STRUCTURED STATEFUL RELATIONAL SCRIPTS WITH A HIGH-QUALITY DATASET FOR MULTI-SHOT AUDIO-VISUAL VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Coherent multi-shot audio-visual video generation requires maintaining consistent identities while depicting evolving states and coordinating sounds across shot boundaries. Although dense captions describe individual shots in detail, they rarely specify which entities persist, how their states change, or which entities and actions produce sounds. To address this limitation, we propose Structured Stateful Relational Scripts (SSRS), a textual supervision representation that separates persistent entity and voice identities from shot-specific visual and auditory states. SSRS explicitly encodes sound-source bindings and temporal relations, while a global audio-visual event sequence connects local actions and sounds, including events that continue across shot boundaries. We further introduce MSAV-Script-100K, a large-scale dataset of curated multi-shot audio-visual videos paired with detailed SSRS annotations, enabling models to learn identity persistence, state transitions, and event continuity from structured supervision. Experiments demonstrate improvements in cross-shot consistency, adherence to specified state changes, and audio-visual alignment, highlighting the effectiveness of our SSRS and MSAV-Script-100K for coherent multi-shot audio-visual generation.

1 INTRODUCTION

Recent advances in text-to-video (T2V) generation have substantially improved visual fidelity, motion realism, and prompt adherence (Kong et al., 2024; Team Wan et al., 2025; Polyak et al., 2024). The field is now moving beyond silent video synthesis toward *joint audio-visual generation*, where speech, sound effects, ambience, music, and visual content are generated in a unified manner (Team Seedance et al., 2025; HaCohen et al., 2026; Low et al., 2025; SII-OpenMOSS Team et al., 2026). This progression naturally leads to the more challenging setting of *multi-shot audio-visual generation*, which requires not only high-quality content within each shot, but also coherent identities, evolving states, ongoing events, and audio-visual correspondence across shot boundaries.

Recent works have explored multi-shot video generation through shot-specific control, cinematic transitions, cross-shot grounding, and long-form narrative modeling (Kara et al., 2025; Wu et al., 2026; Wang et al., 2026a; Meng et al., 2026). However, most focus primarily on the visual stream. Extending multi-shot generation to joint audio-video introduces additional challenges, including speaker identity, dialogue continuation, sound-source correspondence, and cross-modal temporal alignment. While proprietary models such as Seedance 2.0 already demonstrate strong multi-shot audio-video capabilities, open models such as LTX-2 still leave substantial room for improvement (ByteDance Seed Team, 2026; HaCohen et al., 2026; Wei et al., 2026). This gap can be largely attributed to two key bottlenecks. **First, high-quality training data remain limited.** Although recent efforts have advanced the development of long-form cinematic video datasets (Wu et al., 2026; Meng et al., 2026; Chen et al., 2026; Huang et al., 2026), compact, diverse, and high-quality native multi-shot audio-video clips with dense cross-shot and cross-modal supervision remain underexplored for foundation-model training. **Second, existing textual representations fail to sufficiently disentangle the dependencies.** Dense captions may provide rich descriptive information, yet free-form prose often entangles persistent identities, transient states, actions, dialogue, and auditory events within a single context (Tencent Hunyuan Team, 2026). Consequently, the generator must implicitly infer what should persist, what should change, and whether events across different shots represent a continuous process or independent occurrences.

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.

To address the representation bottleneck, we introduce **Structured Stateful Relational Scripts (SSRS)**, a structured textual supervision for multi-shot audio-visual generation. SSRS separates persistent visual and voice identities from shot-specific visual and auditory states, while representing visual evolution, auditory evolution, and audio-visual relations in distinct fields. It further connects local events through a global audio-visual event sequence, explicitly modeling cross-shot continuity. This is important because a *shot boundary is not necessarily an event boundary*: speech, ambience, or other auditory events may continue across visual cuts. SSRS therefore organizes supervision around a simple principle: *what should persist, what should change, and what should connect*.

To support such structured supervision at scale, we introduce **MSAV-Script-100K**, a large-scale, high-quality dataset of compact multi-shot audio-visual clips designed for foundation-model training. Unlike long-form cinematic corpora, MSAV-Script-100K focuses on short but genuine multi-shot sequences with native synchronized audio and dense cross-shot interactions. Each clip is paired with comprehensive audio-visual descriptions and SSRS annotations. We further formulate dense-to-SSRS conversion as source-constrained reorganization, preserving the underlying information while changing only its structure. This enables controlled comparison between dense and structured supervision on the same videos and caption content.

Our contributions are twofold: (1) we propose **SSRS**, a structured supervision representation that explicitly organizes state evolution and audio-visual dependencies across shots; and (2) we introduce **MSAV-Script-100K**, a large-scale native multi-shot audio-visual dataset with SSRS annotations.

2 Related Work

Video and joint audio-visual generation. Video Diffusion Models (Ho et al., 2022) extended image diffusion to temporal synthesis, while VideoCrafter2 (Chen et al., 2024) improved video quality by combining low-quality videos with high-quality synthetic images. Subsequent foundation models, including CogVideoX (Yang et al., 2025), HunyuanVideo (Kong et al., 2024), Wan (Team Wan et al., 2025), and LongCat-Video (Meituan LongCat Team et al., 2025), advanced visual quality, text alignment, and scalable generation through transformer-based architectures and improved training pipelines. Beyond silent video, Movie Gen (Polyak et al., 2024) incorporates video-conditioned audio generation, whereas LTX-2 (HaCohen et al., 2026), and Ovi (Low et al., 2025), MOVA (SII-OpenMOSS Team et al., 2026) jointly model audio and video. Proprietary systems such as Veo 3 (Collins, 2025), Sora 2 (OpenAI, 2025), and Seedance 2.0 (ByteDance Seed Team, 2026) have also demonstrated synchronized audio-video synthesis.

Multi-shot audio-visual generation. Multi-shot visual generation has been explored through hierarchical narrative synthesis in MovieDreamer (Zhao et al., 2025), shot-specific conditioning in ShotAdapter (Kara et al., 2025), transition control in CineTrans (Wu et al., 2026), cross-shot grounding in MultiShotMaster (Wang et al., 2026a), and holistic scene modeling in HoloCine (Meng et al., 2026). Extending to joint audio-video generation, CineDance (Chen et al., 2026) introduces long-form cinematic data with hierarchical dual-modal annotations and shared character and scene anchors. UnityShots (Huang et al., 2026) combines long- and short-term visual memory with audio-visual boundary gating and reference-speaker conditioning. MSAVBench (Wei et al., 2026) provides complementary evaluation of multi-shot audio-video capabilities. Our work complements these efforts with compact native multi-shot training clips and SSRS supervision.

Generation-oriented captioning. Any2Caption (Wu et al., 2025) translates heterogeneous conditions into captions for controllable synthesis, while VC4VG (Du et al., 2025) studies caption design and evaluation for T2V training. For audio-visual perception, video-SALMONN 2 (Tang et al., 2025) and Omni-Captioner (Ma et al., 2026) improve descriptive completeness and faithfulness through preference optimization and tool-assisted annotation, respectively. AVSCap (Wang et al., 2026b) and TCA-captioner (Zhao et al., 2026) emphasize cross-modal event binding. TimeChat-Captioner (Yao et al., 2026) introduces time-aware, structured descriptions of multi-scene videos. Most closely related, Script-a-Video (Tencent Hunyuan Team, 2026) proposes MTSS, factorizing descriptions into information streams connected by identity and temporal links. Building on prior work, SSRS combines global audio-visual context with shared registries for visual entities, voice identities, and audio entities. It distinguishes persistent identities from shot-specific states and voice identities from auditory events, separates shot-local visual evolution, auditory evolution, and cross-modal relations, and connects events across shots through a global audio-visual event sequence.

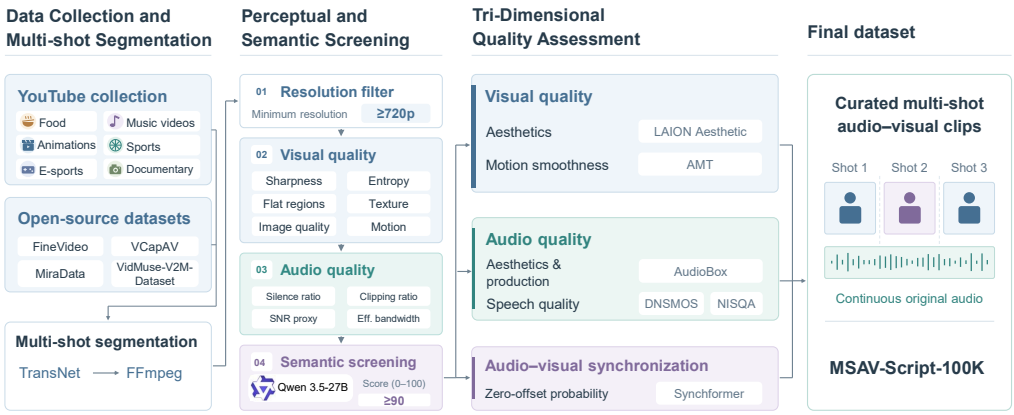

Figure 1: **Overview of the MSAV-Script-100K construction pipeline.** Videos collected from YouTube and existing datasets are segmented into multi-shot clips, followed by perceptual and semantic screening and tri-dimensional assessment of visual quality, audio quality, and audio-visual synchronization, yielding a dataset that preserves consecutive shots and continuous source audio.

3 MSAV-Script-100K

3.1 Data Collection and Multi-Shot Segmentation

Data Collection. Constructing high-quality data for multi-shot audio-visual generation requires native audio, coherent shot sequences, and diverse cross-shot interactions. Existing datasets provide useful sources, but are often designed for other tasks or dominated by cinematic content, with additional limitations in resolution and visual and audio quality that restrict their coverage of diverse, high-quality real-world audio-visual scenarios. We therefore combine videos from existing open datasets, including FineVideo (Farré et al., 2024), MiraData (Ju et al., 2024), VCapAV (Wang et al., 2025), and VidMuse-V2M-Dataset (Tian et al., 2025), with additional high-resolution videos collected from YouTube. For VCapAV (Wang et al., 2025), we use only real videos paired with their original audio tracks, excluding all samples containing synthetic video or audio.

Our YouTube collection covers six categories: *food, music videos, animations, documentaries, sports, and esports*. These sources broaden the range of scenes, subjects, and audio-visual events, including cooking, musical performance, dialogue, narration, athletic actions, and commentated gameplay. They provide examples of recurring entities, evolving actions, and auditory events that persist across visual cuts, forming the source pool for multi-shot segmentation.

Multi-Shot Segmentation. We use TransNet V2 (Souček & Lokoč, 2020) to detect shot boundaries and FFmpeg to split the source videos. We then group consecutive shots into multi-shot clips with durations of approximately 6–12 seconds, preserving their original temporal order and corresponding audio. This retains natural shot transitions and cross-cut audio continuity, providing compact sequences for modeling cross-shot audio-visual dependencies.

This collection and segmentation process yields a pool of approximately 240K multi-shot audio-visual clips, referred to as the *raw data*, which undergo multi-stage quality curation.

3.2 Multi-Stage Quality Curation

Our curation pipeline comprises perceptual and semantic screening followed by tri-dimensional quality assessment, as shown in Figure 1. The first stage removes clips with insufficient resolution, undesirable visual or audio characteristics, or limited semantic content; the second assesses visual, audio, and audio-visual quality jointly across complementary quality dimensions.

Perceptual and Semantic Screening. We first retain clips with a resolution of at least 720p, ensuring spatial quality for subsequent assessment.



Figure 2: **Quality assessment of MSAV-Script-100K against raw data.** The final dataset exhibits higher visual aesthetics, perceptual quality, and motion smoothness (a–c), stronger audio-visual synchronization (d), and improved audio enjoyment, usefulness, and production quality (e, f, h). Audio production complexity (g) characterizes differences in audio composition.

Visual quality filtering. We evaluate six visual metrics: sharpness, image entropy, flat-region proportion, texture statistics, MUSIQ (Ke et al., 2021), and RAFT (Teed & Deng, 2020)-based motion magnitude. We discard clips below P5 in sharpness, entropy, or MUSIQ, above P95 in flat-region proportion or texture score, or outside the P5–P95 motion range, removing degraded, uninformative, near-static, and extreme-motion content.

Audio quality filtering. We screen audio using silence ratio, clipping ratio, an amplitude-based signal-to-noise proxy, and effective bandwidth, which characterize silence, waveform saturation, amplitude contrast, and spectral coverage.

Semantic screening. Since adjacent shots are not necessarily semantically coherent, we use Qwen3.5-27B (Qwen Team, 2026) to score each clip from 0 to 100 for semantic meaningfulness and retain clips scoring at least 90. This visual-only assessment complements the preceding low-level criteria with a content-level measure.

Tri-Dimensional Quality Assessment. Following screening, we assess the remaining clips along three complementary axes: visual quality, audio quality, and audio-visual synchronization, comprising five evaluation criteria for comprehensive multi-modal quality assessment.

Visual quality. We evaluate aesthetics with the LAION Aesthetic Predictor (LAION AI, 2022) and motion smoothness with AMT (Li et al., 2023), covering both frame-level appearance and temporal quality across diverse visual contents and motion patterns in the curated clips.

Audio quality. We use AudioBox (Tjandra et al., 2025) to assess aesthetics and production quality, and DNSMOS (Reddy et al., 2021) and NISQA (Mittag et al., 2021) to evaluate speech quality.

Audio-visual synchronization. We use the Synchformer zero-offset probability (Iashin et al., 2024) to measure temporal alignment between the visual and audio streams. Finally, metric values are converted to within-source percentile ranks and aggregated into five criterion-level subscores, whose equal-weighted average defines the overall quality score.

