CINELONG: CONSISTENT LONG-HORIZON MULTI-SHOT AUDIO-VISUAL GENERATION

Anonymous authors
Paper under double-blind review

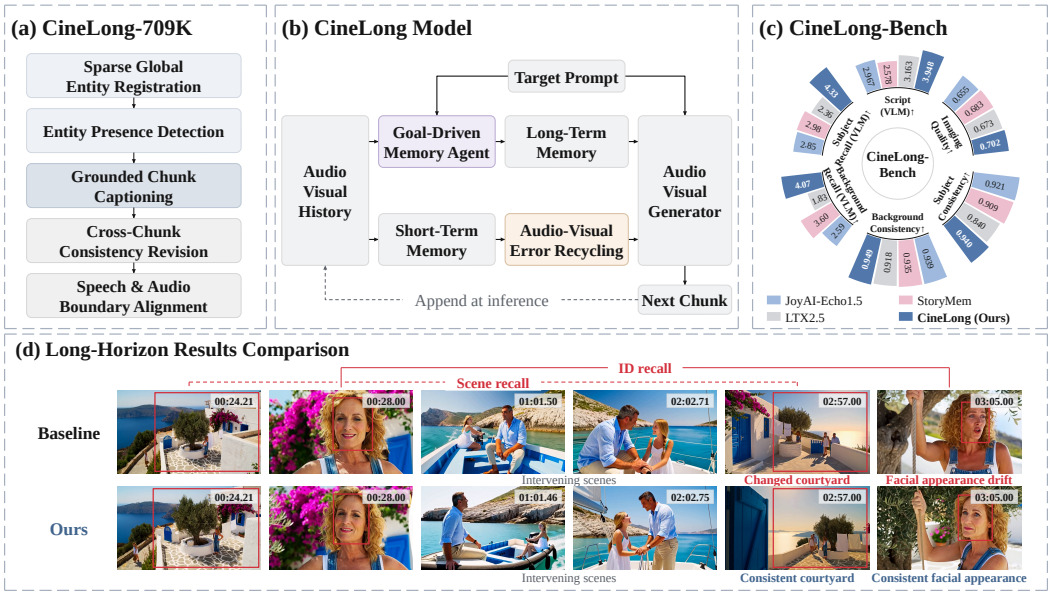


Figure 1: **We propose CINELONG for coherent long-horizon audio-visual generation.** (a) Global-to-local annotation of long videos; (b) long- and short-term memory with goal-driven selection and error recycling; (c) benchmark comparison; (d) long-horizon subject and background recall. More results are available on our project page (`https://cinelong.github.io/project`).

ABSTRACT

Generating coherent long-form audio-visual narratives requires preserving character and scene identities while maintaining the continuity of motion and sound across shots. This calls for training data annotations that connect local events across a narrative and a memory mechanism that retrieves useful context from an expanding, imperfect generation history. To this end, we construct **CINELONG-709K**, a long-video dataset comprising 708,794 videos totaling approximately 20k hours, with individual sequences lasting up to 148 seconds. Its global-to-local annotations link persistent identities to local events, dialogue, and cross-chunk dependencies. Building on this dataset, we propose **CINELONG**, a framework for long-horizon audio-visual generation that combines long-term memory for recalling characters, voices, and scenes with short-term memory for continuing ongoing actions and speech. At inference, a Goal-Driven Memory Agent (GDMA) uses the target prompt to select relevant historical chunks and the most useful keyframe within each of their shots, retaining the associated audio as context. To improve robustness, Audio-Visual Error Recycling (AVER) introduces model-derived visual and acoustic errors into memory during training. On CINELONG-BENCH, CINELONG outperforms existing baselines in subject and background recall, cross-chunk audio continuity, and script adherence, while maintaining competitive visual quality.

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.

1 INTRODUCTION

Recent video diffusion transformers and joint audio-visual generators synthesize detailed scenes with synchronized sound (Peebles & Xie, 2023; Blattmann et al., 2023; HaCohen et al., 2026; MiniMax AI, 2026). The next frontier is long-form audio-visual storytelling, where characters leave and return, locations are revisited, and conversations unfold across many shots. Such narratives demand *long-range consistency*, so that a returning character keeps the same face and voice (Figure 1 (d)), and *short-range continuity*, so that ongoing motion and speech flow across shot boundaries. Since the cost of single-pass generation grows rapidly with duration, a practical route is to generate such narratives as a *long-horizon* process, with each multi-shot chunk conditioned on what came before. Coherence in this regime requires two ingredients that existing work largely lacks: training annotations that connect local events across an entire narrative, and a memory mechanism that retrieves useful context from an ever-growing, error-prone generation history.

