CINEFORCING: STREAMING MULTI-SHOT AUDIOVISUAL GENERATION FOR LONG STORYTELLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans experience the world as a continuous, synchronized stream of sound and vision. While modern generative models excel at synthesizing brief audiovisual clips, they struggle to sustain this continuity over long horizons. During extended generation or scene transitions, models suffer from identity drift: characters alter their appearance and voices lose their distinctive timbre. Tracking visual frames is insufficient, since acoustic identity cannot be recovered from pixels alone. We present **CineForcing**, a memory-conditioned streaming framework for long-horizon audiovisual generation. Instead of carrying an ever-growing buffer of raw tokens, CineForcing summarizes completed shots as compact committed memory states that pair a visual anchor with speech-filtered acoustic evidence from the same shot. Student generation is causal at two timescales: audiovisual blocks are generated sequentially within a shot, while each new shot reads only state committed by completed shots. Memory-conditioned supervised fine-tuning initializes the backbone; parallel audiovisual teacher forcing and two-pass distribution matching then train causal streaming with ground-truth and self-generated *within-shot* prefixes, respectively, under strict-past audiovisual memory. We evaluate CineForcing on 30-shot long-story audiovisual rollouts. On 100 stories comprising 3,000 shots, CineForcing achieves the highest Voice and CAVP scores and the lowest audiovisual desynchronization among the compared methods. It improves SpeechAcc from 0.4224 to 0.7583 over OmniForcing and, with four sampling steps, achieves a measured generation latency of 4.557 seconds per clip at 480p. A matched inference-time ablation shows that audiovisual memory increases SpeechAcc from 0.0037 with both memory inputs zeroed to 0.7583, while achieving the strongest late-shot visual consistency and synchronization among memory variants. Video demos: cineforcing-demos.pages.dev.

1 INTRODUCTION

Interactive cinematic creation requires more than generating a single attractive clip. A movie scene, short drama, or personalized visual story unfolds as a sequence of shots: characters leave and re-enter the frame, cameras cut across viewpoints, dialogue and ambience evolve with the story, and each new shot must remain coherent with what came before. The goal is therefore *streaming multi-shot audiovisual storytelling*: generating synchronized video and sound as the story unfolds, while maintaining both visual identity and acoustic identity over long horizons.

Recent advances in efficient video generation, joint audio-video modeling, and streaming distillation have made high-quality audiovisual generation increasingly practical (HaCohen et al., 2024; 2026; Su et al., 2026). However, extending these models from isolated clips to long multi-shot stories introduces a distinct challenge: the generator must decide what information should persist after each shot is completed. Visual continuity alone is insufficient for audiovisual storytelling, since the appearance of a character and the identity of a speaker are carried by different modalities. Meanwhile, retaining all previous tokens is undesirable for long-horizon streaming generation, where historical context continually grows. This motivates a compact audiovisual state that preserves identity-relevant information across shot boundaries while keeping within-shot generation causal and efficient.

We introduce **CineForcing**, a memory-conditioned streaming framework for long-horizon multi-shot audiovisual generation. Figure 1 illustrates this shot-by-shot storytelling setting. CineForcing

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.

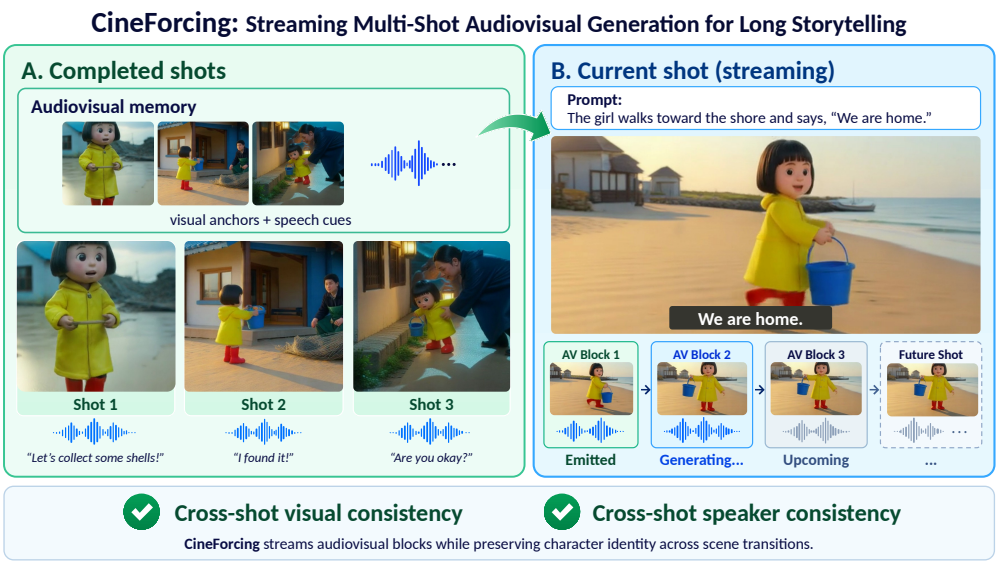


Figure 1: **CineForcing overview.** Completed shots provide paired visual anchors and speech cues for subsequent generation. Conditioned on this audiovisual memory and the current prompt, CineForcing emits audiovisual blocks causally while maintaining character appearance and speaker identity across scene transitions. Waveforms and dialogue illustrate the streaming interface.

organizes generation at two temporal scales. Within each shot, audiovisual blocks are generated causally with a bounded local key/value state. Across shots, completed generations are summarized into at most seven audiovisual memory items containing visual anchors and speech-filtered acoustic evidence. The resulting two-timescale state allows each new block to access both recent causal context and persistent cross-shot information, while memory is updated only when a shot is completed.

We first perform memory-conditioned supervised fine-tuning to adapt the audiovisual backbone to generation with memory context. We then train the streaming student through parallel audiovisual teacher forcing and two-pass self-rollout distribution matching on its own generated within-shot histories. At inference time, CineForcing continually rolls out audiovisual blocks and commits information from completed generated shots back into memory, forming a closed streaming generation loop. We evaluate CineForcing on 30-shot audiovisual stories.

Our contributions are threefold:

- We introduce a **memory-conditioned streaming framework** for multi-shot audiovisual storytelling, where synchronized video and audio are generated in a causal online setting.
- We propose a **two-timescale audiovisual state** that combines bounded within-shot causal context with committed cross-shot audiovisual memory, enabling visual and speaker information to persist across scene transitions.
- We develop a **memory-conditioned streaming training pipeline** using Memory SFT initialization, parallel audiovisual teacher forcing, and two-pass self-rollout distribution matching, and evaluate cross-shot similarity, audiovisual quality, and dialogue fidelity on long-form 30-shot generation.

2 Related Work

Efficient video generation and distillation. Video generators have advanced through latent representations, scalable transformers, and large-scale training (Lin et al., 2025; Blattmann et al., 2023; Menapace et al., 2024; Chen et al., 2024c;b; HaCohen et al., 2024; Wan et al., 2025; Kong et al., 2025; Yang et al., 2025b), alongside efficient autoencoders, attention, and feature caching (Cheng & Yuan, 2025; Xia et al., 2025; Li et al., 2025c; Liu et al., 2025a; Nan et al., 2026; Ghafoorian & Habibian, 2026; Ghafoorian et al., 2026). DMD/DMD2 and related consistency, trajectory-based,

adversarial, and reward-guided distillation reduce sampling cost (Chen et al., 2025a; Lu et al., 2025b; Park et al., 2025; Lv et al., 2025; Ding et al., 2025; Luo et al., 2025b; Kim et al., 2025; Luo et al., 2025a; Hsiao et al., 2024; Yin et al., 2024b; Mei et al., 2024; So et al., 2025; Mao et al., 2025; Liu et al., 2025d; Nie et al., 2026; Xie et al., 2026; Fan et al., 2026; Shao et al., 2026; Yin et al., 2024a; Luo et al., 2023a; Chen et al., 2026b; Huang et al., 2026d; Li et al., 2026a; Dong et al., 2026). These advances supply efficient generators; preserving identity across shot boundaries additionally requires deciding which historical evidence remains available.

Forcing and long-horizon rollout. Diffusion Forcing, CausVid, Self Forcing, and subsequent methods explore noise-conditioned histories, causal distillation, self-generated training contexts, memory selection, and long-rollout stabilization (Yin et al., 2025; Huang et al., 2025b; Liu et al., 2025b; Chen et al., 2026a; Yang et al., 2025a; Zhu et al., 2026a; Savva et al., 2026; Zhou et al., 2025; Sun et al., 2025b; Kodaira et al., 2026; Po et al., 2026; Lu et al., 2026; Chen et al., 2024a; Song et al., 2025; Cui et al., 2025; Yuan et al., 2026; Zhao et al., 2026c;b; Guo et al., 2026b). Context-Matched Distillation examines the teacher's causal information boundary, while Self Gradient Forcing trains historical context representations through future losses (Bandyopadhyay et al., 2026; Zhuang et al., 2026). CineForcing uses self-generated within-shot prefixes with detached historical K/V and fixed strict-past shot memory during distillation; deployment commits memory from generated shots.

Storytelling, multi-shot generation, and identity. Long-video and storytelling methods develop temporal extension, event and transition control, shot assembly, and explicit or implicit memory (Ouyang et al., 2025; Guo et al., 2025; Yang et al., 2025c; Dedhia et al., 2025; Feng et al., 2025; Xiao et al., 2025; Lu et al., 2025a; Liu et al., 2024a; Henschel et al., 2025; Qi et al., 2025; Yan et al., 2025; Li et al., 2025b; Cai et al., 2025; Kara et al., 2025; Tan et al., 2025; Wu et al., 2025b; Dalal et al., 2025; Wang et al., 2025b; Zhai et al., 2025; Huang et al., 2026a; Tian et al., 2026; Xu et al., 2026a; Cui et al., 2026; Wang et al., 2026d; Yesiltepe et al., 2026; Wang et al., 2026a; Zhao et al., 2026a; Girish et al., 2026; Wu et al., 2026b; Ifriqi et al., 2026; Huang et al., 2026c; Ma et al., 2026; Meng et al., 2026; Zhang et al., 2026b;c; An et al., 2026; Wang et al., 2025c; Wu et al., 2026c; Zhang et al., 2025a; Yu et al., 2025; Dou et al., 2026); personalization methods preserve subjects through reference conditioning, representation alignment, or identity objectives (Zhang et al., 2025b; Liu et al., 2025c; Hu, 2024; Jiang et al., 2024; Huang et al., 2025a; Chen et al., 2025b; Yuan et al., 2025; Men et al., 2025; Xue et al., 2026; Yang et al., 2026; Shen et al., 2026; Xu et al., 2026b; Sang et al., 2026; Pan et al., 2026; Zhang et al., 2026d; Wang et al., 2026b; Zhou et al., 2026; Wu et al., 2026a). StoryMem and Context as Memory retain visual evidence, while ShotStream combines local/global caches with intra- and inter-shot self-forcing (Zhang et al., 2025a; Yu et al., 2025; Luo et al., 2026). Our focus is the audiovisual state between shots, including acoustics that visual memory alone cannot provide.

