STORYWEAVER: TOWARDS CONSISTENT AND SYNCHRONIZED MULTI-SHOT AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

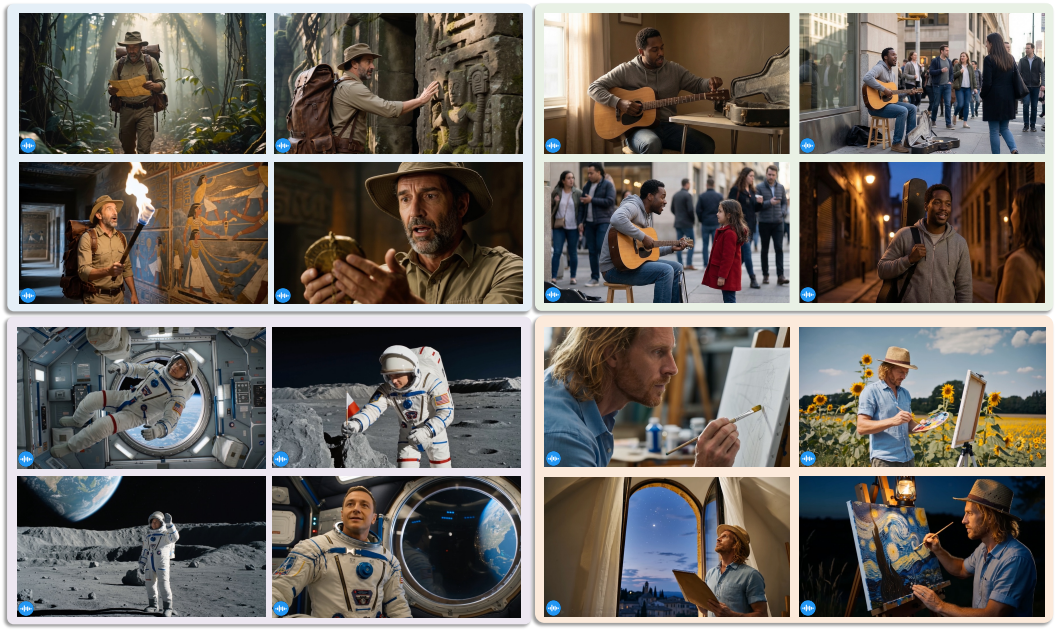

Figure 1: From text prompts, STORYWEAVER generates multi-shot video narratives with synchronized audio. Each story maintains consistent character identity and acoustic continuity across multiple shots and camera perspectives. Audio icons indicate synchronized audio generation.

ABSTRACT

Generating multi-shot videos with synchronized audio remains a significant challenge. While single-shot models perform well, extending them to multi-shot story generation is hindered by consistency degradation across shots. Current approaches typically rely on decoupled paradigms that generate segments independently, inherently limiting long-range coherence and causing identities, voices, and backgrounds to drift over time. Furthermore, existing continuous story frameworks often lack synchronized audio synthesis. To address these limitations, we introduce STORYWEAVER, a framework for consistent multi-shot audio-video generation. Instead of decoupled steps, it adopts a joint generation paradigm conditioned on text and an initial prompt to improve global coherence. To model audio-visual interactions across extended contexts, we propose the Multi-Modal Sink Shot (MSS) mechanism, which couples windowed audio-visual cross-attention for local synchronization with sink-mediated inter-shot self-attention for cross-shot consistency. To extend generation beyond a single denoising window, we introduce Audio-Visual Latent Stitching (AVLS), a train-test matched continuation scheme that operates entirely in the joint audio-video latent space, reducing identity and acoustic drift compared to decode-reencode approaches. Experiments on both video-only and audio-visual benchmarks, supported by audio-specific ablations, adjacent-shot transition metrics, and human preference studies, show that STORY-

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.

WEAVER achieves improved visual-acoustic consistency, character preservation, and temporal coherence over existing methods.

1 INTRODUCTION

The rapid evolution of generative models has fundamentally transformed visual content creation. While recent advancements have propelled video generation to impressive levels of visual fidelity, these efforts predominantly focus on single-shot scenarios (Yang et al., 2024; Liu et al., 2025). However, practical storytelling applications, such as digital filmmaking, automated animation, and long-form content production, inherently demand continuous multi-shot narratives accompanied by synchronized audio. We define this practical yet challenging task as *continuous multi-shot audio-video generation*. Given a text context and an initial frame prompt, it requires synthesizing a coherent sequence of shots where characters, backgrounds, and acoustic features maintain temporal consistency over an extended duration.

Despite this rapid progress, extending single-shot generation capabilities to multi-shot story synthesis is hindered by consistency degradation. Current approaches typically rely on decoupled paradigms that generate individual video segments independently (Chen et al., 2024; Wu et al., 2025; Hu et al., 2024). This separation inherently limits long-range coherence, causing character identities and background details to degrade as the narrative progresses. Furthermore, existing continuous story frameworks are predominantly unimodal (Zhou et al., 2024; Wang et al., 2025b), lacking the capacity for synchronized audio synthesis. Consequently, there remains a critical gap for a unified framework capable of harmonizing visual and acoustic consistency across multiple sequential shots. Compared to video-only multi-shot generation, the audio-visual setting introduces additional challenges: acoustic leakage across shot boundaries is perceptually salient, voice timbre drift is readily noticeable, lip synchronization requires fine-grained temporal alignment, and decode-reencode continuation disrupts both visual and audio streams simultaneously.

To address these limitations, we introduce STORYWEAVER, a joint latent-space audio-video diffusion framework for consistent multi-shot generation. Instead of relying on decoupled generation steps, STORYWEAVER employs a joint paradigm conditioned on textual inputs and an initial visual prompt to improve global coherence throughout the sequence. As shown in Fig. 1, by processing the extended context as a unified representation, the model synchronizes audio and visual elements across continuous shots, maintaining character identity and acoustic continuity across diverse scenarios.

The core of our approach relies on specific architectural designs to handle long-range audio-visual dependencies. We propose the Multi-Modal Sink Shot (MSS) mechanism, which couples windowed audio-visual cross-attention with sink-mediated inter-shot self-attention. This combination enables the model to maintain local audio-visual synchronization while preserving cross-shot identity and acoustic consistency through compact sink token summaries, without the quadratic cost of dense global attention. To extend generation beyond a single denoising window, we develop Audio-Visual Latent Stitching (AVLS), which operates entirely in the joint audio-video latent space: it caches overlapping latent context and clamps it with time-dependent noise at each integration step, avoiding the error accumulation of decode-reencode loops. Crucially, AVLS trains under the same stitched-context interface used at inference, closing the train-test gap that causes drift in standard continuation approaches.

Our main contributions are summarized as follows:

- We formalize the task of continuous multi-shot audio-video generation and construct Weaver-Bench, an evaluation benchmark covering 20 characters across 10 scenarios with metrics for identity, voice, synchronization, and transition quality.
- We propose the Multi-Modal Sink Shot mechanism, a cut-aware dual-stream attention design that uses local audio-visual windows for temporal synchronization and sink-mediated routing for cross-shot identity and acoustic consistency.
- We introduce Audio-Visual Latent Stitching, a train-test matched continuation scheme in the joint audio-video latent space that reduces identity and acoustic drift compared to decode-reencode or clean-prefix approaches.

- We provide comprehensive evaluation including audio-specific ablations, adjacent-shot transition metrics, human preference studies, and supplementary audio-video samples, demonstrating improved performance over existing methods.

2 Related Work

Single-Shot Video Generation. Current single-shot text-to-video (T2V) generation models have achieved highly realistic and temporally consistent synthesis. While initial efforts adapted GANs for video (Abdullah et al., 2024; Chu et al., 2020; Skorokhodov et al., 2022), the field progressed significantly through diffusion models (Blattmann et al., 2023; Chen et al., 2024; Kong et al., 2024; Wan et al., 2025; Yang et al., 2024). Early diffusion-based architectures relied on 3D U-Nets and cascaded generation (Ho et al., 2022; Singer et al., 2022). More recently, Diffusion Transformers (DiTs) (Peebles & Xie, 2023) have become dominant. DiT-based frameworks like Seedance (Gao et al., 2025; Seedance et al., 2025) and open-source models (Kong et al., 2024; Wan et al., 2025; Yang et al., 2024) successfully scale to high-resolution clips. Despite this fidelity in isolated scenes, existing architectures remain confined to single-shot outputs, lacking the structural design to connect multiple shots into cohesive narratives.