Such annotations are hard to obtain because long videos are *locally dense* yet *globally entangled*: each chunk packs several shots, actions, lines, and sounds, while characters return after long absences and utterances run past chunk boundaries. Per-shot annotation (Wang et al., 2023) is accurate but isolated, so a returning character may be renamed and a continuing utterance fragmented. Recent cinematic datasets add sequence-level character and scene lists (Chen et al., 2026), but annotate in a single pass that leaves omissions and contradictions unchecked and cross-chunk dependencies unrecorded. Long-horizon generation calls for annotation that is locally precise yet globally aware.

On the modeling side, the key question is what to remember, since conditioning on the entire growing history is costly and buries the few observations that matter. Prior methods keep recent frames, compressed histories, or keyframe banks (Henschel et al., 2024; Zhang et al., 2025b;a), select history by the current generation state (Li et al., 2026; Ding et al., 2026), or carry speaker cues across shots (Huang et al., 2026; Duan et al., 2026). Yet useful history is target-dependent: the target prompt already states what is needed, such as a character who returns after several scenes and is thus absent from the recent window. History is also imperfect: the model conditions on its own outputs, so visual and acoustic errors accumulate across chunks. Error recycling mitigates this for video (Li et al., 2025b) but leaves the equally error-prone generated audio unaddressed.

To address both gaps, we present **CINELONG**, which couples a long-video annotation paradigm with a goal-driven audio-visual memory. For data, our global-to-local annotation paradigm (Figure 1 (a)) registers persistent characters, objects, and scenes in a global pass, annotates each chunk against this registry with shot-level visual, dialogue, and sound descriptions, and finally revises all chunks at the sequence level to resolve contradictions and mark cross-chunk dependencies such as character re-entries and continuing utterances, yielding **CINELONG-709K**, with 708,794 videos totaling approximately 20k hours and individual sequences lasting up to 148 seconds. For modeling, CINELONG pairs long-term memory (LTM), which recalls characters, voices, and scenes from keyframes and audio of selected past chunks, with short-term memory (STM), which continues ongoing actions and speech from the audio-visual tail of the preceding chunk (Figure 1 (b)). At inference, a Goal-Driven Memory Agent (GDMA) fills LTM by using the target prompt to select relevant chunks and the most informative keyframe in each of their shots, along with the associated audio. During training, Audio-Visual Error Recycling (AVER) injects model-derived visual and acoustic residuals into STM so that the generator learns to tolerate its own imperfect history.

Our contributions are summarized as follows:

- We propose CINELONG, a long-horizon audio-visual generation framework that couples long-term memory for recalling characters, voices, and scenes with short-term memory for continuing ongoing actions and speech.
- We design two key components for CINELONG: the Goal-Driven Memory Agent (GDMA), which selects target-relevant chunks and per-shot keyframes together with their audio at inference, and Audio-Visual Error Recycling (AVER), which trains the generator to tolerate its own imperfect history in both video and audio modalities.
- We construct CINELONG-709K, comprising 708,794 videos (about 20K hours), through a global-to-local annotation paradigm with cross-chunk verification, yielding audio-visual annotations that are precise within each chunk, consistent across chunks, and explicit about

cross-chunk dependencies. We further introduce CINELONG-BENCH, 30 cinematic scripts for evaluating long-horizon coherence.

- Extensive experiments on CINELONG-BENCH show that CINELONG outperforms seven video-only and audio-visual baselines, improving subject recall from 3.55 to 4.33, background recall from 3.60 to 4.07, and audio continuity from 0.393 to 0.486 over the strongest baseline on each metric, while achieving the best VLM-judged script adherence and competitive visual quality (Figure 1 (c)).

