distribution, allowing the DMD direction to remain informative as the causal generator improves.

3.3 LLM-BASED SPEECH PLANNING

Streaming audio–video avatar generation introduces an additional challenge for text conditioning. A conventional text prompt contains the complete speech transcript, but a causal block must determine which portion of the transcript should be spoken at the current time. This alignment is implicit in a bidirectional model, where all text, audio, and video tokens can interact globally. Under causal attention, however, the same prompt leaves the diffusion model responsible for deciding both *what to say* and *when to say it*. The ambiguity becomes particularly problematic in long-form generation and near the boundaries between successive speech segments.

Vorch-Streamer explicitly separates speech planning from audio–video realization. Given the target utterance, the planning LLM predicts a sequence of discrete speech-planning tokens $\mathbf{z} = (z_1, \ldots, z_{T_a})$ at 25 Hz, where each token represents a 40-ms speech-planning unit. The tokens are converted into continuous features using the pretrained Fun-CosyVoice lookup table followed by a learnable projection,

$$\mathbf{e}_t^s = f_{\text{proj}}\left(\mathbf{E}_{\text{LUT}}[z_t]\right). \tag{6}$$

Because the planning sequence and the LTX audio latents share the same 25-Hz temporal resolution, each causal block receives the speech features associated with its own time interval.

We inject these planning features into the audio branch through an additional cross-attention module, and strict the audio tokens to only attend the planning features within the same block. Its output is adaptively fused with the original text-conditioned audio features using a learnable gate,

$$\mathbf{h}_\ell^a \leftarrow \mathbf{h}_\ell^a + \text{sigmoid}(g_\ell)\,\text{Attn}_{\text{speech}}^\ell\left(\mathbf{h}_\ell^a, \mathbf{e}^s\right), \tag{7}$$

where $\mathbf{h}_\ell^a$ is the audio hidden state and g_ℓ is the gate in layer ℓ. The original text cross-attention remains unchanged and continues to describe global semantics and speaking style, while the speech branch supplies an explicit local speaking schedule. This division allows DiT to focus on synchronized audio–video generation rather than inferring speech timing from a global transcript.

During post-training, the causal student is conditioned on the global scene description and aligned speech-planning tokens, while the bidirectional teacher retains its original text-conditioning pathway with the full utterance. This asymmetric conditioning preserves the teacher's pretrained global modeling capability while guiding the student toward causal speech generation.

4 EXPERIMENTS

We conduct quantitative and qualitative comparisons against existing joint and cascaded avatar generation systems, and further ablate the speech-planning module, the Stage 2 and Stage 3 training strategy, and long-horizon Self Forcing. Implemetation details, the ablation study of causal-context selection and more generation results of Vorch-Streamer are provided in Appendix.

4.1 EXPERIMENTAL SETUP

Evaluation protocol. We construct a held-out benchmark containing diverse human identities, languages, speaking styles, backgrounds, and utterance lengths due to there is no existing available benchmark taming long-term audio–video joint generation. No prompt in the benchmark appears in the synthetic training corpus. We compare continuous long-form rollouts targeting approximately two minutes, computing aggregate visual metrics over each complete available rollout and sampling time-dependent metrics every second to expose error accumulation.

The T2AV task requires generating both speech and video directly from text, and Vorch-Streamer is evaluated exclusively in this native setting. The "+TTS" baselines serve as conditional TIA2V references: they receive Qwen3-TTS speech and an LTX2.3-generated first frame as inputs, providing both the target waveform and an appearance anchor and thus not being directly comparable to from-scratch T2AV. We target a nominal two-minute duration while allowing actual lengths to vary according to each system's generation mechanism. Due to OOM constraints, OmniForcing and Hallo-Live are limited to 30-second rollouts, with their temporal curves terminating accordingly.

Baselines. We compare with the bidirectional foundation model LTX2.3 (HaCohen et al., 2026); joint audio–video generator JoyAI-Echo (Li et al., 2026b); streaming avatar models Hallo-Live (Li

Table 1: Quantitative comparison on long-form avatar generation. "Task" distinguishes native joint T2AV generation from TIA2V reference pipelines driven by Qwen3-TTS audio and an LTX2.3-generated first frame. "Stream" indicates causal block-wise output. Best results are shown in bold and second-best results are underlined.