3.3 Dataset Statistics and Quality Analysis

Dataset statistics. MSAV-Script-100K comprises 100,000 multi-shot audio-visual clips, totaling approximately 237.4 hours. Clips span approximately 6–12 seconds, with a mean duration of 8.55 seconds and a median of 7.97 seconds. Automatic shot counting identifies an average of 5.35 shots per clip, with a median of five, providing compact sequences with frequent visual transitions and continuous source audio. The collection combines open-source videos with six categories

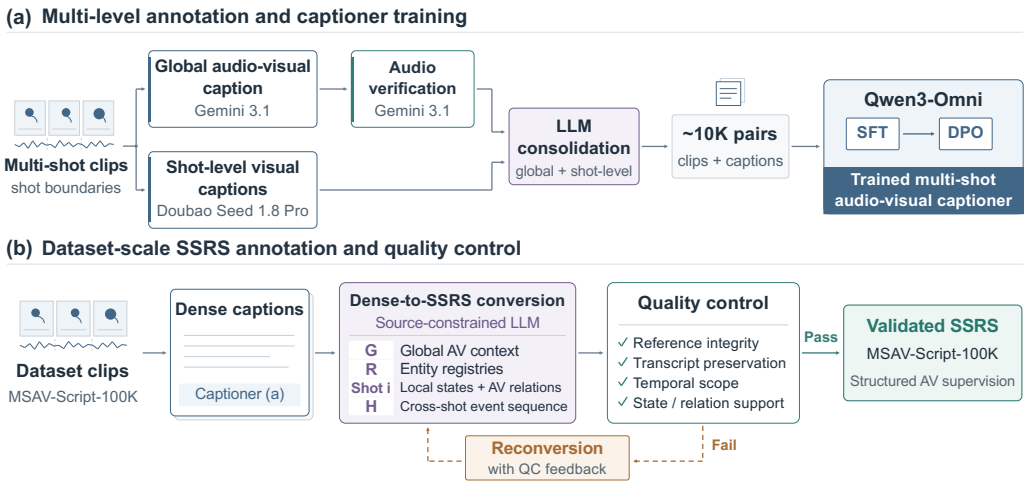

Figure 3: **From content acquisition to structured supervision. (a)** Multi-level expert annotations train a multi-shot audio-visual captioner via SFT and DPO. **(b)** Dense captions for MSAV-Script-100K undergo source-constrained SSRS conversion and quality control.

of YouTube content, covering diverse subjects and soundscapes. Automatic audio tagging detects speech in 54.8% of clips and music in 49.7%, with both tags present in 25.7%. These overlapping categories reflect the coexistence of multiple auditory components within individual clips. Detailed statistics and measurement protocols are provided in Appendix B.

Quality analysis. Figure 2 compares MSAV-Script-100K with raw data in terms of visual quality, audio quality, and audio-visual synchronization. For visual quality, the final dataset achieves higher mean LAION aesthetic and MUSIQ scores, together with better AMT-based motion smoothness than the raw-data reference set. For audio quality, AudioBox reports higher content enjoyment, content usefulness, and production quality, while production complexity is reported separately as a descriptor of audio composition. For audio-visual alignment, the mean Synchformer zero-offset probability increases from 0.18 in the raw-data reference to 0.301 in MSAV-Script-100K, indicating stronger temporal correspondence between the visual and audio streams.

4 Structured Stateful Relational Scripts

Our pipeline separates *content acquisition* from *supervision organization*. We first develop a multi-shot audio-visual captioner that produces comprehensive dense descriptions. We then reorganize these descriptions into Structured Stateful Relational Scripts (SSRS), exposing reusable identities, shot-dependent states, and their audio-visual relations. This separation makes caption structure a distinct design choice rather than conflating it with a stronger description model. Figure 3 summarizes the two stages. Panel (a) illustrates the training pipeline for our multi-shot audio-visual captioner, while panel (b) presents the dense-to-SSRS conversion process.

4.1 Multi-Shot Audio-Visual Captioner

Multi-level annotation. Capturing a multi-shot clip requires both a global account and locally grounded details. Given a clip and its shot boundaries, we use Gemini 3.1 to generate a global audio-visual caption and Doubao Seed 1.8 Pro to describe each shot's visual content. The former provides clip-level context, while the latter supplies shot-specific subjects, actions, scene details, and camera presentation. Gemini 3.1 subsequently re-examines the audio descriptions to check auditory content. Finally, an LLM consolidates these outputs into a comprehensive dense caption, integrating the available visual and auditory information while retaining the shot sequence.

Captioner training. We annotate approximately 10K multi-shot clips with this pipeline and use the resulting video–caption pairs to supervise Qwen3-Omni (Xu et al., 2025). After supervised fine-tuning, we apply direct preference optimization (Rafailov et al., 2023) to obtain our captioner. The

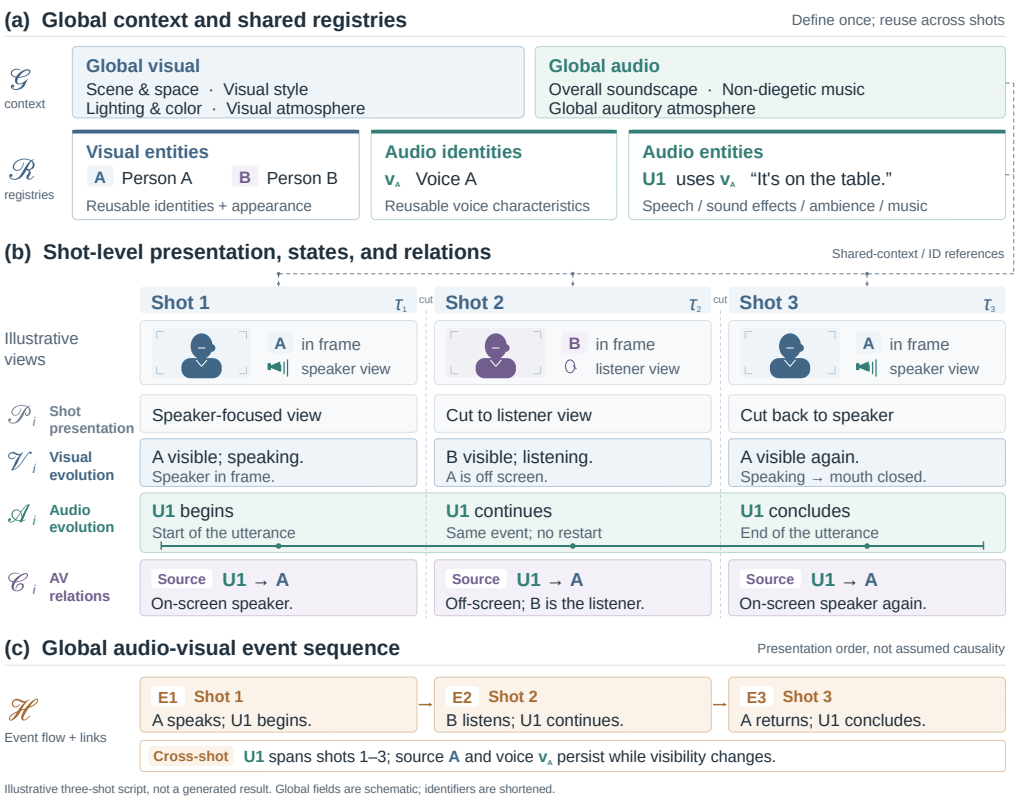

Figure 4: **SSRS distinguishes persistent identities from evolving states.** Global context and shared registries **(a)** are separated from shot-level presentation, states, and relations **(b)**. The global event sequence **(c)** links events across cuts: the same utterance continues despite changing speaker visibility, highlighting that *shot boundaries need not coincide with event boundaries*.

trained model produces dense captions for the dataset; the dense-to-SSRS conversion is performed afterward. Thus, the expert annotation pipeline supplies captioner supervision, whereas SSRS determines how the resulting content is organized for generation.

4.2 Representation Overview

Dense prose often interleaves recurring identities, transient attributes, actions, and sounds. SSRS assigns these factors distinct semantic roles and temporal scopes. For a clip with K shots, we write

$$\mathcal{S} = \left(\mathcal{G}, \mathcal{R}, \{(\tau_i, \mathcal{P}_i, \mathcal{V}_i, \mathcal{A}_i, \mathcal{C}_i)\}_{i=1}^{K}, \mathcal{H}\right), \tag{1}$$

where $\mathcal{G}$ is global context and $\mathcal{R}$ contains shared registries. Each shot specifies its source time interval τ_i, cinematic presentation $\mathcal{P}_i$, visual states and changes $\mathcal{V}_i$, auditory states and changes $\mathcal{A}_i$, and cross-modal relations $\mathcal{C}_i$. The global event sequence $\mathcal{H}$ organizes key events and cross-shot connections. Figure 4 provides a concrete example of this representation. The presentation fields retain transitions, shot size, view angle, camera movement, local lighting, and frame details as specified in the source caption. This preserves cinematic instructions alongside the state–relation factorization; scene layout and visible text are not discarded merely because they do not participate in the main event sequence.

4.3 Global Context and Entity Registries

Global context. Global visual context comprises four complementary aspects: *scene and space* describe the shared environment and spatial layout; *visual style* characterizes the overall aesthetic; *lighting and color* specify shared illumination and the color palette; and *visual atmosphere* captures the overall mood of the scene. Global audio comprises three aspects: *overall soundscape* describes

environmental sounds, action sounds, and non-verbal vocalizations, including their spatial characteristics and layering; *non diegetic music* summarizes the acoustic properties of audience-only scoring; and *global auditory atmosphere* captures the overall auditory mood explicitly.

Shared registries. SSRS organizes reusable content into three registries. The *visual entity registry* defines entity identities and appearance attributes; the *audio identity registry* records voice characteristics; and the *audio entity registry* specifies speech, sound effects, ambience, and music, including content and acoustic properties. Transcribed speech entries preserve the verbatim transcript. Speech entries reference a reusable voice identity when supported by the source, and untranscribed speech is explicitly represented without inventing a transcript. Unique identifiers allow shot descriptions and the global event sequence to reference these entries without repeating their definitions.

4.4 Shot-Level Presentation, States, and Relations

For each shot with source time interval τ_i, SSRS separately represents cinematic presentation, visual evolution, auditory evolution, and cross-modal relations. $\mathcal{P}_i$ describes transitions, shot size, view angle, camera movement, local lighting and visual style, and frame details such as spatial layout and visible text. $\mathcal{V}_i$ records entity positions, poses, expressions, actions, interactions, and supported changes. Actions with unspecified final states are retained without adding unobserved transitions. $\mathcal{A}_i$ references audio entities and describes their occurrence, continuation, repetition, overlap, and local changes, while transcripts and stable acoustic descriptions remain in the registries. $\mathcal{C}_i$ links audio identifiers to visual entities or concise action anchors, specifying source attribution, synchronization, temporal order, duration, overlap, and causation when supported by the source. These relations are distinct: co-occurrence does not imply synchronization, nor does synchronization imply causality. For example, chewing and its associated sound are represented as a visual action and an audio entity, with their synchronization expressed separately in $\mathcal{C}_i$.

4.5 Cross-Shot Event Organization

The global audio-visual event sequence $\mathcal{H}$ organizes key actions, state changes, and auditory events into a main event flow $E_1 \rightarrow E_2 \rightarrow \cdots$. Each node links to local descriptions through existing identifiers and concise action references, with its corresponding shot or shot range. The ordering follows video presentation without assuming causality or chronological story time. Additional source-supported relations specify simultaneity, persistence, cross-shot dependencies, and causal triggers without repeating local details. This organization captures both visual state evolution and auditory continuity across cuts. For example, pouring into the same jar may continue across shots before the jar is repositioned. Likewise, an utterance may continue after a cut from the speaker to a listener, retaining the same speech event and voice identity while the speaker's visibility changes. Thus, shot boundaries need not coincide with event boundaries, and $\mathcal{H}$ explicitly organizes both continuity and intended change across shots.

4.6 Dense-to-Script Conversion and Quality Control

We formulate dense-to-SSRS conversion as source-constrained reorganization. An LLM maps source-caption information to global context, shared registries, shot-level fields, and the global event sequence according to semantic roles and temporal scopes. The conversion is designed to preserve source facts, shot boundaries, source-provided timing, verbatim dialogue and visible text, as well as uncertainty and explicit negation. To reduce redundancy, each fact has a primary location, with identifier-based references avoiding repeated definitions. The conversion specification requires checks for reference integrity, transcript preservation, temporal scope, and unsupported state or relation assertions. These checks help preserve source fidelity and structural consistency, supporting more reliable supervision for multi-shot audio-visual generation.

5 Experiments

We organize the evaluation around two complementary questions: whether SSRS improves general multi-shot audio-visual generation, and whether it improves the specific state and relational controls

Table 1: Comparison on the 150-prompt short-video T2AV subset of MSAVBench. All scores are normalized to $[0, 1]$, where higher values indicate better performance. Both Dense SFT and SSRS SFT are fine-tuned from the pretrained LTX-2.3 model. C-L and I-L denote C-Layout and I-Layout, respectively, while WER-S is computed as $\text{WER-S} = \text{clip}(1 - e, 0, 1)$, where e denotes the applicable transcription error rate (WER or CER). Best and second-best results are highlighted in **bold** and *underlined italics*, respectively.

	Global					Cross-shot				Intra-shot				
Method	Narr↑	Lip↑	Attr↑	Sync↑	VQ↑	C-L↑	VC↑	Mus↑	Spk↑	I-L↑	Cam↑	PQ↑	OCR↑	WER-S↑
Ovi	0.3288	0.0605	0.2864	0.8290	0.5434	0.4333	0.6824	0.4978	0.3094	0.9136	0.5316	0.5733	*0.2222*	0.0065
JavisDiT++	0.4932	0.0084	0.0668	0.8073	0.6499	0.6731	0.6747	0.5055	**0.4651**	0.6288	0.6245	0.4981	0.1379	0.0016
JavisGPT	0.2170	0.0022	0.0101	0.7780	0.3813	0.6400	0.7038	0.4992	0.0659	0.8333	0.4281	0.5575	0.0053	0.0028
ShotStream+HF	0.6844	0.1384	0.2762	0.7863	0.5943	*0.8704*	0.8720	0.4722	0.1860	0.9151	0.6377	**0.6641**	0.1734	0.0033
LTX-2.5	0.7883	0.1242	0.2500	0.7727	0.7269	0.8152	0.8664	0.4673	0.1418	0.8902	0.7529	*0.6416*	**0.2324**	0.7226
Base-LTX-2.3	0.8076	0.1138	0.2061	0.7840	0.8033	0.7301	**0.8905**	0.4666	0.1811	*0.9167*	0.7648	0.6235	0.1829	0.7262
Dense SFT	*0.8222*	*0.2284*	*0.5104*	*0.8450*	*0.8202*	0.8409	*0.8756*	*0.7217*	0.4002	**0.9722**	*0.8027*	0.6244	0.1389	*0.7319*
SSRS SFT	**0.8467**	**0.2468**	**0.5265**	**0.8573**	**0.8376**	**0.9167**	0.8631	**0.7523**	*0.4006*	0.8750	**0.8321**	0.6011	0.2118	**0.7543**

that motivate its design. Public MSAVBench (Wei et al., 2026) comparisons address the former; controlled evaluation and representation ablations address the latter.

5.1 Experimental Setup

Backbone and training controls. Our controlled comparisons use LTX-2.3 with three primary configurations: the pretrained backbone, dense-caption supervised fine-tuning, and SSRS-supervised fine-tuning. Dense and SSRS supervision are derived from the same training clips and source descriptions. To isolate the representation, the two fine-tuned configurations use identical data splits, initialization, trainable parameters, optimization schedules, and numbers of updates.