Audiovisual generation and evaluation. Joint audio-video modeling, video-to-audio synthesis, and audio-conditioned video generation address complementary aspects of semantic and temporal alignment (Wang et al., 2025f; Haji-Ali et al., 2025; Ton et al., 2025; Ruan et al., 2023; Wang et al., 2025e; Mo & Song, 2025; Guan et al., 2025; Li et al., 2025a; Cheng et al., 2025; Hu et al., 2026; Li et al., 2026c; Hyun-Bin et al., 2026; Hong et al., 2026; Zhang et al., 2026a; Simon et al., 2026; Lee et al., 2026; Wang et al., 2026c; HaCohen et al., 2026; Polyak et al., 2025; Kondratyuk et al., 2024; Liu et al., 2026a;b; Wang et al., 2025a; Team et al., 2025; Low et al., 2025; Team et al., 2026). OmniForcing distills joint generation for streaming; Ripple uses recurrent cross-modal memory, while Vorch-Streamer, Omni-LiveAvatar, OmniMate, and U-Mind extend streaming interaction (Su et al., 2026; Ding et al., 2026; Han et al., 2026; Zhu et al., 2026b; Song et al., 2026; Deng et al., 2026). CineDance studies long-form cinematic audiovisual generation, UnityShots combines gated visual memory with speaker-reference conditioning, and DreamID-Omni studies controllable face–voice identity (Chen et al., 2026c; Huang et al., 2026b; Guo et al., 2026a). Duan et al. (2026) describe shot-paired visual and speech memory with causal teacher forcing and rollout distillation. Existing benchmarks cover perceptual quality, composition, narrative continuity, physical sound, and audiovisual or streaming consistency (Liu et al., 2024c; Sun et al., 2025a; Wang et al., 2025g; Han et al., 2025; Wu et al., 2025a; Wang et al., 2025d; Li et al., 2026b; Matsuda et al., 2026; Hua et al., 2026; Huang et al., 2023; 2024; Wei et al., 2026; Liu et al., 2026c). CineForcing focuses on the two-timescale state interface and its training–deployment boundary: bounded within-shot K/V and shot-paired visual and speech-filtered acoustic evidence committed across shots.

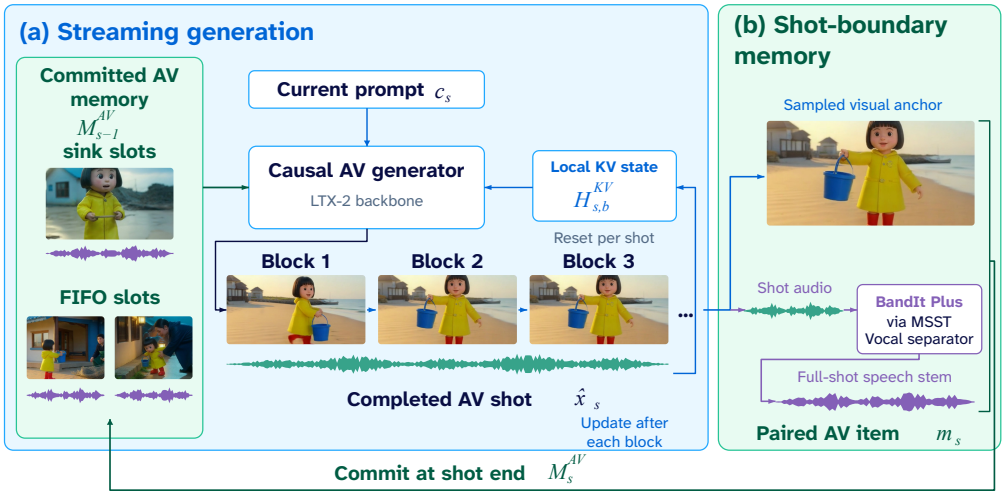


Figure 2: **Streaming generation and shot-boundary memory. (a)** The causal LTX-2 audiovisual generator conditions on the current prompt, committed audiovisual memory, and local K/V state. Local state updates after each block and resets at shot boundaries; committed memory stays fixed within each shot. **(b)** At shot completion, a sampled frame and the full-shot speech stem extracted by BandIt Plus through MSST form a paired memory item. The bounded memory bank retains persistent sink entries and recent entries in FIFO order.

3 METHOD

A camera cut can change the view, while recurring characters should retain their appearance and voice. Streaming generation therefore needs both local continuity and persistent identity evidence. CineForcing separates these roles through two-timescale audiovisual state (Figure 2), then trains the generator to use that state under causal rollout (Figure 3).

3.1 TWO-TIMESCALE STREAMING GENERATION

Recent blocks provide within-shot continuity; completed shots provide evidence that must survive a cut. Let $x_s = (v_s, a_s)$ be shot s of S shots, c_s its prompt, and $x_{s,b} = (v_{s,b}, a_{s,b})$ one of B_s aligned audiovisual blocks. We separate local key/value state $H^{\mathrm{KV}}_{s,b}$ from cross-shot memory M^{AV}_{s-1}:

$$p_\theta(x_{1:S} \mid c_{1:S}) = \prod_{s=1}^{S} \prod_{b=1}^{B_s} p_\theta\big(x_{s,b} \mid c_s, M^{\mathrm{AV}}_{s-1}, H^{\mathrm{KV}}_{s,b}\big)\,.$$

Each shot starts with an empty local cache. After generating block $\hat{x}_{s,b}$, its cached keys and values are updated; only after the shot ends is the shot-paired item $F(\hat{x}_s)$ committed:

$$H^{\mathrm{KV}}_{s,b+1} = \mathcal{U}_{\mathrm{KV}}(H^{\mathrm{KV}}_{s,b}, \hat{x}_{s,b}), \qquad M^{\mathrm{AV}}_{s} = \mathcal{U}_{\mathrm{mem}}(M^{\mathrm{AV}}_{s-1}, F(\hat{x}_s)).$$

Resetting the local cache lets each shot begin a new causal prefix, while committed memory carries completed-shot evidence across the cut. Keeping memory fixed within a shot gives every block the same cross-shot context. Algorithm S.3 gives the inference loop.

3.2 SHOT-PAIRED AUDIOVISUAL MEMORY

Visual anchors preserve appearance evidence, but a speaker's voice requires acoustic context. We pair both modalities from each completed shot in a memory item for causal training and inference:

$$m_s = F(x_s) = (z^v_s, z^a_s, q_s),$$

where q_s is the shot's integer slot in retained memory. The token sequences $z^m_s \in \mathbb{R}^{n^m_s \times d_m}$ have n^m_s memory tokens and native feature width d_m for modality $m \in \{v, a\}$. The visual component is obtained from a sampled frame,

$$z^v_s = \phi_v\left(v_s[f_s]\right),$$

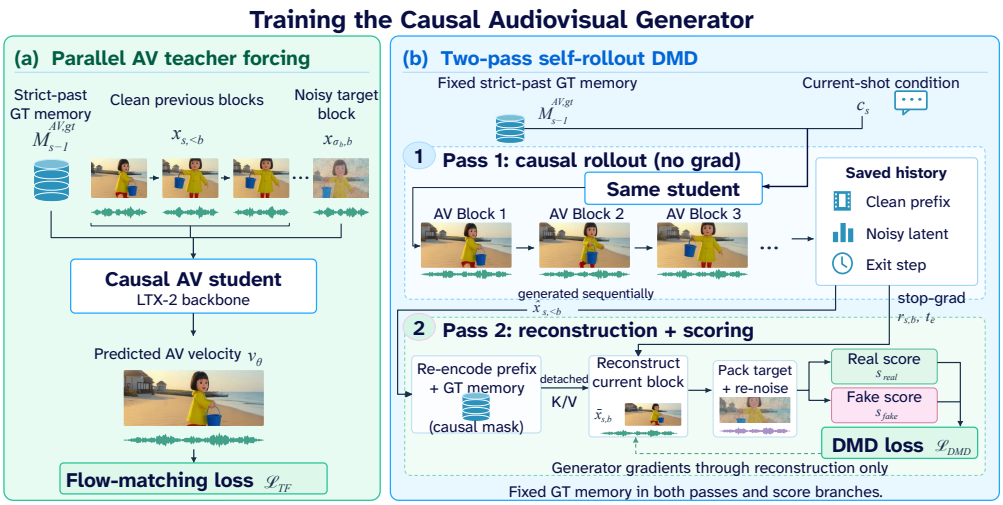


Figure 3: **Causal audiovisual training after Memory SFT. (a)** Parallel teacher forcing predicts rectified-flow velocities from clean prefixes, noisy targets, and strict-past ground-truth memory. **(b)** Two-pass DMD first rolls out the target without gradients, saving clean outputs, exit-noisy latents, and the exit step. Pass 2 reconstructs current blocks using detached historical K/V, then concatenates and re-noises them for real/fake scoring over the complete target window. Both passes and score branches share fixed ground-truth memory; generator gradients flow only through current-block reconstruction. Frames and waveforms schematically represent audiovisual tensors.

while the audio component is constructed from the full speech stem of the shot,

$$z_s^a = \phi_a\left(\mathrm{BandItPlus}(a_s)\right).$$

Here f_s indexes decoded frames, separately from diffusion timesteps; ϕ_v, ϕ_a are the backbone's native encoding/tokenization paths. To focus acoustic memory on speech, BandIt Plus, implemented through the open-source Music-Source-Separation-Training (MSST) toolkit (Solovyev et al., 2023), suppresses music and non-speech interference. We retain the *entire extracted speech stem* and discard the separated music and sound-effect stems.

Pairing identifies a common source shot without requiring temporal alignment: one frame supplies appearance and scene context, while the full speech stem retains acoustic evidence across that shot.

We inject memory tokens into the corresponding video and audio streams of LTX-2. Every current-shot block can attend to this persistent evidence alongside its permitted causal prefix.

Training memory sampling. Memory SFT samples memory context at random, with the intended character specified by the prompt. Causal TF/DMD use strict-past ground-truth memory (Appendix A.2), separately from inference retention.

Inference memory retention. To retain early references and recent context within a bounded bank, we keep at most seven completed shots. The first three occupy persistent *memory sinks*; later shots enter a four-slot FIFO queue, evicting the oldest entry on overflow. Sink shots are not duplicated. Memory updates only at shot boundaries, independently of within-shot K/V handling.

3.3 Training the Causal Audiovisual Generator

Memory-conditioned supervised fine-tuning. The pretrained backbone must learn to use memory alongside the current-shot prompt. We first adapt it through memory-conditioned supervised fine-tuning (Memory SFT), following the context-conditioning principle of Context as Memory (Yu et al., 2025). Randomly sampled memory supplies contextual evidence; the prompt specifies the intended character and shot content. This checkpoint initializes causal teacher forcing to learn blockwise generation, followed by self-rollout distillation to train on the student's own within-shot histories.

Both streaming stages use strict-past ground-truth cross-shot memory. Following the replay-based paradigm of Self-Forcing and SGF (Huang et al., 2025b; Zhuang et al., 2026), distillation reuses a no-gradient rollout for reconstruction and distribution matching.

Audiovisual teacher forcing. To learn causal generation with parallel supervision, we condition each block on clean preceding context. With batch axes omitted, encoded latents are $x_s^v \in \mathbb{R}^{T_v \times H \times W \times C_v}$ and $x_s^a \in \mathbb{R}^{T_a \times C_a}$. T_v, T_a count latent time steps, H, W are latent spatial dimensions, and C_v, C_a are feature widths (audio feature axes are flattened). Blocks are temporal slices of these tensors. For each pair, we sample a shared noise level $\sigma_b \in [0, 1]$ and construct

$$x^m_{\sigma_b,b} = (1 - \sigma_b)x^m_{s,b} + \sigma_b \epsilon^m_b, \qquad \epsilon^m_b \sim \mathcal{N}(0, I), \quad m \in \{v, a\}. \tag{1}$$

Packing clean historical blocks and noisy targets into one forward pass allows parallel supervision with causal visibility. Target block b reads committed audiovisual memory $M^{\mathrm{AV,gt}}_{s-1}$, all preceding clean blocks $x_{s,<b}$, and the current noisy audiovisual block $x_{\sigma_b,b}$. This preserves the information boundary of streaming inference.