Method	#Param.	Task	Stream	FPS ↑	Sync-C ↑	Sync-D ↓	WER ↓
LTX2.3	22B	T2AV		1.83	6.80	<u>7.96</u>	**7.59%**
JoyAI-Echo	22B	T2AV		3.08	2.44	12.91	9.34%
LiveAvatar+TTS	14B	TIA2V	✓	7.58	<u>7.18</u>	8.30	9.21%
SoulX-FlashTalk+TTS	18.9B	TIA2V	✓	5.11	**8.69**	**7.26**	9.21%
OmniForcing (30 s)	19B	T2AV	✓	<u>12.12</u>	0.75	12.22	98.79%
Hallo-Live (30 s)	11.7B	T2AV	✓	11.51	0.58	13.90	94.13%
Vorch-Streamer (ours)	22.8B	T2AV	✓	**27.12**	6.62	8.95	<u>7.92%</u>

et al., 2026a) and OmniForcing (Su et al., 2026); and cascaded TIA2V references LiveAvatar (Huang et al., 2025)+TTS and SoulX-FlashTalk (Shen et al., 2025)+TTS. LTX2.3 is a strong offline T2AV quality reference but cannot emit blocks causally. JoyAI-Echo, OmniForcing, Hallo-Live, and Vorch-Streamer use their native joint T2AV settings. The two conditional references instead generate video from shared Qwen3-TTS (Hu et al., 2026) audio and an LTX2.3-generated first frame, neither of which Vorch-Streamer uses.

Metrics. We measure DiT throughput as generated frames divided by end-to-end wall-clock time on one NVIDIA H200 GPU, reported as FPS. Nine quantitative metrics cover audio–lip synchronization (Sync-C and Sync-D) (Chung & Zisserman, 2016), visual quality (FID and FVD) (Heusel et al., 2017; Unterthiner et al., 2018), VBench (Huang et al., 2024) dynamics (Dynamic Degree) and drift, VBench2 (Zheng et al., 2025) Human Anatomy and Identity, and word error rate (WER). We further assess temporal consistency against the initial frame using ArcFace (Deng et al., 2019) cosine similarity between aligned face embeddings for identity preservation and CLIP (Radford et al., 2021) similarity between full-frame embeddings for global appearance consistency.

4.2 Quantitative Comparison

End-to-end generation quality and efficiency. Table 1 compares Vorch-Streamer with existing joint and cascaded avatar generation systems. Vorch-Streamer reaches 27.12 FPS with a 22.8B-parameter pipeline, making it the only evaluated model that exceeds the 24-FPS real-time threshold in one H200 GPU. Among native T2AV methods, it achieves a WER close to the offline LTX2.3 baseline, whereas directly extending OmniForcing and Hallo-Live to long-form generation leads to severe speech degradation. This highlights the second dilemma: without explicit speech planning, causal blocks cannot reliably determine which part of the global transcript should be generated next. Vorch-Streamer's speech planner preserves transcript accuracy over approximately two minutes, while its synchronization scores remain close to those of the much slower bidirectional LTX2.3, demonstrating that four-step causal inference retains most of the audiovisual quality of the pretrained model. The TIA2V baselines achieve competitive synchronization but use a substantially easier setting, taking the complete target audio and a generated first frame as input. Thus, they neither synthesize speech nor solve causal speech progression, while their inference speeds remain below real-time. More importantly, external audio bypasses the second dilemma by eliminating the need to align global text with local causal context.

Visual quality and human fidelity. Table 2 evaluates perceptual quality and human fidelity over each method's complete available long-form rollout. Vorch-Streamer achieves the strongest overall identity preservation and competitive dynamics, drift, FID, and FVD, while remaining comparable to the best methods in anatomy and perceptual similarity. The TIA2V baselines benefit from first-frame conditioning, which provides a direct appearance and identity reference and leads to higher ArcFace and CLIP scores. Vorch-Streamer instead generates and causally preserves these attributes from text over long horizons. Compared with native streaming T2AV baselines, it consistently achieves stronger visual quality, dynamics, identity preservation, and long-term stability.