Evaluation and metrics. We evaluate text-to-audio-video (T2AV) generation on a fixed, text-only subset of MSAVBench (Wei et al., 2026) comprising 150 short-video prompts. All methods are scored against the same canonical descriptions. We retain 14 metrics that do not require reference media, covering three levels: *global*—narrative coherence (Narr), lip synchronization (Lip), sound attribution (Attr), audio-visual synchronization (Sync), and visual question-answering accuracy (VQ); *cross-shot*—layout consistency (C-Layout), visual consistency (VC), music consistency (Mus), and speaker-timbre consistency (Spk); and *intra-shot*—layout–text alignment (I-Layout), camera adherence (Cam), perceptual audio quality (PQ), text rendering (OCR), and spoken-content fidelity (WER-S). We use Gemini 3.5 Flash for MLLM-based judgments, together with metric-specific expert models and tools. All reported scores lie in $[0, 1]$, with higher values indicating better performance, using metric-specific scoring and normalization rules.

5.2 Comparison on MSAVBench

We compare against representative open-source audio-video and multi-shot generation systems, including Ovi (Low et al., 2025), JavisDiT++ (Liu et al., 2026), JavisGPT (Liu et al., 2025), ShotStream+HF (Luo et al., 2026; Shan et al., 2025), and LTX-2.5 (Lightricks, 2026). Since Ovi, JavisDiT++, and JavisGPT do not natively support multi-shot generation, we follow the MSAVBench protocol (Wei et al., 2026) and generate each shot independently before concatenating them into a multi-shot sequence. We also report the pretrained LTX-2.3 backbone (HaCohen et al., 2026) to disentangle the effects of data adaptation and structured supervision. As shown in Table 1, dense-caption fine-tuning on MSAV-Script-100K improves most metrics over the pretrained backbone, demonstrating the effectiveness of our multi-shot audio-visual training data. Under matched training conditions, SSRS further improves 11 of the 14 metrics over Dense SFT, with consistent gains across all global metrics and notable improvements in cross-shot and instruction-following measures. This indicates that explicitly organizing identities, evolving states, and audio-visual relations provides benefits beyond data adaptation alone. Compared with external systems, SSRS SFT achieves the best performance on 9 of the 14 metrics. Some metrics, including VC, I-Layout, and PQ, remain below Dense SFT, suggesting that stronger structured control does not uniformly improve every aspect of visual consistency and perceptual quality.

Table 2: **Controlled evaluation of state and relational following.** T denotes target satisfaction and P denotes preservation rate. All values are percentages. Δ denotes SSRS SFT minus Dense SFT.

	Dense SFT		SSRS SFT		Δ	
Category	$T\uparrow$	$P\uparrow$	$T\uparrow$	$P\uparrow$	ΔT	ΔP
Visual-state evolution (VS)	20.00	**62.50**	**60.00**	42.50	+40.00	-20.00
Cross-shot audio continuity (AC)	0.00	59.38	**50.00**	**62.50**	+50.00	+3.12
Non-speech source binding (NS)	5.00	45.00	**45.00**	**60.00**	+40.00	+15.00
Cross-shot event order (EO)	0.00	50.00	**40.00**	**55.00**	+40.00	+5.00
On-/off-screen speaker binding (SB)	7.14	25.00	**42.86**	**42.86**	+35.71	+17.86
Overall	7.50	49.38	**48.75**	**52.50**	**+41.25**	**+3.12**

Table 3: Representation ablations under matched training and evaluation settings. All scores are normalized to $[0, 1]$, where higher values indicate better performance.

	Cross-shot Organization			**Audio-Visual Relations**			**Control Metrics**	
Representation	Narr↑	Mus↑	Spk↑	Lip↑	Attr↑	Sync↑	VQ↑	Cam↑
Full SSRS	*0.8467*	**0.7523**	*0.4006*	*0.2468*	**0.5265**	**0.8573**	**0.8376**	*0.8321*
w/o Audio-Visual Relations	**0.8622**	0.7131	**0.4268**	0.2374	*0.5110*	0.8387	0.8137	**0.8420**
w/o Global Registries and Event	0.8289	*0.7168*	0.3801	**0.2530**	0.5003	*0.8477*	*0.8189*	0.8218

5.3 Controlled Evaluation of States and Relations

MSAVBench measures general generation quality but does not directly evaluate the state transitions and event relations targeted by SSRS. We therefore construct a controlled diagnostic set covering five aspects: visual-state evolution (VS), cross-shot audio continuity (AC), non-speech sound-source binding (NS), cross-shot event order (EO), and on-/off-screen speaker binding (SB). We evaluate 40 matched prompt instances for Dense SFT and SSRS SFT. Each output is assessed by Gemini 3.5 Flash with native audio using two target questions and four preservation questions, yielding target satisfaction (T) and preservation (P) rates. We further perform manual verification of the model-based judgments to ensure the accuracy and reliability of the evaluation results. As shown in Table 2, SSRS substantially increases overall T from 7.50% to 48.75% (+41.25 pp), while maintaining comparable preservation (49.38% vs. 52.50%). Target satisfaction improves across all five categories, with gains of 35.71–50.00 pp, while P improves in four categories. The exception is visual-state evolution, where higher target adherence is accompanied by lower preservation. Overall, these results indicate that **SSRS substantially improves adherence to intended state and relational changes without degrading preservation of shared constraints overall**.

5.4 Representation Ablations

We ablate two key components of SSRS under matched training and evaluation settings: the explicit audio-visual relation fields and the global entity registries and event. As shown in Table 3, removing audio-visual relations degrades sound attribution, synchronization, music consistency, and visual-question accuracy, although a few individual metrics slightly improve. Removing global registries and event organization further weakens cross-shot organization and identity-related consistency, indicating the importance of reusable entity references across shots. Overall, the full SSRS provides the most balanced performance across cross-shot organization, audio-visual relations, and control metrics, supporting the complementary roles of explicit relational modeling and global organization.

6 Conclusion

We presented MSAV-Script-100K and Structured Stateful Relational Scripts (SSRS) for multi-shot audio-visual generation. SSRS organizes persistent identities, evolving visual and auditory states, cross-modal relations, and cross-shot event continuity into explicit generation supervision. Experiments on LTX-2.3 show that our dataset improves general multi-shot audio-visual generation, while SSRS further enhances state and relational adherence under matched training conditions. A remaining limitation is that SSRS inherits the coverage and accuracy of its source captions, motivating stronger audio-visual captioning and verification in future work.

AI USE STATEMENT

We used ChatGPT and OpenAI Codex to assist with English polishing, Chinese–English translation, manuscript organization, experimental design discussion, LaTeX editing, figure design, and implementation support. All AI-assisted text, experimental plans, code, and figures were manually reviewed, revised, and verified by the authors before being incorporated into the paper. Generative AI systems that constitute part of our research methodology, including those used for annotation, captioning, and evaluation, are explicitly described in the corresponding sections of the paper. The authors take full responsibility for the final content, claims, experiments, code, and artifacts presented in this work.

ETHICS STATEMENT

MSAV-Script-100K is constructed from publicly accessible video sources, including existing open datasets and publicly available YouTube videos collected for multi-shot audio-visual research. The dataset is intended solely for academic and non-commercial research. We respect the copyright, privacy, and personality rights associated with the source content, and do not intend the dataset for biometric identification, surveillance, or other privacy-invasive applications. Where redistribution of original media is restricted, we will release annotations, metadata, timestamps, and source references rather than redistribute the underlying content without permission. We will also provide a mechanism through which content owners or featured individuals may request removal of relevant samples from future releases. All released materials will be accompanied by clear documentation of their intended research use and applicable licensing conditions. Our annotations, metadata, and derived textual resources will be released under the CC BY-NC-SA 4.0 license, while the redistribution of source videos will follow their original licenses and platform terms.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the dataset construction, multi-shot audio-visual captioning pipeline, SSRS representation, training protocol, evaluation metrics, and controlled diagnostic experiments throughout the main paper and appendix. Implementation details, prompts, data-processing procedures, training configurations, and evaluation protocols are documented to facilitate reproduction. Upon publication, we plan to release the MSAV-Script-100K annotations, SSRS data, training and evaluation code, model checkpoints, and the necessary configuration files, subject to the licensing constraints of the original video sources. These materials are intended to support end-to-end reproduction of our data preparation, SSRS conversion, model training, and evaluation.

REFERENCES

ByteDance Seed Team. Seedance 2.0 official launch. ByteDance Seed, February 2026. URL https://seed.bytedance.com/en/blog/seedance-2-0-official-launch. Official technical release, February 12, 2026.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_VideoCrafter2_Overcoming_Data_Limitations_for_High-Quality_Video_Diffusion_Models_CVPR_2024_paper.html.

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

Eli Collins. Fuel your creativity with new generative media models and tools. Google, May 2025. URL https://blog.google/innovation-and-ai/products/generative-media-models-io-2025/. Official announcement introducing Veo 3, May 20, 2025.

Yang Du, Zhuoran Lin, Kaiqiang Song, Biao Wang, Zhicheng Zheng, Tiezheng Ge, Bo Zheng, and Qin Jin. VC4VG: Optimizing video captions for text-to-video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1124–1138. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.59. URL https://aclanthology.org/2025.emnlp-main.59/.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html.

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. *arXiv preprint arXiv:2606.21661*, 2026.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues, 2024. URL https://arxiv.org/abs/2401.16423.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions, 2024. URL https://arxiv.org/abs/2407.06358.

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*, pp. 28405–28415, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ke_MUSIQ_Multi-Scale_Image_Quality_Transformer_ICCV_2021_paper.html.

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, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

LAION AI. LAION-Aesthetics Predictor V1. GitHub repository, 2022. URL https://github.com/LAION-AI/aesthetic-predictor.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. AMT: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9801–9810, 2023.

Lightricks. LTX-2.5. LTX Documentation, 2026. URL https://docs.ltx.io/models/ltx-2-5. Open-weight joint audio-video model with native multi-shot generation.

Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Jianzhang Gao, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, Jiayi Ji, Fan Zhou, Liang Zheng, Shuicheng Yan, Hao Fei, and Tat-Seng Chua. JavisGPT: A unified multi-modal llm for sounding-video comprehension and generation. In *Advances in Neural Information Processing Systems*, 2025.

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. *arXiv preprint arXiv:2602.19163*, 2026. URL https://arxiv.org/abs/2602.19163.

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

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. In *European Conference on Computer Vision*, pp. 193–209. Springer, 2026.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, Eng Siong Chng, and Xie Chen. Omni-Captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Z091XLyVkJ.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. LongCat-Video technical report. *arXiv preprint arXiv:2510.22200*, 2025. URL https://arxiv.org/abs/2510.22200.

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*, pp. 461–471, 2026.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech*, pp. 2127–2131, 2021. doi: 10.21437/Interspeech.2021-299.

OpenAI. Sora 2 system card. OpenAI, September 2025. URL https://openai.com/index/sora-2-system-card/.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6493–6497, 2021.

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, 2025. URL https://arxiv.org/abs/2508.16930.

SII-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. *arXiv preprint arXiv:2602.08794*, 2026.

Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection, 2020. URL https://arxiv.org/abs/2008.04838.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. URL https://arxiv.org/abs/2506.15220.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 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. *arXiv preprint arXiv:2503.20314*, 2025.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2003.12039.

Tencent Hunyuan Team. Script-a-Video: Deep structured audio-visual captions via factorized streams and relational grounding. *arXiv preprint arXiv:2604.11244*, 2026. URL https://arxiv.org/abs/2604.11244.

Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. VidMuse: A simple video-to-music generation framework with long-short-term modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18782–18793, 2025.

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

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*, pp. 16268–16278, 2026a.

Yanghai Wang, Jiahao Wang, Jiafu Tang, Yuanxing Zhang, Zhe Cao, Hanyan Bian, Zijie Zhang, Weiliang Luo, Zhiyu Pan, Zixuan Dong, Jiaheng Liu, and Zhaoxiang Zhang. AVSCap: Orchestrating audio-visual synergy for omni-modal video captioning. *arXiv preprint arXiv:2607.12820*, 2026b. URL https://arxiv.org/abs/2607.12820.

Yuxi Wang, Yikang Wang, Qishan Zhang, Hiromitsu Nishizaki, and Ming Li. Vcapav: A video-caption based audio-visual deepfake detection dataset. In *Interspeech*, 2025.

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. *arXiv preprint arXiv:2605.20183*, 2026.

Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, and Tat-Seng Chua. Any2Caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379*, 2025. URL https://arxiv.org/abs/2503.24379.

Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. In *International Conference on Learning Representations*, 2026.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. URL https://arxiv.org/abs/2509.17765.

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. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Linli Yao, Yuancheng Wei, Yaojie Zhang, Lei Li, Xinlong Chen, Feifan Song, Ziyue Wang, Kun Ouyang, Yuanxin Liu, Lingpeng Kong, Qi Liu, Pengfei Wan, Kun Gai, Yuanxing Zhang, and Xu Sun. TimeChat-Captioner: Scripting multi-scene videos with time-aware and structural audio-visual captions. *arXiv preprint arXiv:2602.08711*, 2026. URL https://arxiv.org/abs/2602.08711.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. MovieDreamer: Hierarchical generation for coherent long visual sequences. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7d4647db960780c8b5a4e7d9e4a58d68-Abstract-Conference.html.

Chen Zhao, Jiajun Ma, Qilong Huang, Tiehan Fan, Hongyu Li, Zhuoliang Kang, Xiaoming Wei, Jian Yang, and Ying Tai. Temporal and cross-modal alignment for enhanced audiovisual video captioning. In *European Conference on Computer Vision*, pp. 65–85. Springer, 2026.