Writing $\mathcal{C}_{s,b} = (c_s, x_{s,<b}, M^{\mathrm{AV,gt}}_{s-1})$ for this condition, we optimize the rectified-flow objective

$$\mathcal{L}_{\mathrm{TF}} = \sum_{m \in \{v,a\}} \lambda_m \sum_b \left\| W^m_b \odot \left(v^m_\theta(x_{\sigma_b,b}, \sigma_b \mid \mathcal{C}_{s,b}) - [\epsilon^m_b - x^m_{s,b}] \right) \right\|_2^2, \tag{2}$$

Here v^m_θ predicts velocity with the shape of $x^m_{s,b}$; $\lambda_v = 1$ and λ_a weight the modality losses, and W^m_b excludes conditioning and invalid elements. Both predictions receive the joint noisy input $x_{\sigma_b,b} = (x^v_{\sigma_b,b}, x^a_{\sigma_b,b})$; s is suppressed in this notation. The detailed audiovisual visibility masks and the image-conditioned supervision rule are provided in Appendix F.2.

Two-pass self-rollout distribution matching. Teacher forcing uses clean ground-truth histories, whereas inference must continue from the student's own outputs. We address this within-shot mismatch with two-pass self-rollout distribution matching: collect generated within-shot histories, then optimize current-block reconstructions conditioned on those histories. This exposes training to generated prefixes without backpropagating through the full rollout.

In **Pass 1**, the student causally rolls out the current target without gradients. Starting from an empty local cache, it generates audiovisual blocks sequentially and updates the within-shot state after every block. For each block, we store the generated clean output $\hat{x}_{s,b}$, its same-shaped exit-noisy latent $r_{s,b}$, and the sampled denoising exit timestep t_e (a scheduler index, not physical video time). The rollout is conditioned on fixed strict-past audiovisual memory $M^{\mathrm{AV,gt}}_{s-1}$.

In **Pass 2**, the self-generated prefix is re-encoded under the causal mask and reused to reconstruct the current audiovisual block:

$$\widetilde{H}^{\mathrm{KV}}_{s,<b} = C_\theta\left(c_s, \mathrm{sg}(\hat{x}_{s,<b}); \mathrm{sg}(M^{\mathrm{AV,gt}}_{s-1})\right), \tag{3}$$

$$\bar{x}_{s,b} = G_\theta\left(\mathrm{sg}(r_{s,b}), t_e \,\middle|\, c_s, \mathrm{sg}(\widetilde{H}^{\mathrm{KV}}_{s,<b}), \mathrm{sg}(M^{\mathrm{AV,gt}}_{s-1})\right), \tag{4}$$

Here C_θ re-encodes the saved prefix into historical keys and values, G_θ reconstructs a clean block of the same latent shape, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. The bar marks a reconstructed sample, and $\widetilde{H}^{\mathrm{KV}}_{s,<b}$ is the re-encoded prefix cache. In our implementation, the saved rollout, re-encoded historical K/V state, and cross-shot memory are detached, while gradients are propagated through the current-block reconstruction operation.

The reconstructed samples are then re-noised and evaluated by real and fake score branches to obtain the distribution-matching signal for updating the streaming student. The score models operate on the same valid audiovisual target window, while the student itself remains causal during rollout. We alternate generator and fake-score updates following standard distribution-matching training. The score-difference objective, gradient boundaries, and critic update are defined in Appendix F.3.

4 Experiments

We first introduce the evaluation protocol and compare cross-shot consistency, perceptual quality, and dialogue fidelity. We then report streaming efficiency and use memory interventions and audio-supervision studies to examine the sources of these gains.

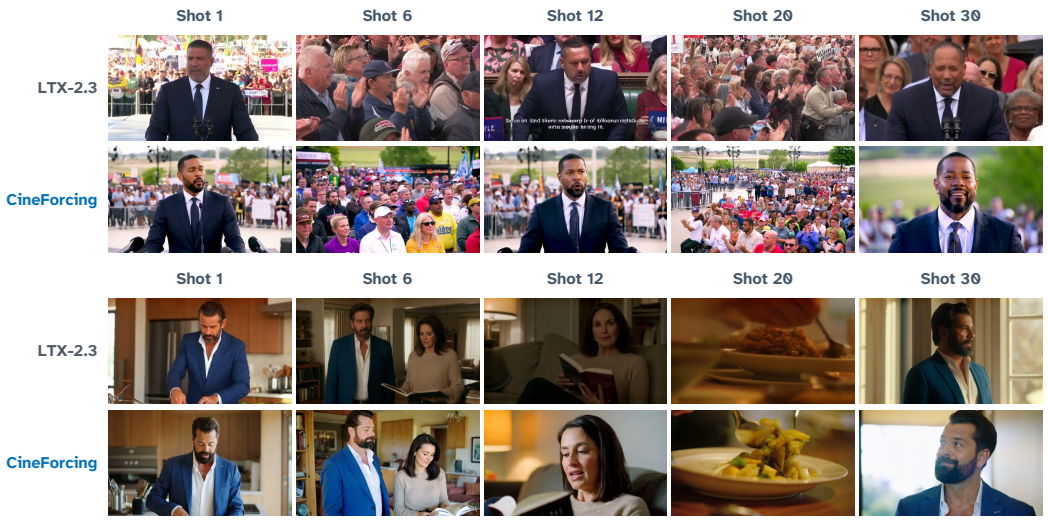


Figure 4: **Qualitative comparison on long-form multi-shot audiovisual storytelling.** We show distant shots from 30-shot rollouts to visualize character appearance and scene evolution over time.

4.1 Experimental Setup

We evaluate 100 stories of 30 shots each (3,000 prompts), spanning specified and original characters, animation and live action, and diverse dialogue, actions, camera motion, ambience, and music. Quality is evaluated on 121-frame 480p clips at 25 FPS (4.84 s), with seed 42 per shot; efficiency is reported separately. Baselines comprise LTX-2.3 (HaCohen et al., 2026), OmniForcing (Su et al., 2026), ShotStream (Luo et al., 2026), and StoryMem (Zhang et al., 2025a); the latter two are video-only, with audio scores marked N/A. OmniForcing generates shots independently; CineForcing uses cross-shot audiovisual memory. Table S.1 summarizes implementation and evaluation settings, and Table S.2 gives training hyperparameters.

Cross-shot consistency is measured by ViCLIP (Wang et al., 2024), Self-CIDS using GroundingDINO, Torchreid, and FaceNet (Liu et al., 2024b; Zhou & Xiang, 2019; Schroff et al., 2015), and CAMPPlus voice similarity (Chen et al., 2024d), which linearly rescales whole-shot embedding cosine similarity from $[-1, 1]$ to $[0, 1]$. Visual quality is assessed by the aesthetic and imaging-quality dimensions of VBench/VBench++ (Huang et al., 2023; 2024), and text–video alignment by CLIP (Radford et al., 2021).

Audio and audiovisual quality are evaluated with AudioBox-Aesthetics (Tjandra et al., 2025), CLAP (Elizalde et al., 2023), ImageBind audiovisual similarity (Girdhar et al., 2023), CAVP (Luo et al., 2023b), and DeSync (Iashin et al., 2024). For scripted dialogue, we transcribe speech with Whisper base.en (Radford et al., 2023) and compute word error rate (WER) against the target dialogue, reporting $\mathrm{SpeechAcc} = \max(0, 1 - \mathrm{WER})$, where higher values indicate closer agreement.

4.2 Main Results

CineForcing improves cross-shot similarity over OmniForcing in ViCLIP (0.7695 to 0.7851), Self-CIDS (0.6624 to 0.6891), and Voice (0.7990 to 0.8121), as shown in Table 1. It attains the highest Voice and CAVP scores and the lowest DeSync among the compared methods. AudioBox-Aesthetics rises from 4.7171 to 5.4261, CLAP from 0.3718 to 0.4318, and ImageBind audiovisual similarity from 0.0929 to 0.2541. SpeechAcc increases from 0.4224 to 0.7583, indicating closer agreement with scripted dialogue. Together, these results support improved audiovisual continuity and speech fidelity; the full table also reports per-clip perceptual quality.

Figure 4 compares distant shots from long-form story rollouts, including shots 1, 6, 12, 20, and 30. Across substantial changes in viewpoint, scene composition, and story context, CineForcing

Table 1: **Quantitative results on long-form audiovisual generation.** Higher is better except for DeSync. Bold and underline indicate the best and second-best values, respectively. ShotStream and StoryMem are video-only methods and have no audio-related scores. CLIP denotes text–video alignment, Voice denotes whole-shot CAMPPlus similarity, and Acc. denotes SpeechAcc.

Method	Visual Metrics					Audio / Audiovisual / Speech Metrics						
	ViCLIP	Self-CIDS	Aesthetic	Imaging	CLIP	Voice	Audio	CLAP	IB-AV	CAVP	DeSync↓	Acc.
LTX-2.3	0.7621	0.5742	0.5585	0.6303	**0.2426**	0.6498	**5.8076**	**0.4573**	**0.2829**	0.7914	0.5674	**0.8300**
ShotStream	0.7688	0.7066	0.5253	**0.7001**	0.1998	N/A	N/A	N/A	N/A	N/A	N/A	N/A
StoryMem	**0.8464**	**0.7443**	**0.6048**	0.6643	0.2368	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OmniForcing	0.7695	0.6624	0.6025	0.6886	0.2313	0.7990	4.7171	0.3718	0.0929	0.7781	0.6275	0.4224
CineForcing	0.7851	0.6891	0.5535	0.6630	0.2396	**0.8121**	5.4261	0.4318	0.2541	**0.7983**	**0.2627**	0.7583

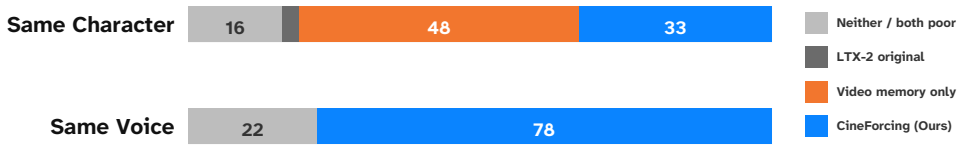


Figure 5: **Human preference study on cross-shot consistency.** Ten participants assess 100 three-candidate comparison sets with method names hidden and order randomized (1,000 responses per question). Appearance and voice/timbre are judged separately. The light-gray segment denotes "neither / both poor".

maintains recognizable character appearance. Table S.3 provides the corresponding story contexts. The quantitative metrics in Table 1 measure consistency over the complete evaluation suite.

Ten participants each assess the same 100 comparison sets, each containing three candidate videos: the original LTX-2 baseline, video-only memory, and CineForcing. With method names hidden and order randomized, they separately judge character appearance and voice/timbre, choosing one candidate or "neither / both poor" (Appendix C). Figure 5 shows a visual–acoustic tradeoff: video-only memory receives more same-character preferences than CineForcing (48% versus 33%), while CineForcing receives 78% of same-voice preferences.

4.3 Streaming Efficiency

Table 2: **Measured generation efficiency at 480p.** Latency: wall-clock seconds per test clip; throughput: generated video seconds per wall-clock second. StoryMem is video-only; the others generate audio and video.

Method	Latency ↓	Thru. ↑
LTX-2.3	96.099	0.0520
StoryMem	163.900	0.0305
OmniForcing	4.027	1.2416
CineForcing	4.557	1.0972

Table 4: **Design study of audio supervision.** Audio loss weighting; higher is better for all metrics.

