FORK-NFT: FORKED-CREDIT NEGATIVE-AWARE FINE-TUNING FOR AUTOREGRESSIVE JOINT AUDIO–VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive joint audio–video generation poses a reinforcement learning problem beyond optimizing complete clips. Each continuation inherits an autoregressive history, while its audio and video streams are tightly coupled: an update may improve one modality while degrading the other or disrupting their synchronization. Consequently, clip-level rewards do not identify which chunks under a shared history should receive video or audio credit. We introduce Fork-NFT, a local reinforcement learning update that branches multiple joint audio–video continuations from a shared prefix and compares them under the same generated history. Modality-specific relative advantages assign video-quality credit to the full continuation after the retained prefix and audio-quality credit to short windows after each fork, while both streams remain jointly generated. To align this local credit with few-step generation, Fork-NFT trains on paired intermediate states recorded from the sampler while re-materializing the causal KV context from each generated prefix under the current policy. We further constrain how cross-modal attention reads the committed history while leaving attention within the active chunk free to adapt. On a 22B eight-step autoregressive model, Fork-NFT improves quality and audio–video consistency over matched baselines on both short- and long-form evaluations, while retaining the same sampler and backbone architecture.

1 INTRODUCTION

Diffusion models have driven rapid progress in visual generation, and recent joint audio–video models extend these capabilities to synthesize sound and vision within a unified denoising process (Liu et al., 2026a; HaCohen et al., 2026; Liu et al., 2026b). By modeling cross-modal interactions, these systems generate clips in which visual events and their sounds are produced together. Most existing systems, however, still denoise a bounded temporal clip with bidirectional attention over the complete audio–video sequence. This full-clip formulation is effective for offline synthesis, but does not natively support causal continuation over long horizons. In video generation, chunk-autoregressive models such as CausVid, Self Forcing, Causal Forcing, and LongLive address this limitation by generating short clips sequentially while conditioning on previously generated context (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026; Yang et al., 2026). OmniForcing extends this paradigm to joint audio–video generation by converting a bidirectional LTX-2 model into a dual-stream autoregressive generator (Su et al., 2026). This chunk-autoregressive formulation enables incremental audio–video generation, but also changes the post-training problem: every generated chunk is both an output to be optimized and part of the context on which future chunks depend.

Reward-based fine-tuning has become an effective way to improve image and video generation. Recent methods adapt policy optimization to diffusion and flow models and improve perceptual quality and semantic alignment (Xue et al., 2025; Liu et al., 2025a; Li et al., 2025), while likelihood-free objectives provide a complementary route when reverse-process likelihoods are unavailable (Zheng et al., 2026; Choi et al., 2026). OmniNFT brings reward-based post-training to complete, bidirectionally generated audio–video clips through modality-specific rewards and optimization controls (Zhang et al., 2026a). Recent autoregressive video methods compare continuations from a shared prefix and localize reward updates to the generated future (Zhang et al., 2026b; He et al., 2026).

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

Together, these advances motivate reward-based post-training for autoregressive joint audio–video generation, where both modalities are generated under a shared causal history.

However, reward-based fine-tuning does not transfer directly to a few-step autoregressive joint audio–video generator. First, the sampler denoises paired audio and video states under its own generated history; independently re-noising a finished sample yields states it never visited under that history. Second, the rewards require different credit horizons. Video-quality scores persist across neighboring segments, so short late forks have little leverage over quality shaped by the prefix. Audio quality varies more locally, yet one chunk contains too little sound to score reliably. A single fork position and credit window cannot serve both modalities equally well. Third, reward updates can change how cross-modal attention reads the committed history, which conditions every later chunk. A constraint limited to the active chunk cannot detect this change.

We introduce Fork-NFT, a forked NFT update that samples multiple joint continuations from the same generated audio–video prefix and assigns their relative rewards to future chunks. It reuses paired sampler states while reconstructing causal caches from the generated prefix under the current policy and inference-time commit-and-evict rules. From each shared prefix, we sample sibling joint continuations. The first fork follows the permanently retained initial chunks (the sink boundary): it assigns video credit to the full future and audio credit to its first few chunks; later forks assign audio credit to short multi-chunk windows. The streams remain coupled during generation, while each reward supervises its modality's token loss. A history-marginal attention regularizer preserves how the reference model reads committed context without directly penalizing within-chunk attention patterns, allowing local audio–video coordination to adapt during reward fine-tuning.

Our contributions are summarized as follows:

(1) We systematically study reward fine-tuning of few-step autoregressive audio–video generators, identifying mismatches between re-noised training inputs and actual sampler states, different temporal requirements for video and audio credit, and shifts in cross-modal attention to generated history.

(2) We propose Fork-NFT, a unified NFT framework that replays paired sampler states with current-policy context, assigns long-horizon video and short-window audio credit through shared-prefix forks, and regularizes cross-modal attention to committed history.

(3) Experiments on JavisBench and long-horizon generation demonstrate improved audio–video quality, consistency, and synchronization over matched autoregressive baselines, while ablations validate the contribution of each design.

2 Related Work

Joint audio–video generation. Early joint diffusion models such as MM-Diffusion couple modality-specific denoisers to generate aligned audio and video within a shared process (Ruan et al., 2023). Recent systems substantially improve generation quality through stronger cross-modal architectures. The Javis family develops unified Transformers with explicit spatio-temporal synchronization, with JavisDiT++ further introducing audio–video preference optimization for perceptual quality, cross-modal consistency, and synchronization (Liu et al., 2026a;b); UniVerse-1 connects specialized audio and video experts through lightweight cross-modal modules (Wang et al., 2025); and LTX-2 adopts an asymmetric dual-stream Transformer coupled by bidirectional cross-attention (HaCohen et al., 2026). Despite their architectural differences, these models generate bounded clips with access to the complete audio–video sequence. OmniForcing extends joint generation to the causal setting by distilling LTX-2 into a dual-stream chunk-autoregressive generator with block-causal attention and self-forced rollouts (Su et al., 2026). We study reward post-training of such generators, where each chunk depends on the past and becomes context for future audio–video generation.

Reinforcement Learning for Visual Generation. Inspired by the success of Group Relative Policy Optimization (GRPO) in improving the reasoning capabilities of large language models, recent studies have extended group-relative reinforcement learning to visual generation (Shao et al., 2024; DeepSeek-AI, 2025). T2I-R1 applies reinforcement learning to autoregressive image generation by jointly optimizing semantic-level reasoning and visual-token generation (Jiang et al., 2025). For diffusion and flow-based generators, Flow-GRPO and DanceGRPO enable group-relative policy

optimization by introducing stochastic sampling trajectories into otherwise deterministic generation processes (Liu et al., 2025a; Xue et al., 2025). DiffusionNFT instead performs reference-relative policy improvement through a forward-process matching objective, avoiding the need to evaluate reverse-transition likelihoods (Zheng et al., 2026); reward-based velocity matching provides a related trajectory-independent formulation (Choi et al., 2026). Reinforcement learning has only recently been extended to joint audio–video generation. OmniNFT introduces modality-specific advantages and regulates optimization interference between the audio and video branches (Zhang et al., 2026a). We build on the forward-process NFT objective and modality-specific supervision, but study them in a chunk-autoregressive joint generator, where reward credit must additionally account for the generated causal history and be localized to the responsible continuation.