APPENDIX CONTENTS

A Data Sources and Multi-Shot Construction 16

B Dataset Statistics 17

C Filtering and Final Selection Details 19

D Qualitative Comparison and Analysis 22

E Details of the Controlled State and Relation Evaluation 24

F Training Details 26

G Preference Optimization of the Captioner 29

H SSRS Conversion and Quality Control 30

I Evaluation Metrics and Implementation Details 33

A Data Sources and Multi-Shot Construction

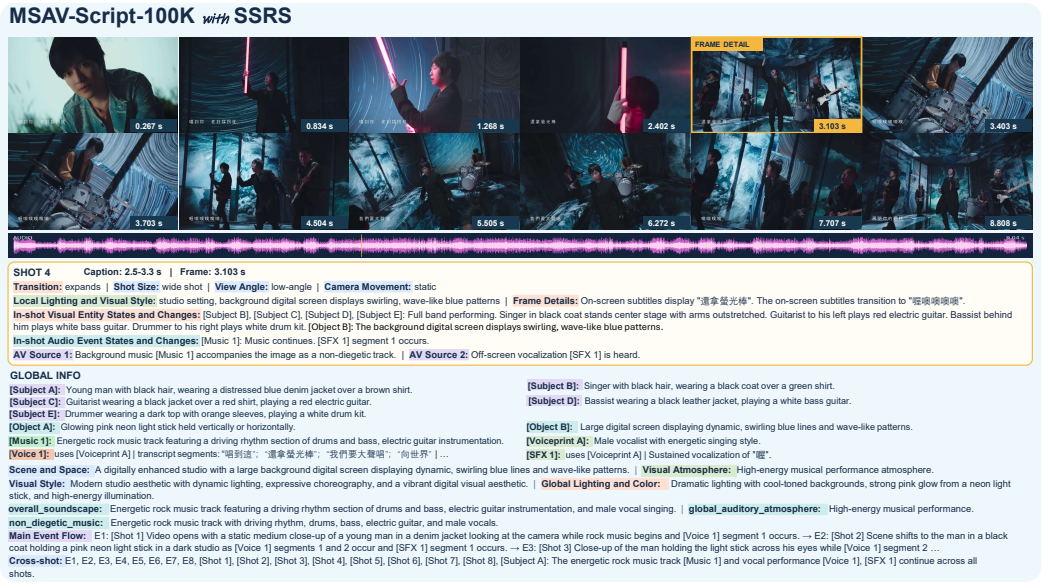

Figure 5: **An example from MSAV-Script-100K with SSRS annotations.** Sampled video frames and the audio waveform are shown together with shared entity and voice registries, global audio-visual context, shot-local states and source relations, and a cross-shot event sequence. The highlighted panel shows a shot-level annotation excerpt.

This appendix details the source collection and temporal construction of MSAV-Script-100K. Appendix B reports dataset composition and temporal, visual, and audio statistics, and Appendix C describes screening and final selection. The processing unit is a contiguous audio-visual clip containing multiple shots, paired with a dense description and its SSRS representation.

A.1 Data Sources and Content Coverage

Our source pool combines FineVideo, MiraData, VCapAV, and VidMuse (VidMuse-V2M-Dataset) with videos collected from YouTube. The YouTube collection covers six categories: food, music videos, animations, documentaries, sports events, and esports events. These categories broaden the range of visual activities and auditory content beyond cinematic dialogue. For example, food preparation provides action-associated sounds, musical performances combine recurring subjects with sustained music, and sports and esports footage includes commentary over changing views. These examples describe the intended coverage of the collection; individual clips need not contain every type of interaction associated with their category.

The collection targets short sequences with naturally occurring shot transitions and their corresponding audio. Native audio preserves the relationship between the original visual editing and the sound track, including cases in which narration, music, ambience, or an utterance continues across a cut. Accordingly, a change of shot does not require a change of speaker, sound source, or auditory event.

A.2 Temporal Construction of Multi-Shot Clips

We use TransNet V2 (Souček & Lokoč, 2020) to detect shot boundaries in source videos. FFmpeg cuts the corresponding audiovisual segments at the resulting clip boundaries. Consecutive shots are grouped into clips of approximately 6–12 seconds. Each clip preserves the original shot order and the audio associated with the same continuous source interval, allowing speech, music, and ambience to continue across visual cuts.

Let $b_0 < b_1 < \cdots < b_K$ denote the boundaries of $K \geq 2$ consecutive shots in the source timeline. A clip spanning $[b_0, b_K)$ has duration $b_K - b_0$, with clip-local shot intervals

$$\tau_i = [b_{i-1} - b_0,\ b_i - b_0), \qquad i = 1, \ldots, K. \tag{2}$$

These timestamps express visual segmentation relative to the clip start. An auditory event may span several intervals τ_i; its boundaries are not required to coincide with visual cuts.

B Dataset Statistics

B.1 Dataset Composition

All statistics in this section refer to the final MSAV-Script-100K. Table 4 summarizes its scale and temporal structure; Tables 5 and 6 report source composition and short-side resolution distribution. Counts and percentages use clips as the counting unit.

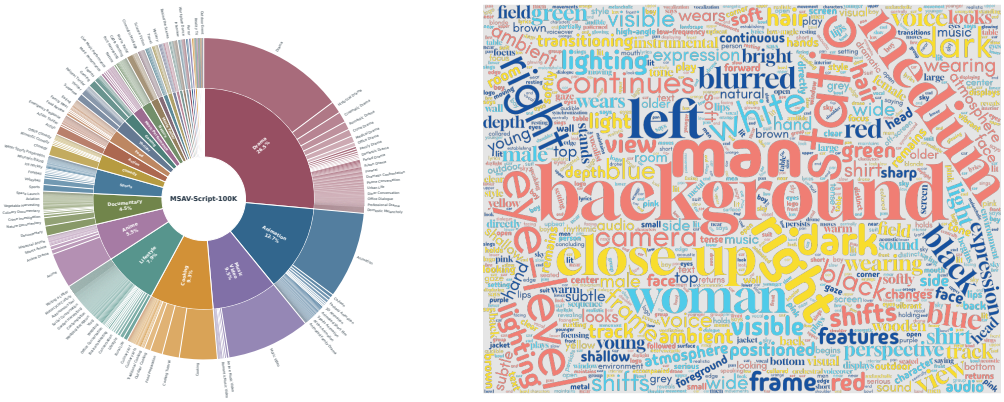

Figure 6: Thematic taxonomy (left) of MSAV-Script-100K and word cloud (right) of SSRS, illustrating the dataset's topic hierarchy and frequently occurring words in its captions.

Table 4: Dataset statistics for the selected 100,000 clips. Audio tags and shot counts are automatic estimates. Audio categories may overlap.

Group	Statistic	Value
Scale	Number of selected clips	100,000
	Total duration in hours	237.426
Sources	Clip counts by source	Table 5
Temporal structure	Mean clip duration (s)	8.547
	Median clip duration (s)	7.966
	Clip duration range (s)	6.083–12.040
	Detected shots per clip: mean, median, and range	5.346 / 5 / 2–75
Visual properties	Short-side resolution	Table 6
Audio content	Automatically tagged clips by sound category	Speech: 54.781% Music: 49.737% Other: 4.821%

Duration measurements. The selected clips contain approximately 237.426 hours of video. Their mean and median durations are 8.547 and 7.966 seconds, respectively, with a range of 6.083–12.040 seconds.

Table 5: Source composition of the final 100,000 clips.

Source	Clips
YouTube collections	74,114
MiraData	23,950
VidMuse	918
FineVideo	518
VCapAV	500
Total	100,000

Table 6: Resolution distribution of MSAV-Script-100K.

Short side (pixels)	Clips	Share (%)
720–1,079	51,832	51.832
1,080–1,439	47,907	47.907
1,440–2,159	147	0.147
$\geq$ 2,160	114	0.114
Total	100,000	100.000

Automatic audio tagging. We apply an AudioSet-trained Audio Spectrogram Transformer (AST) to all 100,000 selected clips.[1] Audio is decoded to mono at 16 kHz and processed in windows of at most 10 seconds, with a 5-second stride and an end-aligned final window. A clip receives a category tag when the maximum sigmoid score over its windows and the category's member classes is at least 0.5. Speech comprises the AudioSet Speech class and its descendants; music comprises Music, Singing, and their descendants. Other sounds comprise the remaining human, animal, natural, object, source-ambiguous, and noise events, excluding Silence and recording-context labels. The tags may overlap, and all percentages use 100,000 clips as the denominator.

These are automatic tagger-positive fractions. At the primary threshold, 18.973% of clips remain unassigned. The estimates are sensitive to the decision threshold: at 0.3, the speech, music, and other-sound fractions are 65.046%, 68.165%, 13.061%, respectively; at 0.7, they are 23.232%, 21.159%, 1.694%, respectively.

Automatic shot counting. The shot-count summaries report automatic estimates of the number of visual segments per clip. A clip with n internal shot boundaries contains $n + 1$ shots. These counts describe visual segmentation; an audio event may continue across multiple shots.

B.2 Temporal, Visual, and Audio Distributions

The 5th and 95th percentiles of clip duration are 7.840 and 11.595 seconds. The automatic shot-count summaries contain 534,571 detected shots, and the 5th and 95th percentiles of shots per clip are 3 and 10. These statistics describe both clip length and the number of visual segments within each clip.

Table 6 summarizes the distribution by the shorter image dimension: 51.832% of clips fall in the 720–1,079-pixel bin, and 48.168% have a short side of at least 1,080 pixels.

The audio categories are reported as overlapping clip-level tags. At the primary threshold of 0.5, 25.672% of clips receive both speech and music tags; this indicates detection of both categories somewhere within a clip, rather than necessarily at the same instant. The 18.973% of clips receiving none of the three tags should not be interpreted as silent clips.

B.3 Comparison with FineVideo (Farré et al., 2024)

Compared with FineVideo (Farré et al., 2024), MSAV-Script-100K exhibits stronger visual quality, audio quality, and audio-visual alignment under the reported automatic metrics, as shown in Figure 7. For visual quality, it achieves higher mean LAION aesthetic (5.471 vs. 4.349) and MUSIQ scores (66.344 vs. 56.517), along with higher AMT-based motion smoothness (0.9855 vs. 0.9789). For audio quality, AudioBox assigns higher mean scores for content enjoyment (5.335 vs. 4.655), content usefulness (6.049 vs. 5.025), and production quality (6.920 vs. 5.869). The mean Synchformer zero-offset probability is also higher (0.301 vs. 0.266), indicating stronger estimated temporal alignment between the visual and audio streams. These advantages are reflected in both the means and medians of all seven metrics, supporting the quality of MSAV-Script-100K as a curated resource for multi-shot audio-visual generation.

[1] Checkpoint: `https://huggingface.co/MIT/ast-finetuned-audioset-10-10-0.4593`.

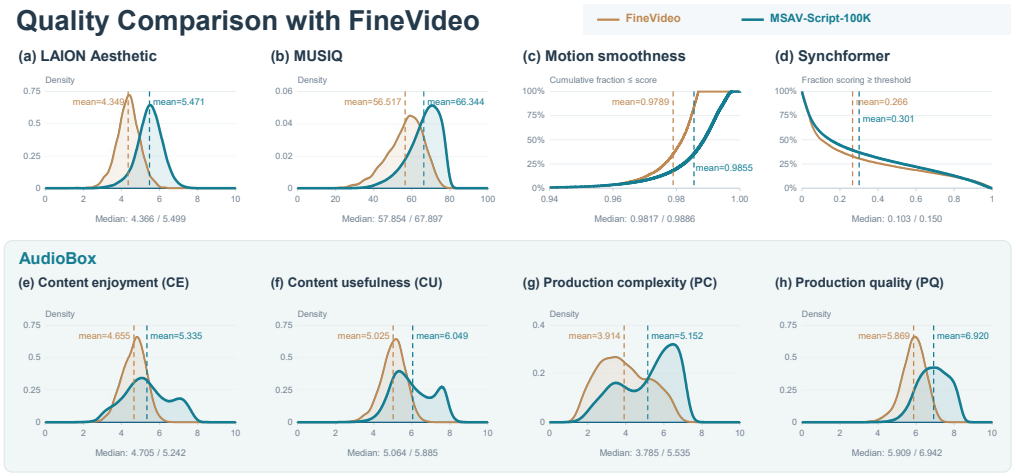

Figure 7: Quality comparison between MSAV-Script-100K and FineVideo.

C Filtering and Final Selection Details

The curation procedure combines perceptual screening, semantic screening, and multi-criterion quality assessment. The following sections distinguish metric-based rejection from the later ranking of retained candidates.

C.1 Resolution and Visual Screening

Resolution and visual screening. We retain clips whose shorter image dimension is at least 720 pixels, treating landscape and portrait orientations consistently. Table 7 summarizes the subsequent six-metric visual filter. Let $P_q(m)$ denote the qth percentile of metric m in the jointly valid reference population used to calibrate the visual filter. Calibration uses 242,109 clips with finite, valid values for all six metrics. Thresholds are computed once on this population and remain fixed through the successive filtering steps; equality at a threshold is retained. A clip must satisfy every applicable retention condition in the table. Percentile thresholds identify the tails of each metric distribution; because the rejected sets can overlap, the overall rejection rate is not the sum of the per-metric tail fractions.

Table 7: Visual screening criteria. Numerical thresholds are rounded for display; filtering uses the stored full-precision values. MUSIQ scores are divided by 100, and RAFT magnitudes are normalized by the evaluation-frame diagonal.

Metric m	Retention condition	Numerical threshold
Sharpness	$m \geq P_5(m)$	10.278426
Image entropy	$m \geq P_5(m)$	5.359885
Flat-region proportion	$m \leq P_{95}(m)$	0.906250
Texture score	$m \leq P_{95}(m)$	1.128293
MUSIQ image quality	$m \geq P_5(m)$	0.391477
RAFT motion magnitude	$P_5(m) \leq m \leq P_{95}(m)$	$[0.004140, 0.046509]$

Spatial descriptors and sampling. Spatial metrics use frames sampled at 1 FPS, with uniform subsampling if more than 64 frames are available. Handcrafted descriptors are computed at the decoded spatial dimensions and averaged over frames. Sharpness is the variance of the grayscale Laplacian. Image entropy is the base-2 Shannon entropy of the 256-bin grayscale histogram. Flat-region proportion is the fraction of complete, non-overlapping 240×240 patches whose Sobel-gradient-magnitude variance is below 800; incomplete border patches are excluded. Texture is gray-level co-occurrence-matrix contrast after quantization to 16 levels, averaged over distance-one offsets at 0, $\pi/4$, $\pi/2$, and $3\pi/4$, using symmetric, normalized matrices.

Learned visual metrics. MUSIQ uses the SPAQ checkpoint. Frames with a long side above 512 pixels are resized to that limit, and the clip score is the mean frame prediction divided by 100. RAFT uses `raft-things.pth` with 20 iterations on adjacent sampled frames at 8 FPS, capped at 256 frames per clip, with the long side limited to 512 pixels. Flow magnitude is averaged over pixels, divided by the resized image diagonal, and then averaged over frame pairs. The motion calculation includes pairs spanning shot transitions, so its upper-tail filter responds to editing cuts as well as within-shot motion.

C.2 Low-Level Audio Screening

We compute four descriptors to characterize audio before learned quality assessment: silence ratio, clipping ratio, an amplitude-based signal-to-noise proxy, and effective bandwidth. These descriptors measure waveform and spectral properties; speech quality and audio aesthetics are evaluated separately in Appendix C.4.