λ_a	ViCLIP ↑	Self-CIDS ↑	Voice ↑
0.10	0.803	0.762	0.712
0.25	0.792	0.748	0.741
0.50	0.786	0.755	0.753
1.00	0.771	0.731	0.766

Table 2 reports measured generation latency and throughput under the evaluated 5-second clips at 480p. CineForcing achieves 4.557 seconds per clip and 1.0972 generated video seconds per wall-clock second, exceeding playback speed. Its block-wise generation emits audiovisual output before shot completion. The reported full-clip latency is distinct from time to first chunk. The adjacent Table 4 reports a separate audio loss-weight sweep.

4.4 Ablation Studies

We isolate the complementary contributions of visual and audio memory through inference-time interventions on the same checkpoint.

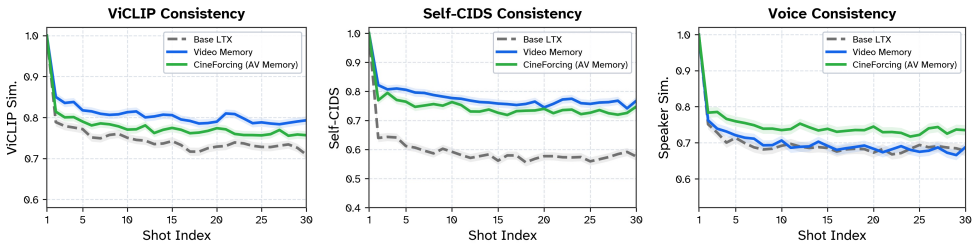


Figure 6: **30-shot memory study.** ViCLIP, Self-CIDS, and speaker similarity for LTX, video-only memory, and audiovisual memory.

Table 3 isolates memory inputs at inference with the same checkpoint, sampler, and memory retention policy for the same story prompts. We zero both memory modalities, only audio memory, or only video memory, and compare them with full AV memory. Video-only memory achieves the highest pairwise ViCLIP (0.7925), but SpeechAcc remains 0.0035, versus 0.0037 with both inputs zeroed. Audio-only memory recovers SpeechAcc to 0.6847. Full AV memory achieves the highest late-shot ViCLIP and SpeechAcc, and the lowest DeSync@4s, supporting the complementary roles of visual and acoustic history.

Table 3: **Inference-time memory ablation.** All rows use the same checkpoint, sampler, and retention policy. Zeroing a modality affects its memory input. Bold indicates the best value in each column. ViCLIP late pairs shot 1 with shots 16–30, averaging these 15 cosine similarities within each story, then across evaluated stories.

Memory input	ViCLIP pairwise ↑	ViCLIP late ↑	SpeechAcc ↑	DeSync@4s ↓
Both modalities zeroed	0.6265	0.5215	0.0037	0.9850
Video only (audio zeroed)	**0.7925**	0.6775	0.0035	0.9339
Audio only (video zeroed)	0.6891	0.5685	0.6847	0.4878
AV memory	0.7851	**0.6797**	**0.7583**	**0.2627**

Figure 6 shows a visual–acoustic tradeoff across the 30-shot rollout. Video-only memory yields higher ViCLIP and Self-CIDS scores, whereas AV memory maintains higher speaker similarity in later shots, highlighting the role of acoustic history in preserving voice consistency across shot transitions.

5 Conclusion

CineForcing formulates streaming multi-shot audiovisual generation as a two-timescale state problem: local key/value state carries the causal prefix within a shot, and audiovisual memory is committed across shots. After memory-conditioned SFT, the student learns causal blockwise generation through audiovisual teacher forcing and then applies two-pass distribution matching to self-generated within-shot prefixes under fixed strict-past memory. Deployment commits memory from generated shots. The evaluation shows improvements over OmniForcing in cross-shot similarity, audiovisual alignment, and dialogue fidelity in the evaluated configurations. Matched memory interventions further show complementary roles for visual and acoustic history, with full audiovisual memory achieving the highest late-shot ViCLIP and lowest DeSync@4s.

Future work could extend CineForcing to complex multi-character scenes, where visual identities, speaker turns, and interactions must remain consistent across shots. Beyond predefined scripts, an agent could continue the narrative, plan subsequent shots, and revise dialogue as generation unfolds. An agent harness could coordinate scriptwriting, audiovisual generation, continuity checks, and editing through persistent story state and iterative feedback, providing a path toward the autonomous production of complete films.

Ethics statement. Long-form audiovisual generation can be misused for impersonation and disinformation; release should therefore include provenance or watermarking guidance, intended-use restrictions, and appropriate access controls. The human preference study collects anonymous judgments without personally identifying information.

Reproducibility statement. The method formulation and streaming algorithms are provided in Sections 3 and F, with Memory SFT initialization described in Sections 3.3 and A.1; evaluation and implementation details are provided in Sections 4, A, and B.

AI use statement. AI tools assisted with manuscript drafting and editing, scientific-figure revision, experiment-design feedback, and interpretation of results. GPT-5 was used for shot-level prompt expansion (Appendix D). The authors take responsibility for the final content.

REFERENCES

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, Chenyang Zhang, Tao Xiang, Fanny Yang, Serge Belongie, and Tian Xie. Onestory: Coherent multi-shot video generation with adaptive memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16173–16184, June 2026.

Hmrishav Bandyopadhyay, Xuanchi Ren, Zijian Huang, Jay Zhangjie Wu, Tianshi Cao, Ruilong Li, Bryan Chu, Sanja Fidler, Yi-Zhe Song, and Zian Wang. Context-matched distillation: Teacher causality for autoregressive video distillation, 2026. URL https://arxiv.org/abs/2608.13391.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, June 2023.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7763–7772, June 2025.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024a. URL https://arxiv.org/abs/2407.01392.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, June 2024b.

Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16185–16195, October 2025a.

Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6441–6451, June 2024c.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context, 2026a. URL https://arxiv.org/abs/2602.06028.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation, 2026b. URL https://arxiv.org/abs/2606.18478.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6099–6110, June 2025b.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Tinglong Zhu, Rongjie Huang, Chong Deng, Qian Chen, Shiliang Zhang, Wen Wang, and Xihao Li. 3d-speaker-toolkit: An open-source toolkit for multimodal speaker verification and diarization, 2024d. URL https://arxiv.org/abs/2403.19971.

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, 2026c. URL https://arxiv.org/abs/2606.09639.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911, June 2025.

Yu Cheng and Fajie Yuan. Leanvae: An ultra-efficient reconstruction vae for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15692–15702, October 2025.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation, 2025. URL https://arxiv.org/abs/2510.02283.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Lol: Longer than longer, scaling video generation to hour. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38132–38142, June 2026.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17702–17711, June 2025.

Bhishma Dedhia, David Bourgin, Krishna Kumar Singh, Yuheng Li, Yan Kang, Zhan Xu, Niraj K. Jha, and Yuchen Liu. Generating, fast and slow: Scalable parallel video generation with video interface networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15385–15394, October 2025.

Xiang Deng, Feng Gao, Yong Zhang, Youxin Pang, Xu Xiaoming, Zhuoliang Kang, Xiaoming Wei, and Yebin Liu. U-mind: A unified framework for real-time multimodal interaction with audiovisual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10874–10886, June 2026.

Yanbo Ding, Zhizhi Guo, Quanyue Song, Yishan He, Zhixiang He, Yongxiang Li, and Yali Wang. Ripple: Real-time streaming audio-video generation with cross-modal recurrent memory, 2026. URL https://arxiv.org/abs/2607.26818.

Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17961–17971, October 2025.

Linwei Dong, Ruoyu Guo, Ge Bai, Zehuan Yuan, Yawei Luo, and Changqing Zou. Guiding distribution matching distillation with gradient-based reinforcement learning, 2026. URL https://arxiv.org/abs/2604.19009.

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

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, Yaofeng Su, Yilang Sun, Haoyu Wang, Zeyue Xue, Songchun Zhang, and Junhao Zhuang. Long-horizon audio-visual generation for persistent stories and interactive worlds, 2026. URL https://arxiv.org/abs/2608.23383.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, June 2023. doi: 10.1109/icassp49357.2023.10095889. URL http://dx.doi.org/10.1109/ICASSP49357.2023.10095889.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41667–41676, June 2026.

Zhanzhou Feng, Qingpei Guo, Xinyu Xiao, Ruihan Xu, Ming Yang, and Shiliang Zhang. Unified video generation via next-set prediction in continuous domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19427–19438, October 2025.

Mohsen Ghafoorian and Amirhossein Habibian. Rehyat: Recurrent hybrid attention for video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40674–40684, June 2026.

Mohsen Ghafoorian, Denis Korzhenkov, and Amirhossein Habibian. Attention surgery: An efficient recipe to linearize your video diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32915–32925, June 2026.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, June 2023.

Sharath Girish, Viacheslav Ivanov, Tsai-Shien Chen, Hao Chen, Aliaksandr Siarohin, and Sergey Tulyakov. Alchemint: Fine-grained temporal control for multi-reference consistent video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23227–23237, June 2026.

Jiazhi Guan, Kaisiyuan Wang, Zhiliang Xu, Quanwei Yang, Yasheng Sun, Shengyi He, Borong Liang, Yukang Cao, Yingying Li, Haocheng Feng, Errui Ding, Jingdong Wang, Youjian Zhao, Hang Zhou, and Ziwei Liu. Audcast: Audio-driven human video generation by cascaded diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10678–10689, June 2025.

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, 2026a. URL https://arxiv.org/abs/2602.12160.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17281–17291, October 2025.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling, 2026b. URL https://arxiv.org/abs/2512.15702.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

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, 2026. URL https://arxiv.org/abs/2601.03233.

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Alper Canberk, Kwot Sin Lee, Vicente Ordonez, and Sergey Tulyakov. Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19373–19385, October 2025.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18858–18868, June 2025.

Menglin Han, Yang Ding, Yulei Lu, Haoran Yu, Xin Ma, Junyi Chen, Zhangkai Ni, Lin Ma, and Yaohui Wang. Vorch-streamer: Extending human audio-visual generation to real-time long-form streaming, 2026. URL https://arxiv.org/abs/2608.05663.

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

Susung Hong, Ira Kemelmacher-Shlizerman, Brian Curless, and Steven M. Seitz. Musicinfuser: Making video diffusion listen and dance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43751–43761, June 2026.

Yi-Ting Hsiao, Siavash Khodadadeh, Kevin Duarte, Wei-An Lin, Hui Qu, Mingi Kwon, and Ratheesh Kalarot. Plug-and-play diffusion distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13743–13752, June 2024.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8153–8163, June 2024.

Teng Hu, Zhentao Yu, Guozhen Zhang, Zihan Su, Zhengguang Zhou, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Harmony: Harmonizing audio and video generation through cross-task synergy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16085–16095, June 2026.

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

Binyuan Huang, Yuning Lu, Weinan Jia, Hualiang Wang, Mu Liu, and Daiqing Yang. Rethinking position embedding as a context controller for multi-reference and multi-shot video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23304–23313, June 2026a.

Chi-Pin Huang, Yen-Siang Wu, Hung-Kai Chung, Kai-Po Chang, Fu-En Yang, and Yu-Chiang Frank Wang. Videomage: Multi-subject and motion customization of text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17603–17612, June 2025a.

Jiehui Huang, Yuechen Zhang, Bin Xia, Jiahao Wang, Xu He, Zhenchao Tang, Meng Chu, Xin Tao, Pengfei Wan, and Jiaya Jia. Unityshots: Memory-driven multi-shot audio-video generation with boundary-aware gating, 2026b. URL https://arxiv.org/abs/2606.21661.

Junjia Huang, Binbin Yang, Pengxiang Yan, Jiyang Liu, Bin Xia, Zhao Wang, Yitong Wang, Liang Lin, and Guanbin Li. Dreamshot: Personalized storyboard synthesis with video diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36736–36746, June 2026c.