Multi-Shot Video Generation. Current literature primarily explores three strategies. Hierarchical frameworks (Bansal et al., 2024; Hu et al., 2024; Long et al., 2024; Wang et al., 2025b; Xie et al., 2024; Zhao et al., 2024) utilize language models to segment narratives into individual prompts for independent shot generation, but the fundamental independence of generation often compromises temporal continuity. Keyframe-centric approaches (Zhou et al., 2024; Huang et al., 2024a; Xiao et al., 2025) generate anchor frames prior to interpolating intermediate video, ensuring alignment at specific timestamps but leaving inter-keyframe synthesis disjointed. Holistic modeling (Esser et al., 2024; Wu et al., 2025; Kara et al., 2025) jointly processes multiple shots: LCT (Guo et al., 2025) introduces interleaved positional embeddings, HoloCine (Meng et al., 2025) develops persistent character memory. These methods are video-only and do not address synchronized audio generation.

Joint Audio-Video Generation. The research community has increasingly focused on end-to-end joint audio-video synthesis (Wang et al., 2025a; Low et al., 2025; Hu et al., 2025; Zhang et al., 2025a; Weng et al., 2025; Liu et al., 2025; Ruan et al., 2023). Methods like MMDisCo (Hayakawa et al., 2024) align pretrained models via cooperative diffusion, yet their reliance on adversarial guidance limits scalability. Dual-stream networks such as JavisDiT (Liu et al., 2025) and MM-Diffusion (Ruan et al., 2023) utilize cross-modal attention but primarily generate ambient audio. MTV (Weng et al., 2025) isolates speech, music, and effects for fine-grained control. Dual-tower designs with rotary positional embeddings (Low et al., 2025; Hu et al., 2025; Zhang et al., 2025a) produce synchronized speech, extended by LTX-2 (HaCohen et al., 2026) and MOVA (Team et al., 2026) to diverse datasets. However, these methods operate in the single-shot regime and do not address multi-shot narrative coherence.

3 Method

We present StoryWeaver, an audio-visual diffusion framework that generates synchronized multi-shot stories as joint latent samples. The model uses flow matching to capture the joint distribution of video and audio latents conditioned on layered text descriptions. We address the temporal alignment requirements and cross-shot blurring inherent in long sequence generation through two complementary mechanisms: the Multi-Modal Sink Shot (MSS) for structured audio-visual attention, and Audio-Visual Latent Stitching (AVLS) for drift-resistant long-form continuation.

3.1 Preliminary

Audio-Visual Diffusion Model. To generate temporally aligned video and audio streams from text descriptions and optional starting frames, current diffusion paradigms operate in a compressed space. Raw video clips V and their corresponding audio waveforms A are transformed into latent codes x^v and x^a using independently pretrained video (Wan et al., 2025) and audio (Shan et al., 2025) autoencoders. Under the flow matching framework, we draw $t \sim \mathcal{U}(0, 1)$ and sample $\epsilon^v, \epsilon^a \sim$

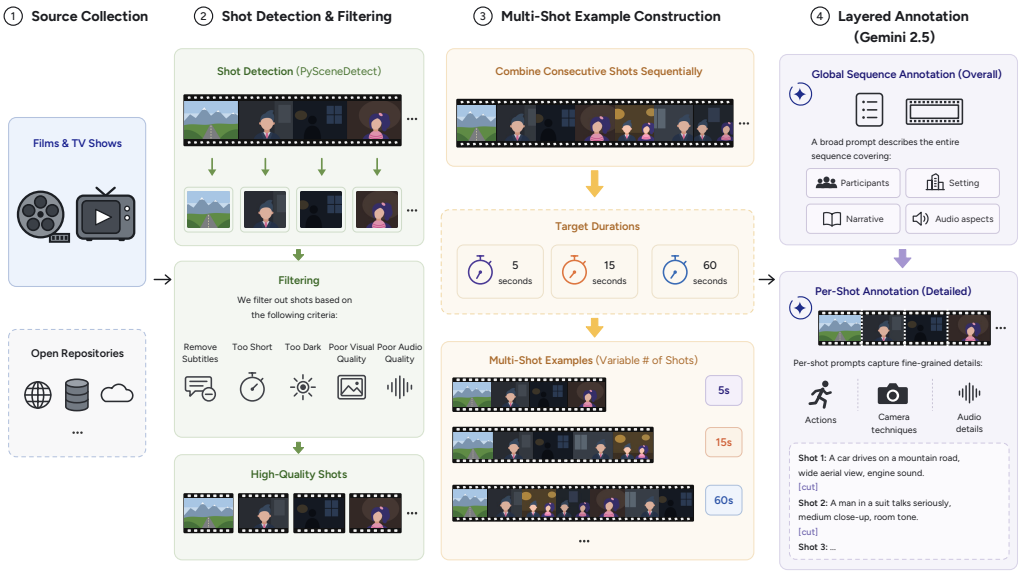

Figure 2: **Data curation workflow.** The pipeline transforms raw films and TV shows into a structured dataset for multi-shot audio-video generation.

$\mathcal{N}(0, I)$. The forward paths are:

$$x_t^v = (1-t)x^v + t\epsilon^v, \quad x_t^a = (1-t)x^a + t\epsilon^a. \tag{1}$$

The target marginal vector fields are $v_t^v = \epsilon^v - x^v$ and $v_t^a = \epsilon^a - x^a$. The network minimizes:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,c,x^v,x^a,\epsilon}\left[\lambda_v \left\|\hat{v}_\theta^v(x_t^v, x_t^a, t, c) - v_t^v\right\|_2^2 + \lambda_a \left\|\hat{v}_\theta^a(x_t^v, x_t^a, t, c) - v_t^a\right\|_2^2\right]. \tag{2}$$

3.2 Data Curation

A key challenge in multi-shot audio-video generation is the scarcity of high-quality datasets with both modalities. We construct a data processing pipeline (shown in Fig. 2) to transform films and TV shows into structured multi-shot training data.

Shot Detection and Selection. We gather film material from online video platforms and apply shot detection (Castellano, 2024) to segment each video. Segments pass through strict filtering: we eliminate subtitles via SubtitleEdit (Bergström, 2024) and exclude clips that are excessively brief, dimly lit, of poor visual quality, or with inadequate audio.

Multi-Shot Example Construction. We combine consecutive shots sequentially, targeting durations of 5, 15, or 60 seconds. Our resulting dataset comprises 350k examples with a managed spread of shot numbers across duration categories.

Layered Annotation. Each multi-shot example receives layered annotations from Gemini 2.5 (Comanici et al., 2025). A broad prompt outlines the entire sequence covering participants, setting, narrative, and audio aspects. Per-shot prompts specify actions, camera techniques, and audio details (Radford et al., 2023). A distinct `[cut]` marker separates per-shot descriptions.

3.3 Multi-Modal Sink Shot

StoryWeaver generates a multi-shot story as a single joint sample in the audio–visual latent space, rather than producing shots independently. As shown in Fig. 3, given N_{shots} shots, the denoising network receives the concatenation of video latents $\{x_{t,i}^v\}_{i=1}^{N_{\text{shots}}}$ and audio latents $\{x_{t,i}^a\}_{i=1}^{N_{\text{shots}}}$ at each step t, together with text conditioning and an optional initial visual prompt. Multi-shot audio–visual synthesis introduces two practical tensions: audio tokens are temporally denser than video

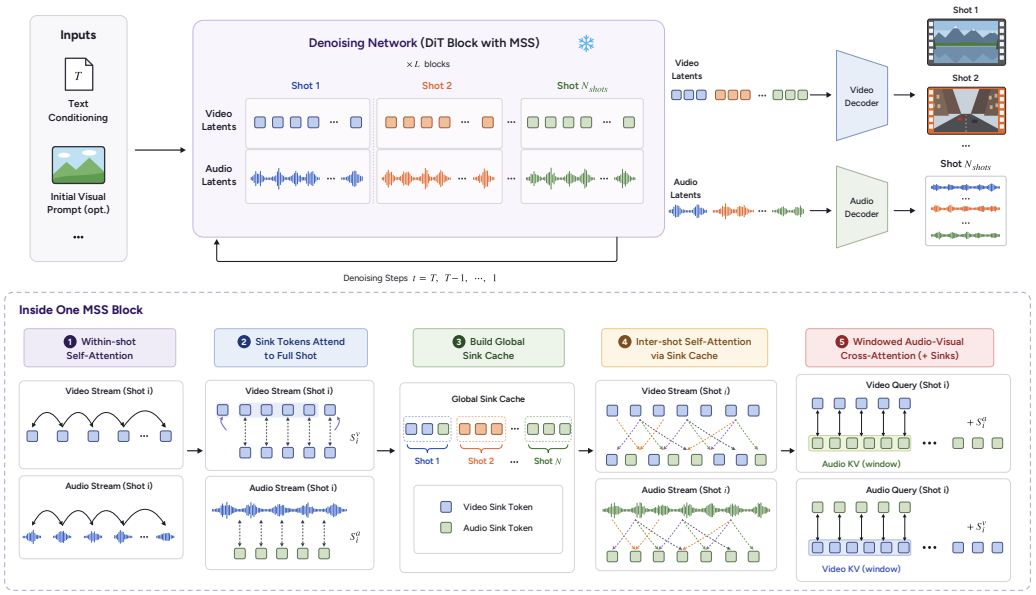

Figure 3: **Workflow of the Multi-Modal Sink Shot.** Within their respective audio and video streams, tokens sequentially process through inter-shot self-attention and windowed audio-visual cross-attention. Sink tokens summarize shot-level appearance and acoustics, enabling cross-shot consistency through sparse routing.

tokens, so cross-modal interaction must respect alignment; meanwhile, dense interactions across all shots are expensive and can blur shot boundaries, which is especially harmful when audio is present because acoustic leakage across shots is perceptually salient. We address both with MSS, which couples windowed audio–visual cross-attention for local synchronization with sink-mediated inter-shot self-attention for efficient global consistency.