The scorer operates on mono floating-point waveforms at the sampling rate of its evaluation input. The 16 kHz mono evaluation proxies have an 8 kHz Nyquist limit, so bandwidth measured on these proxies describes the evaluation signal. Analysis uses 2,048-sample frames with a 512-sample hop; inputs shorter than one frame are zero-padded. All clips entering this stage contain non-silent audio; the silence ratio measures relatively quiet intervals within each clip. For samples y_n and frame RMS values R_t, the descriptors are:

- **Silence ratio:** the fraction of frames satisfying $20\log_{10}(\max(R_t, 10^{-12})/R_{\max}) < -40$, where $R_{\max} = \max_t R_t$ and R_t includes a 10^{-20} stabilizer inside the square root. This relative-amplitude measure is distinct from voice-activity detection.
- **Clipping ratio:** the fraction of samples with $|y_n| \geq 0.99$.
- **Amplitude SNR proxy:** $10\log_{10}(P_{\text{high}}/P_{\text{low}})$, where the two powers are the mean squared amplitudes of samples at or above the 85th percentile and at or below the 15th percentile of $|y_n|$, respectively. The implementation returns zero when $P_{\text{high}} = 0$, and 100 dB when $P_{\text{high}} > 0$ and $P_{\text{low}} \leq 10^{-20}$. This statistic does not separate clean-signal power from noise power.
- **Effective bandwidth:** the first frequency bin at which cumulative power reaches 99% of the time-averaged, one-sided power spectrum, using a Hann window. A zero-energy spectrum yields zero bandwidth.

We use these four audio descriptors to filter out clips with excessive silence, substantial waveform clipping, low amplitude contrast, or overly narrow effective bandwidth. This screening removes clips with undesirable waveform and spectral characteristics before the subsequent learned audio-quality assessment.

C.3 Semantic Screening

We use Qwen3.5-27B (Qwen Team, 2026) to assign a semantic meaningfulness score $s_{\text{sem}}(x) \in [0, 100]$ to each candidate clip x. We retain clips satisfying

$$s_{\text{sem}}(x) \geq 90. \tag{3}$$

This assessment uses video frames only. It provides a content-level criterion beyond low-level visual properties, but does not directly verify speech intelligibility, sound-source correspondence, or audio continuity. These audio-related properties require the complementary audio and audio-visual assessments below.

Scoring criterion and visual inputs. The prompt evaluates visual semantic meaningfulness and editorial coherence. It rewards shots that form a coherent scene, action, event, narrative progression, montage, reaction/cutaway structure, or thematic sequence, and penalizes accidental concatenation or disconnected source fragments. Normal cuts, changes of viewpoint, flashbacks, and intentional discontinuities are acceptable when a plausible narrative, spatial, temporal, causal, thematic, or visual relationship remains. The 90–100 band denotes a clearly meaningful, coherent, and purposeful

sequence. The prompt excludes blur, resolution, exposure, visual aesthetics, acting quality, and audio quality from this semantic judgment.

For scoring with ordered image inputs, we divide a clip of duration D into eight equal-duration intervals and sample one frame at the center of each interval:

$$t_i = \frac{i + 1/2}{8} D, \qquad i = 0, \ldots, 7. \tag{4}$$

The sampled frames are resized while preserving their aspect ratios, with the shorter side capped at 480 pixels. The eight frames are supplied to Qwen3.5-27B in chronological order together with their approximate timestamps, without audio. The model jointly assesses the sequence for meaningful relationships among shots, complementing the preceding frame-level visual quality measurement.

Score output. Inference uses temperature zero, and the model is instructed to return a single integer from 0 to 100. This score is compared directly with the retention threshold above.

C.4 Tri-Dimensional Quality Assessment

After screening, we assess retained clips along three quality axes using the five criteria in Table 8. The visual axis contributes two criteria (LAION aesthetics and AMT motion smoothness), the audio axis contributes two (AudioBox and DNSMOS/NISQA), and audio-visual synchronization contributes one (Synchformer).

Table 8: Five criteria used for quality assessment. Each criterion contributes one fifth of the overall score after within-source percentile conversion and criterion-level aggregation.

Axis	Criterion	Evaluator
Visual	Visual aesthetics	LAION Aesthetic Predictor
Visual	Motion smoothness	AMT
Audio	Audio aesthetics and production quality	AudioBox
Audio	Speech quality	DNSMOS and NISQA
Audio-visual	Temporal synchronization	Synchformer zero-offset probability

Evaluator inputs and clip-level aggregation. Visual aesthetics uses CLIP ViT-L/14 with the LAION head `sa_0_4_vit_l_14_linear.pth`, averaging predictions from the spatial frame samples and dividing by ten. Motion smoothness uses AMT-S with `amt-s.pth`. Up to 32 uniformly spaced valid center indices form consecutive-frame triplets (I_{t-1}, I_t, I_{t+1}). AMT predicts the middle frame from its neighbors; the clip score is one minus the mean absolute reconstruction error divided by 255. This is a sampled approximation rather than an evaluation of every frame.

AudioBox uses `facebook/audiobox-aesthetics` on 16 kHz mono audio in consecutive, non-overlapping 10-second windows. The final short window is zero-padded with a valid-sample mask. Predictions are restored to their native scales and averaged with weights proportional to valid window duration. Its four outputs are Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), and Production Quality (PQ). DNSMOS uses the calibrated OVRL output from `sig_bak_ovr.onnx`, with 16 kHz audio and 9.01-second windows at a one-second stride; short inputs are repeated until a complete window is available. NISQA uses `nisqa.tar`, its checkpoint-defined spectrogram configuration, and a 48 kHz evaluation waveform to produce MOS. Resampling an evaluation proxy does not restore frequency content absent from that proxy.

Synchformer uses checkpoint `24-01-04T16-39-21`, with 25 FPS video, 16 kHz audio, and zero imposed temporal offset. Its raw synchronization score is the softmax probability of the offset class nearest zero. This probability is converted to a percentile before entering the five-criterion average.

Within-source normalization and aggregation. Here, source groups are defined by the original dataset or collection-category label in order to keep data diversity. Raw evaluator outputs have different scales. For metric m, let $n_{s,m}$ be the number of finite, applicable values among candidates with source label s, and let $r_{s,m}(x)$ be the ascending rank of clip x, using average ranks for ties. We

normalize each metric within the source label stored in the candidate metadata:

$$p_{s,m}(x) = \begin{cases} (r_{s,m}(x) - 1)/(n_{s,m} - 1), & n_{s,m} > 1, \\ 0.5, & n_{s,m} = 1. \end{cases} \tag{5}$$

Suppressing the source index, the five subscores are

$$\begin{aligned} c_1(x) &= p_{\mathrm{LAION}}(x), \\ c_2(x) &= p_{\mathrm{AMT}}(x), \\ c_3(x) &= \frac{p_{\mathrm{CE}}(x) + p_{\mathrm{CU}}(x) + p_{\mathrm{PC}}(x) + p_{\mathrm{PQ}}(x)}{4}, \\ c_4(x) &= \frac{p_{\mathrm{DNSMOS\ OVRL}}(x) + p_{\mathrm{NISQA\ MOS}}(x)}{2}, \\ c_5(x) &= p_{\mathrm{sync}}(x). \end{aligned} \tag{6}$$

The four AudioBox axes enter with the same positive sign. Their percentile average is our selection criterion, not an official AudioBox aggregate. The overall quality score is

$$Q(x) = \frac{1}{5} \sum_{j=1}^{5} c_j(x). \tag{7}$$

Thus, equal weighting applies to the five criteria. The total weights of the visual, audio, and synchronization axes are $2/5$, $2/5$, and $1/5$, respectively. Within-source normalization measures a clip's relative standing within its source and reduces the influence of differences in raw score distributions across sources. It does not establish that clips with identical percentiles from different sources have identical absolute quality.

Scope of the quality signals. Each evaluator provides a complementary signal. Speech-quality scores are relevant when speech is present, whereas music, ambience, and sound effects also occur in clips without speech. Synchformer's zero-offset probability is used as a temporal-alignment signal; it does not by itself establish the correctness of every sound-source attribution or cross-shot event relation. Criterion-level scores retain the contributions of these distinct signals alongside the aggregate quality score.

Metric applicability. Speech evaluators run when the VAD record is valid and contains at least one detected speech segment with at least 0.5 seconds of total speech. Samples skipped by this gate receive a neutral percentile of 0.5 for each speech-quality metric, giving $c_4 = 0.5$. Synchformer requires at least five seconds of audio; an explicitly recorded short-input skip likewise receives a neutral synchronization percentile of 0.5. These assignments keep five terms in $Q(x)$ instead of changing their weights from clip to clip. The evaluator records distinguish completed predictions, applicability skips, and processing errors; an applicability skip is not a low-quality prediction.

C.5 Final Selection

After visual and semantic screening and audio assessment, we select 100,000 clips through quality-based ranking within source groups, allocating quotas proportionally to their candidate counts to preserve source diversity. Within each group, clips are ranked in descending order of the aggregate quality score $Q(x)$ defined in Eq. 7, with ties broken by ascending clip identifier. Table 5 reports the source composition of the final dataset.

D Qualitative Comparison and Analysis

We provide representative qualitative comparisons to complement the quantitative evaluation. Figures 8–12 compare SSRS SFT with representative generation systems on multi-shot event progression, while Figures 13 –17 compare Dense SFT and SSRS SFT under matched training conditions. Frames are sampled at the requested shot windows; auditory relations in the controlled cases are additionally evaluated using the native generated audio. **All qualitative examples are provided in the supplementary material.**

Ordered event progression. Figure 8 shows a four-stage fire-lighting sequence: inspecting kindling, placing twigs, producing a flame, and warming beside the fire. SSRS produces a clearer progression from preparation to ignition and the resulting fire state, while competing methods either repeat similar states, introduce the fire prematurely, or drift from the requested event sequence. This case illustrates the benefit of explicitly organizing local state changes through a global event order.

Multi-stage state evolution. Figure 9 considers planting mint through five successive stages. SSRS more clearly assigns different stages to different shots, progressing from preparation to planting, soil treatment, and the final relocation toward light. In contrast, other methods frequently repeat similar views or reveal later-stage content too early. The comparison highlights that SSRS encourages shot-specific states to evolve according to the intended temporal structure rather than merely maintaining topical consistency.

Dialogue-driven shot organization. Figure 10 evaluates a museum conversation involving walking together, a request from the man, and a subsequent response from the woman. SSRS preserves the recurring characters while adapting the shot composition to the evolving dialogue, moving from a shared view to speaker-focused and response-focused views. Other methods tend to repeat similar compositions or provide weaker alignment between narrative roles and shot changes. This demonstrates that cross-shot consistency should preserve identity while still allowing the intended presentation and interaction states to change.

Role-consistent event progression. Figure 11 evaluates whether persistent entity roles are maintained while actions evolve across shots. The requested sequence requires the leader to point, the group to run, and the same leader to finally grasp the flagpole. SSRS better preserves the leader–group roles while following the intended action order, whereas competing methods tend to repeat visually similar states or provide weaker correspondence between entity roles and requested actions. This example illustrates the benefit of jointly organizing persistent identities and cross-shot event evolution.

Fine-grained object-state evolution. Figure 12 considers a two-stage manipulation sequence in which a toolbox is first lowered onto a cart and then secured by tightening its strap. SSRS maintains the toolbox–cart relation while making the strap interaction and the resulting secured state explicit. Although the pretrained LTX-2.3 already captures much of the requested sequence, SSRS more completely realizes the fine-grained object states and interactions, demonstrating improved adherence beyond coarse task completion.

Sound-source binding. Figure 13 isolates the relation between a visual action and its designated sound source. The prompt requires only the right blue bell to be struck in the second shot. SSRS realizes the visible interaction with the specified bell and better preserves the bell configuration across shots, whereas Dense SFT misses the requested strike and later introduces an additional bell. Together with the native-audio evaluation, this case illustrates the benefit of explicitly representing source attribution and audio-visual temporal relations.

Cross-shot object-state persistence. Figure 14 tests whether a yellow backpack remains associated with the green station bench across shot transitions. Dense SFT instead places the backpack on a traveler, changing the intended object–location relation. SSRS preserves the backpack–bench state in the visible wide shots despite an intervening close-up in which the backpack is not observable. This example highlights that temporary visibility changes should not be interpreted as state changes, consistent with SSRS's explicit separation of persistent identity from shot-local visibility and state.

Non-speech sound-source binding. Figure 15 further evaluates source-specific audio-visual grounding. The request requires the right red bag to be unzipped in the second shot with a synchronized zipper sound. SSRS realizes the interaction with the designated red bag while preserving the surrounding scene, whereas Dense SFT leaves both bags largely static. This example shows that SSRS can better associate a non-speech auditory event with its intended visual entity and action. The synchronization of the zipper sound is verified using the native generated audio.

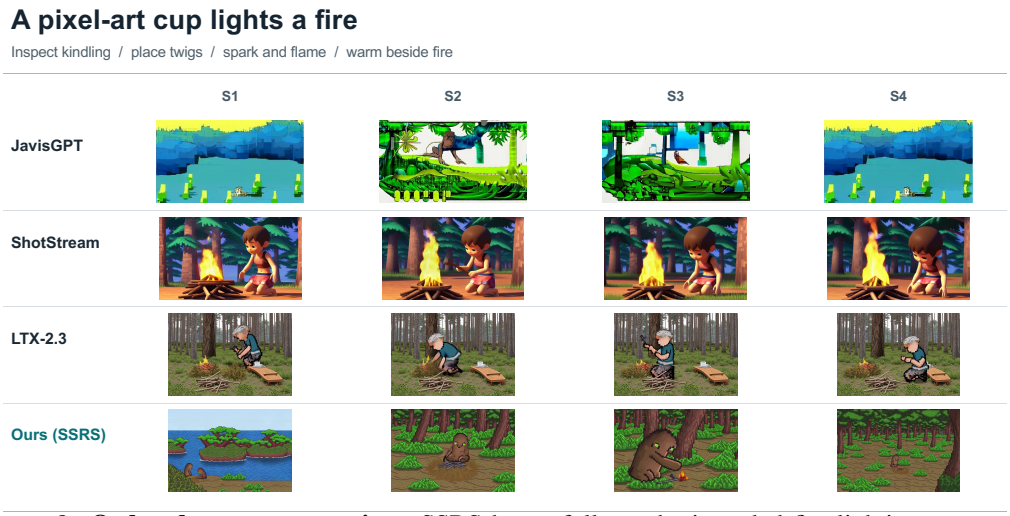

Figure 8: **Ordered event progression.** SSRS better follows the intended fire-lighting sequence, while competing methods exhibit repeated, premature, or weakly ordered states.