Table 2: Visual quality and human-centric fidelity over long-form rollouts.

Method	FID ↓	FVD ↓	Dynamic ↑	Drift ↓	Human Anatomy ↑	Human Identity ↑	ArcFace ↑	CLIP I.Sim. ↑
LTX2.3	115.12	521.74	**0.2941**	0.0776	0.9680	0.8733	0.5679	0.8866
JoyAI-Echo	**110.81**	**354.85**	0.0471	0.0481	0.9710	0.9789	0.6777	0.9075
LiveAvatar+TTS	120.94	457.85	0.2824	0.0301	0.9651	0.9722	0.7387	0.9532
SoulX-FlashTalk+TTS	123.13	448.33	0.2000	**0.0231**	**0.9719**	0.9849	**0.8431**	**0.9552**
OmniForcing (30 s)	167.26	1217.10	0.1529	0.0663	0.8884	0.8012	0.3870	0.8804
Hallo-Live (30 s)	143.18	1007.90	0.0000	0.0811	0.9291	0.8670	0.4652	0.8752
Vorch-Streamer (ours)	119.80	549.99	0.2706	0.0286	0.9233	**0.9996**	0.7534	0.9324

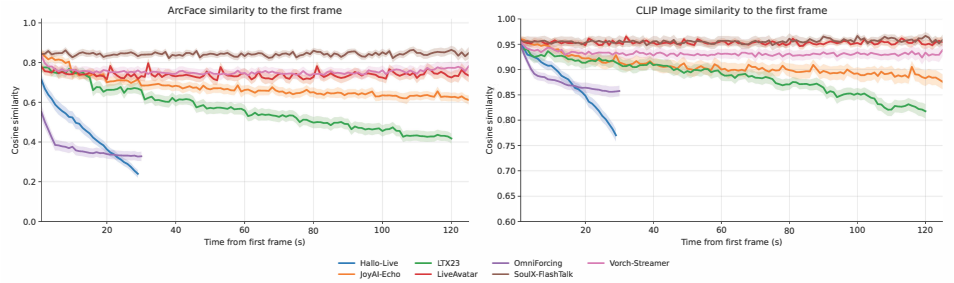

Figure 2: Long-horizon consistency over generation time. Lines and shaded regions show the mean and 95% confidence interval; each curve ends at the method's maximum completed duration.

4.3 Long-Horizon Analysis

Fig. 2 complements the endpoint comparison by tracking appearance consistency throughout each method's actual rollout. We compute ArcFace and CLIP Image cosine similarities between the first frame and each subsequent second, and average the resulting trajectories across samples with 95% confidence intervals. Vorch-Streamer remains stable throughout its approximately two-minute rollout, with no sustained decline in either identity or scene similarity. In contrast, native T2AV baselines show clear degradation over time, while OmniForcing and Hallo-Live decline within their shorter 30-second rollouts. The TIA2V methods remain more stable, but benefit from additional identity and content anchors provided by the first-frame and audio conditions. These results demonstrate that Vorch-Streamer maintains identity and scene appearance throughout long-form native T2AV generation rather than merely postponing degradation.

4.4 Qualitative Comparison

Fig. 3 provides qualitative comparison at the beginning, middle, and end of each rollout. Vorch-Streamer preserves facial identity, clothing, and background over extended generation while maintaining diverse gaze, expressions, hand poses, and object interactions, while its long-term consistency does not come at the cost of static motion. The native T2AV baselines exhibit distinct failure modes: LTX2.3 suffers progressive facial and background corruption, Hallo-Live develops severe color and structural artifacts, and OmniForcing shows frame freezing and substantial degradation within its short rollout. JoyAI-Echo remains relatively coherent but exhibits limited pose variation and noticeable facial degradation. TIA2V methods preserve identity more reliably, but benefit from externally supplied audio and first-frame conditions, making them conditional animation rather than from-scratch T2AV generation. Overall, the qualitative results indicate that Vorch-Streamer provides a stronger balance of long-horizon stability, dynamic motion, and native real-time T2AV generation.