Windowed Audio-Visual Cross-attention. The model maintains two token streams from x_t^v and x_t^a. Let $Q_{i,\tau}^v$ denote video query tokens of shot i at time index τ, and $KV_{i,\kappa}^a$ denote audio key-value tokens at index κ. A deterministic alignment map $\pi(\tau)$ returns the audio index interval covered by the τ-th video latent frame. With window radius Δ, video-to-audio attention is:

$$\text{Attn}^{v\leftarrow a}(Q_{i,\tau}^v, KV_i^a) = \text{Attn}\left(Q_{i,\tau}^v,\ KV_{i,[\pi(\tau)-\Delta,\ \pi(\tau)+\Delta]}^a\right). \tag{3}$$

The symmetric pathway maps each audio index to its corresponding video frame. This two-way windowed exchange encourages synchronized dynamics while reducing interference from temporally distant audio.

MSS introduces M sink tokens per shot for each modality, denoted S_i^v and S_i^a (M=2 in our experiments). Each sink token has the same hidden dimension as regular tokens in its respective stream. The sink tokens are zero-initialized learnable parameters with no positional encoding, making them position-agnostic summaries of their shot. At the start of each sample, sink tokens are initialized from their learned values and then serve as *per-sample hidden states* throughout the denoising process: they persist across all denoising steps t and all transformer blocks within each step, progressively accumulating shot-level information as the latents are refined. This persistence is critical—re-initializing sinks at each step would discard the cross-step context that enables stable identity and voice tracking. Sink tokens participate only in audio-visual and inter-shot attention—they do not attend to text conditioning, keeping them as pure audio-visual context carriers. Each modality maintains its own set: S_i^v summarizes visual appearance (character identity, background style) while S_i^a captures acoustic characteristics (speaker timbre, ambient sound profile). The two sets remain modality-specific and do not mix information until they are concatenated into the global sink cache for cross-shot routing. The video-to-audio pathway becomes:

$$\text{Attn}^{v\leftarrow a}(Q_{i,\tau}^v, KV_i^a) = \text{Attn}\left(Q_{i,\tau}^v,\ [KV_{i,[\pi(\tau)-\Delta,\ \pi(\tau)+\Delta]}^a,\ KV(S_i^a)]\right), \tag{4}$$

Table 1: **Comparison of cross-shot attention mechanisms.**

Method	Modality	Cross-shot Mem.	A-V Local Sync	Leakage Ctrl.	Speaker Mem.
LCT	video	interl. pos. emb.	✗	limited	✗
HoloCine	video	char. memory	✗	limited	✗
Reg. tokens	general	global reg.	✗	✗	✗
MOVA	A-V	✗ (single)	within-shot	✗	✗
MSS	**A-V**	**sink routing**	✓	✓	✓

and analogously for $\text{Attn}^{a\leftarrow v}$. Sink tokens let the model combine strict local synchronization with shot-level acoustic context (speaker timbre, background ambience) without allowing unrelated shots to dominate the local correspondence.

Inter-shot Self-attention. MSS replaces dense inter-shot interactions with sparse routing through sink tokens. Within each shot, full self-attention is maintained inside each modality. Cross-shot communication proceeds through a global sink cache:

$$KV_{\text{sink}} = [KV(S_1^v), KV(S_1^a), \ldots, KV(S_{N_{\text{shots}}}^v), KV(S_{N_{\text{shots}}}^a)]. \tag{5}$$

For a query token in shot i, self-attention in the video stream is computed as:

$$\text{Attn}^v(Q_i^v, KV^v) = \text{Attn}\left(Q_i^v,\ [KV_i^v,\ KV_{\text{sink}}]\right), \tag{6}$$

and similarly for audio. The dominant inter-shot cost scales with $N_{\text{shots}}^2 M(L_v{+}L_a)$ rather than $N_{\text{shots}}^2(L_v{+}L_a)^2$.

Concretely, within each denoising block the MSS update proceeds in five steps: **(i)** within-shot self-attention is applied independently inside each modality, allowing motion and audio to evolve smoothly within each shot; **(ii)** each sink token attends to all local tokens of its shot, aggregating a compact summary of visual appearance or acoustic characteristics; **(iii)** the global sink cache KV_{sink} is assembled by concatenating sink key-value pairs from all shots; **(iv)** shot-local tokens attend to their own tokens together with KV_{sink}, enabling cross-shot identity and voice propagation; **(v)** windowed audio-visual cross-attention aligns the two modalities within temporally local neighborhoods, with sink-derived context appended. The sink cache is not detached from the computation graph, so gradients flow through all five steps during training. This procedure is identical at training and inference time, ensuring that the model's learned cross-shot routing behavior transfers directly to generation without distribution mismatch.

Comparison with Prior Cross-Shot Mechanisms. Table 1 compares MSS with representative cross-shot attention designs. Existing multi-shot methods such as LCT (Guo et al., 2025) and HoloCine (Meng et al., 2025) operate exclusively on video tokens, improving visual identity but lacking audio-aware synchronization and acoustic leakage control. Standard register tokens (Darcet et al., 2024) provide general-purpose global context but are not designed for the cut-aware structure of multi-shot narratives. MOVA (Team et al., 2026) supports joint audio-video generation but is limited to single-shot scenarios. Compared with the representative mechanisms considered in Table 1, MSS is designed to jointly address modality-specific cross-shot memory, temporally local audio-visual synchronization, cut-aware leakage prevention, and speaker/ambience persistence.

3.4 Audio-Visual Latent Stitching

Long multi-shot stories often exceed the temporal span of a single denoising pass. Standard approaches generate the story in several windows, reconstructing pixels and waveforms from the previous window and re-encoding the context for the next. In an audio-visual model, this cycle repeats two autoencoder passes for both modalities, gradually perturbing appearance, speaker timbre, and background acoustics.

Audio-Visual Latent Stitching (AVLS) addresses this with three key design choices that distinguish it from standard autoregressive continuation approaches such as those used in Open-Sora: **(1)** AVLS operates entirely in the *joint audio-video latent space*, caching both video and audio latents simultaneously rather than conditioning on reconstructed frames alone; **(2)** it uses *time-dependent*

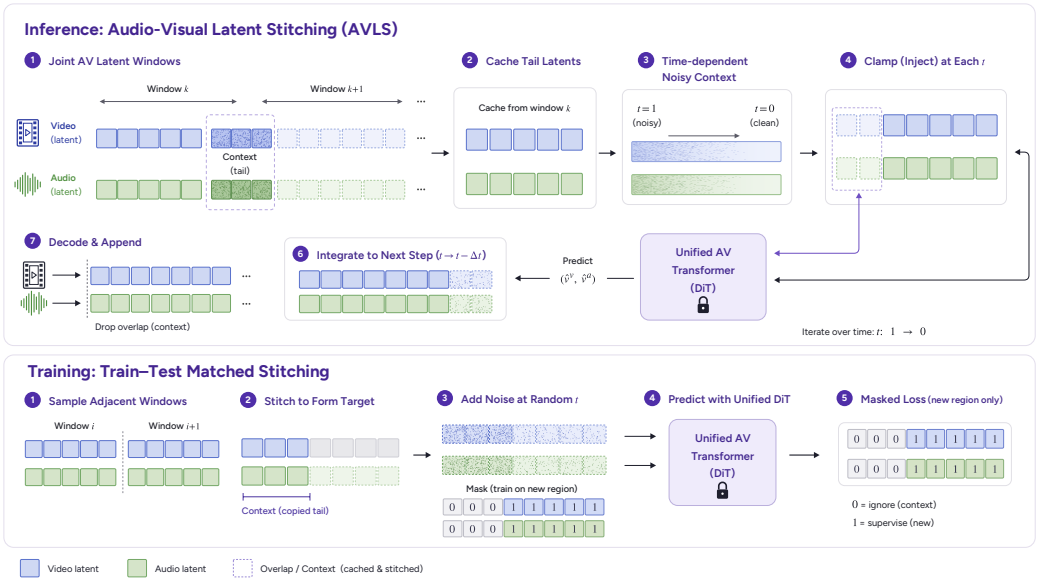

Figure 4: **Workflow of Audio-Visual Latent Stitching (AVLS).** The long sequence is divided into overlapping segments. Video frames and audio waveforms from adjacent windows are independently encoded into the joint audio-video latent space.

noisy context clamping, injecting the cached context at the correct noise level for each integration step rather than providing clean or independently noised prefixes; and **(3)** it employs a *train-test matched interface*, training the model under the same masked stitched-context format used at inference, closing the distribution gap that causes drift in mismatched approaches.