Speaker identity binding. Figure 16 examines whether speech is assigned to the intended recurring character. The request specifies that repairer B, identified by the blue workwear, speaks in the second shot while repairer A remains silent. SSRS better preserves the target character across shots and associates the requested speech with repairer B, whereas Dense SFT exhibits substantial identity drift in the later shot. This case illustrates the role of shared visual and voice identities together with shot-level source relations in maintaining speaker attribution. Speaker activity is verified using the native generated audio.

Cross-shot audio continuity. Figure 17 evaluates whether an ongoing auditory event persists across visual cuts. The request requires a single acoustic-guitar passage to continue through two scene changes. SSRS produces distinct visual views while maintaining the continuous guitar event, whereas Dense SFT remains in nearly the same view before failing to realize the final requested shot. This behavior supports a central motivation of SSRS: visual shot boundaries need not coincide with auditory event boundaries. Audio continuity is verified using the native generated audio.

Overall, these examples show that the gains of SSRS are not limited to visual similarity across shots. By explicitly representing persistent entities, evolving states, audio-visual relations, and cross-shot event organization, SSRS better follows both intended continuity and intended change in multi-shot generation.

E Details of the Controlled State and Relation Evaluation

Diagnostic design. The controlled evaluation is designed to directly probe the stateful and relational behaviors targeted by SSRS, rather than serve as an additional general-purpose benchmark. We evaluate 40 matched prompt instances, producing 40 Dense SFT outputs and 40 SSRS SFT outputs. The instances cover five diagnostic categories: visual-state evolution (VS), cross-shot audio continuity (AC), non-speech sound-source binding (NS), cross-shot event order (EO), and on-/off-screen speaker binding (SB). Because the category sizes are small and unbalanced, category-level results are treated as descriptive diagnostic evidence rather than statistically well-powered benchmark estimates.

Questions and scoring. For each generated video, we predefine two *target questions* and four *preservation questions*. Target questions test whether the requested state transition or relation is realized, whereas preservation questions test whether non-target requirements remain satisfied. Gemini 3.5 Flash observes the generated video together with its native audio and answers the fixed binary questions. Requested content that is absent from the generation is counted as unsatisfied rather than inapplicable.

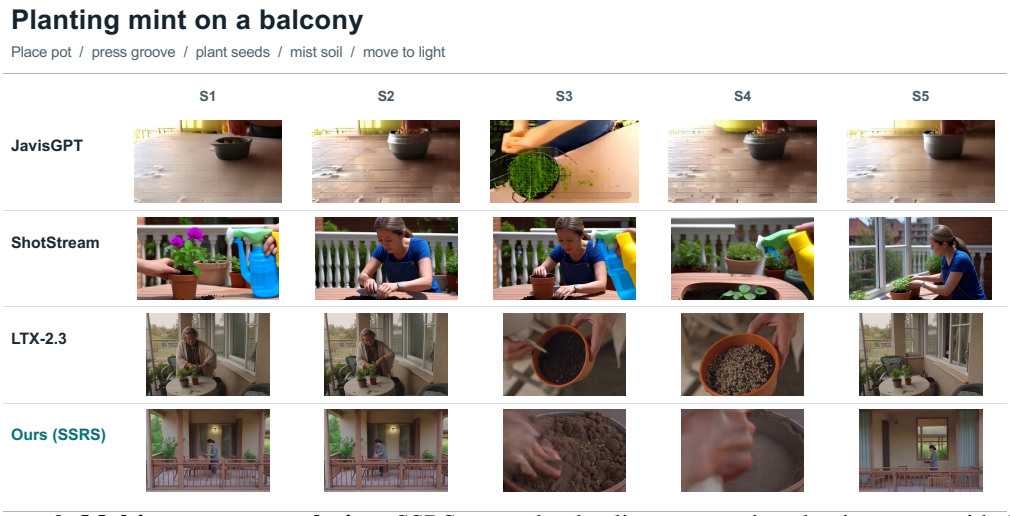

Figure 9: **Multi-stage state evolution.** SSRS more clearly aligns successive planting states with the requested shot sequence, whereas competing methods frequently repeat or collapse stages.

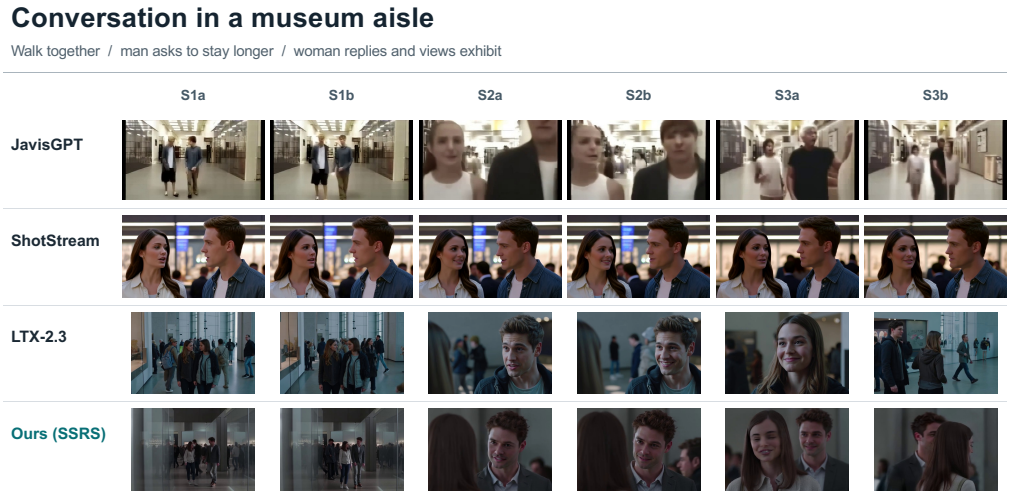

Figure 10: **Dialogue-driven shot organization.** SSRS preserves recurring identities while adapting shot composition to changing speaker and interaction states.

For video i, we define

$$T_i = \frac{1}{2}\sum_{q=1}^{2} \mathbf{1}[y_{iq}^{T} = \text{yes}], \qquad P_i = \frac{1}{4}\sum_{q=1}^{4} \mathbf{1}[y_{iq}^{P} = \text{yes}], \tag{8}$$

where y_{iq}^{T} and y_{iq}^{P} denote the answers to target and preservation questions, respectively. Category-level and overall scores are obtained by averaging T_i and P_i over the corresponding matched prompt instances. Dense SFT and SSRS SFT are evaluated using the same underlying requirements and question sets.

Diagnostic dimensions. VS evaluates whether an entity follows a requested state change while preserving its identity and other shared attributes. AC tests whether an auditory event correctly continues or changes across visual cuts. NS evaluates whether a non-speech sound is assigned to the intended visual source. EO tests the specified ordering of events across shots. SB evaluates whether speech is correctly attributed to an on-screen or off-screen speaker, including cases in which the visible listener differs from the continuing speaker.



Figure 11: **Role-consistent cross-shot event progression.** SSRS preserves the leader identity while following the requested point–run–grasp sequence across shots.

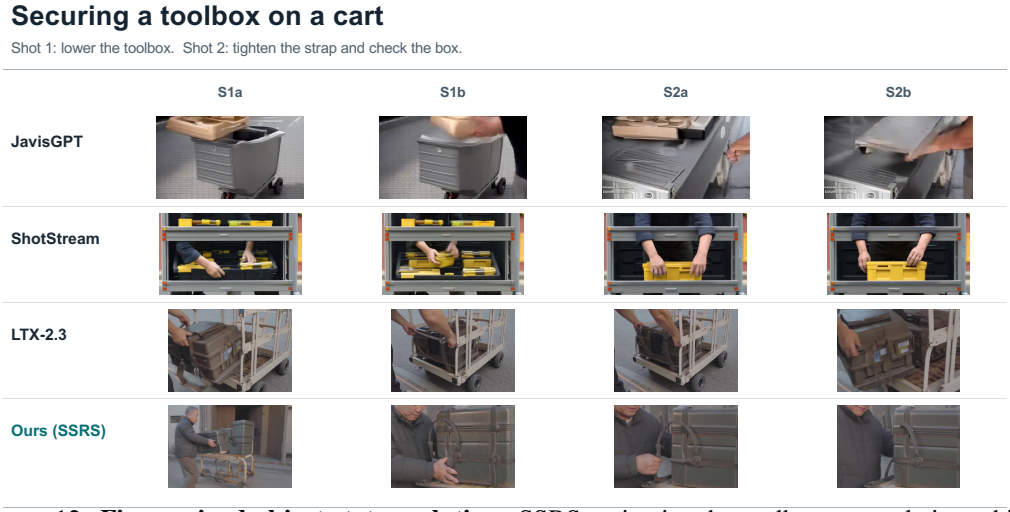

Figure 12: **Fine-grained object-state evolution.** SSRS maintains the toolbox–cart relation while explicitly realizing the requested lowering, strap-tightening, and secured final state.

Additional analysis. SSRS improves target satisfaction in every category: VS by 40.00 pp, AC by 50.00 pp, NS by 40.00 pp, EO by 40.00 pp, and SB by 35.71 pp. Preservation also improves for AC, NS, EO, and SB. VS shows the opposite preservation trend: T increases from 20.00% to 60.00%, whereas P decreases from 62.50% to 42.50%. This suggests that stronger compliance with requested visual-state changes may occasionally perturb non-target visual properties. The overall preservation rate, however, remains comparable and slightly higher for SSRS (52.50% vs. 49.38%).

These results should be interpreted as a controlled diagnostic analysis. In particular, the relatively small numbers of EO and SB instances limit fine-grained statistical conclusions. The primary evidence is therefore the consistent increase in target satisfaction across all five independently defined state and relation categories.

F Training Details

Multi-Shot Audio-Visual Captioner. Table 9 summarizes the key hyperparameters used to fine-tune our Multi-Shot Audio-Visual Captioner. The captioner is initialized from the Qwen3-Omni

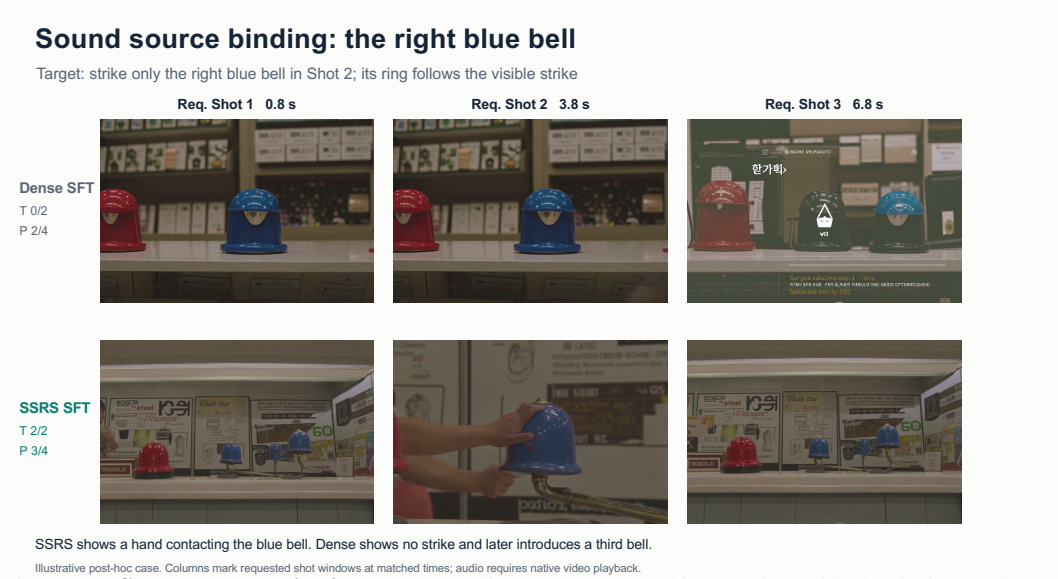

Figure 13: **Sound-source binding.** SSRS realizes the requested interaction with the designated blue bell, whereas Dense SFT misses the strike and exhibits object drift across shots.

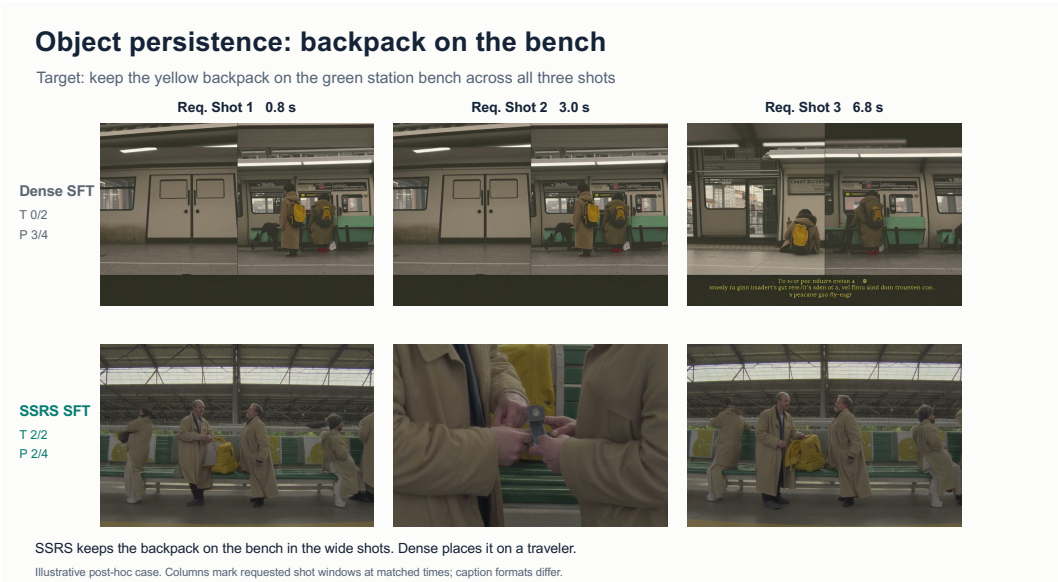

Figure 14: **Cross-shot object-state persistence.** SSRS preserves the backpack–bench relation across shot transitions, whereas Dense SFT incorrectly transfers the backpack to a traveler.

backbone and adapted using LoRA for parameter-efficient fine-tuning. Training is conducted on 8 NVIDIA A100 GPUs with 80 GB memory for three epochs, taking approximately 68 hours. We use the AdamW optimizer with a peak learning rate of 1.0×10^{-4} and a cosine learning-rate schedule. A warmup ratio of 0.05 is employed to stabilize the initial stage of cross-modal alignment learning.

Audio-Visual Diffusion Transformer. We further fine-tune the LTX-2.3 22B audio-visual diffusion transformer using the captions generated by our Multi-Shot Audio-Visual Captioner. Table 10 summarizes the key hyperparameters used to fine-tune our Audio-Visual Diffusion Transformer. The DiT is adapted using LoRA, while the original model parameters remain frozen. LoRA modules are inserted into the query, key, value, and output projections of the attention layers. Both the LoRA rank and scaling parameter are set to 64, and no LoRA dropout is applied.