Reinforcement Learning for Autoregressive Video Generation. CausVid, Self Forcing, Causal Forcing, and LongLive establish chunk-autoregressive video generation with causal histories and few-step sampling; MAGI-1 scales chunk-wise prediction (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026; Yang et al., 2026; Sand.ai et al., 2025). Reward Forcing biases distillation toward high-reward samples (Lu et al., 2026). Astrolabe applies clip-local NFT under a shared KV prefix, while AR-CoPO forks continuations at a pivot and updates that chunk using sequence-level rewards (Zhang et al., 2026b; He et al., 2026). Fork-NFT addresses joint audio–video generation by assigning modality-specific temporal credit and training on states recorded from the deployed sampler.

3 Preliminaries

3.1 Chunk-Autoregressive Joint Audio–Video Flow Matching

Let c be a text condition and $y_n = (y_n^v, y_n^a)$ the synchronized video and audio latents in chunk n. A chunk-autoregressive joint generator factorizes the sequence as

$$p_\theta(y_{0:N-1} \mid c) = \prod_{n=0}^{N-1} p_\theta(y_n \mid h_n, c), \tag{1}$$

where $h_n = y_{0:n-1}$ denotes the previously generated joint audio–video history, summarized by a causal KV cache. For modality $m \in \{v, a\}$, the linear flow-matching path (Lipman et al., 2023) couples the clean target y_n^m with Gaussian noise ϵ_n^m under a denoising time $\sigma \in [0, 1]$:

$$x_{n,\sigma}^m = (1-\sigma)y_n^m + \sigma\epsilon_n^m, \qquad u_n^m = \epsilon_n^m - y_n^m. \tag{2}$$

A joint predictor $f_\theta = (f_\theta^v, f_\theta^a)$ estimates both velocity targets from the noisy audio–video chunk, text condition, and causal history. The two streams exchange information within the active chunk, while cross-chunk dependence is mediated only through h_n. We use n for sequence time and σ for denoising time throughout.

3.2 Negative-Aware Fine-Tuning

Diffusion Negative-Aware Fine-Tuning (DiffusionNFT) encodes reward feedback through implicit positive and negative predictors (Zheng et al., 2026). Although the network predicts a velocity f_θ^m, the NFT loss in our implementation compares clean-latent predictions. Let ξ contain the paired noisy latents, text condition, and causal history. For each modality $m \in \{v, a\}$, write $\hat{y}_\theta^m(\xi) = x_\sigma^m - \sigma f_\theta^m(\xi)$ and define $\hat{y}_{\text{old}}^m$ analogously. The two implicit predictions and the advantage-dependent mixing weight are

$$\hat{y}_\theta^{m,\pm} = \hat{y}_{\text{old}}^m \pm \rho(\hat{y}_\theta^m - \hat{y}_{\text{old}}^m), \qquad p(A) = \tfrac{1}{2}\big[1 + \text{clip}(A/c_A, -1, 1)\big]. \tag{3}$$

Here ρ controls the implicit-policy interpolation and c_A is the advantage-clipping scale. With clean target y^m, the per-modality objective is

$$\ell_{\text{NFT}}^m(\theta; \xi, y^m, A) = p(A)D_m(\hat{y}_\theta^{m,+}, y^m) + (1-p(A))D_m(\hat{y}_\theta^{m,-}, y^m). \tag{4}$$

Here $D_m(z, y) = \text{mean}[(z-y)^2]/\,\text{mean}\,|z-y|$ is normalized per sample for each implicit prediction. Both implicit predictions share one rollout target; their symmetry around the old predictor lets the advantage continuously interpolate between positive and negative feedback. The original NFT construction draws x_σ by forward-noising a generated endpoint as in Eq. (2). We retain the clean-prediction loss but evaluate it on paired states recorded from the few-step autoregressive sampler, as described in Section 4.3.

4 FORK-NFT

Fork-NFT compares joint continuations from a shared autoregressive state. It combines sampler-state replay, current-policy cache re-materialization, modality-specific temporal credit, and history-marginal attention regularization. We first describe the backbone.

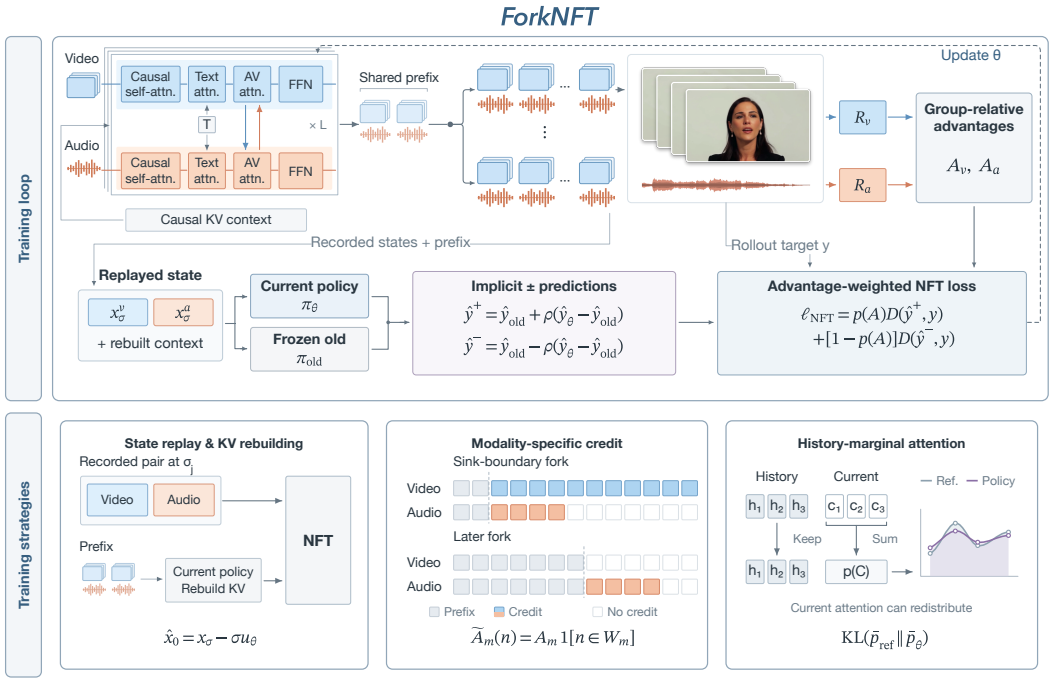


Figure 1: Overview of Fork-NFT. Sibling joint audio–video continuations share a generated prefix. Video-quality credit spans the full continuation after the sink-boundary fork, whereas audio-quality credit covers a short window after either a sink-boundary or later fork. Both updates use recorded paired sampler states and causal caches re-materialized under the current policy; a history-marginal attention loss preserves cross-modal reading of the committed prefix.

4.1 AUTOREGRESSIVE AUDIO–VIDEO BACKBONE AND EXTRAPOLATION

We adapt the bidirectional joint audio–video generator LTX-2.3 (HaCohen et al., 2026) to chunk-autoregressive generation. Each synchronized chunk contains two video latent frames; audio and video are denoised jointly within the chunk, while cross-chunk attention sees only past chunks. During autoregressive adaptation, training mixes clean histories with previously generated chunks and applies the flow-matching loss only to the noisy target chunk. We merge the official eight-step distillation adapter (eight network evaluations per chunk) to obtain the few-step backbone for all rollouts and the matched reward-fine-tuning baselines evaluated below.

For extrapolation, the first two chunks are retained as attention sinks alongside the four most recent chunks. Separate KV caches for video and audio self-attention and both cross-modal directions use modality-specific token quotas. Figure 2(a) illustrates the retained context and its causal attention pattern. Each chunk is denoised for eight steps against this history. We discard KV entries from intermediate denoising steps; a final clean forward pass commits the completed joint chunk, followed by eviction. This keeps noisy states out of future context and bounds the per-chunk cost as generation continues through long sequences of jointly generated audio–video chunks at inference time.