Concretely, we split a long sequence into consecutive latent windows. Each window contains video latents $x^v \in \mathbb{R}^{T_v \times d_v}$ and audio latents $x^a \in \mathbb{R}^{T_a \times d_a}$. Let N^v_{ctx} be the number of video latent frames used as context, with the corresponding audio context length:

$$N^a_{\text{ctx}} = \left| \bigcup_{\tau=1}^{N^v_{\text{ctx}}} \pi(\tau) \right|. \tag{7}$$

After synthesizing window k, we cache its tail latents $x^v_{\text{ctx}} = x^v_{k,\, T_v - N^v_{\text{ctx}}+1:T_v}$ and $x^a_{\text{ctx}} = x^a_{k,\, T_a - N^a_{\text{ctx}}+1:T_a}$. We sample Gaussian noises $\epsilon^v_{\text{ctx}}, \epsilon^a_{\text{ctx}}$ once per window and define the time-dependent noisy context:

$$x^v_{t,\text{ctx}} = (1-t)x^v_{\text{ctx}} + t\epsilon^v_{\text{ctx}}, \quad x^a_{t,\text{ctx}} = (1-t)x^a_{\text{ctx}} + t\epsilon^a_{\text{ctx}}. \tag{8}$$

During sampling, at each integration time t we clamp the current states:

$$y^v_t \leftarrow m^v \odot x^v_{t,\text{ctx}} + (1-m^v) \odot y^v_t, \quad y^a_t \leftarrow m^a \odot x^a_{t,\text{ctx}} + (1-m^a) \odot y^a_t, \tag{9}$$

where m^v, m^a are binary masks selecting the context prefix.

To close the train-test gap, AVLS trains under the same stitched format. From each training example, we sample two adjacent windows with temporal overlap, construct a stitched target by copying the tail latents of window 1 into the prefix of window 2:

$$\bar{x}^v_2 = [x^v_{1,\, T_v - N^v_{\text{ctx}}+1:T_v},\ x^v_{2,\, N^v_{\text{ctx}}+1:T_v}], \quad \bar{x}^a_2 = [x^a_{1,\, T_a - N^a_{\text{ctx}}+1:T_a},\ x^a_{2,\, N^a_{\text{ctx}}+1:T_a}], \tag{10}$$

and supervise with loss masked on the context region:

$$\mathcal{L}_{\text{AVLS}} = \mathbb{E}\Big[\lambda_v \big\|(1-m^v) \odot (\hat{v}^v_\theta - v^v_t)\big\|_2^2 + \lambda_a \big\|(1-m^a) \odot (\hat{v}^a_\theta - v^a_t)\big\|_2^2\Big]. \tag{11}$$

This training procedure encourages the model to continue from the same kind of cached audio and video latents available at inference. After decoding, we drop the duplicated overlap and concatenate only the newly generated part, yielding smooth long stories with stable identities, voices, and background sound.

Table 2: **Quantitative comparison on *ST-Bench*.** The best and runner-up are in **bold** and underlined. Note that the pretrained base model *Wan2.2-T2V* and independently generate each shot without consistency constraints, offering an upper bound reference for single-shot metrics.

Method	Aesthetic Quality↑	Prompt Following↑		Cross-shot Consistency↑	
		Global	Single-shot	Overall	Top-10 Pairs
Wan2.2-T2V	0.6452	0.2174	0.2452	0.3937	0.4248
StoryDiffusion + Wan2.2-I2V	0.6085	0.2288	0.2349	0.4600	0.4732
IC-LoRA + Wan2.2-I2V	0.5704	0.2131	0.2181	0.4110	0.4296
HoloCine	0.5653	0.2199	0.2125	0.4628	0.5005
StoryMem	0.6133	**0.2289**	0.2313	0.5065	**0.5337**
StoryWeaver (Ours)	**0.6318**	0.2207	**0.2466**	**0.5277**	0.5191

4 EXPERIMENTS

4.1 SETUP

We implement STORYWEAVER by extending the 720P variant of MOVA (Team et al., 2026). All training samples are converted into multi-shot sequences with aligned audio following Sec. 3.2. We use three duration settings (5 s, 15 s, and 60 s) with up to 6 shots per sample at 480×832 resolution. We train for 35k steps with a learning rate of 3×10^{-5} on 32 NVIDIA H100 GPUs. MSS uses window radius Δ=5, M=2 sink tokens per modality per shot, and AVLS uses an overlap of 5 latent steps. Full implementation details are provided in Appendix K.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

We conduct experiments across two tracks: video-only narrative evaluation on ST-Bench (Zhang et al., 2025b) and human-centric audio-visual evaluation on WeaverBench. Due to the absence of existing open-source frameworks for holistic multi-shot audio-video generation, we adapt cutting-edge models to construct baselines. We provided the audio-video results in the supplementary videos.

Baselines. For ST-Bench, we compare against: (1) Wan2.2-T2V-A14B applied independently per shot (single-shot reference); (2) StoryDiffusion (Zhou et al., 2024) and IC-LoRA (Huang et al., 2024a) generating keyframes animated by Wan2.2-I2V-A14B (two-stage paradigm); and (3) HoloCine (Meng et al., 2025), a holistic multi-shot model (video-only). For WeaverBench, we extend these paradigms using MOVA (Team et al., 2026) for audio-visual generation: StoryDiffusion + MOVA and IC-LoRA + MOVA (two-stage audio-visual pipelines).

Evaluation Metrics. For ST-Bench, we evaluate aesthetic quality (LAION predictor from VBench (Huang et al., 2024b)), prompt following (ViCLIP (Wang et al., 2023) global and single-shot scores), and cross-shot consistency (ViCLIP pairwise similarity, overall and top-10 pairs). For WeaverBench, we assess 20 characters across 10 situations (200 samples per method) using: visual quality (Q-Align IQA and AES (Wu et al., 2023a)), identity preservation (FaceSIM (Yuan et al., 2025)), audio-visual alignment (ImageBind score (Girdhar et al., 2023), DeSync via Synchformer (Iashin et al., 2024)), and lip synchronization (Sync-C and Sync-D (Chung & Zisserman, 2016)). Human preference results and hyperparameter sensitivity are reported in Appendix I and J.

Quantitative Results. The ST-Bench evaluation is summarized in Table 2. STORYWEAVER achieves the highest Aesthetic Quality and Single-shot Prompt Following among the evaluated multi-shot methods. The model yields an Overall Cross-shot Consistency of 0.5277, improving upon StoryDiffusion and HoloCine. Although StoryMem records higher scores in Global Prompt Following and Top-10 Pairs Consistency, STORYWEAVER provides stronger shot-level text adherence and a more stable balance across the sequence. The WeaverBench evaluation is presented in Table 3. STORYWEAVER achieves the best performance across all reported metrics. These results show that the unified architecture synchronizes audio and video while maintaining character identity across shots.

Table 3: **Comparison of audio-video story generation methods on WeaverBench**. The best results on each metric are **bolded**.

Method	Video Quality		AV-Align		Lip Sync		ID
	IQA↑	AES↑	DeSync↓	IB-Score↑	Sync-C↑	Sync-D↓	FaceSIM↑
StoryDiffusion + MOVA	1.42	1.11	0.552	0.256	8.15	8.02	0.82
IC-LoRA + MOVA	1.36	1.09	0.501	0.251	8.09	7.94	0.79
StoryWeaver (Ours)	**1.57**	**1.17**	**0.446**	**0.303**	**8.33**	**7.66**	**0.87**

Table 4: **Ablation study on *ST-Bench*.** The best and runner-up are in **bold** and underlined.

Method	**Aesthetic Quality↑**	**Prompt Following↑**		**Cross-shot Consistency↑**	
		Global	**Single-shot**	**Overall**	**Top-10 Pairs**
MOVA	0.5894	0.2122	0.2196	0.4628	0.4125
MOVA + MSS	0.6021	0.2156	0.2249	0.5032	0.4997
MOVA + AVLS	0.6160	0.2199	0.2281	0.4719	0.5022
StoryWeaver (Ours)	**0.6318**	**0.2207**	**0.2466**	**0.5277**	**0.5191**

4.3 Ablation Study

We evaluate individual contributions of MSS and AVLS on ST-Bench (Table 4). Integrating MSS improves Overall Cross-shot Consistency from 0.4628 to 0.5032 and Top-10 Pairs from 0.4125 to 0.4997, confirming that windowed audio-visual cross-attention with inter-shot self-attention helps align the narrative structure. Independently applying AVLS raises Aesthetic Quality from 0.5894 to 0.6160 and improves prompt following, demonstrating that latent-space continuation prevents error accumulation at shot boundaries. The complete StoryWeaver framework achieves the highest performance across all metrics.

