VORCH-DIRECTOR: INTERACTIVE WORLD STORY MODEL VIA NOISE-AWARE ERROR RECTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive continuation provides a natural path toward minute-scale audio-visual generation by repeatedly extending a short-window generator conditioned on previously generated video and audio. However, models are trained on *clean* ground-truth histories, while inference relies on their own generated histories, where accumulated errors cause identity drift, over-smoothing, and audio-visual desynchronization. Recent methods reduce this mismatch by reusing prediction *residuals* as synthetic corruption, but we observe that the effectiveness of residual correction critically depends on the flow-matching noise level σ at which residuals are produced. We propose **VORCH-DIRECTOR**, a *noise-level-aware* residual correction strategy that associates each residual with its originating noise level and injects residuals from matched noise regimes during training. By aligning injected errors with the denoising process, VORCH-DIRECTOR produces more realistic autoregressive histories while retaining efficient teacher-forcing training. Built on the audio-visual LTX-2 diffusion transformer, VORCH-DIRECTOR further introduces task embeddings to distinguish historical video, reference images, and target video, enabling unified conditioning for long-horizon generation. Together with a clean conditioning sink and mixed-task training, VORCH-DIRECTOR supports multi-shot, multi-subject, reference-guided audio-visual long-video generation. We evaluate VORCH-DIRECTOR on ST-Bench and introduce a new long-horizon audio-visual benchmark with metrics for quality drift and long-range consistency. Extensive experiments demonstrate improved stability and audio-visual fidelity over strong baselines.

1 INTRODUCTION

Diffusion transformers (DiTs) (Peebles & Xie, 2023) have rapidly advanced text-to-video generation. Trained at massive scale, models such as Wan (Wan et al., 2025), HunyuanVideo (Kong et al., 2024), and CogVideoX (Yang et al., 2024) generate several seconds of high-fidelity video from text, while recent systems additionally synthesize synchronized audio and video (Team Seedance et al., 2026). Real-world media, however, is minute-scale, multi-shot, and audio-visual. Characters must preserve their identities across distant scene transitions, speakers must maintain synchronized voice and lip motion after camera cuts, and background music should continue seamlessly over long temporal horizons (Meng et al., 2025; Wei et al., 2026; Joy Future Academy (JD) and collaborators, 2026). Since a single forward pass cannot cover such sequences, current systems typically rely on *autoregressive continuation*: generating long videos segment by segment while conditioning each segment on previously generated video and audio (Chen et al., 2024; Huang et al., 2025; Liu et al., 2025).

Although autoregressive continuation naturally extends a strong short-window generator to arbitrary lengths, it introduces train–test mismatches that accumulate throughout the rollout. The first mismatch arises from *teacher forcing*. During training, the conditioning history is encoded from clean ground-truth video and audio, whereas at inference it consists entirely of the model's own predictions, which inevitably accumulate identity drift, blur, motion inconsistency, and audio-visual misalignment. As these imperfect predictions are recursively reused as future context, generation quality progressively deteriorates over long horizons (Chang et al., 2026; Wang et al., 2025). We

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

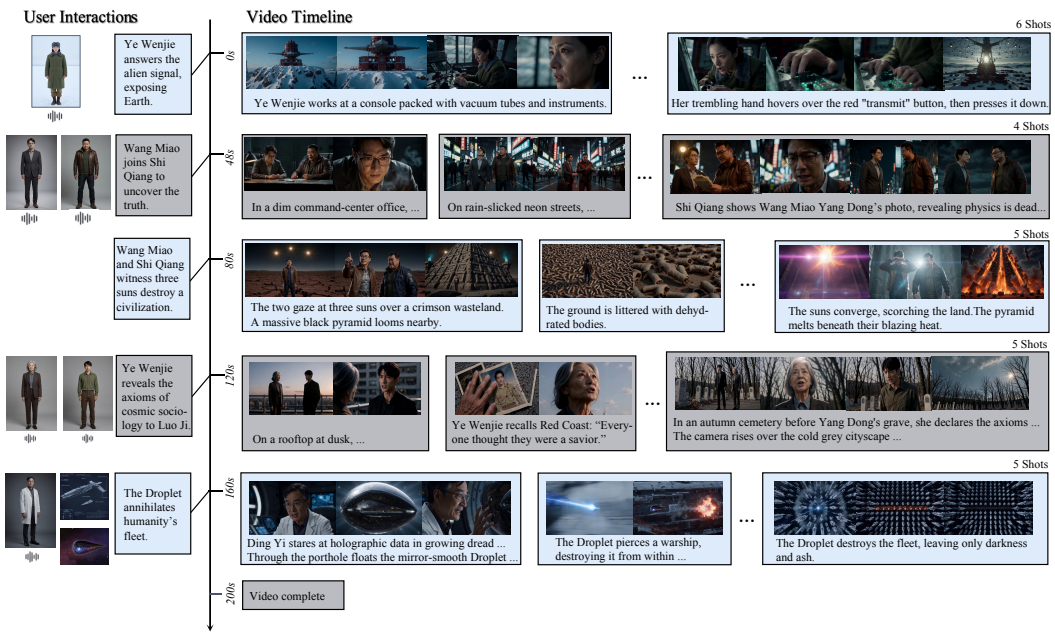

Figure 1: VORCH-DIRECTOR generates minute-scale, multi-shot, audio-visual video by autoregressive continuation. Given a few reference images that fix subject identity, our model extends an audio-visual video over dozens of shots with camera cuts, keeping character appearance, scene, and speaker voice consistent while emitting synchronized audio.

argue that this history mismatch, rather than context length itself, is the primary obstacle to minute-scale audio-visual generation.

Existing methods mitigate this problem from complementary perspectives. Memory and retrieval approaches (Wei et al., 2026; Zhou et al., 2026; Huang et al., 2026; Meng et al., 2026) improve *what* historical information is recalled, but still assume the recalled context is clean. Self-rollout methods (Huang et al., 2025; Liu et al., 2025; Zheng et al., 2026) instead train on self-generated histories, matching the inference distribution at the cost of expensive multi-step rollouts. Matrix-Game (Wang et al., 2026) recently proposed a lightweight alternative by injecting prediction residuals into the conditioning history, allowing standard teacher-forcing training while exposing the model to inference-like errors. However, existing residual reuse overlooks a key property of diffusion models: prediction residuals are inherently dependent on the flow-matching noise level at which they are produced.