4.2 EMPIRICAL MOTIVATION

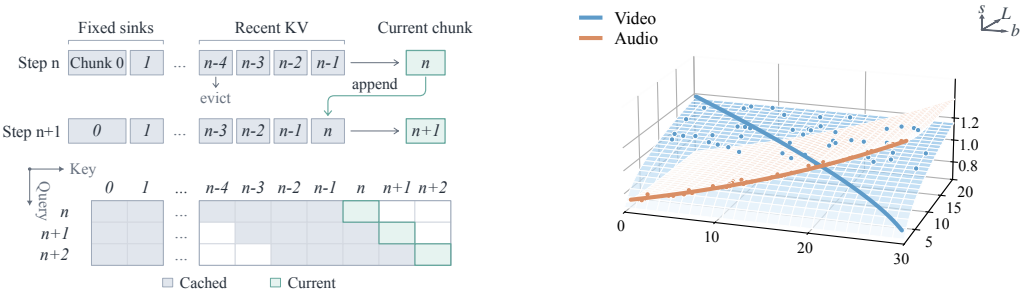
Sampler states. Each denoising input in a few-step autoregressive generator comes from paired audio and video trajectories under a generated prefix. Independently re-noising their endpoints lowers

audio quality and cross-modal correspondence (Table 4). Reward learning should instead use visited paired states and reconstruct causal context from that prefix.

Reward credit. Variation across a rollout pool need not yield informative comparisons among siblings sharing a prefix. Table 1a shows that only 1.9% of VideoAlign video-quality (VQ) variance lies within sibling groups, versus 41.3% for AudioBox perceptual quality (PQ). Forked advantages use these within-group differences; between-group variation also reflects prompts, prefixes, and fork positions. The scored continuation thus matters, especially for video.

Figure 2b fits within-sibling reward dispersion s to fork index b and scored duration L as $s \propto L^{\alpha} e^{\beta b}$. Video position and duration vary together, so its length-only fit ($\alpha_v = 0.177$) describes the observed trajectory, not an independent position effect. Audio varies across both dimensions: $(\alpha_a, \beta_a) = (-0.034, 0.0246)$ indicates little dependence on window length but greater dispersion at later forks. This motivates full-future video credit from the sink boundary and short-window audio credit at multiple positions; the shared sink prefix receives no reward credit.

Committed history. On identical inputs and committed history, we compare policy and reference post-RoPE attention. Historical keys receive most attention in both cross-modal directions. After pooling current keys, the history marginal accounts for 40.6% and 55.4% of full attention KL, respectively (Table 1b). History attention mass falls in 25 of 30 audio-to-video chunks and all 30 video-to-audio chunks. Current-only attention penalties miss these changes; we therefore regularize the history marginal while allowing within-chunk attention to adapt.



(a) Autoregressive extrapolation

(b) Fork-dependent reward variation

Figure 2: Autoregressive extrapolation and fork geometry. **(a)** Two sink chunks and four recent chunks condition generation; the matrix shows causal query–key visibility. **(b)** Fitted reward-dispersion surfaces with surface-sampling markers. The axis triad denotes fork chunk index b, scored continuation length L (seconds), and within-sibling reward standard deviation s, normalized by each surface's mean.

Table 1: Diagnostics motivating temporal credit and history regularization. **(a)** Reward-variance decomposition over 508 shared-prefix sibling groups. **(b)** Diagnostic RL checkpoint versus frozen backbone on a shared cache; history KL share is summed history-marginal KL divided by summed full attention KL.

(a) Reward variance decomposition

Scorer	Between groups	Within group
VideoAlign-VQ	98.1%	**1.9%**
AudioBox-PQ	58.7%	**41.3%**

(b) Attention to committed history

Cross-modal direction	$a \to v$	$v \to a$
History attention mass	81.7%	73.9%
History KL share	40.6%	55.4%

4.3 Joint State Replay and Context Re-Materialization

NFT is commonly evaluated on states obtained by forward-noising completed samples. A few-step autoregressive generator instead denoises video and audio together under a causal history assembled from its earlier chunks. We train on the paired states encountered during this generation process.

For sibling continuation g, we record the two input latents before each denoising step. At chunk n and noise level σ, the NFT input and clean-latent prediction are

$$\begin{aligned} \xi^{g}_{n,\sigma} &= \big(x^{v,g}_{n,\sigma}, x^{a,g}_{n,\sigma}, \mathcal{C}^{g}_{<n}(\theta), c\big), \\ \hat{y}^{m,g}_{n,\sigma} &= x^{m,g}_{n,\sigma} - \sigma f^{m}_{\theta}(\xi^{g}_{n,\sigma}), \quad m \in \{v, a\}. \end{aligned} \tag{5}$$

Both latents come from the same sampler step and share its noise level. Equation (4) compares $\hat{y}^{m,g}_{n,\sigma}$ with the clean endpoint $y^{m,g}_{n}$ of that continuation.

Rollouts store completed chunks and paired denoising states, not KV tensors. During training, a no-gradient current-policy pass at $\sigma = 0$ re-materializes the four prefix caches under the same commit-and-evict rules as inference. The trainable, behavior-weight, and reference-weight predictors use this shared, detached cache. The latter two have frozen weights during each update, but their predictions remain conditioned on the current-policy cache.

4.4 Forked Temporal Credit Assignment

A full-rollout reward mixes contributions from many chunks. We instead compare sibling joint continuations from one prefix and assign their reward differences to modality-specific windows.

Let N be the number of chunks and n_{sink} the first chunk after the fixed sink prefix. We fork at the pivot $b = n_{\text{sink}}$ and at selected later chunks. Siblings share chunks before b and independently generate the joint audio–video continuation from b onward. The sink-boundary fork assigns video credit to the full continuation and audio credit to its first K_a chunks, where K_a is the audio credit-window length. Later forks use the same audio window. The credited windows are

$$\begin{aligned} \mathcal{W}_v(n_{\text{sink}}) &= \{n_{\text{sink}}, \dots, N-1\}, \\ \mathcal{W}_v(b) &= \varnothing, \quad b > n_{\text{sink}}, \\ \mathcal{W}_a(b) &= \{b, \dots, b + K_a - 1\}. \end{aligned} \tag{6}$$

The shared sink chunks provide causal context but receive no reward credit.

On each credited video or audio span, we combine scorer outputs standardized over the collected pool to obtain $R^{g}_{m,b}$ for modality m. Among siblings with the same prompt and fork state, its advantage and chunk-level assignment are

$$\begin{aligned} A^{g}_{m,b} &= \frac{R^{g}_{m,b} - \bar{R}_{m,b}}{s_{m,b} + \epsilon}, \\ \widetilde{A}^{g}_{n,m} &= \mathbf{1}\{n \in \mathcal{W}_m(b)\} A^{g}_{m,b}, \quad m \in \{v, a\}. \end{aligned} \tag{7}$$

Here $\bar{R}_{m,b}$ and $s_{m,b}$ are the sibling-group mean and standard deviation. Chunks with zero advantage for both modalities only advance the causal cache. For a chunk credited to one modality, both NFT terms are evaluated: the other modality has $p(0) = 1/2$ in Eq. (4), yielding a symmetric term centered on the behavior prediction, together with the reward-independent regularizers. We optimize one adapter over the two fork types.