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

Yushi Huang, Xiangxin Zhou, Ruoyu Wang, Chi Zhang, Jun Zhang, and Tianyu Pang. Reinforcing few-step generators via reward-tilted distribution matching, 2026d. URL https://arxiv.org/abs/2605.26108.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models, 2024. URL https://arxiv.org/abs/2411.13503.

Oh Hyun-Bin, Yuhta Takida, Toshimitsu Uesaka, Tae-Hyun Oh, and Yuki Mitsufuji. Pavas: Physics-aware video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14481–14491, June 2026.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, April 2024. doi: 10.1109/icassp48485.2024.10448489. URL http://dx.doi.org/10.1109/ICASSP48485.2024.10448489.

Tariq Berrada Ifriqi, John Nguyen, Karteek Alahari, Jakob Verbeek, and Ricky T. Q. Chen. Flowception: Temporally expansive flow matching for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16185–16195, June 2026.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6689–6700, June 2024.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28405–28415, June 2025.

Jisoo Kim, Wooseok Seo, Junwan Kim, Seungho Park, Sooyeon Park, and Youngjae Yu. V.i.p. : Iterative online preference distillation for efficient video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17235–17245, October 2025.

Akio Kodaira, Tingbo Hou, Ji Hou, Markos Georgopoulos, Felix Juefei-Xu, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29200–29210, June 2026.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024. URL https://arxiv.org/abs/2312.14125.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Junwon Lee, Juhan Nam, and Jiyoung Lee. Hear what matters! text-conditioned selective video-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36680–36690, June 2026.

Haoyu Li, Tingyan Wen, Lin Qi, Zhe Wu, Yihuang Chen, Xing Zhou, Lifei Zhu, Xueqian Wang, and Kai Zhang. 1.x-distill: Breaking the diversity, quality, and efficiency barrier in distribution matching distillation, 2026a. URL https://arxiv.org/abs/2604.04018.

Tingle Li, Siddharth Gururani, Kevin J. Shih, Gantavya Bhatt, Sang-gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu. Benchmarking single-factor physical video-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1939–1949, June 2026b.

Xinjie Li, Ziyi Chen, Xinlu Yu, Iek-Heng Chu, Peng Chang, and Jing Xiao. Co-speech gesture video generation with implicit motion-audio entanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11384–11394, June 2025a.

You Li, Dewei Zhou, Fan Ma, Fu Li, Dongliang He, and Yi Yang. Foleydirector: Fine-grained temporal steering for video-to-audio generation via structured scripts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29254–29264, June 2026c.

Zhuoling Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. Longdiff: Training-free long video generation in one go. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17789–17798, June 2025b.

Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, and Li Yuan. Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17778–17788, June 2025c.

Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, Cha Chen, Yiran Fei, Lezhi Li, Yinfei Yang, Yizhou Sun, and Kai-Wei Chang. Stiv: Scalable text and image conditioned video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16249–16259, October 2025.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm - open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6190–6200, June 2024a.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7353–7363, June 2025a.

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, 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, 2026b. URL https://arxiv.org/abs/2602.19163.

Kaiqi Liu, Haoxuan Zeng, Jingqi Liu, Jiacong Fang, Ziqi Cai, Yunyao Mao, Henglin Liu, Yu Sheng, Shuchen Weng, and Boxin Shi. Streamav-bench: A comprehensive benchmark for streaming audio-video generation, 2026c. URL https://arxiv.org/abs/2608.26336.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time, 2025b. URL https://arxiv.org/abs/2509.25161.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, October 2025c.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024b. URL https://arxiv.org/abs/2303.05499.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, June 2024c.

Yunpeng Liu, Boxiao Liu, Yi Zhang, Xingzhong Hou, Guanglu Song, Yu Liu, and Haihang You. See further when clear: Curriculum consistency model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18103–18112, June 2025d.

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

Chen-Yi Lu, Md Mehrab Tanjim, Ishita Dasgupta, Somdeb Sarkhel, Gang Wu, Saayan Mitra, and Somali Chaterji. Skald: Learning-based shot assembly for coherent multi-shot video creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17859–17868, October 2025a.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J. Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16818–16829, October 2025b.

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 (CVPR)*, pp. 34385–34397, June 2026.

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

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 48855–48876. Curran Associates, Inc., 2023b. doi: 10.52202/075280-2121. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/98c50f47a37f63477c01558600dd225a-Paper-Conference.pdf.

Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. Shotstream: Streaming multi-shot video generation for interactive storytelling, 2026. URL https://arxiv.org/abs/2603.25746.

Yihong Luo, Tianyang Hu, Yifan Song, Jiacheng Sun, Zhenguo Li, and Jing Tang. Adding additional control to one-step diffusion with joint distribution matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4009–4018, October 2025a.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17719–17728, October 2025b.

Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K. Wong, Yu Qiao, and Ziwei Liu. Dual-expert consistency model for efficient and high-quality video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14983–14993, October 2025.

Yukuo Ma, Cong Liu, Junke Wang, Junqi Liu, Haibin Huang, Zuxuan Wu, Chi Zhang, and Xuelong Li. Tempomaster: Efficient long video generation via next-frame-rate prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30414–30424, June 2026.

Xiaofeng Mao, Zhengkai Jiang, Fu-yun Wang, Jiangning Zhang, Hao Chen, Mingmin Chi, Yabiao Wang, and Wenhan Luo. Osv: One step is enough for high-quality image to video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12585–12594, June 2025.

Ryosuke Matsuda, Keito Kudo, Haruto Yoshida, Nobuyuki Shimizu, and Jun Suzuki. Slvmeval: Synthetic meta evaluation benchmark for text-to-long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7784–7794, June 2026.

Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M. Patel, and Peyman Milanfar. Codi: Conditional diffusion distillation for higher-fidelity and faster image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9048–9058, June 2024.

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21181–21191, June 2025.

Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, and Sergey Tulyakov. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7038–7048, June 2024.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, Yanhong Zeng, Xing Zhu, Yujun Shen, and Huamin Qu. Holocine: Holistic generation of cinematic multi-shot long video narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 461–471, June 2026.

Shentong Mo and Yibing Song. Foley-flow: Coordinated video-to-audio generation with masked audio-visual alignment and dynamic conditional flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28912–28921, June 2025.

Kepan Nan, Wangbo Zhao, Penghao Zhou, Jun Li, Zhenheng Yang, Jian Yang, and Ying Tai. Accelerating autoregressive video diffusion via history-guided cache and residual correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43740–43750, June 2026.

Weili Nie, Julius Berner, Nanye Ma, Chao Liu, Saining Xie, and Arash Vahdat. Transition matching distillation for fast video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4645–4655, June 2026.

Wenqi Ouyang, Zeqi Xiao, Danni Yang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Tokensgen: Harnessing condensed tokens for long video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18197–18206, October 2025.

Panwang Pan, Jingjing Zhao, Yuchen Lin, Chenguo Lin, Chenxin Li, Hengyu Liu, Tingting Shen, and Yadong Mu. Id-crafter: Vlm-grounded online rl for compositional multi-subject video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36627–36637, June 2026.

Geon Yeong Park, Sang Wan Lee, and Jong Chul Ye. Inference-time diffusion model distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4049–4058, October 2025.

Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43727–43739, June 2026.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. URL https://arxiv.org/abs/2410.13720.

Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Mask^2dit: Dual mask-based diffusion transformer for multi-scene long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18837–18846, June 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.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 (CVPR)*, pp. 10219–10228, June 2023.

Shen Sang, Tiancheng Zhi, Tianpei Gu, Jing Liu, and Linjie Luo. Lynx: Towards high-fidelity personalized video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9192–9202, June 2026.

Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in minecraft, 2026. URL https://arxiv.org/abs/2602.22208.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Shitong Shao, Yufei Gu, and Zeke Xie. Fastlightgen: Fast and light video generation with fewer steps and parameters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2104–2114, June 2026.

Liao Shen, Wentao Jiang, Yiran Zhu, Jiahe Li, Tiezheng Ge, Zhiguo Cao, and Bo Zheng. Identity-preserving image-to-video generation via reward-guided optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27281–27290, June 2026.

Christian Simon, Masato Ishii, Wei-Yao Wang, Koichi Saito, Akio Hayakawa, Dongseok Shim, Zhi Zhong, Shuyang Cui, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji. Echoes over time: Unlocking length generalization in video-to-audio generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15840–15849, June 2026.

Junhyuk So, Jiwoong Shin, Chaeyeon Jang, and Eunhyeok Park. Pcm : Picard consistency model for fast parallel sampling of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23313–23322, June 2025.

Roman Solovyev, Alexander Stempkovskiy, and Tatiana Habruseva. Benchmarks and leaderboards for sound demixing tasks. arXiv preprint arXiv:2305.07489, 2023. Code: https://github.com/ZFTurbo/Music-Source-Separation-Training.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion, 2025. URL https://arxiv.org/abs/2502.06764.

Quanyue Song, Yishan He, Yanbo Ding, Zhixiang He, Yongxiang Li, Caigui Jiang, and Zhi Zhi Guo. Omnimate: Open-ended real-time streaming audio-visual generation for interactive avatars, 2026. URL https://arxiv.org/abs/2607.23023.

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, 2026. URL https://arxiv.org/abs/2603.11647.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416, June 2025a.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, Siyu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7364–7373, June 2025b.

Jiangtong Tan, Hu Yu, Jie Huang, Jie Xiao, and Feng Zhao. Freepca: Integrating consistency information across long-short frames in training-free long video generation via principal component analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27979–27988, June 2025.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao

Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-omni technical report, 2025. URL https://arxiv.org/abs/2512.16776.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Shiqi Jiang, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. Mova: Towards scalable and synchronized video-audio generation, 2026. URL https://arxiv.org/abs/2602.08794.

Jiahao Tian, Chenxi Song, Wei Cheng, and Chi Zhang. Free-lunch long video generation via layer-adaptive o.o.d correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1973–1982, June 2026.

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, 2025. URL https://arxiv.org/abs/2502.05139.

Tri Ton, Ji Woo Hong, and Chang D. Yoo. Taro: Timestep-adaptive representation alignment with onset-aware conditioning for synchronized video-to-audio synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14228–14237, October 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts, 2025a. URL https://arxiv.org/abs/2509.06155.

Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, Peter Vajda, Niraj K. Jha, and Xiaoliang Dai. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2578–2588, June 2025b.

Hongjun Wang, Lin Liu, Jianguo Li, and Tao Lin. Dual-granularity memory for efficient video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38016–38026, June 2026a.

Jiahao Wang, Hualian Sheng, Sijia Cai, Weizhan Zhang, Caixia Yan, Yachuang Feng, Bing Deng, and Jieping Ye. Echoshot: Multi-shot portrait video generation, 2025c. URL https://arxiv.org/abs/2506.15838.

Jiahao Wang, Hualian Sheng, Sijia Cai, Yuxiao Yang, Weizhan Zhang, Caixia Yan, Bing Deng, and Jieping Ye. Anyid: Ultra-fidelity universal identity-preserving video generation from any visual references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12808–12817, June 2026b.

Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with lmm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18869–18880, June 2025d.

Juncheng Wang, Chao Xu, Cheng Yu, Lei Shang, Zhe Hu, Shujun Wang, and Liefeng Bo. Synchronized video-to-audio generation via mel quantization-continuum decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3111–3120, June 2025e.

Kai Wang, Tao Zhou, Jiayi Lei, Jing Wang, Jinman Zhao, Weiguo Pian, Yuan Cheng, Yapeng Tian, Peng Gao, Bin Fu, Yihao Liu, Dimitrios Hatzinakos, and Yuewen Cao. Hear what you see: Video-to-audio generation with diffusion transformer and semantic-temporal alignment-ranked direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43396–43406, June 2026c.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. Multishotmaster: A controllable multi-shot video generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16268–16278, June 2026d.