Our first contribution is therefore a σ-*aware* residual injection strategy. Residuals generated under different noise levels correspond to different error regimes: high-σ residuals primarily describe coarse structural deviations, whereas low-σ residuals capture fine-detail errors. Reusing them indiscriminately either under-corrupts or over-corrupts the conditioning history. We instead associate each residual with its originating noise level and inject only residuals sampled from matching noise regimes, producing training histories that more faithfully resemble inference-time contexts while retaining the efficiency of teacher forcing.

Beyond imperfect history, we identify a second source of train–test mismatch that has received little attention: *positional extrapolation*. Existing continuation models typically distinguish historical context, reference images, and target video through positional layouts, causing positional indices to grow continuously as generation proceeds. In contrast, training is almost always performed on short clips with much smaller positional ranges. Consequently, long-horizon inference requires the model to extrapolate positional encodings far beyond those observed during training, introducing another form of exposure bias. To eliminate this mismatch, we explicitly represent the role of each token using lightweight *task embeddings* that distinguish historical video, reference images, and target video. This decouples semantic roles from positional encoding, allowing positional indices to be

reset for every newly generated segment while preserving temporal consistency. As a result, the positional distributions remain identical between training and inference regardless of rollout length.

We integrate these two techniques into the LTX-2 audio-visual DiT together with a short clean conditioning sink, a unified prompt template, frame-rate-aligned positional encodings, and mixed-task training spanning continuation, image reference, subject reference, and audio-visual continuation. A single model thereby supports multi-shot, multi-subject, reference-guided, audio-visual long-video generation.

The main contributions of this paper are summarized as follows:

- We propose VORCH-DIRECTOR, a *σ-aware* residual injection strategy that matches prediction residuals to the corresponding flow-matching noise level, substantially mitigating exposure bias during autoregressive diffusion generation.
- We introduce a unified conditioning interface based on lightweight task embeddings, which explicitly distinguish historical video, reference images, and target video, enabling positional encoding reuse across continuation steps.
- We build an LTX-2-based audio-visual long-video generation framework that combines the proposed training strategy with mixed-task training, supporting multi-shot, multi-subject, reference-guided audio-visual generation within a single model.
- We evaluate our approach on ST-Bench and further introduce a longer audio-visual benchmark with two long-horizon metrics, *Quality Drift* and *Anchor-relative Consistency*, which better characterize progressive degradation than conventional average or adjacent-shot evaluations.

2 RELATED WORK

Autoregressive Long Video Generation Diffusion Transformers scaled text-to-video to high fidelity in systems such as Wan, HunyuanVideo, and CogVideoX (Wan et al., 2025; Kong et al., 2024; Yang et al., 2024; Ma et al., 2025; Wang et al., 2024), but a single quadratic-attention pass cannot reach minute scale. The field therefore adopts *autoregressive continuation*, synthesizing a long video chunk by chunk with each segment conditioned on previous context. Diffusion Forcing (Chen et al., 2024) unifies next-token prediction with full-sequence diffusion via per-token noise levels, and later work pursues real-time streaming through causal generators, distillation, and attention sinks (Zheng et al., 2026; Liu et al., 2025), or inference-time trajectory and hierarchical-denoising tricks for scale (Park et al., 2026; Zou et al., 2026; Ji et al., 2026). Across all of these the recurring failure is consistency drift and error accumulation as identity and background degrade with length; we target this failure at the training level rather than proposing a new streaming schedule, and our correction is compatible with such schedules.

Exposure Bias and Error-Residual Trainingtextbf The root cause of drift is exposure bias: models are trained on clean ground-truth history yet run on self-generated history at inference (Chang et al., 2026; Wang et al., 2025). Self Forcing (Huang et al., 2025) and distillation recipes (Zheng et al., 2026) close the gap by rolling the generator out during training, which is effective but adds multi-step generation to every step. A lighter family instead reuses the model's own *errors*: Matrix-Game 3.0 (Wang et al., 2026) injects prediction residuals from an online buffer into the context of an interactive world model, Stable Video Infinity (Li et al., 2025) recycles self-generated errors as supervision, and corruption-aware training (Maduabuchi et al., 2025) shows that controlled input corruption improves robustness. VORCH-DIRECTOR belongs to this error-reuse family, but where prior injections apply residuals at a single, noise-level-agnostic strength, we make the correction *noise-level aware*: each residual is stored with the flow-matching level σ at which it arose, and to corrupt the history for a step at level σ we sample residuals of a matching level and scale them to that regime. This σ-matched strength control, together with a clean conditioning sink and audio-visual task embeddings, is what adapts residual reuse into a stable correction for long audio-visual continuation.

Multi-Shot, Reference-Guided, and Audio-Visual Generation Story-level generation must keep recurring subjects, scenes, and, once sound is present, speaker voice consistent across many shots. One line composes independently generated shots: storyboard-then-animate pipelines such as Sto-

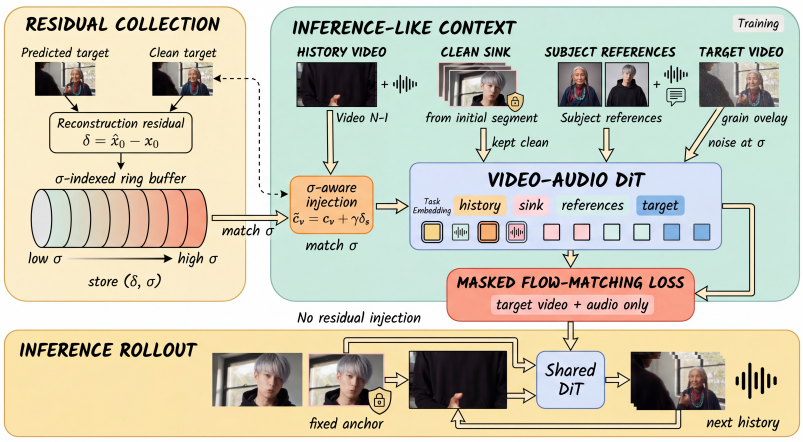

Figure 2: **Overview of VORCH-DIRECTOR.** During training the model predicts velocities on noised target tokens; the resulting reconstruction residuals are collected into a buffer *tagged by their noise level* σ. When forming a later step, residuals of a *matching* σ are sampled and injected into the conditioning-history tokens with strength γ, so the history carries inference-like error at the right intensity for the current denoising level. A short clean clip is appended as a stable *sink*, and subject reference images plus audio conditioning are assembled in context, each tagged by a task embedding. The flow-matching loss is applied only to the target video and audio tokens.