MSS Sub-module Analysis. To further understand the complementary roles within MSS, we progressively enable its components (Table 5). Windowed audio-visual cross-attention alone improves consistency by better associating visual events with temporally aligned audio. Adding sink context to the cross-attention pathway yields substantially larger gains, as shot-level summaries of speaker timbre and appearance statistics help maintain coherence beyond local windows. Sink-mediated inter-shot self-attention without local alignment primarily benefits cross-shot consistency but less so prompt following, confirming that inter-shot routing alone cannot replace local audio-visual alignment. The full MSS design achieves the best results across all metrics, demonstrating that the two mechanisms are complementary. AVLS continuation strategy comparisons are in Appendix E.

Table 5: Ablation on MSS sub-modules over *ST-Bench*. Best and runner-up in **bold** and underlined.

Method	**Aesthetic Quality↑**	**Prompt Following↑**		**Cross-shot Consistency↑**	
		Global	**Single-shot**	**Overall**	**Top-10 Pairs**
MOVA	0.5894	0.2122	0.2196	0.4628	0.4125
+ Windowed A-V Cross-Attn	0.5958	0.2141	0.2237	0.4816	0.4518
+ Windowed A-V Cross-Attn + Sink Context	0.6096	0.2150	0.2345	0.5127	0.4989
+ Sink-mediated Inter-shot Self-Attn	0.5979	0.2138	0.2218	0.4971	0.4863
Full MSS	**0.6318**	**0.2207**	**0.2466**	**0.5277**	**0.5191**

5 Conclusion

This paper introduces StoryWeaver, a framework that addresses the consistency challenges in continuous multi-shot audio-video generation. By adopting a joint generation paradigm conditioned on text and an initial prompt, our model synchronizes visual and acoustic elements across extended narrative sequences. The Multi-Modal Sink Shot mechanism models long-range audio-visual dependencies through local windowed attention and sink-mediated cross-shot routing, while Audio-Visual Latent Stitching prevents temporal drift through train-test matched latent-space continuation. Experimental results, supported by audio-specific ablations, transition metrics, and human preference studies, confirm that StoryWeaver achieves improved visual-acoustic consistency, temporal coherence, and character retention compared to existing methods.

AI USE STATEMENT

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

REFERENCES

Md Tahmeed Abdullah, Sejuti Rahman, Shafin Rahman, and Md Fokhrul Islam. Vae-gan3d: Leveraging image-based semantics for 3d zero-shot recognition. *Image and Vision Computing*, 147: 105049, 2024.

Hritik Bansal, Yonatan Bitton, Michal Yarom, Idan Szpektor, Aditya Grover, and Kai-Wei Chang. Talc: Time-aligned captions for multi-scene text-to-video generation. *arXiv preprint arXiv:2405.04682*, 2024.

Niklas Bergström. SubtitleEdit: Open source subtitle editor. `https://github.com/SubtitleEdit/subtitleedit`, 2024. Version 4.0.1.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Brandon Castellano. Pyscenedetect: Python-based video scene detection, 2024. URL `https://github.com/Breakthrough/PySceneDetect`.

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*, pp. 7310–7320, 2024.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech*, pp. 3830–3834, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

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*, pp. 15180–15190, 2023.

Google. Nano banana: AI image generation with character consistency. `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/`, 2025.

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*, pp. 17281–17291, 2025.

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

Akio Hayakawa, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Mmdisco: Multi-modal discriminator-guided cooperative diffusion for joint audio and video generation. *arXiv preprint arXiv:2405.17842*, 2024.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*, 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. *arXiv preprint arXiv:2511.21579*, 2025.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024a.

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

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, 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 Computer Vision and Pattern Recognition Conference*, pp. 28405–28415, 2025.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, pp. 2350–2354, 2019.

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

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025.

Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*, pp. 468–485. Springer, 2024.

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

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10219–10228, 2023.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3626–3636, 2022.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

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

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

Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human efforts. *International Journal of Computer Vision*, 133(6):3083–3104, 2025b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Shuchen Weng, Haojie Zheng, Zheng Chang, Si Li, Boxin Shi, and Xinlong Wang. Audio-sync video generation with multi-stream temporal control. *arXiv preprint arXiv:2506.08003*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023a.

Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. *arXiv preprint arXiv:2508.11484*, 2025.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023b.

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. In *The Fourteenth International Conference on Learning Representations*, 2025.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv:2408.11788*, 2024.

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

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 Computer Vision and Pattern Recognition Conference*, pp. 12978–12988, 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. *arXiv preprint arXiv:2511.03334*, 2025a.

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

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

A MSS AND AVLS ALGORITHM DETAILS

We provide pseudocode for the MSS forward pass within a single denoising block (Algorithm 1) and for the full sampling procedure with persistent sink tokens and AVLS (Algorithm 2).

Algorithm 1 MSS Forward Pass (Single Denoising Block)

Require: Video tokens X_i^v, audio tokens X_i^a, sink tokens S_i^v, S_i^a for $i = 1, \ldots, N$, alignment map π, window radius Δ
Ensure: Updated tokens and sink states
// Step (i): Within-shot self-attention
for each shot $i = 1, \ldots, N$ **do**
  $X_i^v \leftarrow \mathrm{SelfAttn}(X_i^v)$; $X_i^a \leftarrow \mathrm{SelfAttn}(X_i^a)$
end for
// Step (ii): Sink tokens attend to shot-local tokens
for each shot $i = 1, \ldots, N$ **do**
  $S_i^v \leftarrow \mathrm{CrossAttn}(Q{=}S_i^v,\ KV{=}X_i^v)$
  $S_i^a \leftarrow \mathrm{CrossAttn}(Q{=}S_i^a,\ KV{=}X_i^a)$
end for
// Step (iii): Build global sink cache
$C \leftarrow \mathrm{Concat}[KV(S_1^v), KV(S_1^a), \ldots, KV(S_N^v), KV(S_N^a)]$
// Step (iv): Shot-local tokens attend to local + global cache
for each shot $i = 1, \ldots, N$ **do**
  $X_i^v \leftarrow \mathrm{Attn}(Q{=}X_i^v,\ KV{=}[X_i^v;\ C])$
  $X_i^a \leftarrow \mathrm{Attn}(Q{=}X_i^a,\ KV{=}[X_i^a;\ C])$
end for
// Step (v): Windowed audio-visual cross-attention
for each shot i, video index τ **do**
  $X_{i,\tau}^v \leftarrow \mathrm{Attn}(Q{=}X_{i,\tau}^v,\ KV{=}[X^a_{i,[\pi(\tau)\pm\Delta]};\ KV(S_i^a)])$
end for
for each shot i, audio index κ **do**
  $X_{i,\kappa}^a \leftarrow \mathrm{Attn}(Q{=}X_{i,\kappa}^a,\ KV{=}[X^v_{i,[\pi^{-1}(\kappa)\pm\Delta]};\ KV(S_i^v)])$
end for
return updated X_i^v, X_i^a, S_i^v, S_i^a

In Algorithm 1, sink tokens are passed by reference: their updated values after each block carry over to the next block within the same denoising step, and across denoising steps within the same window. In Algorithm 2, sink tokens are re-initialized from learned parameters at the start of each generation window, but persist throughout all T steps and L blocks within that window. This makes them per-sample hidden states that accumulate shot-level context during generation, analogous to hidden states in recurrent models but operating within the iterative denoising framework.

B BENCHMARK AND ANNOTATION DETAILS

ST-Bench. ST-Bench is the public multi-shot video storytelling benchmark released with StoryMem (Zhang et al., 2025b). It contains 30 story scripts with 1–4 scenes each and 300 five-second shot prompts in total (8–12 shots per script). The scripts were generated with GPT-5 under a structured instruction covering characters, actions, scenes, mood, and camera language. Each script provides a story overview, shot-level video prompts, first-frame prompts for keyframe-based baselines, and `cut` indicators that distinguish a new scene from a smooth continuation. We evaluate on the full released benchmark under its official protocol. ST-Bench is video-only, so we use it solely for visual narrative evaluation; none of our audio or audio-visual synchronization claims rest on it.