Training is distributed over two compute nodes, each equipped with eight NVIDIA H800 GPUs with 80 GB memory, resulting in a total of 16 GPUs. Each GPU processes one sample per optimization

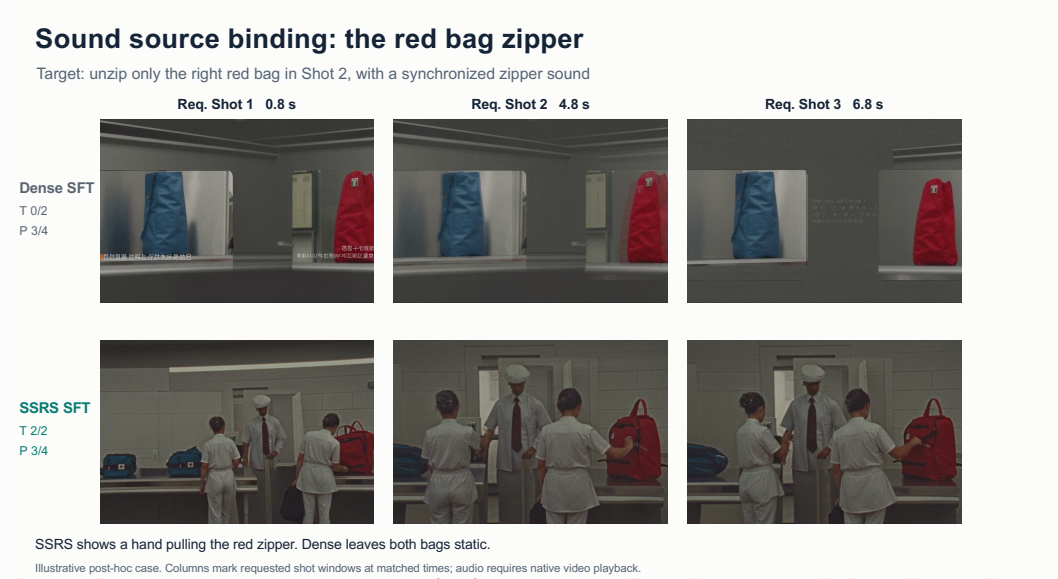

Figure 15: **Non-speech sound-source binding.** SSRS realizes the requested zipper interaction on the designated red bag, whereas Dense SFT leaves both bags largely static. Zipper synchronization is verified using the native generated audio.

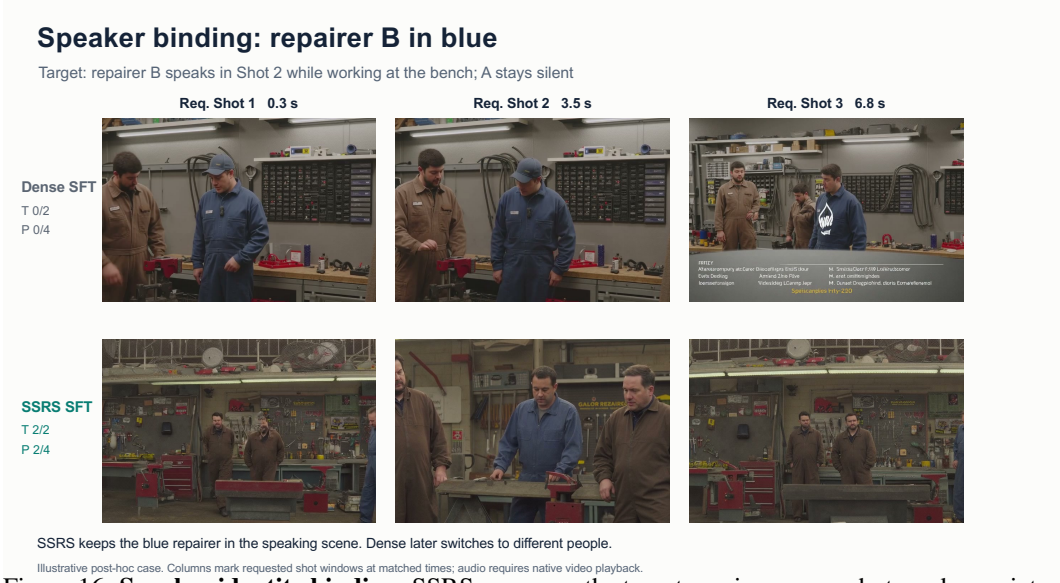

Figure 16: **Speaker identity binding.** SSRS preserves the target repairer across shots and associates the requested speech with repairer B, whereas Dense SFT exhibits identity drift. Speaker activity is verified using the native audio.

step, giving an effective global batch size of 16 without gradient accumulation. We use AdamW with a learning rate of 1.0×10^{-4}, a linear learning-rate schedule, and a maximum gradient norm of 1.0. Training is performed in BF16 precision with gradient checkpointing enabled.

Each training sample contains 121 video frames at a spatial resolution of 768×512. Rather than enforcing a fixed frame rate during training, we uniformly sample 121 frames over the complete duration of each source clip. Therefore, the effective sampling frame rate is adaptively determined by the source duration. The temporally aligned audio segment is encoded over the same interval. This strategy retains the complete temporal structure of clips with slightly different durations while maintaining a fixed tensor shape for distributed training.

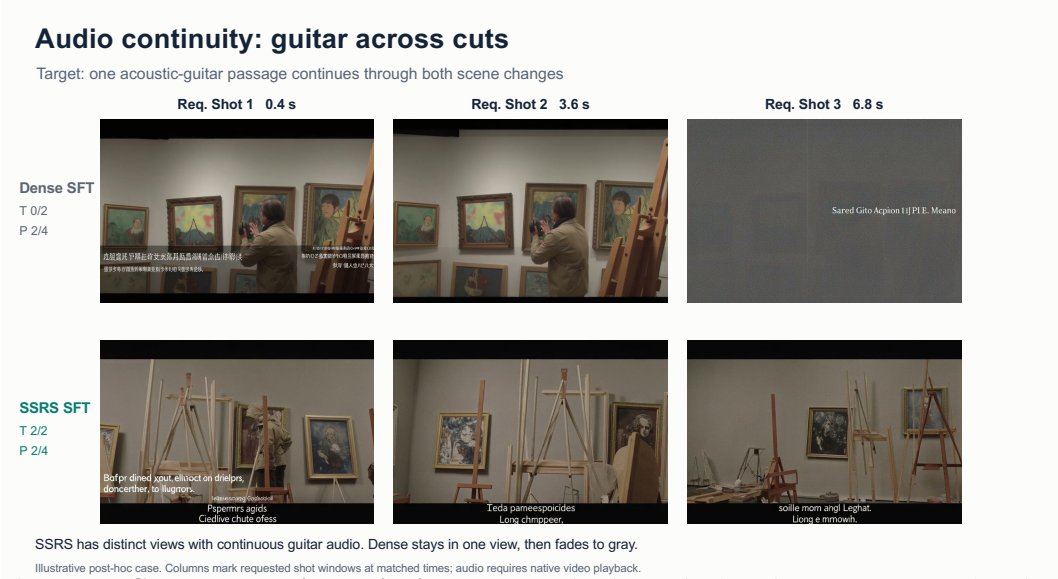

Figure 17: **Cross-shot audio continuity.** SSRS maintains a single guitar event across changing visual views, illustrating that auditory events can persist across shot boundaries, whereas Dense SFT fails to realize the requested multi-shot progression.

Table 9: Training hyperparameters of the Multi-Shot Audio-Visual Captioner.

Hyperparameter	**Configuration**
Backbone	Qwen3-Omni
GPU Usage	$8\times$ A100 80 GB
Batch Size per GPU	1
Gradient Accumulation	2
Effective Batch Size	16
Training Time	68 hours
Training Epochs	3
Peak Learning Rate	1.0×10^{-4}
Learning Rate Scheduler	Cosine
Optimizer	AdamW
Fine-tuning Type	LoRA
Warmup Ratio	0.05

To reduce repeated computation, text conditions and audio-video VAE latents are encoded once and cached before DiT optimization. The training stage consequently loads only the cached conditions and latents, avoiding repeated execution of the text encoder and audio-video VAE encoders. The cache is verified for completeness, identity consistency, and finite tensor values before optimization.

G Preference Optimization of the Captioner

Preference-pair construction. We perform direct preference optimization (DPO) after supervised fine-tuning (SFT) of the Qwen3-Omni-based captioner. For each multi-shot clip in the preference-data subset, the SFT captioner samples 20 dense captions from the same audio-visual input. Sampling uses a temperature of 0.7, top-p of 0.9, top-k of 50, and a maximum of 2,048 generated tokens. We then provide Gemini 3.1 Pro with these candidates and the clip's expert-constructed dense caption as a textual reference. The judge is asked to identify a preferred caption and three dispreferred captions exhibiting, respectively, the most severe visual errors, audio errors, and audio-visual alignment or temporal-logic errors. This produces up to three preference pairs per clip, each sharing the same preferred response but targeting a different failure mode. Unlike the initial annotation

Table 10: Training hyperparameters of the LTX-2.3 audio-visual DiT.

Hyperparameter	Configuration
Backbone	LTX-2.3 22B
GPU Usage	2×8 H800 80 GB
Batch Size per GPU	1
Gradient Accumulation	1
Effective Batch Size	16
Training Resolution	768×512
Frames per Sample	121
Training Frame Sampling	Adaptive full-duration
Audio Processing	Temporally aligned
Maximum Training Steps	20,000
Learning Rate	1.0×10^{-4}
Learning Rate Scheduler	Linear
Optimizer	AdamW
Maximum Gradient Norm	1.0
Numerical Precision	BF16
Gradient Checkpointing	Enabled
Fine-tuning Type	LoRA
LoRA Rank	64
LoRA Alpha	64
LoRA Dropout	0.0
LoRA Target Modules	Q/K/V/Output projections
Text Condition	Precomputed
Audio-Video Latents	Precomputed

pipeline, this preference-selection step compares captions against the reference text; the judge does not directly re-inspect the clip.

Across four multi-shot subsets, 3,781 clips yielded 11,341 pair records. Each record contains the clip's video and audio paths, the captioning instruction, a preferred response, and one dispreferred response. The preference data are used only to refine the captioner that produces dense descriptions; the subsequent dense-to-SSRS conversion is a separate stage.

Optimization details. We initialize DPO from the merged SFT checkpoint of Qwen3-Omni-30B-A3B and use that SFT model as the reference policy. Training uses the standard sigmoid DPO objective (Rafailov et al., 2023) with $\beta = 0.1$. We apply LoRA to the language model with rank 8, scaling factor 16, and zero LoRA dropout, while freezing the vision tower and multimodal projector. Training runs for three epochs, or 534 optimizer steps, on 32 H800 GPUs across four nodes with DeepSpeed ZeRO-3 and bfloat16 precision. The per-device batch size is one preference pair, with two gradient-accumulation steps, giving an effective batch size of 64 pairs. We use a learning rate of 10^{-4}, a cosine schedule, a 0.05 warm-up ratio, and a maximum sequence length of 22,000 tokens. The resulting captioner is used to annotate the larger clip collection before SSRS conversion.

H SSRS Conversion and Quality Control

We convert dense audio-visual captions into SSRS through a typed JSON intermediate representation, followed by deterministic text rendering. The conversion is designed to reorganize source information into explicit entities, states, and relations while preserving the source content and avoiding unsupported additions. This section describes the representation, conversion procedure, and automated quality-control mechanisms.

Canonical JSON representation. The intermediate representation contains global visual and audio attributes, a visual entity registry, an audio identity registry, an audio entity registry, an ordered shot sequence, and a global event sequence. Visual entities store their categories and appearance attributes. Audio identities describe reusable voice characteristics, while audio entities specify speech,

sound effects, ambience, and music, including their content and acoustic properties. Transcribed speech entries preserve the source transcript, and speech entries reference reusable voice identities when supported by the source. Speech without an available transcript is represented through its transcription status rather than an invented utterance. Non-linguistic vocalizations are distinguished through a subtype.

Each shot records available temporal boundaries, camera attributes, frame details, visual and audio states, and within-shot relations. The global event sequence connects these entries through temporal and relational descriptions, including simultaneity, persistence, cross-shot continuity, and source-supported triggering relations. Registries provide reusable definitions, whereas shot-level fields and the global event sequence describe their contextual states and interactions.

The schema specifies objects, arrays, strings, enumerated values, and nullable references. Visual entities, audio identities, audio entities, shots, and global events use separate identifier namespaces: `v1`, `p1`, `a1`, `s1`, and `e1`, respectively. Identifiers must be unique within their namespaces, and references must resolve to entries of the expected type. The supported audio entity types are `speech`, `sfx`, `ambience`, and `music`. Source evidence and quality-control metadata are retained for auditing but excluded from the rendered training caption.

Language-model conversion. We use `Qwen3.5-35B-A3B` to produce the structured JSON. The conversion protocol comprises system instructions, an output schema, an initial conversion prompt, and a feedback-based repair prompt. The source caption serves as the reference for conversion: the model is instructed to preserve its stated facts, dialogue, temporal information, negation, and uncertainty. Registries define reusable entities, while shot-level fields describe their local states and relationships.

Relations are included only when supported by the source. The co-occurrence of an entity and a sound does not by itself establish a sound-source binding, and temporal succession does not by itself establish a triggering relation. The conversion instructions require unspecified information to remain unspecified through optional or nullable fields, and require uncertainty expressed in the source to be retained.

Deterministic text rendering. A rule-based renderer traverses the JSON in a fixed order, maps internal identifiers to readable labels, and formats the global, registry, shot-level, and event-sequence fields into SSRS text. Field labels follow the selected language convention. Rendering does not invoke another language model. For a fixed renderer version and configuration, the same JSON therefore produces the same textual representation. This separates structured content generation from text formatting and makes the rendering stage reproducible.

Automated quality control and repair. Quality control combines structural validation, source-string consistency checks, and targeted rules for common conversion errors. Structural checks cover JSON validity, schema compliance, identifier uniqueness, reference integrity, and relation-endpoint validity. Source-consistency checks examine transcript correspondence and agreement with unambiguous shot-count and temporal anchors extracted from the source caption. Additional rules flag selected audio-type inconsistencies and violations of the separation between reusable registry definitions and shot-level relationships. Evidence strings are checked against the source text and associated with source locations when possible.

Validation findings provide feedback for the repair stage. The repair prompt directs the model to address the reported violations while preserving source-supported content. Repaired outputs undergo the same validation checks.

H.1 Representative Dense-to-SSRS Examples

The following examples are compact, human-verified pairs. They preserve the facts in the displayed dense captions while omitting implementation-only metadata and evidence spans for readability.