Xihua Wang, Xin Cheng, Yuyue Wang, Ruihua Song, and Yunfeng Wang. Vaflow: Video-to-audio generation with cross-modality flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11777–11786, October 2025f.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b7bfab38ed694b43e8c20c14f6c0e900-Paper-Conference.pdf.

Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13629–13638, June 2025g.

Yujie Wei, Yujin Han, Zhekai Chen, Yongming Li, Kaixun Jiang, Zhihang Liu, Quanhao Li, Zhiwu Qing, Xiang Wang, Zhen Xing, Ruihang Chu, Lingyi Hong, Yefei He, Junjie Zhou, Junqiu Yu, Yang Shi, Difan Zou, Kai Zhu, Shiwei Zhang, Yingya Zhang, Yu Liu, Xihui Liu, and Hongming Shan. Msavbench: Towards comprehensive and reliable evaluation of multi-shot audio-video generation, 2026. URL https://arxiv.org/abs/2605.20183.

Mingyang Wu, Ashirbad Mishra, Soumik Dey, Shuo Xing, Naveen Ravipati, Hansi Wu, Binbin Li, and Zhengzhong Tu. Consid-gen: View-consistent and identity-preserving image-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1853–1863, June 2026a.

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28984–28994, June 2025a.

Xiaoxue Wu, Xinyuan Chen, Yaohui Wang, and Yu Qiao. Shotdirector: Directorially controllable multi-shot video generation with cinematographic transitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2079–2089, June 2026b.

Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models, 2026c. URL https://arxiv.org/abs/2508.11484.

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23989–24000, June 2025b.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15982–15993, October 2025.

Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Yang Zhao, Shanchuan Lin, Jiepeng Cen, Zhibei Ma, Alan Yuille, and Lu Jiang. Videoauteur: Towards long narrative video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19163–19173, October 2025.

Qingsong Xie, Zhenyi Liao, Chen Chen, Zhijie Deng, and Haonan Lu. Logcd: Local-to-global consistency distillation for few-step image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8954–8964, June 2026.

Qianxun Xu, Chenxi Song, Yujun Cai, and Chi Zhang. Switchcraft: Training-free multi-event video generation with attention controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29136–29145, June 2026a.

Xuancheng Xu, Yaning Li, Sisi You, and Bing-Kun Bao. Smrabooth: Subject and motion representation alignment for customized video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16130–16141, June 2026b.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. Stand-in: A lightweight and plug-and-play identity control for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23314–23324, June 2026.

Xin Yan, Yuxuan Cai, Qiuyue Wang, Yuan Zhou, Wenhao Huang, and Huan Yang. Long video diffusion generation with segmented cross-attention and content-rich video data curation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3184–3194, June 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation, 2025a. URL https://arxiv.org/abs/2509.22622.

Yuhang Yang, Fan Zhang, Huaijin Pi, Ailing Zeng, Shuai Guo, Guowei Xu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Gloria: Consistent character video generation via content anchors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36724–36735, June 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025b. URL https://arxiv.org/abs/2408.06072.

Zuhao Yang, Jiahui Zhang, Yingchen Yu, Shijian Lu, and Song Bai. Versatile transition generation with image-to-video diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16981–16990, October 2025c.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40256–40265, June 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet,

J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1505. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/54dcf25318f9de5a7a01f0a4125c541e-Paper-Conference.pdf.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, June 2024b.

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 (CVPR)*, pp. 22963–22974, June 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval, 2025. URL https://arxiv.org/abs/2506.03141.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988, June 2025.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model, 2026. URL https://arxiv.org/abs/2603.04379.

Shangjin Zhai, Zhichao Ye, Jialin Liu, Weijian Xie, Jiaqi Hu, Zhen Peng, Hua Xue, Danpeng Chen, Xiaomeng Wang, Lei Yang, Nan Wang, Haomin Liu, and Guofeng Zhang. Stargen: A spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26822–26833, June 2025.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1950–1960, June 2026a.

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, 2025a. URL https://arxiv.org/abs/2512.19539.

Ke Zhang, Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Endless world: Real-time 3d-aware long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18386–18395, June 2026b.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 659–669, June 2026c.

Yiheng Zhang, Zhaofan Qiu, Zunxu Liu, Yingwei Pan, Ting Yao, and Tao Mei. Evoid: Reinforced evolution for identity-preserving video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41870–41880, June 2026d.

Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magicmirror: Id-preserved video generation in video diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14464–14474, October 2025b.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4245–4257, June 2026a.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation, 2026b. URL https://arxiv.org/abs/2605.15141.

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed kv-memory for consistent long video generation, 2026c. URL https://arxiv.org/abs/2603.21366.

Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Nan Duan, and Xiangyu Zhang. Taming teacher forcing for masked autoregressive video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7374–7384, June 2025.

Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch, 2019. URL https://arxiv.org/abs/1910.10093.

Zijian Zhou, Shikun Liu, Haozhe Liu, Haonan Qiu, Zhaochong An, Weiming Ren, Zhiheng Liu, Xiaoke Huang, Kam-Woh Ng, Tian Xie, Xiao Han, Yuren Cong, Hang Li, Chuyan Zhu, Aditya Patel, Tao Xiang, and Sen He. Scaling zero-shot reference-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9253–9262, June 2026.

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, 2026a. URL https://arxiv.org/abs/2602.02214.

Lunjie Zhu, Xingtong Ge, Fangyu Lin, Yi Zhang, Zhening Liu, Mengfei Li, Yumeng Zhang, Guanglu Song, Yu Liu, and Jun Zhang. Omni-liveavatar: Minute-level real-time streaming joint audio-video avatar generation, 2026b. URL https://arxiv.org/abs/2608.13602.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Self gradient forcing: Native long video extrapolation, 2026. URL https://arxiv.org/abs/2607.20368.

APPENDIX

A Implementation Details **26**
A.1 Training configuration 26
A.2 Training targets and historical context 27
A.3 Distillation and denoising schedules 28

B Metric Implementation Details **28**

C Human Evaluation Details **29**

D Qualitative Prompt Summaries **29**

E Release, Safeguards, Assets, and Licenses **30**

F Additional Method Details and Algorithms **31**
F.1 Memory Construction and Attention Visibility 31
F.2 Teacher-Forcing Masks and Supervision 31
F.3 Two-Pass Recovery and Score Branches 31
F.4 Tail-Aware Video Fake-Score Regression 32
F.5 Training and Inference Pseudocode 33

A Implementation Details

CineForcing uses the LTX-2 audiovisual backbone and the block-causal runtime conventions of OmniForcing (Su et al., 2026). Memory-conditioned SFT adapts the pretrained backbone to memory context, followed by parallel audiovisual teacher forcing and two-pass self-rollout DMD. Table S.1 summarizes the implementation and evaluation settings. Causal TF/DMD use 97-frame targets; the quality evaluation length is listed below. The inference memory bank retains three persistent sink shots and four recent non-sink shots.

Speech-memory extraction. We use BandIt Plus through MSST with model type `bandit` and checkpoint `model_bandit_plus_dnr_sdr_11.47.chpt`. The separator extracts speech, music, and effects at 44,100 Hz, resampling the input when required. The returned speech is resampled back to the original waveform rate, trimmed or padded to the original sample count, and restored to the input channel layout before native audio encoding. The public filter returns a zero waveform when the output RMS is below 0.004.

Table S.1: **Implementation details.** Cross-shot memory settings refer to causal streaming training and inference; Memory SFT context is listed separately.

Item	Setting
Backbone	LTX-2 audiovisual backbone
LTX-2.3 baseline	Updated checkpoint from the LTX-2 model family
Streaming architecture	Block-causal LTX-2 student and runtime conventions follow OmniForcing (Su et al., 2026)
Training stages	Memory-conditioned SFT, then parallel AV teacher-forced causalization, then fixed-target two-pass self-rollout DMD
Memory SFT context	Randomly sampled memory and a prompt specifying the intended character
Teacher-forcing context	Ground-truth causal prefixes and memory extracted from completed ground-truth shots
Memory during DMD	Pre-encoded strict-past memory from completed training shots; fixed during the current target's two passes
Pass 1	No-gradient causal AV rollout of the current five-block target; save clean outputs, exit-noisy latents, and exit step; do not update cross-shot memory
Pass 2	Parallel reconstruction from detached saved values and the same fixed strict-past memory under the causal target mask
Historical K/V gradient	Detached during current-block reconstruction
Cross-shot memory gradient	Fixed pre-encoded inputs during both DMD passes
Score-model context	Real and fake branches read the same valid target window and fixed memory
Retained memory states	Causal TF/DMD: $1 \leq K \leq 7$ strict-past ground-truth shots; inference: at most seven generated shots
Memory selection at inference	First three shots as persistent sinks; four most recent non-sink shots in FIFO order
Visual memory	One sample-keyed frame per retained shot
Audio memory	Full-shot speech stem extracted by BandIt Plus through MSST
Prompt suite	100 stories, 3,000 shots, 30 shots/story
Evaluation resolution	480p (quality and efficiency)
Quality evaluation length	121 decoded frames per shot at 25 FPS (4.84 s)
Checkpoint / sampling	CineForcing interventions: `step0500`; CineForcing and OmniForcing: four steps in Table 1
Seed policy	Same story scripts and seed 42 per shot across methods
Efficiency benchmark	CineForcing measured latency: 4.557 s per clip; throughput: 1.0972 video seconds per wall-clock second
Release assets	Code, checkpoints, and reproduction instructions are planned for release

Inference-time memory intervention. The implementation replaces the conditioned video-memory latents, audio-memory latents, or both with same-shaped zero tensors before generation. Slot metadata and positional coordinates are retained; the intervention changes memory content rather than removing tokens or applying an attention mask.

A.1 Training configuration

The pretrained audiovisual backbone first undergoes Memory SFT with randomly sampled memory context and prompt-specified character identity. The resulting memory-conditioned transformer

initializes causal teacher forcing. The TF5K continuation described here starts from this transformer and an existing 1000-update teacher-forcing LoRA adapter, with a fresh optimizer and a reset step counter. The selected TF5K adapter contains 5000 additional updates and initializes the DMD generator. DMD initializes a separate fake-score adapter. The 1000 inherited updates and 5000 continuation updates both belong to causal teacher forcing; neither count denotes the preceding Memory SFT stage. Table S.2 summarizes the optimizer and adaptation settings for these two stages.

Table S.2: Hyperparameters of the TF5K training configuration. DMD updates the generator and fake-score critic once each per global step, using separate batches. The distributed DMD configuration specifies 32 workers.

Setting	Teacher-forcing continuation	DMD
Optimizer	AdamW	Separate AdamW optimizers
Learning rate	2×10^{-5}	2×10^{-5} for both branches
Adam (β_1, β_2)	$(0.9, 0.999)$	$(0.85, 0.999)$
Weight decay	0.01	0.01
Gradient clipping norm	1.0	3.0
Learning-rate schedule	100 warmup updates, then constant	Constant, no warmup
Per-worker batch / accumulation	$1/1$	$1/1$ per branch
Batch per optimizer update	32	32 as configured
Selected checkpoint updates	5000 additional updates	500 or 1000, by variant
LoRA rank / alpha / dropout	$256/256/0$	$256/256/0$ per adapter
Computation precision	BF16	BF16; FP32 FSDP reductions
Activation checkpointing	Enabled	Enabled
Video / audio loss weights	$1/1$	$1/1$
Random seed	42	42; rank-local seed $42 + r$
Checkpoint interval	1000 updates	100 global steps

Here r denotes the zero-based distributed rank. In these TF and DMD configurations, the backbone remains frozen while LoRA adapts attention query, key, value, and output projections. The DMD adapters cover video and audio self-attention, their text cross-attention, and both audiovisual cross-attention directions. The fake-score branch enables the critic adapter; the frozen real-score branch uses the same backbone with that adapter disabled. DMD disables generator exponential moving averaging.