4.5 History-Marginal Cross-Modal Attention Regularization

Cross-modal attention both reads committed history and coordinates current-chunk tokens. A penalty normalized only over current keys misses shifts between these regions. We preserve the reference model's history use while allowing attention within the active chunk to adapt.

For a query q in either cross-modal direction, let H_q contain the visible keys from committed chunks and C_q the visible keys from the active chunk. After rotary positional embeddings (RoPE), we compute the attention distribution $p_\theta(\cdot \mid q)$ over their joint, causally masked support. Its history marginal retains each historical key but pools the current keys into one entry:

$$\bar{p}_\theta(\cdot \mid q) = \left(\{p_\theta(j \mid q)\}_{j \in H_q}, \sum_{j \in C_q} p_\theta(j \mid q) \right). \tag{8}$$

We form the reference distribution with frozen weights using the same paired input state, shared current-policy cache, and visible-key mask, and average the forward KL over cross-modal directions, layers, heads, and current-chunk queries:

$$\mathcal{L}_{\text{hist}} = \mathbb{E}_{d,\ell,h,q}[D_{\text{KL}}(\text{sg}[\bar{p}_{\text{ref}}(\cdot \mid q)] \parallel \bar{p}_\theta(\cdot \mid q))] \,. \tag{9}$$

Here sg stops gradients through the reference distribution. The objective constrains both the total attention assigned to history and its distribution among historical keys. Indeed, for each query the chain rule gives

$$\begin{aligned} D_{\text{KL}}(p_{\text{ref}} \| p_\theta) = {} & D_{\text{KL}}(\bar{p}_{\text{ref}} \| \bar{p}_\theta) \\ & + P_{\text{ref}}(C_q) D_{\text{KL}}(p_{\text{ref}}(\cdot \mid C_q) \| p_\theta(\cdot \mid C_q)) \,. \end{aligned} \tag{10}$$

Our loss retains the first term and leaves redistribution among current-chunk keys unpenalized. When no history is visible, the marginalized distributions contain only the current-key entry and this term vanishes. Let $\mathcal{P}_v$ and $\mathcal{P}_a$ aggregate the modality-wise NFT losses over credited chunks. The auxiliary term $\mathcal{R}$ contains a squared-error penalty between current and frozen-reference clean-latent predictions. The joint update minimizes

$$\mathcal{L} = \mathcal{P}_v + \mathcal{P}_a + \mathcal{R} + \lambda_{\text{hist}} \mathcal{L}_{\text{hist}}. \tag{11}$$

5 Experiments

We evaluate Fork-NFT along three dimensions: short-form audio–video quality on JavisBench, long-horizon generation, and the core design choices behind the local NFT update.

5.1 Experimental Setup

Autoregressive backbone and optimization. We use LTX-AR, the chunk-autoregressively trained and distilled backbone described in Section 4.1, with bounded-cache extrapolation at inference. All reinforcement-learning variants start from the same 22B checkpoint and use the same eight-NFE sampler. We freeze the backbone and optimize a single low-rank adaptation (LoRA) adapter with the single-stage forked-credit NFT update of Section 4.4. Sink-boundary forks score video quality over the full continuation and audio quality over the first four chunks; later forks score audio quality over a four-chunk window. Both geometries are mixed during training, with paired sampler-state replay, current-policy cache re-materialization, and history-marginal cross-modal attention regularization.

Reward models. The video reward combines HPSv3 with the visual-quality and motion components of VideoAlign, while the audio reward combines AudioBox-PQ and CLAP (Ma et al., 2025; Liu et al., 2025b; Tjandra et al., 2025; Wu et al., 2022). Each scorer is standardized over the collected rollout pool before composition, and the composite video and audio rewards are normalized separately within each group of prefix-shared sibling continuations. This produces the modality-specific advantages defined in Section 4.4. Reward weights are fixed across runs.

Evaluation protocols. Our primary evaluation uses the full JavisBench benchmark and its official evaluator (Liu et al., 2026a;b). We report video and audio quality (VQ/AQ), text and audio–video consistency, synchronization, and the aggregate JavisScore; the ablations additionally use AV-Align. We also evaluate 20-second outputs on matched prompts (Table 3). VQ, AQ, DeSync, and JavisScore cover the full long-form output; other metrics follow the evaluator's sampling protocol. All systems within each comparison use matched prompts and generation settings.

Comparisons. Table 2 compares audio-first and video-first cascades (Yariv et al., 2023; Jeong et al., 2023; 2024; Xing et al., 2024; Zhang et al., 2024; Cheng et al., 2025), unified bidirectional models (Liu et al., 2026a;b; Wang et al., 2025; HaCohen et al., 2026), and autoregressive systems. The primary controlled comparison is the autoregressive block: LTX-AR, an OmniNFT adaptation (Zhang et al., 2026a) reimplemented on LTX-AR, and Fork-NFT share the same starting checkpoint, sampler, output specification, and evaluation pipeline. Published results from other model families provide broader context; the autoregressive block isolates the reward-learning comparison.

5.2 Main Results

Audio–video quality. Table 2 compares reward learning on the same autoregressive backbone. Relative to LTX-AR, Fork-NFT improves video and audio quality, cross-modal consistency, and synchronization. Compared with the OmniNFT adaptation, it attains similar VQ but higher AQ, AV-IB, AVH, and JavisScore at the same inference time. The largest separation is in joint quality: JavisScore rises from 0.192 to 0.205 and AV-IB from 0.225 to 0.241. Text-alignment scores remain close, with a slightly lower CLAP. The gains thus primarily reflect better audio–video quality and correspondence.

Table 2: Main results on the full JavisBench benchmark. Missing entries were not reported. Best results are **bold** and second-best results are underlined. (↑: higher is better; ↓: lower is better).

Model	AV-Quality		Text-Consistency				AV-Consistency		AV-Synchrony		Runtime
	VQ↑	AQ↑	TV-IB↑	TA-IB↑	CLIP↑	CLAP↑	AV-IB↑	AVH↑	DeSync↓	JavisScore↑	
T2A + A2V											
TempoTkn	—	—	0.084	—	0.205	—	0.139	0.122	1.532	0.103	20s
TPoS	—	—	0.201	—	0.229	—	0.124	0.129	1.493	0.095	19s
T2V + V2A											
ReWaS	—	—	—	0.123	—	0.280	0.110	0.104	1.071	0.079	17s
See & Hear	—	—	—	0.129	—	0.263	0.160	0.143	1.099	0.112	25s
FoleyCrafter	—	—	—	0.149	—	0.383	0.193	0.186	0.952	0.151	16s
MMAudio	—	—	—	0.160	—	0.407	0.198	0.182	0.849	0.150	15s
T2AV											
JavisDiT	1.291	4.478	0.263	0.143	0.302	0.391	0.197	0.179	1.039	0.154	30s
UniVerse-1	1.357	4.839	0.272	0.111	0.309	0.245	0.104	0.098	0.929	0.077	13s
JavisDiT++	1.462	5.049	**0.282**	0.164	**0.316**	0.424	0.198	0.184	0.832	0.159	**10s**
LTX-2	2.038	**5.197**	0.272	**0.170**	0.311	0.412	0.232	0.223	0.569	0.192	75.9s
LTX-2.3	2.374	5.141	0.278	0.167	0.312	**0.432**	**0.245**	**0.251**	**0.472**	**0.222**	62.9s
Autoregressive T2AV											
LTX-AR	2.326	4.919	0.275	0.150	0.309	0.352	0.223	0.209	0.597	0.186	12.7s
LTX-AR + OmniNFT	3.101	4.963	0.274	0.159	0.312	0.360	0.225	0.214	0.582	0.192	12.8s
Ours	**3.127**	5.042	0.274	0.161	0.313	0.359	0.241	0.233	0.569	0.205	12.8s
RL Δ vs. LTX-AR	+0.801	+0.123	-0.001	+0.011	+0.004	+0.007	+0.018	+0.024	-0.028	+0.019	—

Long-horizon quality. Table 3 evaluates 20-second outputs from the same models. Fork-NFT leads on the full-output metrics VQ, AQ, JavisScore, and DeSync; relative to the OmniNFT adaptation, AQ increases from 4.9990 to 5.2736 and JavisScore from 0.2782 to 0.2969. AV-IB and AVH show the same ordering. TA-IB and CLAP improve, while TV-IB and CLIP remain close to the strongest matched baseline. The short-form quality and synchronization gains therefore persist as autoregressive generation continues.