4.5 Ablation Study

4.5.1 Effect of Speech Planning

We compare the proposed speech-planning features with three alternatives offering different degrees of temporal localization. As shown in Table 3, conditioning each block on the complete utterance performs poorly because the causal generator lacks a reliable signal for which part to speak next.



Figure 3: Qualitative comparison on two long-form T2AV prompts at the beginning, middle, and end of each rollout. OmniForcing and Hallo-Live end at 30 seconds. LiveAvatar and SoulX-FlashTalk are TIA2V methods, whose references is given by Qwen3-TTS audio and an LTX2.3 first frame; other methods generate both modalities from text.

Table 3: Ablation of speech planning for T2AV.

Speech condition	WER ↓	Sync-C ↑	Sync-D ↓
Complete long utterance	184%	5.58	9.66
Segmented short text	62.77%	6.76	8.98
Window text RoPE	65.46%	4.75	9.19
LLM speech plan (ours)	7.92%	6.62	8.95

Splitting the utterance into short segments decreases WER but introduces boundary inconsistencies: when the text condition switches, the cached audio context still corresponds to the previous segment, often causing incorrect speech progression or repetition. The "Window text RoPE" variant applies timestamp RoPE into each text token and further restricts each audio token to its corresponding text window, reducing this ambiguity. However, it only biases the attention map and still requires the DiT to infer the next content from limited context, leaving consecutive word repetitions. In contrast, the LLM speech planner converts the utterance into an explicit block-aligned sequence of speech units, substantially reducing WER while maintaining comparable synchronization. These results show that temporal restrictions on text attention alone are insufficient; explicit speech planning is crucial for accurate long-form causal T2AV generation.

4.5.2 Effect of Stage 2 and Stage 3

We ablate Stage 2 and Stage 3 to examine the role of causal initialization and long-horizon Self Forcing. For a fair "Without Stage 2" comparison, we first adapt the bidirectional DiT to the LLM speech planner, then skip causal forcing and directly initialize Stage 3 from this checkpoint. As shown in Table 4, omitting Stage 2 severely reduces motion while only slightly affecting WER and synchronization. Although the resulting low-motion rollout may exhibit favorable Drift and identity scores, these metrics are less meaningful when the model largely stops moving. Stage 2 is therefore essential for converting the bidirectional model into a causal generator while preserving meaningful motion for long-horizon rollout. Stage 3 provides a complementary benefit by adapting the causal

Table 4: Ablation of Stage 2 and Stage 3 on continuous long-form rollouts. Without Stage 2, the bidirectional DiT is first flow-matching fine-tuned to use the introduced LLM planner, after which causal audio–video forcing is skipped and Stage 3 is applied. Without Stage 3, the Stage 2 checkpoint is evaluated using 20-step CFG denoising; the full model uses four denoising steps per block.

Variant	FVD ↓	Sync-C ↑	Sync-D ↓	Dynamic ↑	Drift ↓	Human Identity ↑	WER ↓
Without Stage 2	515.30	6.89	8.73	0.0824	0.0238	1.0000	11.16%
Without Stage 3 (20-step CFG)	442.99	6.07	9.18	0.3059	0.0312	0.9939	9.17%
Stage 2 + Stage 3 (ours)	549.99	6.62	8.95	0.2706	0.0286	0.9996	7.92%

Table 5: Ablation of long-horizon self forcing.

Training horizon	FVD ↓	Dynamic ↑	Drift ↓	Human Identity ↑	ArcFace ↑	CLIP-I ↑
Stage 2 only (20-step CFG)	442.99	0.3059	0.0312	0.9939	0.7478	0.9445
First five blocks	556.32	0.5059	0.0739	0.9021	0.5800	0.8757
Random five blocks	461.95	0.1647	0.0698	0.9878	0.6869	0.8861
Last five blocks	443.87	0.2000	0.1368	0.6995	0.4437	0.7895
Full horizon (ours)	549.99	0.2706	0.0286	0.9996	0.7534	0.9324

model to efficient few-step inference. Compared with the Stage 2 checkpoint, Self Forcing enables four-step inference while improving speech synchronization, WER, drift, and identity preservation. Together, the two stages establish a causal generator that maintains dynamic motion and remains accurate and stable under efficient long-form rollout.