Cross-shot semantic similarity. The Cross-shot Consistency columns on ST-Bench (Tables 2, 4, 5 and 13) follow the official protocol. For Overall, we extract a ViCLIP (Wang et al., 2023) embedding for each generated shot, compute the cosine similarity for every unique pair of shots, and average the pairwise scores. Top-10 Pairs first selects the ten shot pairs whose *prompt* embeddings are most similar and then averages the corresponding video-embedding similarities, so it focuses on shots

Algorithm 2 Sampling with Persistent Sink Tokens and AVLS

Require: Text conditions c, number of windows W, steps T, context overlap N^v_{ctx}
Ensure: Generated video $\hat{V}$, audio $\hat{A}$

for each window $k = 1, \dots, W$ **do**
  Sample $y^v_1, y^a_1 \sim \mathcal{N}(0, I)$
  Init sinks: $S^v_i \leftarrow S^{v,\text{learned}}_i$, $S^a_i \leftarrow S^{a,\text{learned}}_i$
  if $k > 1$ **then**
    Retrieve cached context $x^v_{\text{ctx}}, x^a_{\text{ctx}}$ from window $k-1$
    Sample fixed noise $\epsilon^v_{\text{ctx}}, \epsilon^a_{\text{ctx}} \sim \mathcal{N}(0, I)$
  end if
  for each step $t = 1, \frac{T-1}{T}, \dots, \frac{1}{T}, 0$ **do**
    if $k > 1$ **then** ▷ AVLS clamping
      $y^v_t \leftarrow m^v \odot [(1-t)\, x^v_{\text{ctx}} + t\, \epsilon^v_{\text{ctx}}] + (1-m^v) \odot y^v_t$
      $y^a_t \leftarrow m^a \odot [(1-t)\, x^a_{\text{ctx}} + t\, \epsilon^a_{\text{ctx}}] + (1-m^a) \odot y^a_t$
    end if
    for each DiT block $\ell = 1, \dots, L$ **do** ▷ sinks persist
      $y^v_t, y^a_t, S^v_i, S^a_i \leftarrow \text{MSS}_\ell(y^v_t, y^a_t, S^v_i, S^a_i, c)$
    end for
    $y_{t-1/T} \leftarrow y_t - \frac{1}{T}\, \hat{v}_\theta(y^v_t, y^a_t, t, c)$ ▷ Euler step
  end for
  Cache $x^v_{\text{ctx}} \leftarrow y^v_{0,\text{tail}}$, $x^a_{\text{ctx}} \leftarrow y^a_{0,\text{tail}}$
  Decode and append non-overlapping region to $\hat{V}, \hat{A}$
end for
return $\hat{V}, \hat{A}$

that are expected to be semantically related. This score measures cross-shot semantic similarity. It does not directly measure identity preservation, narrative causality, acoustic consistency, or transition quality, and it can reward repetitive shots. We evaluate those properties separately on WeaverBench: FaceSIM for identity, SpkSim for voice continuity, IB-Score for cross-modal correspondence, and DeSync and Sync-C/D for temporal synchronization. Similarly, AES and Q-Align measure frame- or clip-level quality and may understate perceptual differences between methods, which is why we also report a blind human study (Section I).

WeaverBench. WeaverBench is designed to evaluate multi-shot audio-video generation with a focus on human-centric scenarios. It consists of 20 synthetic characters generated using NanoBanana (Google, 2025), each with a reference face image that serves as the identity anchor for both generation and FaceSIM evaluation. Characters are placed in 10 distinct situations (e.g., interview, outdoor walk, classroom presentation, cafe conversation), yielding 200 evaluation samples (20 characters $\times$ 10 situations). Each prompt specifies 3–4 shots with per-shot scene descriptions separated by `[cut]` markers, following the same layered annotation format used in training (Sec. 3.2). The protocol was fixed before evaluation: NanoBanana generates the 20 character references from predefined appearance descriptions, and every character is evaluated in all 10 situations, so no items were selected after observing results. The situations span indoor and outdoor settings as well as space and science-fiction environments.

The benchmark covers diverse audio types within its human-centric scope: dialogue and speech across shots, ambient sound (cafe noise, street sounds, indoor reverb), foley effects (footsteps, door sounds, object interactions), and background music. All scenarios involve at least one character with a visible face, enabling FaceSIM-based identity evaluation. Audio-visual alignment is assessed through DeSync, Sync-C, Sync-D, and IB-Score as described in the main text. We will publicly release WeaverBench, including the prompts, character references, scenario definitions, annotation schema, and evaluation code.

Layered annotation. The layered annotation serves as the text condition at both training and inference. The global annotation describes sequence-level participants, setting, narrative, and audio context. Per-shot annotations specify local action, camera, dialogue, and sound, and `[cut]` markers define shot boundaries. A representative WeaverBench prompt (cafe conversation) is:

> **Global:** A young woman with shoulder-length black hair, a mustard-yellow coat, and a calm, warm voice meets a colleague in a quiet neighborhood cafe to discuss her finished short film. Maintain her appearance and voice across all shots. The audio bed contains low cafe chatter, occasional cup sounds, an espresso machine, and soft background jazz.
>
> **Shot 1:** Wide establishing shot. The woman enters the cafe, looks around, and walks toward her colleague by the window. The doorbell rings, footsteps cross the wooden floor, and the cafe ambience and soft jazz remain audible. `[cut]`
>
> **Shot 2:** Medium two-shot at the table. She places a folder beside her coffee, smiles, and says in the same warm voice, "I finally finished the film." A cup lightly touches its saucer while the espresso machine sounds briefly in the background. `[cut]`
>
> **Shot 3:** Close-up of the woman listening, then laughing softly and replying, "Let's show it tonight." Keep her face, clothing, and voice unchanged; preserve the same cafe ambience and music without carrying over the previous dialogue.

Baseline inputs. StoryDiffusion + MOVA and IC-LoRA + MOVA are keyframe-first, shot-wise pipelines. StoryDiffusion or IC-LoRA first generates a keyframe for each shot, and MOVA then jointly generates the video and audio of that shot conditioned on the keyframe and the shot description. "Two-stage" thus refers to keyframe-based story construction, not to decoupled audio-video synthesis. All methods receive the same global description and ordered shot content through their native interfaces: StoryWeaver receives the global description and the full `[cut]`-separated sequence jointly, while the keyframe-based baselines receive the same information through their stage-specific inputs. The comparison therefore matches semantic content but not the input interface, and no baseline lacks global or shot-level information.

C Detailed Comparison of MSS with Related Mechanisms

The main text provides a concise comparison of MSS with prior cross-shot attention designs (Table 1). Here we elaborate on the structural differences that distinguish MSS from superficially similar mechanisms.

MSS vs. register tokens. Register tokens (Darcet et al., 2024) are general-purpose learnable tokens appended to the input sequence, designed to absorb global information and reduce artifact tokens in vision transformers. They are task-agnostic, shared across all spatial positions, and carry no structural awareness of temporal boundaries or modality distinctions. In contrast, MSS sink tokens are *shot-aware*: each shot maintains its own set of sink tokens (S_i^v, S_i^a) that are updated by attending to shot-local tokens, producing shot-specific summaries rather than a single global context. Furthermore, sink tokens are *modality-specific*: video sinks and audio sinks capture different characteristics (visual appearance vs. acoustic timbre) and only merge in the global cache for cross-shot routing. Finally, MSS routing is *cut-aware*: the windowed cross-modal attention restricts audio-video interaction to temporally aligned neighborhoods, preventing acoustic leakage that register tokens have no mechanism to control. A further difference is persistence. Register tokens are discarded after each forward pass, whereas MSS sinks carry their state across transformer blocks and denoising steps within a generation window (Section A) and form the path through which identity and acoustic information is routed across shots. Table 6 summarizes these differences.

Table 6: **Register tokens vs. MSS sink tokens.**

	Register tokens (Darcet et al., 2024)	**MSS sinks**
Purpose	absorb representation artifacts in image ViTs	carry identity and acoustic state across shots
Organization	global	per shot and per modality
State	discarded after each forward pass	persist across blocks and denoising steps within a window
Routing	ordinary global attention	shot-scoped aggregation, then sparse global sink routing
Modality and boundaries	image-only; no cut structure	joint audio-video; cut-aware local alignment

MSS vs. HoloCine persistent character memory. HoloCine (Meng et al., 2025) maintains a persistent character memory that stores visual identity features and propagates them across shots. This design is effective for video-only consistency but cannot naturally extend to audio for two reasons. First, HoloCine's memory operates on a single token stream (video), providing no mechanism to summarize or route acoustic information such as speaker timbre or ambient sound profiles. Naively appending audio tokens to the same memory would conflate visual and acoustic features, losing the modality-specific structure that is critical when audio and video autoencoders operate at different temporal resolutions. Second, HoloCine does not restrict cross-shot interaction to temporally aligned windows, so extending it to audio-video would risk cross-shot acoustic leakage—a perceptually salient artifact where speech or sound events from one shot bleed into adjacent shots. MSS avoids both issues through its dual-stream design: modality-specific sinks maintain separate visual and acoustic summaries, while windowed cross-modal attention enforces temporal alignment before any audio-video information exchange.