Table 3: Long-horizon audio–video results. Higher is better except for DeSync.

Model	AV-Quality		Text-Consistency				AV-Consistency		AV-Synchrony	
	VQ ↑	AQ ↑	TV-IB ↑	TA-IB ↑	CLIP ↑	CLAP ↑	AV-IB ↑	AVH ↑	DeSync ↓	JavisScore ↑
LTX-AR	7.4792	4.9784	**0.4014**	0.1340	0.3015	0.4268	0.3934	0.3483	0.0938	0.2757
LTX-AR + OmniNFT	7.4888	4.9990	0.3967	0.1322	**0.3069**	0.4418	0.3905	0.3537	0.0869	0.2782
Ours	**7.6179**	**5.2736**	0.3997	**0.1368**	0.3051	**0.4476**	**0.4162**	**0.3742**	**0.0806**	**0.2969**

Qualitative behavior. Figure 3 compares four matched prompts spanning live-action dialogue and stylized action. The upper row of each pair comes from the autoregressive backbone; the lower row uses the same backbone after Fork-NFT training. Displaying video frames alongside the corresponding audio waveform makes changes in both streams visible on a shared timeline. In the outdoor conversation and library scenes, the samples show how people and their sound activity evolve over the clip; the cartoon and Minecraft examples probe the same joint update with different visual structure. These examples complement the quality and correspondence metrics in Tables 2 and 3 with concrete audio–video sequences.



Figure 3: Qualitative audio–video examples. Each prompt shows video frames and audio waveforms before (top) and after (bottom) reward fine-tuning. Prompts are abbreviated.

5.3 Ablation Studies

Core method ablations. Table 4 isolates the three designs under a common initialization. In the shared-geometry variant, video follows audio: both use early and late forks with the same short reward window. Independent re-noising falls below the no-RL backbone on VQ, AQ, AV-IB, and JavisScore. Shared fork geometry improves AQ but not VQ or JavisScore; modality-specific windows recover both. Removing history-marginal KL leaves VQ and AQ nearly unchanged but lowers AV-Align from 0.375 to 0.351 and JavisScore from 0.205 to 0.196. These patterns separate sampler-state fidelity, temporal reward credit, and cross-modal history use.

Table 4: Core ablations on full JavisBench. T: state replay; M: modality-specific forks; H: history-marginal attention KL.

	Method design			**JavisBench**				
Variant	T	M	H	VQ↑	AQ↑	AV-IB↑	AV-Align↑	JavisScore↑
LTX-AR (no RL)	×	×	×	2.326	4.919	0.223	0.339	0.186
Shared fork geometry	✓	×	✓	2.310	4.983	0.230	0.365	0.183
w/o sampler-state replay	×	✓	✓	2.294	4.681	0.219	0.340	0.162
w/o history-marginal KL	✓	✓	×	3.119	**5.042**	0.235	0.351	0.196
Fork-NFT (full)	✓	✓	✓	**3.127**	**5.042**	**0.241**	**0.375**	**0.205**

6 Conclusion

We presented Fork-NFT for reward fine-tuning of few-step autoregressive audio–video generators. Paired sampler-state replay and current-policy context reconstruction align NFT updates with the deployed generation process. Shared-prefix forks assign distinct temporal credit to video and audio, while history-marginal attention regularization stabilizes cross-modal use of prior chunks without fixing within-chunk attention. Across short and long outputs, Fork-NFT improves joint quality and synchronization over matched autoregressive baselines. Together, these results show why effective reward learning in this setting must account for both the generated history and the different time scales of audiovisual feedback throughout long autoregressive audio–video sequences.

Reproducibility Statement

Sections 4 and 5 describe the training objective and evaluations. The appendix provides implementation details, supporting derivations, the evaluation protocol, and additional controlled analyses.

Ethics Statement

Improved audiovisual generation can support creative tools, but can also facilitate impersonation and misleading synthetic media. Preference and quality scorers may encode biases and do not measure factuality, consent, or misuse risk. Any release and subsequent human evaluation should account for data and model licenses, participant consent, and the disclosure of generated content.

Use of Large Language Models

We used GPT-6 and Claude Opus 5 to assist with literature review, discussion of the method and its mathematical presentation, interpretation of experimental results, figure design, translation, drafting, and language editing. The reported measurements came from our experiments, not from these models. The authors take responsibility for the technical content, references, claims, and final manuscript.

References

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.15322.

Jaemoo Choi, Wei Guo, Yuchen Zhu, Arash Vahdat, Molei Tao, Julius Berner, and Yongxin Chen. Scaling reinforcement learning for diffusion models via velocity matching. *arXiv preprint arXiv:2608.23664*, 2026. URL https://arxiv.org/abs/2608.23664.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. URL https://arxiv.org/abs/2601.03233.

Dailan He, Guanlin Feng, Xingtong Ge, Yi Zhang, Bingqi Ma, Guanglu Song, Yu Liu, and Hongsheng Li. AR-CoPO: Align autoregressive video generation with contrastive policy optimization. *arXiv preprint arXiv:2603.17461*, 2026. URL https://arxiv.org/abs/2603.17461.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.08009.

Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound: Audio reactive video generation with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2309.04509.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. *arXiv preprint arXiv:2407.05551*, 2024. URL https://arxiv.org/abs/2407.05551.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level cot. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.00703.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. MixGRPO: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025. URL https://arxiv.org/abs/2507.21802.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online rl. In *Advances in Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=oCBKGw5HNf.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://arxiv.org/abs/2501.13918.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2503.23377.

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

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.04678.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_HPSv3_Towards_Wide-Spectrum_Human_Preference_Score_ICCV_2025_paper.html.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 2023. URL https://arxiv.org/abs/2212.09478.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. URL https://arxiv.org/abs/2505.13211.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

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. URL https://arxiv.org/abs/2603.11647.

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

Daquan Wang, Weimin Zuo, Anran Li, Lai Chen, Xianlong Liao, Deying Zhou, Zilong Yin, Xing Dai, Dongmei Jiang, and Guohao Yu. UniVerse-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025. URL https://arxiv.org/abs/2509.06155.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *arXiv preprint arXiv:2211.06687*, 2022. URL https://arxiv.org/abs/2211.06687.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2402.17723.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. URL https://arxiv.org/abs/2505.07818.

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

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. *arXiv preprint arXiv:2309.16429*, 2023. URL https://arxiv.org/abs/2309.16429.

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*, 2025. URL https://arxiv.org/abs/2412.07772.