2 RELATED WORK

Long video generation. StreamingT2V introduces short-term conditioning and a long-term appearance anchor for autoregressive video extension (Henschel et al., 2024). CausVid distills a bidirectional diffusion model into a causal generator for low-latency streaming (Yin et al., 2025), while FramePack compresses historical frames according to their importance within a fixed context budget (Zhang et al., 2025b). DiTCtrl enables coherent multi-prompt generation through attention sharing (Cai et al., 2025). For narrative consistency, StoryMem conditions successive shots on a dynamically updated keyframe bank (Zhang et al., 2025a), while TetherMem separates subject and scene queries to balance identity preservation with scene progression (Li et al., 2026). Stable Video Infinity (SVI) further introduces error-recycling fine-tuning to improve robustness to imperfect generated context (Li et al., 2025b). However, these approaches do not jointly retrieve visual and acoustic evidence for long-horizon audio-visual generation. Our work uses a Goal-Driven Memory Agent to retain the associated audio in long-term memory alongside short-term audio-visual context to support both distant recall and immediate continuation.

Joint audio-video generation. JavisDiT introduces hierarchical spatiotemporal priors to synchronize jointly generated video and audio (Liu et al., 2025). LTX-2 couples asymmetric audio and video streams through bidirectional cross-attention (HaCohen et al., 2026), while MiniMax-H3 supports native multi-shot generation with jointly generated audio and multimodal conditioning (MiniMax AI, 2026). MAVIN introduces boundary-aware attention and identity-aware propagation for controllable multi-shot audio-visual narratives (Liu et al., 2026). UnityShots combines gated long- and short-term visual memory with a reference speaker token to maintain cross-shot consistency (Huang et al., 2026), while JoyAI-Echo1.5 aggregates historical visual evidence and speaker cues from full-shot audio for persistent appearance and voice identity (Duan et al., 2026).While these models advance cross-shot consistency, selecting useful context in history are still central challenges for long-horizon generation. Our work combines goal-driven historical selection with long-term memory containing keyframes and associated chunk audio, while short-term memory preserves recent motion and acoustic state.

3 DATA CONSTRUCTION

3.1 CINELONG-709K DATASET

Long-video generation requires annotations that both ground local events and keep identities consistent across distant shots. We construct CINELONG-709K through a *global-to-local annotation pipeline with cross-chunk revision*. As shown in Figure 2, a shared entity registry guides per-chunk descriptions, which are then revised against the complete sequence and aligned with the audio track.

Data sources and chunking. We collect source videos from public datasets, including Condensed Movies (Bain et al., 2020), Short-Films20K (Ghermi et al., 2026), and OpenHumanVid (Li et al., 2025a), as well as publicly available videos from online platforms such as YouTube. After filtering, CINELONG-709K contains 708,794 videos ranging from 17 to 148 seconds (about 102 seconds on average), totaling roughly 20K hours. We first detect shot boundaries with TransNetV2 (Soucek & Lokoc, 2020), then randomly partition each video into consecutive, non-overlapping temporal chunks. For each chunk, we record the shots it overlaps and their temporal extents within the chunk. Each chunk corresponds to one generation step and can span multiple shots; chunk boundaries need not coincide with shot boundaries. The chunks average 16.8 seconds and 4.6 shots, while their temporal ordering preserves character re-entry and actions or utterances that continue across chunks.

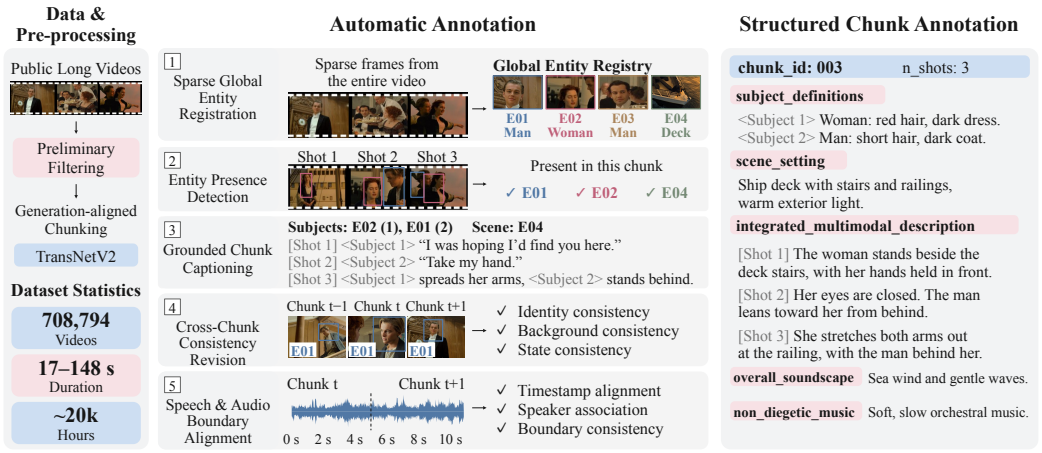


Figure 2: **Data construction pipeline of CINELONG-709K.** Global registration builds a shared entity registry that anchors per-chunk presence grounding and captioning; sequence-level revision then reconciles captions across chunks, and a final stage aligns audio to chunk boundaries.