4.5.3 Effect of Long-Horizon Self Forcing

To isolate the benefit of training on complete causal trajectories, we compare Stage 2 with Self Forcing applied to the first, random, or last few blocks, as well as the proposed full-horizon training. As shown in Table 5, supervising only a small portion of the rollout leads to unbalanced behavior: early-block training increases motion but causes substantial appearance drift, while late-block training is particularly unstable because the supervised states depend on earlier, unoptimized self-generated context. Random-block training provides broader coverage but still exhibits noticeably higher drift than full-horizon training. In contrast, full-horizon Self Forcing achieves the strongest overall long-term preservation, with the lowest drift and highest identity and ArcFace similarity. Although it does not maximize motion or FVD, it provides the best balance between dynamics and temporal consistency, demonstrating that the complete inference trajectory should be optimized rather than only a sparse subset of rollout positions.

5 Conclusion

We presented Vorch-Streamer, a post-training framework that transforms a pretrained bidirectional audio–video diffusion model into a causal generator for real-time, long-form text-to-audio-video avatar generation. Causal audio–video forcing establishes stable block-wise generation, while long-horizon Self Forcing aligns the model with its inference-time rollout distribution. An LLM-based speech planner is introduced to decouple speech planning from audio-video generation, ensuring coherent speech progression under limited causal context. Together with bounded global–local context and four-step block denoising, these components enable efficient streaming without memory growth with sequence length. Experiments on long-term generations show that Vorch-Streamer achieves real-time performance while maintaining competitive visual quality, audio–lip synchronization, speech accuracy, and long-term identity preservation. Unlike conditional TIA2V methods, it requires neither an input audio track nor a first frame, directly addressing text-driven speech progression in causal generation. These results demonstrate that causal forcing and inference-aligned long-horizon training provide an effective path for adapting large bidirectional audio–video foundation models to interactive, long-form avatar generation.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, implementing methods, and assisting with translation, clean and reformat dataset. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, and designing or providing feedback on research methodology or experiments, and formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, supporting qualitative and thematic data analysis, and interpreting results are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, and editing a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Vorch-Streamer is designed for real-time avatar audio-video generation, which raises considerations regarding potential misuse for deepfakes or non-consensual identity replication. To mitigate these risks, we emphasize that the framework should be deployed with strict adherence to consent-based usage policies and watermarking mechanisms for generated content. Our method relies on pretrained bidirectional models and does not involve new human subject data collection; however, developers and users are encouraged to implement safeguards against bias in speech and visual representation, ensuring fair and respectful deployment across diverse demographics. We advocate for transparent disclosure when synthetic media is used in public-facing applications to maintain trust and accountability.

REFERENCES

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Lianghua Huang, Zhi-Fan Wu, Yupeng Shi, Wei Wang, Mengyang Feng, Junjie He, Chen-Wei Xie, Yu Liu, Jingren Zhou, Ang Wang, et al. Wan-streamer v0. 2: Higher resolution, same latency. *arXiv preprint arXiv:2607.04443*, 2026a.

Lianghua Huang, Zhifan Wu, Wei Wang, Yupeng Shi, Mengyang Feng, Junjie He, Chenwei Xie, Yu Liu, Jingren Zhou, Ang Wang, et al. Wan-streamer v0. 1: End-to-end real-time interactive foundation models. *arXiv preprint arXiv:2606.25041*, 2026b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026c.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

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

Haoran Li, Fredreic Li, Shichen Ma, Jie Huang, Yijun Liu, Jiaqi Shi, and Yanwen Ma. Joyai-echo: Pushing the frontier of long audio-visual generation. 2026b.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