ryDiffusion followed by an image-to-video model (Zhou et al., 2024; Wan et al., 2025) enforce consistency only at sparse keyframes. Holistic and memory-based systems instead couple shots directly: HoloCine (Meng et al., 2025) models a whole scene with window cross-attention and sparse inter-shot attention, while shot-by-shot methods carry explicit memory, e.g. StoryMem (Zhang et al., 2025), the subject-reconstruction memory of Memento (Wei et al., 2026), and further memory/cache designs (Zhou et al., 2026; Luo et al., 2025). Reference- and subject-driven conditioning is likewise central to keeping identity fixed across cuts, which our four-way training mixture supports through subject reference images and a subject-IP task. Extending to sound raises the bar to audio quality, speaker identity, and audio-visual synchronization: JoyAI-Echo (Joy Future Academy (JD) and collaborators, 2026) targets minute-scale audio-visual stories with a cross-modal memory bank, UnityShots (Huang et al., 2026) turns a single-shot audio-visual diffusion model on the same LTX-2 backbone into a multi-shot generator via memory slots and boundary gating, and others address music-motion, multi-voice, and talking-avatar settings (Chen et al., 2026; Guo et al., 2026; Ye et al., 2026; Cui et al., 2026). These systems obtain consistency mainly by deciding *what* history to recall or by adding memory and gating modules; VORCH-DIRECTOR is orthogonal, operating at the training level to make the base continuation model robust to imperfect recalled or generated history, and jointly generating synchronized audio. We evaluate on the multi-shot story benchmark ST-Bench (Zhang et al., 2025; Wei et al., 2026) against StoryDiffusion+Wan, StoryMem, HoloCine, and Memento, and on a longer audio-visual benchmark with audio metrics and long-horizon degradation measures.

3 METHOD

We target autoregressive audio-visual continuation with *no train-inference gap*: the conditioning history the model sees during training should carry the same kind of imperfection, at the same intensity, that it will condition on at inference. Following the error-residual idea of Matrix-Game (Wang et al., 2026), we reuse the model's own prediction residuals as the corrupting signal. Our key change is to make this correction *noise-level aware*: the strength of the injected residual is controlled by, and matched to, the flow-matching noise level σ of the current step (Section 3.3). We further stabilize long rollouts with a short clean conditioning *sink* (Section 3.4), and support multi-subject, multi-shot, audio-visual generation through task-embedded in-context conditioning and a unified multi-task corpus (Section 3.2).

Table 1: **Task-embedding identifiers.** Each in-context token is tagged by a role-specific task embedding e_τ; the target is $\tau_{\text{tgt}}{=}0$ and the conditioning roles take distinct, non-conflicting ids. History video and audio share one id; the i-th reference image and the i-th reference audio are index-aligned so they are bound one-to-one. We refer to the roles by symbol throughout.

Symbol	Conditioning role	Task id
τ_{tgt}	Target video / audio (generated)	0
τ_{hist}	History cond. video & audio (shared)	1
τ_{sink}	Clean conditioning sink	2
$\tau_{\text{ref}}^{(i)}$	i-th subject reference image	$2{+}i$
$\tau_{\text{aud}}^{(i)}$	i-th subject reference audio (paired)	$2{+}N{+}i$

3.1 Overview and Preliminaries

Problem formulation. A single forward pass cannot span minutes, so we generate a long audio-visual video *autoregressively*: the video is split into fixed-length segments, and segment t is generated conditioned on the previously produced video and audio. At inference the target of segment t is re-encoded and becomes the history of segment $t{+}1$, so the model conditions on its own output. This exposes a train-inference gap in the *quality* of the conditioning history: during training the history is encoded from *clean* ground-truth video and audio (teacher forcing), whereas at inference the same slots are filled by the model's own previous outputs, which already carry identity drift, blur, and audio-visual misalignment. Fed back segment after segment, these errors compound. Our goal is to make the training-time history carry the same kind of imperfection, at the same intensity, as the self-generated history seen at inference.

Flow-matching backbone. We build on LTX-2, an audio-visual diffusion transformer (DiT) that generates video and audio jointly in a shared token space with rectified flow. Let x_0 denote a clean latent (video or audio). Given noise $\epsilon \sim \mathcal{N}(0, I)$ and level $\sigma \in [0, 1]$, the forward interpolation and velocity target are

$$x_\sigma = (1-\sigma)\, x_0 + \sigma\, \epsilon, \qquad u = \epsilon - x_0, \tag{1}$$

and the network v_θ predicts u. The clean latent is recovered from any x_σ by the one-step estimate

$$\hat{x}_0 = x_\sigma - \sigma\, v_\theta(x_\sigma, c), \tag{2}$$

where c collects all conditioning tokens.

Task-embedded in-context conditioning. A segment is generated from several conditioning signals: a text prompt p, the previous segment's video and audio (history latents c_v, c_a), a short clean reference clip r, and N subject reference images $\{s_i\}_{i=1}^N$. All visual and audio conditions are VAE-encoded, patchified, and concatenated with the noised target tokens into a single sequence; text p enters by cross-attention. Every token carries a learned *task embedding* e_τ added to it, where the role identifier τ ranges over the roles in Table 1. The generated target video and audio share τ_{tgt}; the continuation history video and its audio share a single id τ_{hist} (the two modalities of the same history carry the same tag); the clean sink uses a distinct τ_{sink} so the model can tell the stable anchor apart from the (corrupted) recent history. Because a sample may carry several subjects, the N reference images take *indexed* ids $\tau_{\text{ref}}^{(i)}$ and their paired reference audios take *index-aligned* ids $\tau_{\text{aud}}^{(i)}$, so image i and audio i of the same subject are bound one-to-one. The latents form

$$\begin{aligned} z = \big[\; & \underbrace{s_1, \ldots, s_N}_{\text{ref. img } (\tau_{\text{ref}}^{(i)})},\ \underbrace{a_1, \ldots, a_N}_{\text{ref. aud } (\tau_{\text{aud}}^{(i)})}, \\ & \underbrace{c_v, c_a}_{\text{history } (\tau_{\text{hist}})}\ ,\ \underbrace{r}_{\text{sink } (\tau_{\text{sink}})}\ ,\ \underbrace{x^v, x^a}_{\text{target } (\tau_{\text{tgt}})}\ \big], \end{aligned} \tag{3}$$

with a mask m marking conditioning ($m{=}1$) versus target ($m{=}0$) tokens. Conditioning tokens are inserted clean at $\sigma{=}0$ and only the target is noised.