Benchmark source sequences are excluded before training-data construction. Annotation fields and filtering details are provided in Appendix A.

Audio-visual quality filtering. We estimate optical flow with RAFT (Teed & Deng, 2020) to identify and discard near-static candidate segments. We further filter candidates using perceptual quality thresholds for both modalities, assessing visual aesthetics with the improved aesthetic predictor (Schuhmann, 2022) and audio quality with Audiobox-Aesthetics (Tjandra et al., 2025).

Stage 1: Global registration. Qwen3.6-27B (Qwen Team, 2026) processes sparsely sampled frames from the full video and registers people, objects, and scenes, each with a canonical ID and concise attributes. Recurring entities are thus referenced through the same registry entry in every chunk rather than described independently.

Stages 2–3: Local grounding and captioning. For each chunk, Qwen3.6-27B matches visible entities to the registry and records their states and relations in a presence table. Qwen3-Omni-30B-A3B (Xu et al., 2025) then captions the chunk from its audio and video, conditioned on the presence table and the registry, describing shots, actions, dialogue, and sound. Decoupling grounding from captioning separates entities registered across the full video from those visible in the current chunk, discouraging captions that mention registered but absent entities.

Stage 4: Cross-chunk revision. A sequence-level pass revisits the ordered captions together with the registry and presence tables. It resolves identity and scene inconsistencies, separates genuine state changes (e.g., a change of clothing) from contradictory descriptions, and labels continuity links between adjacent chunks (e.g., an action spanning a boundary).

Stage 5: Boundary-aligned audio. We obtain word-level timestamps with WhisperX (Bain et al., 2023) and use Qwen3-Omni-30B-A3B to attribute each utterance to a registered speaker and to the chunk interval(s) it occupies. Utterances crossing a chunk boundary are tagged with continuation markers, and non-speech sounds are assigned to the chunks in which they occur. This stage extends continuity from visual identity to the audio track.

3.2 CINELONG-BENCH

CINELONG-BENCH evaluates long-horizon audio-visual generation with character re-entry, voice consistency, and cross-chunk continuation. It contains 30 cinematic scripts, each organized into 10 consecutive generation chunks, yielding 300 chunks and approximately 144 seconds per generated chain. The scripts include scene transitions, returning characters, and cross-boundary dialogue or ambience. All benchmark source sequences are excluded before training-data construction.