A.2 Training targets and historical context

Each causal TF/DMD training target contains 97 decoded video frames at 25 FPS and 1280×736 resolution (approximately 3.88 s). Encoding produces 13 video-latent frames and 98 audio-latent frames. Each video-latent frame has a 23×40 spatial grid, or 920 tokens. The five video blocks have temporal lengths $1|3|3|3|3$ in latent frames; their audio counterparts have lengths $2|24|24|24|24$. The configured probability of conditioning on the current shot's first frame is 0.6. The 480p efficiency setting is separate from these training targets. Teacher forcing additionally uses an empty-prompt probability of 0.1 and shifted logit-normal timestep sampling with maximum shift 12 and 1000 training timesteps. Video and audio targets receive equal loss weights.

For causal TF and DMD, historical context contains one to seven distinct, strictly preceding ground-truth shots. Teacher forcing samples memory-count buckets in proportion to their sizes. Prepared DMD inputs retain the last available K preceding shots in chronological order, where K is the number of memory slots. Each slot pairs one visual-latent frame with the full speech-filtered audio of that historical shot. The visual frame is selected deterministically from the sample key. Speech separation uses MSST/Bandit with a 44.1 kHz separator input; the teacher-forcing configuration uses a minimum speech-mask ratio of 0.02, minimum root-mean-square amplitude of 0.004, and zero background gain. Memory tokens precede the current target and use slot-centered positions with offset 500 and inter-slot stride 50. Teacher forcing uses clean historical memory, with memory noise, degradation, and dropout disabled; current-shot context noise is also zero.

A.3 DISTILLATION AND DENOISING SCHEDULES

For DMD score evaluation, an integer timestep is sampled uniformly from $\{20, \ldots, 950\}$ and mapped to a noise level by the native LTX scheduler with maximum shift 2.05. The real-score classifier-free guidance scales are 3 for video and 4 for audio; spatiotemporal guidance is disabled. Each rank uniformly samples a rollout exit step. A first pass generates the current five-block target without gradients; a second pass reconstructs gradients at the sampled exit from saved, detached values. Historical keys and values remain detached. Both passes condition on the same fixed, pre-encoded ground-truth memory from strictly preceding shots.

The eight-step schedule uses timestep indices $[1000, 994, 988, 981, 975, 909, 725, 422]$ and noise levels

$$\boldsymbol{\sigma}_8 = [1, 0.99375, 0.9875, 0.98125, 0.975, \\ 0.909375, 0.725, 0.421875, 0].$$

The four-step schedule uses indices $[1000, 750, 500, 250]$ and

$$\boldsymbol{\sigma}_4 = [1, 0.9626806624, 0.8965547058, 0.7473050411, 0].$$

Here $\boldsymbol{\sigma}_8 \in \mathbb{R}^9$ and $\boldsymbol{\sigma}_4 \in \mathbb{R}^5$ list the scalar noise levels in descending denoising order, including the terminal zero. This endpoint does not require another denoiser evaluation. Four and eight therefore count denoising evaluations per generated block, rather than total evaluations per shot. The corresponding demonstration configurations use student classifier-free guidance 1 and spatiotemporal guidance 0.

B METRIC IMPLEMENTATION DETAILS

We evaluate cross-shot consistency, visual quality, text alignment, and speech fidelity, following the evaluation categories described by Duan et al. (2026). The definitions below apply to Table 1; SpeechAcc is WER-based, not word-level recall.

Cross-shot visual consistency. ViCLIP measures the cosine similarity between shot-level video embeddings, averaged over valid within-story shot pairs. For ViCLIP late, shot 1 is paired with each shot from 16 through 30; these 15 cosine similarities are averaged within each story, then across evaluated stories. Self-CIDS compares recurring roles across shots: GroundingDINO detects characters, Torchreid provides body features, and FaceNet provides face features. Comparisons use shots in which the corresponding role is visible. Detection and feature quality therefore affect this score.

Acoustic consistency. For shots i, j with CAMPPlus embeddings $e_i, e_j \in \mathbb{R}^{d_{\text{spk}}}$, where d_{spk} is the embedding width, the reported similarity is

$$S_{ij}^{\text{voice}} = \frac{1}{2}\left(\frac{e_i^\top e_j}{\|e_i\|_2 \|e_j\|_2} + 1\right).$$

Voice averages these similarities across evaluated shot pairs. The inputs are whole-shot audio, rather than isolated utterances of a matched speaker; background sounds and speaker changes can consequently influence the score. This evaluation embedding is separate from the BandIt Plus extraction used to construct the generator's memory.

Quality and audiovisual alignment. Aesthetic and imaging quality use VBench/VBench++, while CLIP measures alignment with the corresponding text. AudioBox-Aesthetics evaluates audio quality, CLAP compares text and audio, and ImageBind and CAVP compare visual and acoustic content. DeSync uses Synchformer to estimate audiovisual temporal offsets on approximately 4-second audiovisual segments, aggregated across shots; lower is better. DeSync@4s in the inference-time memory ablation uses this same metric and protocol as DeSync in Table 1. These metrics complement cross-shot similarity by assessing the quality and alignment of each generated output.

Scripted speech. Whisper base.en transcribes each dialogue-containing shot. Given substitution, deletion, and insertion counts $n_{\text{sub}}, n_{\text{del}}, n_{\text{ins}}$ and $n_{\text{ref}} > 0$ reference words, $\text{WER} = (n_{\text{sub}} + n_{\text{del}} + n_{\text{ins}})/n_{\text{ref}}$. We report $\max(0, 1 - \text{WER})$, averaged over the evaluated dialogue shots. Shots without reference dialogue do not contribute to SpeechAcc.

C Human Evaluation Details

All ten participants evaluated the same 100 comparison sets. Each set contained three candidate videos from the original LTX-2 baseline, video-only memory, and CineForcing, shown in randomized order with method names hidden. Participants could replay the videos and answered two single-choice questions: (1) Which output best preserves character appearance across shots? (2) Which output best preserves voice or timbre across shots? For each question, they chose one of the three candidates or the "neither / both poor" option when none was satisfactory. Figure 5 reports each option's share of the 1,000 responses to the corresponding question (10 participants $\times$ 100 sets), including the neither/both-poor responses. The two questions yield 2,000 selections in total. Only anonymous preference choices enter the analysis.

D Qualitative Prompt Summaries

Table S.3 summarizes the story-level prompts in our qualitative suite, including the rally and home-routine examples shown in Figure 4. For multi-shot generation, we use GPT-5 to expand each story-level prompt into 30 shot-level prompts with explicit visual, dialogue, action, emotion, camera, background, sound-effect, and music conditions.

Table S.3: **Qualitative prompt summaries.** Story-level prompts used for the qualitative rollouts.

Short label	Prompt summary
Calm domestic home routine	A middle-aged man and an early-middle-aged woman move through a warm modern home across a full day, cooking, reading, setting the table, dining, and stepping outside at night. The prompt stresses stable visual identity across rooms and time, warm male narration, gentle female speech, quiet voiceover, piano ambience, domestic sound effects, and a calm reflective emotional arc.
Public campaign rally speech	A middle-aged male speaker in a navy suit delivers a long campaign-style public address at a large outdoor rally, alternating podium shots, crowd cutaways, media views, applause, and chant responses. The prompt emphasizes a resonant baritone speaker identity, visible lip motion, sustained rhetorical delivery, and crowd audio that evolves from attentive reaction to collective movement energy.
Intimate room diary monologue	A young woman records a quiet evening diary-like monologue in a softly lit bedroom or study, moving between a curtain, desk lamp, phone, old photo, notebook, and writing desk. The sequence tests long-range preservation of a single female speaker and visual identity while the emotional arc moves from uncertainty and memory toward a concrete decision to begin again.
Warehouse confrontation and chase	A Chinese woman confronts a Chinese man at gunpoint in a dim warehouse, forcing a tense truth-revealing standoff that transitions into corridors, an exterior yard chase, and a car escape. The prompt combines two speaking identities, industrial lighting, emergency-light intrusions, tactical blocking, restrained anger, wary male replies, and a shift from static confrontation to kinetic pursuit.
Masked vigilante rooftop rescue	A masked vigilante superhero patrols a stormy city rooftop, intercepts a dockside hostage convoy, chases criminals through rain-soaked industrial roads, rescues captives near a warehouse, and ends at dawn after a tense alliance with another masked figure. The prompt stresses armored visual identity, dark rain-soaked action continuity, comm-channel support, low gravelly male vocal identity or internal narration, and long-horizon action staging.

E RELEASE, SAFEGUARDS, ASSETS, AND LICENSES

We plan to release CineForcing code and model checkpoints with reproduction instructions and intended-use notes. The training and evaluation data are not released. Because long-form audiovisual generation can be misused for synthetic media impersonation or disinformation, the release package should include provenance or watermarking guidance, use restrictions, and access controls when appropriate. We also avoid releasing unreleased evaluation data or identifiable source assets as part of the public package.

F Additional Method Details and Algorithms

F.1 Memory Construction and Attention Visibility

Latent dimensions in Section 3.3 precede tokenization; memory widths d_m describe token features. For one attention head, $Q \in \mathbb{R}^{N_q \times d_h}$ and $K_{\text{att}}, V_{\text{att}} \in \mathbb{R}^{N_k \times d_h}$ are query, key, and value matrices, where N_q, N_k count tokens and d_h is the head width. A visibility mask has shape $N_q \times N_k$, with zero for permitted pairs and $-\infty$ otherwise before softmax. H^{KV} collects the retained key/value matrices across layers, heads, and attention paths; its superscript distinguishes this cache from the latent height H. K_{att} is distinct from the memory count K.

For the causal streaming stages, the prepared memory cache stores a sample-keyed visual frame and the full available speech stem extracted by BandIt Plus through MSST from each selected shot. Audio length is variable. The two modalities share shot provenance but need not represent the same timestamp. Their encoders are the model's latent encoding and tokenization paths rather than external identity encoders.

The DMD memory cache selects the last K distinct strict-past ground-truth shots with $1 \leq K \leq 7$ and orders them from oldest to newest. Video and audio positions are centered separately at each slot's coordinate. At inference, $\mathcal{U}_{\text{mem}}$ preserves the first three completed shots and keeps the latest four non-sink shots in a FIFO queue, ordered chronologically. This retains at most seven distinct shots without duplicating sink entries.

In the causal streaming stages, memory queries read memory only. During causal student recovery, target queries read all retained memory and the permitted causal target prefix. During score evaluation, target queries read all memory and the complete valid target window. These rules apply across the audiovisual attention paths; there is no additional mask restricting cross-modal memory attention to tokens from the same slot. Same-shot pairing is a property of memory construction, not a timestamp-aligned attention constraint.