3.2 Unified Multi-task Data Construction

Our training corpus covers text-, image-, video-, and audio-conditioned generation. We collect same-source shots, filter low-quality or hard-cut clips, index recurring subjects with face/person embeddings, and use a multimodal captioner (Doubao) to produce video, audio, and joint audio-visual captions. Each source is split into a history (Video-1) and target (Video-2) pair and rendered into the structured prompt template; the complete construction pipeline is shown in the supplementary material.

Task mixture. The corpus is a balanced mixture (25% each) of four task types: **continuation** (video history only), **continuation+image** (history plus subject-reference images), **subject-IP** (reference-image-driven generation), and **audio-visual continuation** (history video, history audio, and reference images). Sampling all four jointly lets the same weights serve pure continuation, reference-image identity control, and synchronized audio-visual generation. The σ-aware residual injection of Section 3.3 is applied to the continuation history stream (τ_{hist}) across these tasks.

Prompt construction. We use a fixed slot-based template with explicit image/audio identifiers and an outer instruction wrapper, so identity, voice, background sound, and per-shot actions are represented uniformly across tasks (full templates in the supplementary material).

3.3 σ-Aware Error-Residual Injection

Collecting residuals with their noise level. At every training step the model already predicts velocities on the noised target tokens; from Eq. (2) this yields a reconstruction $\hat{x}_0$ and a residual $\delta = \hat{x}_0 - x_0$ on each target token, which is exactly the error such a token would carry if it were fed back as history. We collect these residuals into a fixed-size ring buffer, but crucially we *store each residual together with the noise level σ at which it was produced*. Only target-token residuals are collected, and a mixed top-norm/random selection keeps a fraction of each batch (larger-norm residuals, which correspond to harder errors, plus random ones for diversity).

Matching residual strength to the current noise level. The central design choice is how strongly, and with which residuals, to perturb the conditioning history. We observe that the useful magnitude of the correction depends on the noise level: a residual produced at a high σ (coarse, low-detail regime) is not interchangeable with one produced at a low σ (fine-detail regime). We therefore make the injection *σ-aware*. When perturbing the history for a step whose target noise level is σ, we draw residuals from the buffer whose stored level σ' satisfies $|\sigma' - \sigma| \le \tau_\sigma$ (falling back to the nearest available level if none qualifies), and inject them into the conditioning-history tokens (τ_{hist}):

$$\tilde{c}_v \;=\; c_v \;+\; \gamma\,\delta_s, \qquad \delta_s \sim \text{Buffer}\big(\,|\sigma' - \sigma| \le \tau_\sigma\,\big), \tag{4}$$

where the injection strength γ is sampled per step from a range $[\gamma_{\min}, \gamma_{\max}]$. This ties the corruption applied to the history to the same noise regime the model is currently learning to denoise, so the history "looks wrong" in a way that is consistent with the current step rather than at an arbitrary, mismatched intensity. In our final model we use a comparatively strong correction, $\gamma \sim \mathcal{U}(0.9, 1.2)$, with tolerance τ_σ=0.05; injection is applied to the history tokens only and never to the target, reference images, or the clean sink.

3.4 Conditioning Stabilization

Injecting error into the whole history reproduces inference-time corruption, but if *every* history token is perturbed the model loses any reliable cue and identity drifts. We therefore keep one uncorrupted anchor. Let h be the true history and $r = h_{[0:\Delta]}$ its leading Δ-second window; the injection of Eq. (4) is applied only to the main history stream and never to r:

$$\begin{aligned} \tilde{c}_v &= c_v + \gamma\,\delta_s \;\; (\tau_{\text{hist}}), \\ r &= \text{VAE}(h_{[0:\Delta]}) \;\; (\tau_{\text{sink}},\ \text{no injection}). \end{aligned} \tag{5}$$

The sink r carries a distinct task embedding $e_{\tau_{\text{sink}}} \neq e_{\tau_{\text{hist}}}$, so the model can tell the clean anchor from the corruptible history and learns the residual map

$$x_0 \;\approx\; \mathcal{G}\big(\,\tilde{c}_v,\; r,\; \{s_i\},\; p\,\big), \tag{6}$$

i.e. it reconstructs the clean target x_0 by leaning on the clean sink r and *correcting* the corrupted history $\tilde{c}_v$ rather than propagating its error. For multi-subject, multi-shot generation we additionally supply N clean subject reference images $\{s_i\}_{i=1}^{N}$ as global (non-temporal) conditioning tokens with subject-reference task ids, so one model can place named subjects and re-establish their appearance after camera cuts.

3.5 Training Objective

At each step we sample a clip and its task from the four-way mixture of Section 3.2, assemble the sequence of Eq. (3), apply σ-aware residual injection (Eq. (4)) to the history after a short warmup, and noise only the target. The model is supervised by a masked flow-matching loss over both modalities,

$$\begin{aligned}\mathcal{L} \;= \mathbb{E}_{x_0,\epsilon,\sigma}\Big[\big\| v_\theta^v(\tilde{z}_\sigma, p) - (\epsilon^v - x_0^v) \big\|^2 \\ +\; \lambda_a \big\| v_\theta^a(\tilde{z}_\sigma, p) - (\epsilon^a - x_0^a) \big\|^2 \Big],\end{aligned} \tag{7}$$

where $\tilde{z}_\sigma$ is the assembled sequence with the residual-injected history $\tilde{c}_v$, clean sink, references, and audio held at $\sigma{=}0$, and noised targets x_σ^v, x_σ^a; λ_a balances the audio term. The loss is applied only where $m{=}0$, so the perturbed history receives no gradient and the supervision signal stays clean while the *context* distribution is made inference-like.

4 Experiments

4.1 Implementation Details

Vorch-Director fine-tunes the 22B LTX-2 audio-visual DiT with Gemma-3 text encoding at 480p and 24 fps, using the balanced four-way task mixture and full fine-tuning. Buffer settings, the 1-s clean sink, and inference procedure are reported in the supplementary material.

4.2 Benchmarks and Metrics

Unified metric suite. To make comparisons fair, we re-score every method on every benchmark with one evaluator and identical settings. We report cross-shot consistency (ViCLIP, Self-CIDS, ARC, and Reappear), video quality (Aesthetic and Imaging), text consistency (CLIP-T), and speech (Voice and character accuracy). Implementation details and language-specific scoring choices are given in the supplementary material.