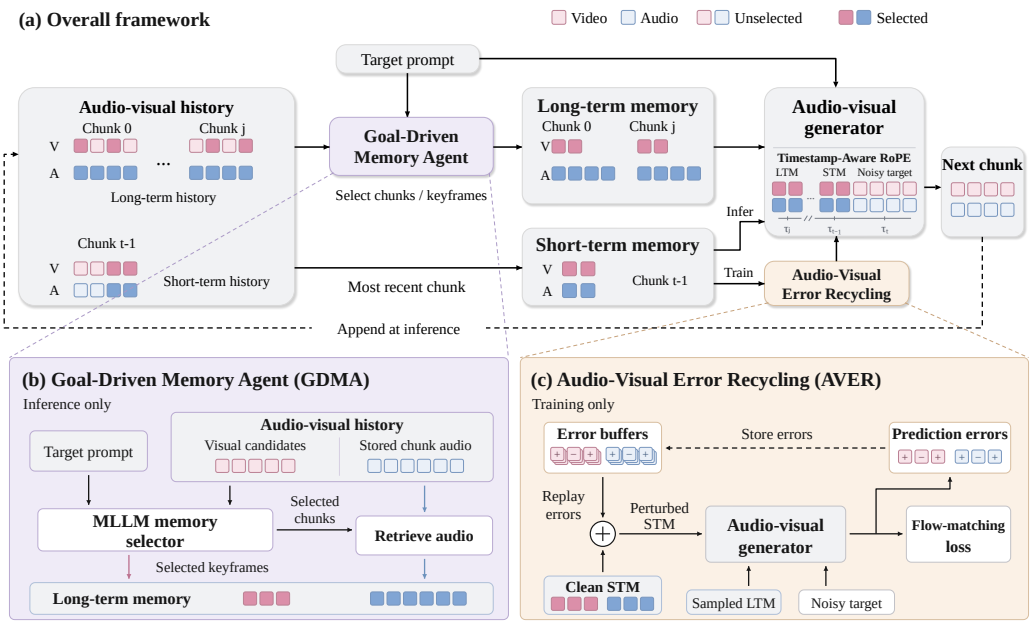


Figure 3: **Overview of CINELONG.** (a) The generator conditions on the target prompt, long-term memory (LTM) of selected historical chunks, and short-term memory (STM) from the preceding chunk; LTM is randomly sampled in training and selected by GDMA at inference. (b) GDMA selects chunks and per-shot keyframes according to the target prompt, together with their audio. (c) During training, AVER replays audio-visual prediction errors into STM.

4 CINELONG

Building on CINELONG-709K (Section 3), we propose CINELONG, a framework for coherent long-form audio-visual generation that recalls characters, voices, and scenes from distant history while continuing ongoing motion and speech. As shown in Figure 3, Section 4.1 first formulates the problem. Section 4.2 introduces the long- and short-term audio-visual memory that conditions the generator. Section 4.3 presents the Goal-Driven Memory Agent (GDMA), which selects long-term memory at inference, and Section 4.4 describes Audio-Visual Error Recycling (AVER), which makes the generator robust to imperfect history during training.

4.1 PROBLEM FORMULATION

Let $Y_t = (V_t, A_t)$ denote the t-th multi-shot chunk with aligned video and audio, p_t its target prompt, and $\mathcal{H}_t = \{Y_0, \ldots, Y_{t-1}\}$ the history. Conditioning on the full history $\mathcal{H}_t$ is computationally infeasible due to the quadratic growth of attention cost with t, yet truncating it undermines cross-shot consistency. We therefore approximate it with two memories:

$$p_\theta(Y_t \mid p_t, \mathcal{H}_t) \approx p_\theta(Y_t \mid p_t, \mathcal{M}_t^{\mathrm{L}}, \mathcal{M}_t^{\mathrm{S}}), \tag{1}$$

where the long-term memory (LTM) $\mathcal{M}_t^{\mathrm{L}}$ holds evidence from a set $\mathcal{S}_t$ of at most K chunks selected from the history, and the short-term memory (STM) $\mathcal{M}_t^{\mathrm{S}}$ holds the audio-visual information of the most recent chunk Y_{t-1}. At inference, each generated chunk $\hat{Y}_t$ is appended to the history, so subsequent chunks are conditioned on the model's own, possibly imperfect, outputs.

We build on a audio-visual Diffusion Transformer (Peebles & Xie, 2023) operating on the video and audio latents $x_t = (x_t^v, x_t^a)$ of Y_t with the flow-matching objective (Lipman et al., 2022).

4.2 LONG- AND SHORT-TERM AUDIO-VISUAL MEMORY

As shown in Figure 3 (a), we split the history into two complementary memories: LTM draws sparse keyframes and full audio from selected distant chunks to recall characters, voices, and scenes, while

STM keeps a dense tail of the most recent chunk to continue the motion and speech in progress. Both are injected via in-context conditioning (Tan et al., 2024; Cai et al., 2026): memory latents are concatenated with the noisy target tokens into a single sequence, so the generator attends to them jointly through its existing self-attention without extra modules.

Long-Term Memory. LTM is $\mathcal{M}_t^{\mathrm{L}} = \{m_j\}_{j\in\mathcal{S}_t}$, where each selected chunk j contributes $m_j = (\{k_{j,s}\}_{s=1}^{S_j}, a_j, \tau_j)$, consisting of one keyframe $k_{j,s}$ for each of its S_j shots, the full chunk audio a_j, and its timestamps τ_j. Per-shot keyframes preserve every camera view of a character or location at low token cost, while the full audio retains voices, ambience, and music. During training, we sample $\mathcal{S}_t$ uniformly from the history, with random frames per shot, so that the generator learns to use arbitrary, noncontiguous evidence. At inference, the historical chunks and their per-shot keyframes are selected by GDMA based on the target prompt p_t (Section 4.3).