MSS vs. LCT long-context attention. LCT (Guo et al., 2025) enables multi-shot generation by introducing interleaved positional embeddings into the MMDiT architecture, allowing all tokens across shots to interact via standard self-attention. While effective for video-only settings, this approach has quadratic cost in the total sequence length $N_{\text{shots}}(L_v + L_a)$, which becomes prohibitive when audio tokens (typically much denser than video tokens) are included. LCT also lacks explicit boundary control: all tokens can attend to all others, providing no mechanism to prevent cross-shot leakage. MSS addresses both issues through sparse sink-mediated routing, which reduces the inter-shot cost to $O(N_{\text{shots}}^2 M(L_v + L_a))$ with $M \ll L_v, L_a$, and through windowed cross-modal attention that confines audio-video interaction to local temporal neighborhoods.

D AUDIO QUALITY EVALUATION

To complement the audio-visual synchronization metrics in the main text, we evaluate pure audio quality using three additional metrics (Table 7): Fréchet Audio Distance (FAD) (Kilgour et al., 2019) measures distributional distance between generated and real audio; CLAP score (Wu et al., 2023b) evaluates semantic alignment between generated audio and text descriptions; and SpkSim computes cross-shot speaker embedding cosine similarity using ECAPA-TDNN (Desplanques et al., 2020), quantifying voice timbre consistency across shots.

Table 7: **Audio quality metrics on WeaverBench.** Best results are **bolded**.

Method	FAD↓	CLAP↑	SpkSim↑
StoryDiff. + MOVA	166.25	0.285	0.48
IC-LoRA + MOVA	139.76	0.307	0.53
StoryWeaver (Ours)	**84.41**	**0.339**	**0.78**

STORYWEAVER achieves the lowest FAD (84.41 vs. 139.76 and 166.25), indicating that the joint generation process produces audio closer to the real distribution than the keyframe-first pipelines, in which MOVA jointly generates each shots video and audio from an independently generated keyframe. The CLAP improvement (0.339 vs. 0.307 and 0.285) shows that the generated audio better matches the text descriptions, benefiting from the shared conditioning and cross-modal attention in MSS. The most pronounced gap appears in SpkSim (0.78 vs. 0.53 and 0.48), confirming that the keyframe-based baselines suffer from significant voice timbre drift across shots—each shot's audio is generated without awareness of the previous shot's speaker characteristics. In contrast, STORYWEAVER's sink-mediated cross-shot routing and AVLS latent continuation maintain speaker identity throughout the multi-shot sequence.

Interpreting absolute FAD values. FAD values are not directly comparable across evaluation protocols: the absolute value depends on the embedding model, the reference distribution, the number of samples, clip length, sample rate, and silence handling. Our evaluation uses multi-shot film-domain clips that mix speech, ambience, and foley, so the values in Table 7 should be compared within this protocol rather than against thresholds reported for other audio benchmarks. Within this protocol, STORYWEAVER reduces FAD by 39.6% relative to IC-LoRA + MOVA and by 49.2% relative to

StoryDiffusion + MOVA. The same ordering holds for CLAP and SpkSim, and participants in the blind study prefer the audio fidelity of StoryWeaver in 62.85–71.38% of comparisons (Table 12).

E AVLS Continuation Strategy Comparison

We isolate the effect of each AVLS design choice by comparing five continuation strategies on WeaverBench (Table 8), all using the same MSS backbone.

Table 8: **Comparison of continuation strategies on WeaverBench.** All methods use the same MSS backbone and differ only in how adjacent generation windows are connected. The best results are **bolded**.

Continuation Strategy	Decode-Reencode?	Joint A-V Latent?	Noisy Context?	Train-Test Match?	FaceSIM↑	DeSync↓	Sync-C↑	IB-Score↑
Decode-Reencode	yes	no	no	no	0.69	0.592	7.54	0.231
Generated-frame Conditioning	partial	no	no	no	0.73	0.557	7.74	0.249
Clean Latent Prefix	no	yes	no	no	0.76	0.532	7.68	0.269
Noisy Latent Prefix	no	yes	yes	no	0.81	0.507	7.93	0.281
AVLS (Ours)	no	yes	yes	yes	**0.87**	**0.446**	**8.33**	**0.303**

The decode-reencode baseline, which reconstructs pixels and waveforms at every window boundary, performs worst across all metrics (FaceSIM 0.69, DeSync 0.592), confirming that repeated autoencoder passes accumulate reconstruction error in both modalities. Generated-frame conditioning partially alleviates this by conditioning on rendered frames rather than re-encoding, but still operates outside the joint latent space, leaving audio without proper continuity (Sync-C 7.74). Moving to the joint latent space with a clean prefix improves identity preservation (FaceSIM 0.76) but introduces a distribution mismatch: the clean context is inconsistent with the noisy state of the generation tokens, limiting synchronization gains. Adding time-dependent noise to the latent prefix (Noisy Latent Prefix) further improves all metrics, with FaceSIM reaching 0.81 and DeSync dropping to 0.507, demonstrating that noise-level consistency matters for the flow matching framework. The full AVLS, which additionally matches the training interface to the inference procedure, achieves the best results across all metrics. The gap between Noisy Latent Prefix and AVLS (FaceSIM 0.81→0.87, DeSync 0.507→0.446, Sync-C 7.93→8.33) highlights that closing the train-test distribution gap is essential for high-quality long-form continuation.

Joint audio-video stitching. The video and audio context regions contain different numbers of latents but cover the same physical interval. With the default overlap of $N^v_{\text{ctx}}=5$ video latent frames (0.83 s), the alignment map gives $N^a_{\text{ctx}} \approx 21$ audio latent steps (Section K). Both contexts are clamped at the same solver time t, so the stitched prefix remains temporally aligned across modalities.

F Long-Horizon Generation

To test generation beyond 60 s, we evaluate 100 cases at 120 s and 180 s from the same WeaverBench prompt subset, increasing the numbers of shots and generation windows proportionally (Table 9).

Table 9: **Long-horizon generation on WeaverBench.** The 120 s and 180 s rows use 100 cases from the same prompt subset.

Duration	FaceSIM↑	SpkSim↑	DeSync↓
60 s	0.87	0.78	0.446
120 s	0.85	0.74	0.457
180 s	0.84	0.74	0.468

From 60 s to 180 s, FaceSIM decreases from 0.87 to 0.84 and SpkSim from 0.78 to 0.74, while DeSync increases from 0.446 to 0.468. Over this horizon, quality degrades gradually rather than abruptly. This cross-window continuity comes from AVLS rather than from the sink tokens storing the full history: sinks route information only within a generation window and are re-initialized at the start of each window (Section A), while AVLS passes overlapping joint audio-video latents from one window to the next.

G TRANSFER TO ANOTHER JOINT AUDIO-VIDEO BACKBONE

To examine whether MSS and AVLS depend on MOVA, we apply them to LTX-2 (HaCohen et al., 2026). LTX-2 natively generates synchronized audio and video but does not provide hierarchical shot conditioning, explicit cut control, or inter-shot memory. We compare three LTX-2 variants (Table 10): the pretrained model applied zero-shot with simple latent stitching across windows, LTX-2 fine-tuned on our multi-shot data without MSS and AVLS, and LTX-2 fine-tuned on the same data with MSS and AVLS.

Table 10: **Applying MSS and AVLS to LTX-2 on WeaverBench.** The two fine-tuned LTX-2 variants use the same training data.

Backbone	Multi-shot adaptation	DeSync↓	FaceSIM↑
LTX-2	zero-shot, simple latent stitching	0.461	0.64
LTX-2	fine-tuned without MSS/AVLS	0.472	0.73
LTX-2	fine-tuned with MSS + AVLS	0.456	0.82
MOVA (STORYWEAVER)	fine-tuned with MSS + AVLS	**0.446**	**0.87**

Between the two fine-tuned LTX-2 variants, MSS and AVLS raise FaceSIM from 0.73 to 0.82 and reduce DeSync from 0.472 to 0.456. The identity gain is clear, whereas the synchronization gain is small: zero-shot LTX-2 with simple stitching already reaches a DeSync of 0.461. Because the LTX-2 adaptation used a short training schedule, we treat this result as preliminary evidence that the identity benefit transfers to another joint backbone, not as evidence about synchronization. The controlled MOVA experiments (Tables 4 and 8) remain the main evidence for both components.

H MULTIPLE CHARACTERS AND SCENE CHANGES

Multiple characters. WeaverBench centers on one main reference character. To probe multi-character generation, we run a targeted two-character test with matched prompts and seeds and report per-character FaceSIM and per-speaker similarity (Table 11). M does not assign one sink to each entity: the $2MN$ sink states of an N-shot window are distributed summaries, and every shot retains its full local token sequence and per-shot text.

Table 11: **Two-character test.** The two-character settings use matched prompts and seeds.