Long-range identity metrics. Average and adjacent-shot similarity can hide gradual identity collapse. We therefore use an anchor-based *Anchor-relative Consistency* (ARC, higher better), which compares every later appearance with the first reliably detected subject shot,

$$\mathrm{ARC} = \tfrac{1}{N}\textstyle\sum_t \mathrm{sim}(f_{\text{anchor}}, f_t), \tag{8}$$

and *Reappearance Consistency* (Reappear, higher better), which applies the same comparison only to non-adjacent shot pairs (gap > 1). Details of the identity embeddings and aggregation appear in the supplementary material.

Benchmarks. We evaluate on three benchmarks. **ST-Bench** is a multi-shot story benchmark without audio, so the speech metrics do not apply. **UnityShots** is a multi-shot benchmark that we run in both text-to-video (T2V) and reference-to-video (R2V) settings. Our own **long audio-visual benchmark** comprises subject-referenced, minute-scale multi-shot generation with synchronised speech, music, and sound effects; it is the only one of the three that exercises audio and minute-scale drift jointly.

4.3 Quantitative Comparisions

Comparison on ST-Bench Table 2 reports ST-Bench under the unified suite. Vorch-Director leads all four cross-shot consistency metrics by a wide margin, most visibly on ViCLIP (0.7887 against a best baseline of 0.5900) and on the two long-range measures ARC and Reappear. The baselines remain competitive on video quality and text consistency, where per-frame scores do not reward temporal stability.

Table 2: **ST-Bench comparison** under our unified metric suite. All rows are computed by the same evaluator under identical settings; no published numbers are mixed in. ST-Bench has no audio track, so the speech metrics are omitted. Best per column in **bold**.

	Cross-Shot Consistency				Video Quality		Text Consist.
Method	ViCLIP↑	Self-CIDS↑	ARC↑	Reappear↑	Aesthetic↑	Imaging↑	CLIP-T↑
StoryMem	0.5645	0.4816	0.4034	0.3940	**0.5992**	0.6350	0.2815
HoloCine	0.4666	0.4871	0.3329	0.3399	0.5557	0.5178	0.2945
Memento	0.5258	0.4883	0.4003	0.3978	0.5979	0.6291	**0.3063**
JoyAI-Echo	0.5900	0.4567	0.3961	0.4114	0.5068	**0.6696**	0.2840
VORCH-DIRECTOR (ours)	**0.7887**	**0.6348**	**0.6570**	**0.6648**	0.6423	0.5950	0.27444

Table 3: **UnityShots comparison** under the unified metric suite, with T2V in the top block and R2V in the bottom block. Visual-only baselines emit no audio track, so their speech metrics are "n/a". Best per column within each block in **bold**.

	Cross-Shot Consistency				Video Quality		Text	Speech	
Method	ViCLIP↑	Self-CIDS↑	ARC↑	Reappear↑	Aesthetic↑	Imaging↑	CLIP-T↑	Voice↑	Acc↑
Text-to-video (T2V)									
StoryMem	0.4997	0.4847	0.3743	0.3742	0.5415	0.5853	**0.2818**	n/a	n/a
HoloCine	0.7079	**0.6211**	0.6048	0.5978	0.4941	0.3413	0.2696	n/a	n/a
Memento	0.5356	0.4947	0.4402	0.4125	0.5343	0.5771	0.2789	n/a	n/a
JoyAI-Echo	0.5497	0.4603	0.4020	0.4065	0.3850	**0.6436**	0.2761	0.6193	0.4084
VORCH-DIRECTOR (ours)	**0.7335**	0.6031	**0.6544**	**0.6605**	**0.5533**	0.5872	0.2538	**0.6358**	**0.4720**
Reference-to-video (R2V)									
DreamID-Omni	0.6770	0.5368	0.5453	0.5581	**0.5913**	**0.7232**	**0.2671**	0.6191	0.3722
VORCH-DIRECTOR (ours)	**0.7422**	**0.6332**	**0.6861**	**0.6756**	0.5820	0.6625	0.2551	**0.6550**	**0.4582**

Comparison on UnityShots Table 3 reports UnityShots in the T2V and R2V settings. In T2V VORCH-DIRECTOR leads on ViCLIP, ARC, and Reappear; HoloCine, which generates a whole scene jointly rather than shot by shot, is the one baseline that comes close on cross-shot consistency and slightly exceeds us on Self-CIDS (0.6211 vs. 0.6031), while paying for it in video quality (Imaging 0.3413). In R2V the reference images give the model an explicit identity target to hold throughout the rollout, and VORCH-DIRECTOR leads the strongest R2V baseline on all four consistency metrics and on both speech metrics, while trailing it on single-frame Imaging.

Long Audio-Visual Comparison Our long audio-visual benchmark is the only setting in which audio and minute-scale drift are exercised together. We compare against JoyAI-Echo, the strongest publicly runnable audio-visual system, and against three visual-only multi-shot systems — Memento, StoryMem, and HoloCine — which we run on the same multi-shot prompts and score identically, leaving their speech metrics as "n/a" (Table 4).

VORCH-DIRECTOR leads all four cross-shot consistency metrics on this benchmark. Against the audio-visual baseline the margins are large (ViCLIP 0.6869 vs. 0.4922; Self-CIDS 0.6032 vs. 0.4561; ARC 0.5034 vs. 0.4026; Reappear 0.6049 vs. 0.4622), and VORCH-DIRECTOR also leads on both speech metrics, improving voice consistency from 0.7244 to 0.7533 and speech accuracy from 0.7102 to 0.7811. The strongest visual-only competitor is Memento, which VORCH-DIRECTOR still exceeds on all four consistency measures, most clearly on the two long-range ones (ARC 0.5034 vs. 0.3913; Reappear 0.6049 vs. 0.5976), while ARC separates the two by the widest relative margin because it is the measure most sensitive to slow drift away from the subject's first appearance. VORCH-DIRECTOR additionally attains the best Aesthetic of all five systems. JoyAI-Echo retains the best single-frame Imaging (0.7028 vs. 0.6766) and the best CLIP-T (0.2500 vs.

Table 4: **Long audio-visual benchmark** under the unified metric suite. All rows are generated from the same multi-shot prompts and scored by the same evaluator. Memento, StoryMem, and HoloCine emit no audio track, so their speech metrics are "n/a" and they solve a strictly easier problem than the audio-visual systems. VORCH-DIRECTOR is the σ-aware configuration with the clean sink. Best per column in **bold**.