F.2 Teacher-Forcing Masks and Supervision

Let $B_{b,\text{clean}}$ and $B_{b,\text{noisy}}$ be the token sets for the clean and noisy copies of audiovisual block b. $\mathcal{S}(B)$ lists the target blocks visible to queries in B, and j indexes key blocks. In addition to clean strict-past committed memory, the visible target sets are

$$\mathcal{S}(B_{b,\text{clean}}) = \{B_{j,\text{clean}} : j \leq b\},$$
$$\mathcal{S}(B_{b,\text{noisy}}) = \{B_{j,\text{clean}} : j < b\} \cup \{B_{b,\text{noisy}}\}.$$

The same temporal boundary governs video self-attention, audio self-attention, video-to-audio attention, and audio-to-video attention. Each video/audio block pair shares σ_b, with modality noise $\epsilon_b^m \sim \mathcal{N}(0, I)$. Clean context and committed memory are excluded from the flow-matching prediction loss. In image-conditioned generation, video frame 0 is also clean conditioning and receives no prediction loss. Audio has no analogous clean first frame, so all valid target audio tokens remain supervised.

In the flow objective, I is the identity covariance after flattening each modality, $\odot$ denotes elementwise multiplication, and $\|\cdot\|_2^2$ sums squared latent elements. The binary W_b^m is broadcast to the prediction shape. Noise ϵ_b^m has the same shape as the clean latent block.

Algorithm S.1 summarizes teacher forcing over a ground-truth shot sequence: shot s reads only earlier-shot memory, and $F(x_s)$ is committed after supervision.

F.3 Two-Pass Recovery and Score Branches

The DMD target comprises five paired blocks, with 13 video-latent frames and 98 audio-latent frames in total. Pass 1 samples one exit step t_e and saves each block's final clean output and exit-noisy latent. It rolls out only the current target without gradients, using pre-encoded strict-past ground-truth memory. It neither generates another shot nor updates cross-shot memory.

Pass 2 packs saved clean prefixes and exit-noisy values for parallel reconstruction under the causal recovery mask. Historical K/V produced by re-encoding is detached before conditioning reconstructed

blocks, so that the generator update affects current-block reconstruction only. Later-block losses therefore do not backpropagate through historical K/V construction. Saved rollout values and cross-shot memory inputs are also detached, as shown in Section 3.3.

The logical real-score branch uses the frozen backbone without the critic adapter, whereas the fake-score branch enables its trainable adapter:

$$s_{\text{real}} = D_{\phi=0}, \qquad s_{\text{fake}} = D_{\phi}.$$

Here D_ϕ is the score-branch operator, ϕ denotes trainable critic-adapter parameters, and $\phi = 0$ disables that adapter. Both logical branches receive the same valid audiovisual target window and are bidirectional within it. The fake-score update uses detached generator samples, separately from the generator update. Algorithm S.2 summarizes these alternating updates.

Distribution-matching objective. We use the score-difference formulation of DMD (Yin et al., 2024b;a) on the reconstructed audiovisual target $\bar{x}_s$. Its reconstruction follows the two-pass view of self-forcing described by SGF (Zhuang et al., 2026), with historical K/V detached as specified above. Let $\mathcal{C}_s = (c_s, \text{sg}(M_{s-1}^{\text{AV,gt}}))$ be the condition shared by the score branches. At noise level $\sigma \in (0, 1)$, form $y_\sigma = (1-\sigma)\bar{x}_s + \sigma\epsilon$, where $\bar{x}_s$ concatenates the reconstructed blocks temporally and ϵ is same-shaped standard Gaussian noise in each modality. Denote the conditional real and fake score estimates by $s_{\text{real}}(y_\sigma, \sigma \mid \mathcal{C}_s)$ and $s_{\text{fake}}(y_\sigma, \sigma \mid \mathcal{C}_s)$. The distribution-matching gradient has the form

$$\nabla_\theta \mathcal{L}_{\text{DMD}} \simeq \mathbb{E}\left[\sum_{m\in\{v,a\}} \lambda_m \left(\frac{\partial \bar{x}_s^m}{\partial \theta}\right)^\top \text{sg}(\Delta_\sigma^m)\right], \qquad \Delta_\sigma^m = w_m(\sigma)(1-\sigma)\big(s_{\text{fake}}^m - s_{\text{real}}^m\big). \quad (5)$$

The scores in Eq. 5 are evaluated at the same noisy joint target and condition, with the same modality shape as $\bar{x}_s^m$. Δ_σ^m is the detached update signal in these latent coordinates. For the Jacobian, flatten each modality to D_m elements and the trainable student parameters to P elements, giving size $D_m \times P$. The scalar weights $w_m(\sigma)$ collect noise-level weighting and gradient normalization; the expectation is over rollout samples, score noise levels, and Gaussian noise. The sign is fake minus real for gradient descent. For rectified-flow velocity prediction, a branch's clean estimate is $\hat{x}_0 = y_\sigma - \sigma v(y_\sigma, \sigma)$, where v is that branch's velocity prediction (not decoded video), giving the score $s_{\text{score}} = -(y_\sigma + (1-\sigma)v)/\sigma$. The subscript 0 denotes zero noise. Thus velocity outputs are converted to the same score convention before taking their difference.

A stop-gradient surrogate makes the gradient boundary explicit:

$$\mathcal{L}_{\text{DMD}}^{\text{sur}} = \frac{1}{2}\sum_{m\in\{v,a\}} \frac{\lambda_m}{N_m} \left\| W^m \odot [\bar{x}_s^m - \text{sg}(\bar{x}_s^m - \Delta_\sigma^m)] \right\|_2^2, \quad (6)$$

where the binary mask W^m has the shape of $\bar{x}_s^m$ after broadcasting, and $N_m = \sum_i W_i^m$ counts valid scalar elements (the index i traverses the flattened tensor); masking and mean reduction are applied to the gradient in Eq. 5 accordingly. The score signal is detached in the generator update. The fake branch is trained separately on detached student outputs with rectified-flow targets $\epsilon - \text{sg}(\bar{x}_s)$, while the real branch remains frozen.

F.4 Tail-Aware Video Fake-Score Regression

The fake-score branch regresses $\epsilon - u$, where u is a detached clean student sample from its critic-update rollout and ϵ is the noise used to corrupt it. Let $e_j^v \geq 0$ be the channel-averaged squared velocity error for video token j, and let $\mathcal{A}$ contain the valid, supervised video tokens. Define the selected token count k and the index set $\mathcal{I}_k$ of the k largest errors as

$$k = \max(1, \lceil \rho |\mathcal{A}| \rceil), \qquad \mathcal{I}_k = \underset{j\in\mathcal{A}}{\overset{k}{\text{arg topk}}}\, e_j^v.$$

The video fake-score loss is

$$\mathcal{L}_{\text{fake}}^v = (1-\beta)\,\text{Mean}_{j\in\mathcal{A}}\, e_j^v + \beta\,\text{Mean}_{j\in\mathcal{I}_k}\, e_j^v, \qquad \rho = 0.001, \quad \beta = 0.01.$$

Here ρ is the selected fraction, β is the tail-loss mixing weight, and Mean is the arithmetic average over the indexed set. Only the video fake-score reduction changes; audio retains mean regression, and the generator's DMD loss is unchanged.

F.5 Training and Inference Pseudocode

Memory SFT initializes the student before the algorithms below: causal teacher forcing adapts this checkpoint, DMD continues the teacher-forced student on a current target with fixed precomputed memory, and inference commits generated-shot memory. They retain separate intra-shot H^{KV} and cross-shot M^{AV} states.

Algorithm S.1 Parallel Audiovisual Teacher-Forced Causalization

Require: Shot sequence $\{(c_s, x_s)\}_{s=1}^{S}$, block-causal student G_θ initialized from Memory SFT, memory operator F, update $\mathcal{U}_{\mathrm{mem}}$

Initialize ground-truth committed memory $M_0^{\mathrm{AV,gt}} \leftarrow \emptyset$
for $s = 1$ to S **do**
    Split the video/audio latents of x_s into aligned macro-chunks and sample shared σ_b for each pair
    Pack clean/noisy copies and apply the four audiovisual masks in Appendix F.2
    Predict all noisy chunks conditioned on c_s and clean $M_{s-1}^{\mathrm{AV,gt}}$
    Exclude clean context and memory from prediction loss; in image-conditioned mode also exclude video frame 0, but supervise all valid target audio tokens
    Update G_θ with $\mathcal{L}_{\mathrm{TF}}$
    $M_s^{\mathrm{AV,gt}} \leftarrow \mathcal{U}_{\mathrm{mem}}(M_{s-1}^{\mathrm{AV,gt}}, F(x_s))$
end for

Algorithm S.2 Two-Pass Self-Rollout DMD with Tail-Aware Video Critic

Require: Current-shot condition c_s, fixed strict-past memory $M_{s-1}^{\mathrm{AV,gt}}$, teacher-forced student G_θ, logical score branches $D_{\phi=0}$ and D_ϕ, $\mathcal{U}_{\mathrm{KV}}$

Pass 1: current-shot self rollout (no gradient)

Initialize $H_{s,1}^{\mathrm{KV}} \leftarrow \emptyset$ and sample one exit step t_e
for $b = 1$ to 5 paired target blocks **do**
    Generate $\hat{x}_{s,b}$ conditioned on c_s, $\mathrm{sg}(M_{s-1}^{\mathrm{AV,gt}})$, and $H_{s,b}^{\mathrm{KV}}$; save $(r_{s,b}, \hat{x}_{s,b}, t_e)$
    $H_{s,b+1}^{\mathrm{KV}} \leftarrow \mathcal{U}_{\mathrm{KV}}(H_{s,b}^{\mathrm{KV}}, \hat{x}_{s,b})$
end for
▷ The current DMD path does not generate another shot or update M^{AV}.

Pass 2: differentiable reconstruction and generator update

Reconstruct $\bar{x}_s \leftarrow G_\theta(\mathrm{sg}(r_s), t_e \mid c_s, \mathrm{sg}(\widetilde{H}_s^{\mathrm{KV}}), \mathrm{sg}(M_{s-1}^{\mathrm{AV,gt}}))$ using recovery masks that detach clean-history K/V during the forward
Re-noise the fixed target window; memory queries read memory, and target queries read all memory and the complete target window
Evaluate $D_{\phi=0}$ and D_ϕ on the same input and accumulate $\mathcal{L}_{\mathrm{DMD}}$
Update G_θ using $\mathcal{L}_{\mathrm{DMD}}$

Alternating fake-score update

Obtain detached fake samples with the critic-update rollout; re-noise them and compute per-token regression errors
Update D_ϕ with tail-aware video loss in Appendix F.4 and mean audio loss

Algorithm S.3 Streaming Rollout with Two-Level State

Require: Streaming conditions $c_1, \ldots, c_S$, student G_θ, $F, \mathcal{U}_{\mathrm{KV}}, \mathcal{U}_{\mathrm{mem}}$

Initialize committed memory $M_0^{\mathrm{AV}} \leftarrow \emptyset$
for $s = 1$ to S **do**
  Read committed history M_{s-1}^{AV} and initialize $H_{s,1}^{\mathrm{KV}} \leftarrow \emptyset$
  for $b = 1$ to B_s **do**
    Generate $\hat{x}_{s,b} \leftarrow G_\theta(c_s, M_{s-1}^{\mathrm{AV}}, H_{s,b}^{\mathrm{KV}})$
    Emit the decoded audiovisual block $\hat{x}_{s,b}$
    $H_{s,b+1}^{\mathrm{KV}} \leftarrow \mathcal{U}_{\mathrm{KV}}(H_{s,b}^{\mathrm{KV}}, \hat{x}_{s,b})$
  end for
  Assemble generated shot $\hat{x}_s \leftarrow \{\hat{x}_{s,1}, \ldots, \hat{x}_{s,B_s}\}$
  Commit $M_s^{\mathrm{AV}} \leftarrow \mathcal{U}_{\mathrm{mem}}(M_{s-1}^{\mathrm{AV}}, F(\hat{x}_s))$; preserve the first three shots and the latest four non-sink shots
end for

Ensure: Audiovisual stream $\{\hat{x}_s\}_{s=1}^{S}$