H.1.1 EXAMPLE 1

Dense caption. The video contains two shots. From 0.0–1.0 s, a static, eye-level medium shot shows a boy at center stage in a green grasshopper costume, with a girl in a leaf costume dimly visible to his right. Warm stage lighting illuminates a white fence, a prop tree, and a shimmering curtain. The boy looks forward and says, "I'm a grasshopper," with visible lip synchronization. The recording is dry and close, with little reverberation. From 1.0–6.0 s, a static, eye-level medium close-up shows a couple in the audience against softly blurred spectators. A man in a dark blue shirt looks toward the stage and says, "I told you." He turns to the smiling woman in a floral dress and continues, "He is the cutest one." The woman gives a soft verbal response whose words are not provided. The man then looks back toward the stage while the woman continues looking forward. The lighting and overall atmosphere are warm, intimate, and affectionate, and the close, dry recording continues.

SSRS caption. **Global Visual.** Scene and space: a stage area and an audience area. Lighting and atmosphere: warm lighting is present across the sequence; the audience shot has a warm, intimate, and affectionate atmosphere.

Global Audio. Overall soundscape: close, dry dialogue recording with little reverberation. Global auditory atmosphere: not specified. Non-diegetic music: not specified.

Visual Entity Registry. [Subject A] A boy wearing a green grasshopper costume. [Subject B] A girl wearing a leaf costume. [Subject C] A man wearing a dark blue shirt. [Subject D] A woman wearing a floral dress. [Object A] A white fence. [Object B] A prop tree. [Object C] A shimmering curtain.

Audio Identity Registry. [Voiceprint A] Voice identity used by [Voice 1]. [Voiceprint B] Shared voice identity used by [Voice 2] and [Voice 3]. [Voiceprint C] Voice identity used by [Voice 4].

Audio Entity Registry. [Voice 1] uses [Voiceprint A]; transcript: "I'm a grasshopper." [Voice 2] uses [Voiceprint B]; transcript: "I told you." [Voice 3] uses [Voiceprint B]; transcript: "He is the cutest one." [Voice 4] uses [Voiceprint C]; transcript: unavailable; a soft verbal response.

Shot 1 (0.0–1.0 s). Presentation: static, eye-level medium shot with warm stage lighting. Frame details: [Subject A] is at center stage; [Subject B] is dimly visible to his right. [Object A], [Object B], and [Object C] are present on the stage. Visual state and change: [Subject A] looks forward and speaks. Audio state: [Voice 1] occurs. Relations: [Subject A] is the visible source of [Voice 1], and the visible lip motion of [Subject A] is synchronized with [Voice 1].

Shot 2 (1.0–6.0 s). Presentation: static, eye-level medium close-up with warm lighting; spectators in the background are softly blurred. Frame details: [Subject C] and [Subject D] are shown in the audience. Visual state and change: [Subject C] looks toward the stage and says [Voice 2], turns toward [Subject D] and says [Voice 3], and then looks back toward the stage. [Subject D] smiles, gives [Voice 4], and continues looking forward. Audio state: [Voice 2] precedes [Voice 3], followed by [Voice 4]. Relations: [Subject C] is the visible source of [Voice 2] and [Voice 3]. [Subject D] is the visible source of [Voice 4].

Global Event Sequence. E1: [Shot 1] [Subject A] says [Voice 1] on stage. → E2: [Shot 2] [Subject C] says [Voice 2] and [Voice 3], after which [Subject D] gives [Voice 4]. The visual setting changes from the stage to the audience, while the close, dry dialogue recording continues.

H.1.2 EXAMPLE 2

Dense caption. The video contains four shots in a bright, rustic outdoor cooking setting. From 0.0–1.0 s, a static, slightly low-angle close-up shows a left hand holding a green bell pepper on a wooden cutting board while a right hand slices it with a metal chef's knife; the rhythmic blade impacts are clearly audible. From 1.0–4.8 s, a static, eye-level medium shot reveals a man in a dark camouflage-patterned jacket behind a table of vegetables, bowls, and a terracotta jug. He continues slicing the green pepper as the chopping sound continues, accompanied by a low wind rumble. From 4.8–6.0 s, a static close-up shows a basket containing red pepper rings and halves while the wind continues. From 6.0–8.0 s, pale yellow pepper slices are added to the same basket, producing a crisp rustling sound as they settle; the wind remains audible. Bright natural sunlight and an earthy, tranquil atmosphere persist throughout.

SSRS caption. **Global Visual.** Scene and space: a rustic outdoor cooking setting containing a cutting board, a table of vegetables and bowls, a terracotta jug, and a basket. Lighting and atmosphere: bright natural sunlight and an earthy, tranquil atmosphere persist throughout.

Global Audio. Overall soundscape: rhythmic blade impacts, a low wind rumble, and the crisp rustling of pepper slices settling. Global auditory atmosphere: not specified. Non-diegetic music: not specified.

Visual Entity Registry. [Subject A] A man wearing a dark camouflage-patterned jacket. [Object A] A green bell pepper. [Object B] A metal chef's knife. [Object C] A wooden cutting board. [Object D] A table containing vegetables, bowls, and a terracotta jug. [Object E] A basket containing red pepper rings and halves. [Object F] Pale yellow pepper slices.

Audio Identity Registry. None.

Audio Entity Registry. [SFX 1] Rhythmic blade impacts associated with slicing the green pepper on the cutting board. [Ambience 1] A low wind rumble. [SFX 2] Crisp rustling produced as the pale yellow pepper slices settle in the basket.

Shot 1 (0.0–1.0 s). Presentation: static, slightly low-angle close-up in bright natural sunlight. Visual state and change: a left hand holds [Object A] on [Object C] while a right hand repeatedly slices it with [Object B]. Audio state: [SFX 1] occurs repeatedly. Relations: [SFX 1] is associated with the repeated slicing action involving [Object B], [Object A], and [Object C].

Shot 2 (1.0–4.8 s). Presentation: static, eye-level medium shot. Frame details: [Subject A] is behind [Object D]. Visual state and change: [Subject A] continues slicing [Object A] with [Object B] on [Object C]. Audio state: [SFX 1] continues; [Ambience 1] is present. Relations: [SFX 1] remains associated with the ongoing slicing action; [Ambience 1] is present in the outdoor scene.

Shot 3 (4.8–6.0 s). Presentation: static close-up. Visual state: [Object E] contains red pepper rings and halves. Audio state: [Ambience 1] continues. Relations: [Ambience 1] remains present in the outdoor scene.

Shot 4 (6.0–8.0 s). Presentation: not further specified in the source caption. Visual state and change: [Object F] is added to [Object E] and settles. Audio state: [SFX 2] occurs while [Ambience 1] continues. Relations: the addition and settling of [Object F] produces [SFX 2]; [Ambience 1] remains audible.

Global Event Sequence. E1: [Shot 1–Shot 2] [Object A] is sliced, with [SFX 1] continuing across the cut. → E2: [Shot 3] [Object E] is shown while [Ambience 1] continues. → E3: [Shot 4] [Object F] is added to [Object E], producing [SFX 2]. [Ambience 1] persists from [Shot 2] through [Shot 4].

I Evaluation Metrics and Implementation Details

We evaluate multi-shot audio-visual generation using 14 metrics adapted from MSAVBench. Table 11 summarizes their measurement procedures and normalization. We describe the implementation-specific choices below to make the resulting scores interpretable and reproducible.

Evaluation dimensions. The metrics cover three complementary levels. Global metrics assess narrative coherence (Narr), lip–speech synchronization (Lip), speaker–visual attribution (Attr), general audiovisual synchronization (Sync), and realization of requested visual attributes (VQ). Cross-shot metrics assess spatial consistency (C-Layout), visual consistency (VC), music consistency (Mus), and speaker identity consistency (Spk). Local metrics assess within-shot spatial relationships (I-Layout), camera instruction adherence (Cam), perceptual audio quality (PQ), on-screen text fidelity (OCR), and spoken-content fidelity (WER). VQ measures visual attribute realization rather than aesthetic quality. VC averages five components: subject appearance, background, style, illumination, and color.

Visual judging and protocol adaptations. We use `gemini-3.5-flash` for visual judgments. Direct visual evaluation uses an 8-fps silent video proxy with a maximum width of 640 pixels, while expert audiovisual metrics retain the original media. VQ questions are constructed from the canonical evaluation prompt before inspecting generated videos and are shared across models evaluated

Table 11: Definitions and computation of the 14 metrics in our MSAVBench-derived evaluation. All normalized scores lie in $[0, 1]$; higher is better.

Metric	Dimension	Measurement	Normalization
Global metrics			
Narr	Narrative coherence	Binary visual judgments of requested events and their narrative order.	$B(\mathcal{Q}_{\mathrm{narr}})$.
Lip	Lip–speech synchronization	SyncNet synchronization confidence c and signed frame offset o for matched speaking tracks.	$0.7C(c/6) + 0.3C(1 - \lvert o\rvert/8)$; aggregated across tracks using visual-duration weights.
Attr	Speaker attribution	Match diarized audio speakers to active visual speaker tracks.	$\sum_k d_k \mathbb{I}[m_k] / \sum_k d_k$, where d_k is speech duration and m_k indicates a successful assignment.
Sync	Audiovisual synchronization	Synchformer estimates signed delays δ_{first} and δ_{last} for the first and last feature blocks, in seconds.	$C(1 - d/2)$, where $d = (\lvert\delta_{\mathrm{first}}\rvert + \lvert\delta_{\mathrm{last}}\rvert)/2$.
VQ	Visual realization	Answer a shared set of visual multiple-choice questions constructed from the evaluation prompt.	Number of correct answers divided by the total number of questions.
Cross-shot metrics			
C-Layout	Cross-shot layout	Tool-assisted binary judgments of spatial relationships across aligned shots.	$B(\mathcal{Q}_{\mathrm{cross}})$.
VC	Visual consistency	Compare subject appearance using localized DINOv2-L features, background using subject-removed CLIP ViT-L/14 features, and style using CSD-ViT-L features. Assess illumination and color with Gemini binary judgments.	Mean of five components: $C((1 + \bar{c})/2)$ for subject and background, $C(\bar{c})$ for style, and $B(\mathcal{Q})$ for illumination and color.
Mus	Music consistency	Extract accompaniment with Demucs and compare MuQ embeddings, tempo, and beat intervals across audio windows. Tempo and beats are estimated with All-In-One.	$C(0.55S_{\mathrm{emb}} + 0.25S_{\mathrm{bpm}} + 0.20S_{\mathrm{beat}})$; components are defined below.
Spk	Speaker consistency	Extract W2V-BERT speaker embeddings and compare all unordered window pairs belonging to the same diarized speaker.	Map pairwise cosine similarity using $C((1 + \cos)/2)$, average within each speaker, and aggregate speaker scores by speech duration.
Local metrics			
I-Layout	Intra-shot layout	Tool-assisted binary judgments of object positions and spatial relationships within each shot.	Average $B(\mathcal{Q}_{s,\mathrm{layout}})$ across evaluated shots.
Cam	Camera adherence	Binary judgments of adherence to requested framing, viewpoint, and camera motion.	Average $B(\mathcal{Q}_{s,\mathrm{cam}})$ across evaluated shots.
PQ	Audio production quality	Use the production-quality output q of `facebook/audiobox-aesthetics`, with a nominal rating scale of 1–10.	$C((q - 1)/9)$.
OCR	Visual text fidelity	Match OCR detections to requested text strings using normalized-string similarity r_t and the selected detection's confidence c_t.	$\frac{1}{T}\sum_{t=1}^{T} r_t(0.5 + 0.5c_t)$, where T is the number of requested text targets.
WER-S	Spoken-content fidelity	Compare ASR output with requested transcripts using word error rate, or character error rate for Chinese.	$1 - \min(E, 1)$, where $E = (N_{\mathrm{sub}} + N_{\mathrm{del}} + N_{\mathrm{ins}})/N_{\mathrm{ref}}$.

Notation. $C(x) = \min(1, \max(0, x))$; $B(\mathcal{Q}) = N_{\mathrm{yes}}/(N_{\mathrm{yes}} + N_{\mathrm{no}})$. $\bar{c}$ denotes mean cosine similarity. For WER, N_{sub}, N_{del}, N_{ins}, and N_{ref} denote the numbers of substitutions, deletions, insertions, and reference units, respectively. The reported WER-column score is normalized speech fidelity, rather than raw error rate.

Music components. For retained music-window embeddings $\mathbf{z}_i$,

$$S_{\mathrm{emb}} = \frac{2}{n(n-1)} \sum_{i<j} C\left(\frac{1 + \cos(\mathbf{z}_i, \mathbf{z}_j)}{2}\right),$$

$$S_{\mathrm{bpm}} = C\left(1 - \frac{\sigma_b}{\max(1, \mu_b)}\right), \qquad S_{\mathrm{beat}} = C\left(1 - \frac{\sigma_\tau}{\mu_\tau + 10^{-6}}\right).$$

Here, μ_b and σ_b are the population mean and standard deviation of positive BPM estimates. Each τ_i is the mean inter-beat interval of a window containing at least two detected beats; μ_τ and σ_τ are the population mean and standard deviation of these window-level means. S_{beat} measures interval consistency, rather than beat-phase alignment.

Audio windows. Spk uses 4-s windows with a default stride of 4 s and a maximum of eight windows, retaining speaker–window chunks containing at least 1 s of speech. For clips longer than 32 s, the eight window starts are evenly distributed from 0 to $D - 4$, where D is clip duration. Mus uses 4-s windows with a maximum of four windows and discards decoded trailing segments of at most 0.5 s.

OCR matching. Text preprocessing applies NFKC normalization, lowercasing, and punctuation/whitespace filtering. String matching uses a substring-match rule with score 1 and otherwise computes `difflib.SequenceMatcher.ratio()`. The best string match determines r_t and its associated detection confidence c_t.

on that prompt. Shot boundaries are obtained from visual observations by the judge, rather than the original benchmark's full shot-detection and iterative-correction pipeline. Music and speaker consistency use fixed audio windows, which need not coincide with visual shot boundaries.

Score normalization. All normalized scores in Table 11 lie in $[0, 1]$, with higher values indicating better performance. Binary rubric scores average applicable yes/no judgments. Expert outputs use metric-specific mappings rather than a shared rescaling rule. In particular, Lip combines synchronization confidence and frame offset, whereas Sync penalizes delay measured in seconds. OCR uses confidence-weighted string similarity. For spoken-content fidelity, the normalized score is $1 - \min(E, 1)$, where E is WER, or CER for Chinese. These mappings differ from some descriptions in the original protocol; raw and normalized quantities should therefore not be interchanged.