	Cross-Shot Consistency				Video Quality		Text	Speech	
Method	ViCLIP↑	Self-CIDS↑	ARC↑	Reappear↑	Aesthetic↑	Imaging↑	CLIP-T↑	Voice↑	Acc↑
StoryMem	0.4782	0.5362	0.4284	0.4831	0.5356	0.6797	0.2267	n/a	n/a
HoloCine	0.4203	0.4909	0.3651	0.3705	0.5274	0.4322	0.2207	n/a	n/a
Memento	0.6724	0.5951	0.3913	0.5976	0.5468	0.6797	0.2299	n/a	n/a
JoyAI-Echo	0.4922	0.4561	0.4026	0.4622	0.5615	**0.7028**	**0.2500**	0.7244	0.7102
VORCH-DIRECTOR (ours)	**0.6869**	**0.6032**	**0.5034**	**0.6049**	**0.6306**	0.6766	0.2468	**0.7533**	**0.7811**

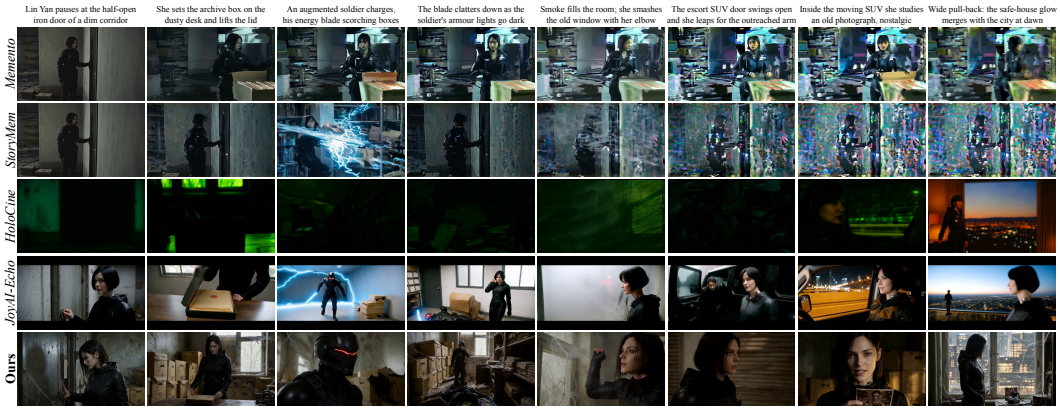

Figure 3: **Qualitative comparison over a three-minute rollout.** Eight frames span the rollout. Memento and StoryMem accumulate corruption, HoloCine is consistently under-exposed, JoyAI-Echo loses identity, while VORCH-DIRECTOR preserves the protagonist and follows the scene changes.

0.2468), the two metrics that score frames independently and therefore do not reward temporal stability. One caveat applies when reading the visual-only rows: those systems emit no audio and thus solve a strictly easier problem, and they produce noticeably less inter-frame motion than VORCH-DIRECTOR, which inflates any similarity-based consistency score in their favour — making the consistency margins reported here conservative rather than optimistic.

4.4 QUALITATIVE COMPARISON

Figure 3 shows the complementary failure modes: Memento and StoryMem accumulate corruption, HoloCine is poor from the first shot, and JoyAI-Echo loses identity. VORCH-DIRECTOR remains recognisable and artefact-free while following the prompt across the full rollout.

5 CONCLUSION

We presented VORCH-DIRECTOR, a noise-level-aware residual correction strategy for autoregressive audio-visual generation. Matching residuals to the flow level, together with a clean conditioning sink and task-embedded multimodal context, reduces the train–test gap without multi-step rollout training. On ST-Bench and our long audio-visual benchmark, VORCH-DIRECTOR improves cross-shot consistency and speech fidelity over strong baselines. The ablation results in the supplementary material show that both the matched residual and clean sink are necessary; extending this correction to audio history is future work.

REFERENCES

Shuning Chang, Weihua Chen, Jiasheng Tang, Hao Xu, Zeyu Zhang, Hangjie Yuan, Yu Lu, and Ruigang Niu. Towards error-free long video generation. *arXiv preprint arXiv:2606.22370*, 2026.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024.

Yuheng Chen, Teng Hu, Yuji Wang, Qingdong He, Zhucun Xue, Qianyu Zhou, Jason Li, Lizhuang Ma, Jiangning Zhang, and Dacheng Tao. Cinedance: Towards next-generation multi-shot long-form cinematic audio-video generation. *arXiv preprint arXiv:2606.09639*, 2026.

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.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. Dreamid-omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

Jiehui Huang, Yuechen Zhang, Bin Xia, Jiahao Wang, Xu He, Zhenchao Tang, Meng Chu, and Xin Tao. Unityshots: Memory-driven multi-shot audio-video generation with boundary-aware gating. *arXiv preprint arXiv:2606.21661*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Longbin Ji, Xiaoxiong Liu, Junyuan Shang, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. Videoar: Autoregressive video generation via next-frame and scale prediction. *arXiv preprint arXiv:2601.05966*, 2026.

Joy Future Academy (JD) and collaborators. Joyai-echo: Pushing the frontier of long audio-visual generation. `https://github.com/jd-opensource/JoyAI-Echo`, 2026. Technical report. GitHub: jd-opensource/JoyAI-Echo.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, and Jianwei Zhang. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Xiangyang Luo, Qingyu Li, Xiaokun Liu, Wenyu Qin, Miao Yang, Meng Wang, Pengfei Wan, Di Zhang, Kun Gai, and Shao-Lun Huang. Filmweaver: Weaving consistent multi-shot videos with cache-guided autoregressive diffusion. *arXiv preprint arXiv:2512.11274*, 2025.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

Chika Maduabuchi, Hao Chen, Yujin Han, and Jindong Wang. Corruption-aware training of latent video diffusion models for robust text-to-video generation. *arXiv preprint arXiv:2505.21545*, 2025.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, and Yanhong Zeng. Holocine: Holistic generation of cinematic multi-shot long video narratives. *arXiv preprint arXiv:2510.20822*, 2025.

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Li, 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.

Jangho Park, Geon Yeong Park, Gihyun Kwon, and Jong Chul Ye. Flowlong: Inference-time long video generation via manifold-constrained tweedie matching. *arXiv preprint arXiv:2605.20910*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, and Jianxiao Yang. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jing Wang, Fengzhuo Zhang, Xiaoli Li, Vincent YF Tan, Tianyu Pang, Chao Du, Aixin Sun, and Zhuoran Yang. Error analyses of auto-regressive video diffusion models: A unified framework. *arXiv preprint arXiv:2503.10704*, 2025.