Guohui Zhang, Xiaoxiao Ma, Jie Huang, Hang Xu, Hu Yu, Siming Fu, Yuming Li, Zeyue Xue, Lin Song, Haoyang Huang, Nan Duan, and Feng Zhao. OmniNFT: Modality-wise omni diffusion reinforcement for joint audio-video generation. *arXiv preprint arXiv:2605.12480*, 2026a. URL https://arxiv.org/abs/2605.12480.

Songchun Zhang, Zeyue Xue, Siming Fu, Jie Huang, Xianghao Kong, Yue Ma, Haoyang Huang, Nan Duan, and Anyi Rao. Astrolabe: Steering forward-process reinforcement learning for distilled autoregressive video models. *arXiv preprint arXiv:2603.17051*, 2026b. URL https://arxiv.org/abs/2603.17051.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. FoleyCrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024. URL https://arxiv.org/abs/2407.01494.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.16117.

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

A Implementation Details

Autoregressive backbone. Our LTX-AR backbone converts LTX-2.3 into a chunk-autoregressive joint audio–video generator. Audio and video share an eight-step denoising schedule within each chunk. Two initial sink chunks and four recent chunks provide causal context; separate self-attention and cross-modal KV caches obey modality-specific token budgets. After denoising, a clean forward pass commits the completed chunk to the caches before eviction. We train one LoRA adapter while keeping the backbone frozen.

Rollout and replay. Each forked group shares a generated prefix and contains eight independently sampled joint continuations. At the start of each round, the current adapter is fixed as the behavior policy for rollout collection. Optimization then takes 25 minibatch updates from this buffer before the behavior policy is refreshed for the next round; the buffer is not regenerated between updates. The reference weights remain frozen. We retain the final clean audio and video latents and the paired input states recorded immediately before each sampler step; the two modalities use the same recorded noise-level index. The rollout does not store KV tensors. During optimization, a no-gradient current-policy clean pass re-materializes the four caches from the recorded prefix under the same commit-and-evict rules as inference. A chunk outside both credited windows advances these caches but contributes no training loss. The sink-boundary fork credits video over its complete continuation and audio over its first four chunks; later forks credit audio over their first four chunks.

Prompt pool and collected data. Training prompts are drawn from OpenHumanVid. Each prompt is a structured caption with a visual description, a transcribed speech line, and an ambient-sound description, so that the video and audio reward models each receive a modality-appropriate conditioning text; the audio rewards use the speech and sound fields, and the video rewards use the visual field. Captions exceeding the CLAP text encoder's token limit are removed before collection, because such prompts silently return no CLAP score and would otherwise leave holes in the audio reward. One collection round draws 128 prompts and, for each prompt, four fork positions with eight sampled continuations each, giving 4,096 continuations per round. Every rollout is generated autoregressively for 32 chunks (20.5 s) with no teacher forcing, so the prefix that a forked group shares is itself produced by the current policy rather than taken from data.

Reward and objective. Let z_{pool} denote standardization over the collected rollout pool. The modality rewards are composed from the scorers used in the main text:

$$\begin{aligned} R_v &= \lambda_{\mathrm{HPS}} z_{\mathrm{pool}}(\mathrm{HPSv3}) + \lambda_{\mathrm{VQ}} z_{\mathrm{pool}}(\text{VA-VQ}) + \lambda_{\mathrm{MQ}} z_{\mathrm{pool}}(\text{VA-MQ}), \\ R_a &= \lambda_{\mathrm{PQ}} z_{\mathrm{pool}}(\text{AudioBox-PQ}) + \lambda_{\mathrm{C}} z_{\mathrm{pool}}(\mathrm{CLAP}). \end{aligned} \tag{12}$$

The composite rewards are then normalized separately among continuations of the same prefix to obtain the video and audio advantages. At a recorded paired state, both NFT terms compare clean-latent predictions with that continuation's own endpoint, as in Eq. (4). The joint objective also includes the reference-prediction penalty collected in $\mathcal{R}$, together with history-marginal cross-modal attention regularization (Eq. (11)).

B Properties of the Forked Objective

The state distribution seen by the update. Consider chunk n of a behavior-policy rollout and let $j = 0, \ldots, T-1$ index its Euler denoising steps. Write $X_{n,j} = (X^v_{n,j}, X^a_{n,j})$ for the paired input to the denoiser, Y_n for the completed joint chunk, and $P_{<n}$ for its generated prefix. With $\Delta\sigma_j = \sigma_{j+1} - \sigma_j$, the actual sampler obeys

$$X^m_{n,j} = Y^m_n - \sum_{k=j}^{T-1} \Delta\sigma_k \, f^m_{\mathrm{old}}(X^v_{n,k}, X^a_{n,k}; \mathcal{C}_{\mathrm{old}}(P_{<n}), c, \sigma_k), \quad m \in \{v, a\}. \tag{13}$$

This is a telescoping form of the sampler update. Each state depends on earlier coupled audio–video predictions and the causal cache, rather than only on its final endpoint. Forward re-noising instead constructs $\widetilde{X}^m_{n,j} = (1-\sigma_j)Y^m_n + \sigma_j \epsilon^m$ with newly sampled noise. With freshly sampled noise, the re-noised state generally differs from the recorded state even if the sampler follows a straight path.

Table 5: Final training configuration. Rewards are standardized over the collected pool and then normalized within each sibling group; all remaining weights are held fixed across the runs reported in the main text.

Setting	Final Fork-NFT run
Backbone and trainable weights	LTX-AR (22B); one LoRA adapter, rank 32 ($\alpha = 64$), applied to all attention projections, feed-forward layers, and cross-modal bridges
Sampler and causal history	8 NFE; two sink and four recent chunks
Sibling group; audio credit window	8 continuations; 4 chunks (2.6 s)
Fork positions; rollout length	Four positions per prompt at chunk $b \in \{2, 10, 18, 26\}$ of a 32-chunk rollout (20.5 s); the sink-boundary fork b=2 additionally carries the video credit
Optimizer, batch, and update count	AdamW, learning rate 3×10^{-5}; effective batch 64 continuations (8 GPUs $\times$ gradient accumulation 8); 25 minibatch updates per collected buffer of 4,096 continuations
Reward coefficients in Eq. (12)	$\lambda_{\mathrm{HPS}} = \lambda_{\mathrm{VQ}} = \lambda_{\mathrm{MQ}} = 1$; $\lambda_{\mathrm{PQ}} = 1.5$, $\lambda_{\mathrm{C}} = 0.5$
NFT mixing and regularizer weights	$\beta_{\mathrm{mix}} = 1$; advantage clip 5.0; reference-prediction penalty 2×10^{-2}; history-marginal KL 0.2
Precision and compute	bf16 with gradient checkpointing; 8×H100 (80 GB) per round

During training we evaluate the recorded $X_{n,j}$ with $\mathcal{C}_\theta(P_{<n}) = \mathrm{Evict}(\mathrm{Commit}_\theta(P_{<n}))$. Thus the inputs follow the behavior sampler while the cache projections use the current policy. For fixed $P_{<n}$, this cache is exactly what the current-policy inference procedure would construct from that prefix. Reusing the behavior cache would instead evaluate the new predictor with stale key/value projections. If f_θ is differentiable with respect to the cache, its local discrepancy is

$$f_\theta(X, \mathcal{C}_{\mathrm{old}}(P)) - f_\theta(X, \mathcal{C}_\theta(P)) = J_{\mathcal{C}} f_\theta(X, \mathcal{C}_\theta(P))[\mathcal{C}_{\mathrm{old}}(P) - \mathcal{C}_\theta(P)] + o(\|\mathcal{C}_{\mathrm{old}}(P) - \mathcal{C}_\theta(P)\|). \tag{14}$$

Current-policy re-materialization removes this cache-projection discrepancy for the recorded prefix; it does not assert that a behavior-policy rollout is a fresh rollout of θ.