Setting	Sequences	Per-character FaceSIM↑	Per-speaker similarity↑
1 character, M=2	200	0.87	0.78
2 characters, M=2	100	0.81	0.76
2 characters, M=4	100	0.83	0.75

With two characters and M=2, identity similarity drops from 0.87 to 0.81 and speaker similarity from 0.78 to 0.76. Increasing M to 4 raises FaceSIM to 0.83, while speaker similarity stays at a similar level (0.75). Thus M=2 is not limited to a single entity, but adding sinks alone does not resolve multi-speaker generation. Crowded multi-character and multi-speaker scenes remain a limitation.

Scene changes. ST-Bench stories contain 1–4 scenes, so the visual evaluation is not restricted to single-scene sequences. In multi-shot generation, hard cuts are intentional: the goal is to preserve persistent attributes such as identity and voice, avoid audio leakage across cuts, and follow the prompted scene changes. Per-shot prompts specify these changes, while the sinks route persistent attributes across shots. The sink cache is additional context, not a shared scene representation that forces all shots to look or sound alike. The blind human study evaluates transition naturalness directly (Table 12).

I HUMAN PREFERENCE STUDY

We conduct a blind human preference study with 30 participants. Each participant evaluates 60 comparison pairs (20 prompts $\times$ 3 baseline comparisons), for a total of 1,800 judgments. For each pair, two generated multi-shot videos with audio are presented side by side in randomized left-right order; method names are hidden throughout. Participants wear headphones and use a standardized display setup. All participants have experience with video or audio content but are not authors of this work. For each pair, participants select the preferred video across seven dimensions: prompt adherence, video quality, audio-video synchrony, lip sync accuracy, audio fidelity, voice consistency, and transition naturalness. Table 12 reports win-rates of STORYWEAVER against each baseline.

Table 12: **Human preference study results (win-rate %).** Each cell reports the percentage of times StoryWeaver is preferred over the corresponding baseline. 30 participants, blind evaluation.

Comparison	Prompt↑	Video Quality↑	A/V Sync↑	Lip Sync↑	Audio Fidelity↑	Voice Consist.↑	Transition↑
vs MOVA	71.47	54.93	55.96	48.56	62.85	78.94	88.31
vs StoryDiffusion + MOVA	63.85	59.21	59.82	56.04	68.24	72.18	76.57
vs IC-LoRA + MOVA	68.22	61.88	63.77	57.83	71.38	81.19	74.92

STORYWEAVER is consistently preferred across all baselines and most dimensions. The strongest advantages appear in transition naturalness (74–88% win-rate) and voice consistency (72–81%), confirming that MSS and AVLS effectively address the cross-shot continuity challenges that the baselines struggle with. Prompt adherence also shows clear gains (63–71%), reflecting the benefit of joint generation over independently produced shots. Audio fidelity is also preferred against all three baselines (62–71%). Lip sync win-rates are more moderate (48–58%), with the comparison against MOVA being near parity (48.56%), which is expected since both methods share the same underlying audio-video diffusion backbone. The strongest overall margins are against MOVA in transition (88.31%) and voice consistency (78.94%), highlighting that single-pass generation without explicit cross-shot mechanisms produces noticeable drift over multi-shot sequences.

J HYPERPARAMETER SENSITIVITY

We study the sensitivity of STORYWEAVER to its three key hyperparameters on ST-Bench (Table 13), varying one at a time while keeping the others at default.

Sink token count M. Using only $M{=}1$ sink token per modality per shot leads to a notable drop across all metrics, particularly in aesthetic quality (0.5411) and cross-shot consistency (Overall: 0.4739), suggesting that a single token is insufficient to capture the diversity of shot-level visual and acoustic characteristics. Increasing to $M{=}4$ yields slightly higher aesthetic quality and single-shot prompt following than the default $M{=}2$, but at the cost of lower cross-shot consistency. At $M{=}8$ the trend reverses: prompt following degrades, likely because an excessive number of sink tokens introduces redundancy that dilutes the global cache signal. The default $M{=}2$ offers the best overall balance.

Window radius Δ. A small window restricts audio-visual interaction too aggressively, hurting aesthetic quality (0.5788) and prompt adherence. The medium (default) setting achieves the strongest consistency and prompt following. A large window slightly improves global prompt following but reduces consistency, consistent with the observation that overly broad cross-modal windows can blur shot boundaries.

AVLS overlap length. Performance is relatively stable across overlap lengths. The 3-step setting shows a small drop in prompt following, while 7 steps performs comparably to the default 5 steps with marginal differences. This suggests that AVLS is robust to the exact overlap length, as long as sufficient context is provided for identity and acoustic continuity.

Table 13: **Hyperparameter sensitivity on ST-Bench.** We vary one hyperparameter at a time while keeping others at default.

Hyperparameter	Value	Aesthetic Quality↑	Prompt Following↑		Cross-shot Consistency↑	
			Global	Single-shot	Overall	Top-10 Pairs
Sink tokens M	1	0.5411	0.2187	0.2254	0.4739	0.4306
	2 (default)	0.6318	**0.2207**	0.2466	**0.5277**	**0.5191**
	4	**0.6322**	0.2197	**0.2483**	0.5106	0.4985
	8	0.6204	0.2089	0.2173	0.5257	0.5009
Window radius Δ	small	0.5788	0.2094	0.2273	0.5126	0.5025
	medium (default)	**0.6318**	0.2207	**0.2466**	**0.5277**	**0.5191**
	large	0.6279	**0.2236**	0.2413	0.5209	0.5162
AVLS overlap	3 steps	0.6306	0.2159	0.2307	0.5144	0.5172
	5 steps (default)	0.6318	**0.2207**	**0.2466**	**0.5277**	0.5191
	7 steps	**0.6324**	0.2193	0.2386	0.5239	**0.5198**

K More Implementation Details

All experiments follow the same training and inference setting. We build on top of the 720P MOVA backbone and train on multi-shot sequences with durations of 5s, 15s, and 60s. The maximum number of shots per training sample is 6, and the spatial resolution is 480×832. Training uses AdamW with a base learning rate of 3×10^{-5}. The global batch size is 32, implemented with 32 GPUs and a per-GPU batch size of 1. Video is sampled at 24 fps, and audio is processed at 16 kHz.

The text condition is organized around character and scene descriptions. Each training video corresponds to a story centered on a main character whose appearance remains consistent across shots while the scene description varies. For each shot, the text input combines character appearance and scene description. Model parameters are sharded with Fully Sharded Data Parallelism, while token sequences are partitioned along the temporal axis with Context Parallelism. Each device holds a contiguous time range of video latents together with aligned audio latents. Sink tokens are replicated and synchronized across context-parallel ranks.

Attention kernels are designed around the two MSS structures. We avoid padding by packing per-shot token ranges into variable-length segments and compute attention with FlashAttention-3 (Dao et al., 2022). Windowed audio-visual cross-attention gathers contiguous audio neighborhoods for each video index, keeping the attention footprint bounded by the window radius. At inference time, we use 30 sampling steps with a classifier-free guidance scale of 7.

Temporal alignment and generation windows. Video is sampled at 24 fps and the video autoencoder has a temporal stride of 4, giving 6 video latent frames per second; audio is resampled to 25 latent steps per second. The alignment map $\pi(\tau)$ converts between these two time bases, returning the audio latent indices that cover the physical interval of the τ-th video latent frame; the windowed cross-attention then learns the actual synchronization within this neighborhood. At inference, each denoising window spans 15 s (90 video and 375 audio latents), and window boundaries need not coincide with shot cuts. The 5, 15, and 60 s durations in Section 4.1 refer to training examples, not to the inference window. AVLS reuses 5 video latent frames (0.83 s, about 21 audio latent steps) between adjacent windows, so a 60 s result uses five windows and is cropped to 60 s.

Sink token count. M counts sink tokens per modality and per shot, so a window with N shots holds $2MN$ sink states. With M=2, a six-shot window contains 24 sink tokens.

Matched component study. The component study in Table 4 is fully matched. MOVA, MOVA + MSS, MOVA + AVLS, and StoryWeaver start from the same 720P MOVA initialization and use the same 350K-example training set, 35K training steps, learning rate of 3×10^{-5}, global batch size of 32 on 32 H100 GPUs, 480×832 resolution, and inference settings (30 sampling steps, classifier-free guidance scale 7). Only the indicated MSS and AVLS components differ. The MOVA row is therefore MOVA fine-tuned on our multi-shot data, not the zero-shot pretrained model.

Training data provenance and release. The training corpus was collected from online video platforms (e.g., YouTube) and processed through shot detection, subtitle removal, and visual and audio quality filtering (Section 3.2). Source videos will not be redistributed. To support reproduction within this constraint, we will release the processing code, annotation prompts, training configurations, evaluation code, checkpoints, and permissible source metadata.

L MORE QUALITATIVE RESULTS

We provide additional qualitative results in Fig. 5. The examples cover varied real-world settings. STORYWEAVER maintains improved identity stability and produces clear clips with natural body gestures.



Figure 5: **More Qualitative Results.** The examples illustrate improved character consistency and narrative continuity across diverse scenes generated by STORYWEAVER from text prompts.