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.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, and Peiyu Wang. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Xuan Wei, Longbin Ji, Guan Wang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Qingqi Hong. Memento: Reconstruct to remember for consistent long video generation. *arXiv preprint arXiv:2606.14667*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, and Guanyu Feng. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Zhen Ye, Xu Tan, Aoxiong Yin, Hongzhan Lin, Guangyan Zhang, Peiwen Sun, Yiming Li, Chi-Min Chan, Wei Ye, Shikun Zhang, and Wei Xue. Talker-t2av: Joint talking audio-video generation with autoregressive diffusion modeling. *arXiv preprint arXiv:2604.23586*, 2026.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory. *arXiv preprint arXiv:2512.19539*, 2025.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, and Ming-Yu Liu. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-cong Chen. Videomemory: Toward consistent video generation via memory integration. *arXiv preprint arXiv:2601.03655*, 2026.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems, arXiv preprint arXiv:2405.01434*, 2024.

Kai Zou, Dian Zheng, Hongbo Liu, Tiankai Hang, Bin Liu, and Nenghai Yu. Hiar: Efficient autoregressive long video generation via hierarchical denoising. *arXiv preprint arXiv:2603.08703*, 2026.

A BACKBONE ARCHITECTURE

The LTX-2 backbone is a diffusion transformer (not a U-Net). A causal video VAE compresses video by 32×32 spatially and $8\times$ temporally into 128-channel latents; a latent patchifier maps one latent voxel to one token, projected to a hidden size of 4096 for video and 2048 for audio. The denoiser is a stack of transformer blocks; each block applies video self-attention, text cross-attention, and bidirectional audio↔video cross-attention, all AdaLN-modulated by the flow timestep, with RMSNorm QK-normalization and 3D (video) / 1D (audio) axial RoPE, using full (bidirectional) attention. Text is injected through cross-attention via a learned-register connector; conditioning frames, reference images, and history enter *in context* as clean tokens ($\sigma = 0$) marked by a conditioning mask; the role of each conditioning item is added through a learned task embedding. Training uses rectified-flow velocity prediction with the interpolation $x_\sigma = (1 - \sigma)x_0 + \sigma\epsilon$ and target $\epsilon - x_0$. We use the 22B task-embedding checkpoint.

B σ-AWARE ERROR BUFFER

The error buffer is a GPU-resident ring buffer of the model's target-token residuals $\delta = \hat{x}_0 - x_0$, collected where the video loss mask is true. Each stored token records both its residual vector and the flow-matching level σ at which it was produced.

Push. After each forward pass, residuals from the current batch are selected by a `mixed_top_random` rule (half from the largest-norm tokens, half random) keeping a fraction (25%) of the batch, then written to the ring buffer together with their σ. Optional per-token L2-norm clipping is available but disabled by default.

σ-matched sampling. When injecting into a history at target level σ, the buffer restricts candidates to stored tokens with $|\sigma' - \sigma| \leq \tau_\sigma$; if none qualify it falls back to the single nearest-σ token. Residuals are then drawn uniformly (with replacement) from the candidate set and injected as $\tilde{c}_v = c_v + \gamma\,\delta_s$ into the continuation-history tokens (τ_{hist}) only, excluding reference images, the clean sink, and the target.

Defaults. Buffer size $1{,}048{,}576$; minimum fill $16{,}384$; warmup 100 steps; injection probability 1.0; σ-match tolerance $\tau_\sigma = 0.05$; per-step strength $\gamma \sim \mathcal{U}(0.9, 1.2)$; injection restricted to τ_{hist}.

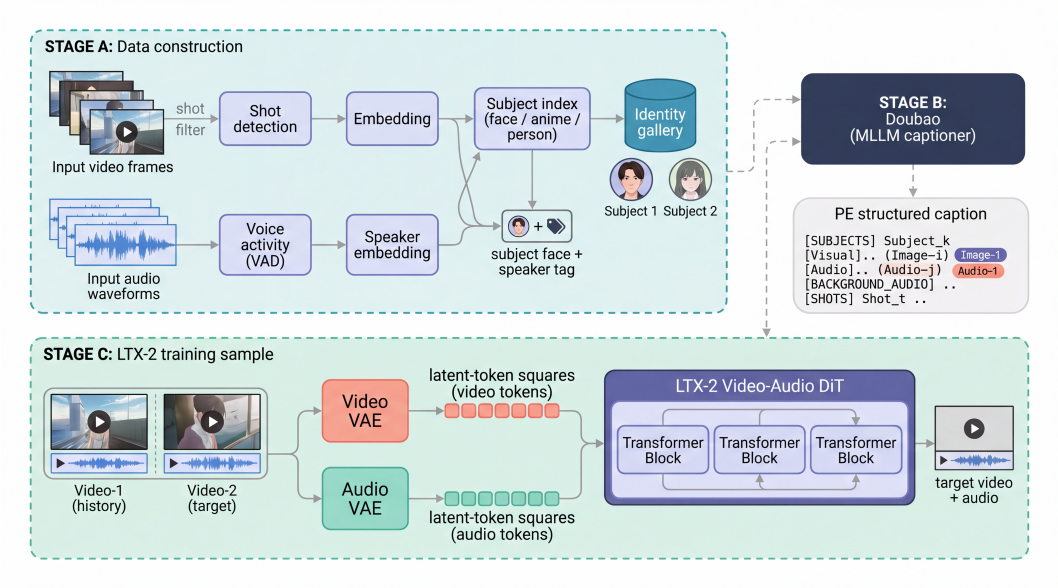

Figure 4: **Data-construction pipeline.** Same-source videos are filtered and indexed by subject; a multimodal captioner produces video, audio, and joint audio-visual captions; each source is split into history and target pairs and rendered into the four-task mixture.

C Data Construction

D Task Layout and Prompt Template

Conditioning items each carry a task embedding; the identifiers are listed in Table 1 of the main paper and a position type. Subject reference images (τ_{ref}) use a global (non-temporal) position. In the continuation-with-clean-sink task, the conditioning is ordered as [main continuation history video/audio, τ_{hist}] then [clean sink clip of length Δ, τ_{sink}, appended]; the sink reads only the first $\lceil \text{fps} \cdot \Delta \rceil$ frames of the true history, padded so the frame count is $\equiv 1 \pmod 8$ for the causal VAE. The residual injection of Appendix B is applied to the τ_{hist} history stream, never to the τ_{sink} sink. Target audio-visual tokens (τ_{tgt}) are noised and supervised; all conditioning tokens are inserted at $\sigma = 0$.