Xiang Lyu, Yuxuan Wang, Tianyu Zhao, Hao Wang, Huadai Liu, and Zhihao Du. Build llm-based zero-shot streaming tts system with cosyvoice. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2. IEEE, 2025.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Tien-Tsin Wong, Yuan-Fang Li, and Cunjian Chen. Consistent and controllable image animation with motion diffusion models. In *Computer Vision and Pattern Recognition*, pp. 7288–7298, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Le Shen, Qian Qiao, Tan Yu, Ke Zhou, Tianhang Yu, Yu Zhan, Zhenjie Wang, Ming Tao, Shunshun Yin, and Siyuan Liu. SoulX-FlashTalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation, 2025. URL https://arxiv.org/abs/2512.23379.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation. *arXiv preprint arXiv:2603.11647*, 2026.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, et al. Streamavatar: Streaming diffusion models for real-time interactive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10887–10897, 2026.

Linrui Tian, Qi Wang, and Bang Zhang. Streamchar: Long-horizon streaming character audio-video generation with decoupled orchestration. *arXiv preprint arXiv:2605.25659*, 2026.

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

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pp. 1–20, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

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, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

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.

A Appendix

A.1 Discussion and Comparison with TIA2V Models

Vorch-Streamer targets unified text-to-audio-video (T2AV) generation, where a global textual description serves as the primary external condition for jointly generating temporally coordinated audio and video. This differs from text-and-image-to-audio-video (TIA2V) models, which typically take a reference image and a driving audio signal, with the latter often synthesized by an external TTS model before video generation. Thus, TIA2V follows a cascaded audio-to-video paradigm, whereas Vorch-Streamer jointly models audio and video within a unified DiT framework without requiring pre-generated audio or a reference image.

Speech planning is not driving audio. The speech planning tokens in Vorch-Streamer should not be regarded as an implicit driving audio signal. Driving audio is a dense acoustic condition that specifies phonetic realization, speaker characteristics, prosody, and temporal structure, thereby strongly constraining subsequent video generation. In contrast, speech planning tokens only provide high-level information about *what* to say and *when* to speak. The final acoustic realization, including speaker timbre and prosody, is still generated by the DiT conditioned on the global text and preceding audio-video context. Thus, speech planning serves as a lightweight semantic and temporal planning mechanism rather than a substitute for pre-generated audio.

No reference-image dependency. TIA2V models use a reference image to anchor character identity, appearance, and the initial scene state. Vorch-Streamer instead generates these properties directly from text, together with subsequent motion and scene evolution. This enables a pure text-conditioned T2AV setting, while avoiding the need for a visually compatible reference image. Conversely, TIA2V benefits from the stronger identity and appearance constraints provided by such an input, making the two paradigms complementary.

Joint audio-video generation. The unified formulation also allows Vorch-Streamer to model audio events beyond speech, such as clapping, footsteps, object impacts, and ambient sounds, together with their corresponding visual events. More generally, jointly generating the two modalities provides a direct mechanism for learning temporal correspondences between speech and lip motion, physical actions and their associated sounds, and scene events and ambient audio. In contrast, in a cascaded TIA2V pipeline, the downstream video generator is conditioned on an audio signal that has already been determined upstream.

Overall, the key distinction lies not simply in the input modality, but in how information is allocated between external conditions and the generative model. TIA2V explicitly provides strong visual and acoustic conditions, which is beneficial when preserving a specific identity, appearance, or prescribed speech performance. Vorch-Streamer instead generates the complete audio-video stream from text, with speech planning providing only lightweight semantic and temporal guidance while the final multimodal realization is jointly determined by the model.

Table 6: Long-horizon preservation over each method's available rollout. We report the score over the final 10-second window and, in parentheses, its change from the initial 10-second window.