The effect of replacing recorded states can also be expressed at the update level. Fix a sampled endpoint, advantage, prefix, and noise level, and let $g_\theta(X)$ denote the resulting NFT gradient with the same current-policy cache in both evaluations. If g_θ is locally L_g-Lipschitz in the paired latents, then under a coupling of the recorded and re-noised inputs,

$$\left\|\mathbb{E}[g_\theta(X_{n,j})] - \mathbb{E}[g_\theta(\widetilde{X}_{n,j})]\right\| \le L_g \, \mathbb{E}\left[\|X_{n,j} - \widetilde{X}_{n,j}\|\right]. \tag{15}$$

The expectation includes the sampled rollouts and the re-noising variables. Equation (15) does not order the quality of two trained models; it identifies the extra input displacement introduced when a clean endpoint substitutes for the states actually visited by the sampler. Replay eliminates that displacement relative to the collected trajectory.

What the shared-prefix baseline removes. For a fixed prompt and fork state s_b, write the reward of sibling g as $R_m^g = R_m(s_b, a^g)$. Its group-relative advantage is

$$A_m^g = \frac{R_m^g - \overline{R}_m}{s_m + \epsilon}. \tag{16}$$

If a term $c(s_b)$ is identical across all siblings, replacing every R_m^g with $R_m^g + c(s_b)$ leaves both the numerator and the group standard deviation unchanged. Hence Eq. (16) removes a shared additive prefix term exactly. It does not assume that the prefix has no effect on the continuation: comparisons are made conditional on that prefix, while the credited window determines which generated chunks receive the resulting advantage.

The distinction between variation across fork states and variation among their continuations follows from the law of total variance:

$$\mathrm{Var}(R_m) = \mathbb{E}_{s_b}[\mathrm{Var}(R_m \mid s_b)] + \mathrm{Var}_{s_b}[\mathbb{E}(R_m \mid s_b)]\,. \tag{17}$$

For conditionally independent sibling continuations, their sample variance with denominator $G - 1$ estimates the first term conditional on s_b, not the second. Group centering therefore removes between-prefix shifts, but cannot create a learning signal when different continuations at a chosen fork and

scoring window receive nearly identical rewards. This is why the temporal extent of the video and audio credit windows is part of the method, rather than an incidental choice of how to normalize rewards.

Implicit NFT predictions at zero advantage. The two predictions in Eq. (3) obey the exact identity $\hat{y}_\theta^{m,+} + \hat{y}_\theta^{m,-} = 2\hat{y}_{\text{old}}^m$. When $A = 0$, $p(A) = 1/2$ and the NFT term averages the two clean-prediction errors. This is a symmetric, reward-neutral term, not a zero loss; the actual implementation normalizes each error separately as specified in Eq. (4). A modality with zero credit on a chunk can therefore still receive this symmetric term and the reward-independent regularizers. Chunks with zero credit for both modalities are skipped by the loss and used only to re-materialize causal context.

History-marginal attention KL. For one current-chunk query, let P and Q be the reference and policy attention distributions over visible historical keys H and current keys C. Let $\bar{P}$ and $\bar{Q}$ retain each key in H but merge all of C into one entry. Expanding the current-key contribution to the KL divergence gives

$$\sum_{j \in C} P_j \log \frac{P_j}{Q_j} = P(C) \log \frac{P(C)}{Q(C)} + P(C) D_{\mathrm{KL}}(P(\cdot \mid C) \,\|\, Q(\cdot \mid C)) \,. \tag{18}$$

The historical-key contributions are unchanged by merging. Therefore

$$D_{\mathrm{KL}}(P \| Q) = D_{\mathrm{KL}}(\bar{P} \| \bar{Q}) + P(C) D_{\mathrm{KL}}(P(\cdot \mid C) \,\|\, Q(\cdot \mid C)) \,. \tag{19}$$

The second term is nonnegative, so marginalization yields a weaker, precisely characterized constraint. Let $\alpha = P(H)$ and $\beta = Q(H)$. Decomposing the remaining marginal term once more gives

$$D_{\mathrm{KL}}(\bar{P} \| \bar{Q}) = D_{\mathrm{KL}}(\mathrm{Ber}(\alpha) \,\|\, \mathrm{Ber}(\beta)) + \alpha D_{\mathrm{KL}}(P(\cdot \mid H) \,\|\, Q(\cdot \mid H)) \,. \tag{20}$$

The first term penalizes a change in total attention mass assigned to committed history; the second penalizes its redistribution among historical keys. Neither term specifies which current-chunk key receives the remaining mass.

This selectivity also holds at the gradient level. With the reference probabilities fixed, the loss for one query is

$$\mathcal{K}(Q) = \sum_{h \in H} P_h \log \frac{P_h}{Q_h} + P(C) \log \frac{P(C)}{Q(C)}, \qquad Q(C) = \sum_{j \in C} Q_j. \tag{21}$$

Every current key has the same ambient derivative $\partial \mathcal{K} / \partial Q_j = -P(C)/Q(C)$ for $j \in C$. Consequently, a perturbation that only reallocates probability within the active chunk, with $\sum_{j \in C} \delta Q_j = 0$ and $\delta Q_h = 0$ for $h \in H$, has zero directional derivative. In fact, it leaves $\mathcal{K}$ exactly unchanged. By contrast, full attention KL contains the additional conditional term in Eq. (19) and penalizes that same reallocation. The proposed regularizer therefore targets historical cross-modal routing while permitting the active chunk's internal attention pattern to adapt; it is not a constraint on value vectors or generated outputs.

C Evaluation Protocol

Matched comparisons. The controlled short-form comparison evaluates LTX-AR, its OmniNFT adaptation, and Fork-NFT with the same autoregressive initialization, prompts, sampler, and output specification on the full JavisBench benchmark. The official evaluator supplies the metrics in Table 2. Long-horizon evaluation uses 128 matched prompts and 20.2-second joint outputs at 832×480 with 8 NFE, paired by clip identity across arms and generated from the same seed. Audio is transcoded to 16 kHz mono before scoring, and prompts are truncated to 75 BPE tokens for the ImageBind text tower, so that both scorers receive inputs within their supported ranges. VQ, AQ, DeSync, and JavisScore cover the complete output; TA-IB, AV-IB, and the audio component of AVH follow the evaluator's 30% audio-subsampling protocol. Because VideoAlign logits shift systematically with the number of sampled frames, VQ and MQ are compared only across arms evaluated under an identical frame budget, and are not compared against values reported elsewhere.

D Additional Controlled Analyses

Temporal credit geometry. Table 6 compares measured video and audio credit horizons. The modality-specific setting gives the strongest balance of video quality, audio quality, and audio–video alignment. In the collected pools, audio reward dispersion increases at later fork positions, whereas longer early continuations make video sibling differences clearer.

Table 6: Measured credit-horizon variants and their evaluation scores.

Video / audio credit	VQ↑	AQ↑	AV-Align↑	JavisScore↑
Full future / 4 chunks (ours)	3.127	5.042	0.375	0.205
4 chunks / 4 chunks	2.310	4.983	0.365	0.183
Full future / full future	3.125	4.936	0.343	0.191

Which attention constraint matters. Table 7 extends the on/off ablation in Table 4. History-marginal KL yields the highest AV-IB, AV-Align, and JavisScore among the compared constraints. Unlike full-attention KL, it preserves the mass and distribution assigned to committed history without constraining attention within the current chunk (Eq. (21)).