Heterogeneous conditions are serialized into a structured VLM prompt of the form `[Task] / [Conditions] / [User Instruction] / [Output]`, with a per-task variant that names each visual stream (e.g. "Video-1 is the main conditioning video; Video-2 is a clean short reference clip") and lists the reference-image descriptions. To place clips of different frame rates on a common time axis, the 3D RoPE positions of every conditioning and target token are assigned by true frame rate.

E Inference Details

At test time we generate segment by segment. The first segment is produced from the subject references $\{s_i\}$ and prompt p_1. For segment $t \geq 2$ we condition on the references, the clean sink r (length Δ) taken from the start of the running context, the previous segment's generated video and audio as the history $(c_v^{(t)}, c_a^{(t)})$, and prompt p_t, each tagged by its task embedding as in training. Starting from Gaussian noise on the target tokens we integrate the rectified-flow ODE

$$z_{\sigma-\Delta\sigma} = z_\sigma - v_\theta(z_\sigma, p_t)\,\Delta\sigma \tag{9}$$

over a decreasing σ schedule (conditioning tokens held at σ=0), decode the target latents, re-encode them as the history of segment t+1, and repeat with a camera cut. Crucially, *no* residual injection

Configuration	Cross-Shot Consistency				Video Quality		Text	Speech	
	ViCLIP↑	Self-CIDS↑	ARC↑	Reappear↑	Aesthetic↑	Imaging↑	CLIP-T↑	Voice↑	Acc↑
(a) Noise strategy: replace the injected σ-aware residual, keeping the clean sink									
w/ Gaussian context noise	0.5980	0.4733	0.4597	0.5642	0.6283	0.6438	0.2195	0.7115	0.7831
(b) Sink token: remove the clean conditioning sink of length Δ									
w/o clean sink	0.6378	0.5609	0.4891	0.5739	0.6288	0.6552	0.2224	0.7395	0.7954
w/ clean sink + σ-aware residual (ours)	**0.6869**	**0.6032**	**0.5434**	**0.6049**	**0.6416**	**0.6766**	**0.2468**	**0.7533**	**0.8433**

Table 5: **Ablation studies** on the long audio-visual benchmark. Removing either the clean sink or σ-aware residual lowers every metric; replacing model-shaped residuals with Gaussian context noise is the larger degradation. Best per column in **bold**.

is applied at inference: the self-generated history already carries the imperfection that the σ-aware training taught the model to correct.

F Comparison and Ablation Variants

All variants share the identical LTX-2 backbone, data mixture, and inference pipeline, differing only in how the conditioning history is treated.

Clean history. Standard teacher forcing with the clean history latent c_v; no perturbation.

Gaussian context noise. A structure-agnostic control mixing the history latent with isotropic noise,

$$c_v' = (1-\sigma_g)\, c_v + \sigma_g\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad \sigma_g \sim \mathcal{U}(0.25, 0.40), \tag{10}$$

testing whether robustness comes from *any* perturbation or specifically from model-shaped error.

σ-agnostic residual injection. The Matrix-Game-style buffer of Wang et al. (2026): residuals are pooled and sampled without regard to their noise level and injected at a fixed strength, $\tilde{c}_v = c_v + \gamma\, \delta_s$, $\delta_s \sim \mathrm{Buffer}$. This is the design our σ-matched sampling and strength control depart from.

Common settings. All variants use the 22B task-embedding LTX-2 backbone, a Gemma-3 text encoder (max length 4096), 480p, 24 fps, full fine-tuning at learning rate $1\mathrm{e}{-5}$.

Ablation results. The clean sink and σ-aware residual are complementary. Replacing the residual with Gaussian context noise is the larger degradation, while removing the sink also lowers every metric.

G Evaluation Protocol

Published protocols differ across methods, so we re-score every method with one evaluator. Cross-shot metrics are ViCLIP, Self-CIDS, ARC, and Reappear; video quality is Aesthetic and Imaging; text consistency is CLIP-T; speech uses speaker verification and character accuracy $\mathrm{Acc} = \max(0, 1 - \mathrm{CER})$. Self-CIDS and the two long-range metrics use person crops and combined body/face embeddings; ARC compares later appearances with the first reliable anchor, while Reappear restricts comparisons to non-adjacent shots. Mandarin speech is scored with character CER after simplifying traditional characters, and pairwise scores are averaged per sample before averaging across samples. The long benchmark adds 3D-Speaker voice consistency and Whisper speech accuracy.

H PROMPT CONSTRUCTION.

Because a single model must serve continuation, subject-reference, and audio-visual tasks, we do not feed free-form captions. Instead we construct every prompt with a fixed *prompt template* that factorizes a clip into slots and binds each recurring entity to its reference images and reference audio by explicit identifiers (`Image-`i, `Audio-`j). A multimodal captioner fills the slots, so identity, voice, background sound, and per-shot action are described in a uniform, machine-parseable structure rather than as prose.

Prompt template

```
[SUBJECTS]
Subject_k: [Visual] {appearance} (Reference images: Image-a, ...)
[Audio] {voice/timbre} (Reference audio: Audio-b)
[BACKGROUND_AUDIO]
{ambient / music description}
[SHOTS]
Shot_t: {camera, action, subtitles, [speech], <BGM>, «SFX»}
```

The filled prompt template is embedded into an outer *instruction wrapper* that names the role of each in-context stream and issues the generation instruction, so the same interface covers continuation, subject-reference, and audio-visual tasks.

Instruction wrapper

```
[Task]
Video continuation with a clean sink and subject references.
[Conditions]
Video-1: main conditioning history (video+audio).
Video-2: clean short reference clip from Video-1.
{N subject reference images}; prompt template above.
[Instruction] {filled prompt template}
[Output] Generate the next segment, faithful to the conditions.
```

I AI USE STATEMENT

In this work, we used generative AI tools solely to improve the readability, grammar, and overall presentation of the manuscript. The AI tools were not used to generate research ideas, develop the methodology, design or conduct experiments, analyze or interpret experimental results, or produce any scientific claims. All AI-assisted text was carefully reviewed and revised by the authors. The authors take full responsibility for the final content of this work, including all scientific claims, methodology, experimental results, and conclusions.