Method	Sync-C ↑	Sync-D ↓	Dynamic ↑	Drift ↓	Human Anatomy ↑	Human Identity ↑
LTX2.3	6.21 (−18.51%)	8.46 (+14.98%)	0.1647 (−61.11%)	0.0336 (−32.93%)	0.9589 (−1.21%)	0.9595 (+0.07%)
LiveAvatar+TTS	7.04 (−3.57%)	8.38 (+2.80%)	0.2941 (−3.85%)	0.0234 (−17.54%)	0.9707 (+0.32%)	0.9960 (+1.12%)
SoulX-FlashTalk+TTS	8.79 (+1.65%)	7.25 (−1.55%)	0.2118 (+28.57%)	0.0202 (−8.99%)	0.9690 (+0.16%)	0.9956 (+1.22%)
JoyAI-Echo	2.24 (−16.30%)	12.90 (+0.11%)	0.0353 (−50.00%)	0.0355 (+68.78%)	0.9868 (+2.19%)	0.9874 (−1.13%)
OmniForcing (30 s)	0.46 (−31.94%)	13.52 (+2.40%)	0.2235 (+280.00%)	0.0096 (−78.35%)	0.9117 (+3.44%)	1.0000 (+13.04%)
Hallo-Live (30 s)	0.29 (−53.94%)	15.58 (+12.05%)	0.0000 (−100.00%)	0.0243 (−30.41%)	0.9000 (−3.76%)	0.9633 (−0.70%)
Vorch-Streamer (ours)	6.67 (−0.49%)	8.88 (+0.47%)	0.2588 (+10.00%)	0.0221 (−5.48%)	0.9284 (−0.04%)	1.0000 (+0.06%)

A.2 Implementation details

We initialize Vorch-Streamer from the 22B-parameter LTX2.3 model and train on 80K synthetic 12–21-second avatar clips at 768×512 and 24 FPS.

Stage 2 mixes 10% Teacher Forcing with 90% Diffusion Forcing and runs for 6,000 steps on 64 NVIDIA H200 GPUs using Fully Sharded Data Parallel (FSDP), bfloat16 mixed precision, gradient checkpointing, a global batch size of 64, and a $1e-4$ learning rate.

Stage 3 initializes the causal student from Stage 2, performs self-rollouts over the full 12–21-second horizon, and freezes the original bidirectional LTX2.3 as the real-score teacher for distribution matching distillation. It runs for 1,000 iterations on 32 H200 GPUs with a global batch size of 32, a 6:1 critic-to-generator update ratio, and a $1e-5$ learning rate for both the generator and fake-score model.

Inference uses four denoising steps per approximately one-second audio–video block and, unless stated otherwise, the ablation-selected $3{+}1$ causal window: three persistent prefix blocks plus the latest preceding block.

A.3 Additional Experiments

A.3.1 Additional Long-Horizon Analysis

We evaluate long-horizon behavior through both endpoint changes and complete temporal trajectories. Table 6 compares the final 10-second window with the initial 10-second segment. Vorch-Streamer exhibits minimal degradation over its approximately two-minute rollout while maintaining strong dynamics, demonstrating that long-term stability does not come at the cost of static motion. In contrast, native T2AV baselines show substantially larger degradation in synchronization, dynamics, or drift, with OmniForcing and Hallo-Live already exhibiting severe degradation within their shorter 30-second rollouts. The TIA2V references maintain strong endpoint metrics, but rely on external audio and a first-frame condition and therefore are not directly comparable to text-driven T2AV generation. Overall, Vorch-Streamer achieves the best balance of speech synchronization, sustained motion, and long-term consistency among T2AV methods.

A.3.2 Impact of Context Selection

We evaluate context selection at inference time by fixing the model and varying the causal attention window, denoted as $P{+}R$, where P and R are the numbers of persistent prefix and recent blocks, respectively. Although the tested settings show comparable short-term quality, their long-horizon stability differs substantially. Removing the persistent prefix ($0{+}3$) leads to severe degradation, while retaining prefix blocks alleviates this failure. The proposed $3{+}1$ setting achieves the strongest long-term preservation, with the lowest drift and highest ArcFace and CLIP-I similarity. These results suggest that three persistent prefix blocks provide stable global appearance anchors, while restricting recent context to the immediately preceding block limits error accumulation from self-generated history. Thus, the $3{+}1$ context is particularly important for reliable long-form inference, with its main benefit lying in long-horizon preservation rather than uniformly improving all quality metrics.