Table 7: Cross-modal attention regularization ablations.

Attention constraint	AV-IB↑	AV-Align↑	JavisScore↑
None	0.235	0.351	0.196
Current-chunk KL	0.237	0.349	0.196
Full-attention KL	0.238	0.353	0.198
History-marginal KL (ours)	0.241	0.375	0.205

E Additional Qualitative Examples

We provide matched 20.2-second generations to complement the quantitative long-horizon results. Each prompt is shown with temporally aligned video frames, audio waveform, and spectrogram for the LTX-AR backbone without RL and Fork-NFT.

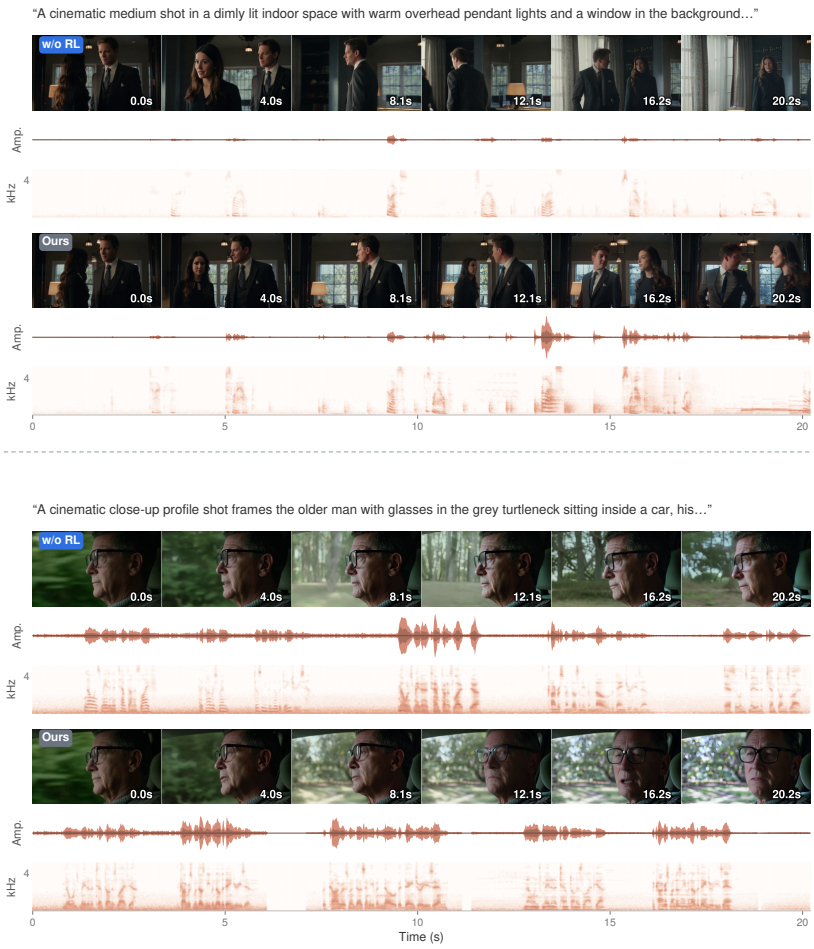


Figure 4: Additional long-horizon audio–video examples (A). For each prompt, the upper sequence shows LTX-AR without RL and the lower sequence shows Fork-NFT; frames, waveform, and spectrogram share the 20.2-second timeline.

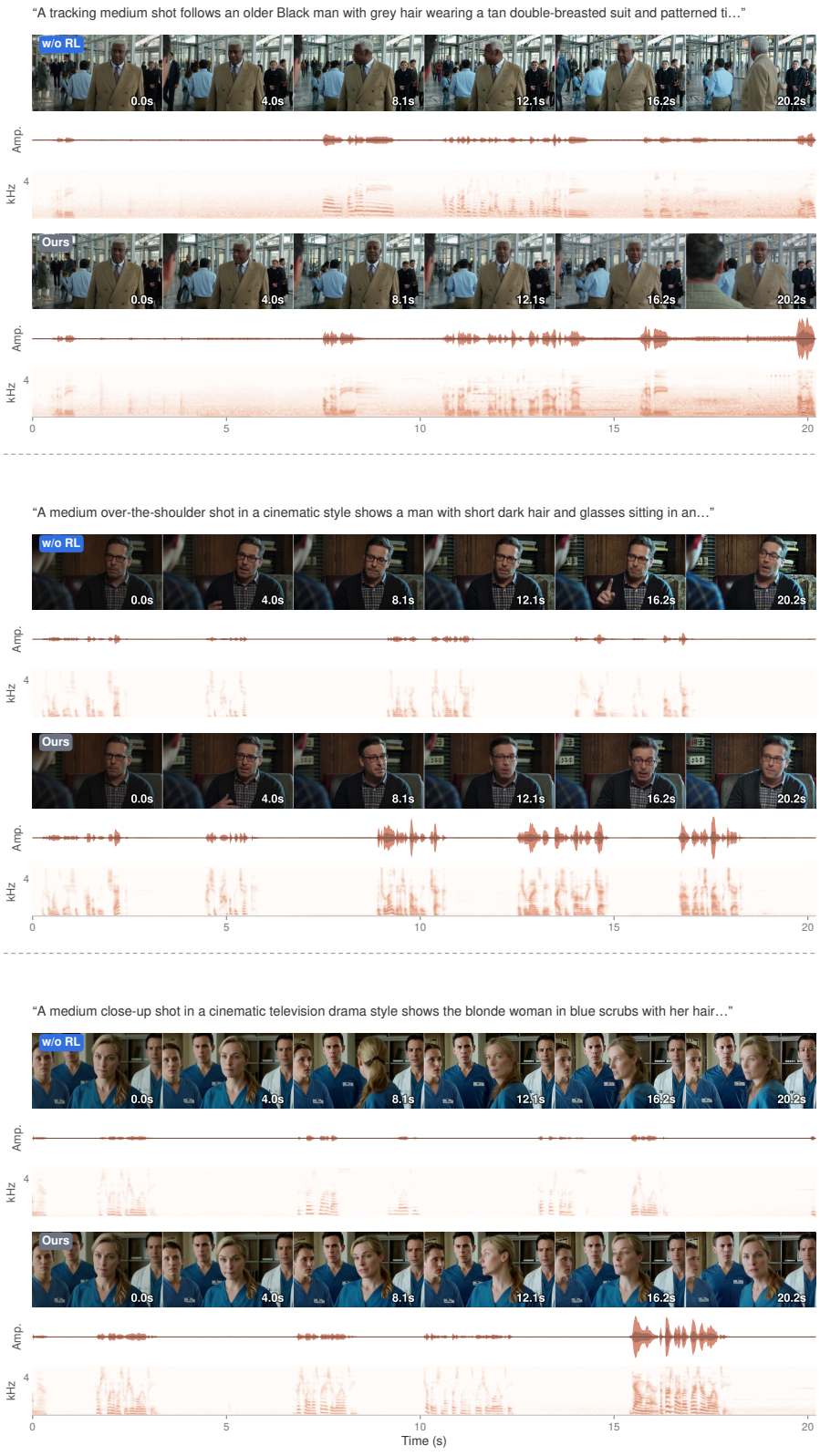


Figure 5: Additional long-horizon audio–video examples (B), displayed with the same matched layout as Figure 4.