Table 7: Inference-time ablation of causal-context selection. Following Table 5, FVD, VBench Dynamic Degree and Drift, and VBench2 Human Identity are computed over the available rollout. ArcFace and CLIP-I report the mean similarity of the final 10-second window to the first frame.

Context	FVD ↓	Dynamic ↑	Drift ↓	Human Identity ↑	ArcFace ↑	CLIP-I ↑
1+3	544.13	0.2941	0.0383	0.9328	0.5882	0.9082
0+3	501.70	0.2235	0.1589	0.6921	0.2759	0.7533
3+3	498.93	0.2706	0.0377	0.9427	0.6386	0.9150
3+1 (ours)	549.99	0.2706	0.0286	0.9996	0.7534	0.9324

A.3.3 More Showcases of Vorch-Streamer

As shown in Fig. 5, we provide representative showcases generated by Vorch-Streamer across diverse characters, scenes, and interactions. Our method generates dynamic and coherent audio-video content with synchronized speech, facial expressions, and body motions, while consistently preserving character identity and scene appearance throughout long-horizon real-time inference. These results demonstrate that Vorch-Streamer maintains temporal coherence and visual consistency over long generation horizons, supporting high-quality long-form streaming audio-video generation.

A.3.4 Extreme-Long-Horizon Inference

To further evaluate the extreme-long-horizon stability of Vorch-Streamer, we conduct an extreme-long-horizon inference experiment with a duration of up to 2 hours. We evaluate the model on 40 test cases in our evaluation set, where Qwen3.5-27B continuously generates speech utterances as the speech input throughout the entire 2-hour rollout. This setting substantially extends beyond the training horizon and provides a stringent test of the model's ability to maintain audiovisual consistency over prolonged inference.

A key property of Vorch-Streamer is that the generation of a distant block does not intrinsically become more difficult as the rollout proceeds. With our windowed context design and explicit speech planning, the 7200-th block, for example, attends to the same type of context as the 7-th block, namely the first three blocks and the most recent preceding block. This bounded and structured context prevents the effective conditioning scope from growing with the sequence length, enabling Vorch-Streamer to preserve scene appearance and character identity over extremely long horizons. Nevertheless, as the number of generated blocks increases, the probability of encountering a deviation somewhere over the entire rollout naturally accumulates. In a small number of cases, we therefore observe short-lived scene drift, which is subsequently recovered as generation proceeds, rather than a progressively increasing tendency for later blocks to fail.

As shown in Fig. 4, we visualize representative frames from several test cases at 0, 30, 60, 90, and 120 minutes, together with the corresponding ArcFace and CLIP Image Similarity curves throughout the 2-hour generation. In most cases, the generated results maintain consistent character identity and scene appearance throughout the rollout, while preserving active and coherent visual dynamics rather than converging to static outputs. The similarity metrics remain largely stable as the inference horizon increases, indicating that the occasional deviations are limited in duration and do not lead to persistent drift. These results demonstrate that Vorch-Streamer can sustain stable and dynamic audiovisual generation over an extreme 2-hour inference horizon, with limited degradation in identity and scene consistency.



Figure 4: Extreme-long-horizon inference of **Vorch-Streamer** over a 2-hour rollout. We visualize representative frames at 0, 30, 60, 90, and 120 minutes for three test cases, together with the corresponding ArcFace and CLIP image similarity to the first frame throughout the rollout. Vorch-Streamer largely preserves character identity and scene appearance over the entire 2-hour horizon while maintaining active visual dynamics.



Figure 5: Real-time long-form text-to-audio-video generation with **Vorch-Streamer**. Conditioned on a global caption and long speech (left), our model causally and jointly streams video and synchronized speech at 27 FPS, without externally supplied audio or a reference first frame. Frames sampled at 0, 30, 60, and 90 seconds and near the end of three approximately two-minute rollouts demonstrate sustained identity and scene consistency together with expressive facial, gestural, and object motion; the blue waveforms visualize the continuously generated audio.