Short-Term Memory. LTM cannot convey the motion or speech in progress at a chunk boundary. STM therefore keeps a dense, temporally aligned audio-visual tail of the preceding chunk:

$$\mathcal{M}_t^{\mathrm{S}} = \left(x_{t-1}^{v}[-L_{\mathrm{STM}}^{v}:],\, x_{t-1}^{a}[-L_{\mathrm{STM}}^{a}:]\right), \tag{2}$$

where L_{STM}^{v} and L_{STM}^{a} denote the numbers of retained video and audio latent frames. STM serves only as a conditioning input and does not enforce visual continuity, so the generator can still start a new shot when the target prompt calls for a cut.

Timestamp-Aware RoPE. To distinguish memory from target tokens within one sequence, we propose Timestamp-Aware RoPE (TA-RoPE), which sets the temporal RoPE indices (Su et al., 2024) of video and audio tokens by their timestamps. Target tokens take the timestamps of chunk t, STM tokens those immediately preceding it, and LTM tokens their own timestamps in τ_j. The resulting offsets reflect temporal distance, so STM is treated as the adjacent past to continue and LTM as distant evidence to recall.

4.3 Goal-Driven Memory Agent

Deciding what enters LTM is difficult, as the history grows without bound and the observations a target needs may lie arbitrarily far in the past. Existing memory banks maintain a fixed-size buffer updated during generation (Zhang et al., 2025a), permanently discarding evicted observations, so a character returning after several scenes cannot be recalled. We instead retain the full history and let GDMA retrieve from it on demand, using the target prompt as the goal to select both *which chunks* enter LTM and *which frame* represents each of their shots (Figure 3 (b)).

Candidate construction. When a chunk enters the history, we build a compact candidate set $\mathcal{C}_{j,s}$ for each of its shots by removing near-duplicate frames (CLIP-L (Radford et al., 2021)) and low-quality ones (HPSv3 (Ma et al., 2025)). Keeping multiple views per shot lets the same shot supply different evidence for different targets.

Goal-driven selection. Given the target prompt p_t and all shot-indexed candidates, an MLLM memory selector reasons about character appearance and visibility, scene layout and viewpoint, and target-relevant objects. It scores the relevance of each historical chunk and, for every shot of the selected chunks, picks the frame that best matches the prompt, since the highest-quality frame may not show the face, prop, or viewpoint the prompt requests:

$$\left(\mathcal{S}_t, \{k_{j,s}\}\right) = \mathrm{MLLM}\big(p_t, \{\mathcal{C}_{j,s}\}_{j<t}\big), \tag{3}$$

where $\mathcal{S}_t$ holds the top-K scoring chunks and $k_{j,s} \in \mathcal{C}_{j,s}$ is the keyframe for shot s of chunk $j \in \mathcal{S}_t$. GDMA fills LTM with these keyframes and the full audio and timestamps of their chunks. As a training-free module, it plugs directly into the generator and keeps characters, voices, and scenes consistent throughout long-horizon generation, even after long absences.

4.4 Audio-Visual Error Recycling

Training conditions on ground-truth history, whereas inference conditions on generated history whose visual and acoustic errors accumulate across chunks. Inspired by SVI (Li et al., 2025b), which recycles prediction errors for video, we further introduce Audio-Visual Error Recycling (AVER), which injects the model's own video and audio errors into STM during training (Figure 3 (c)):

$$\widetilde{x}_{\mathrm{STM}}^{q} = x_{\mathrm{STM}}^{q} + b_q\, e_q, \quad b_q \sim \mathrm{Bernoulli}(p_q),\ q \in \{v, a\}, \tag{4}$$

Table 1: **Main comparison on CINELONG-BENCH. Bold**: best. Underline: second best. †: video-only. –: not applicable. Full metrics are reported in Table. 4.

Method	Cross-shot Consistency				Script Adherence				LSE-C↑
	BGR↑	SBR↑	VC↑	AC↑	S-VLM↑	Cut↑	SR↑	CLIP↑	
ShotStream† (Luo et al., 2026)	2.83	2.86	–	–	1.737	–	–	0.247	–
StoryMem† (Zhang et al., 2025a)	3.60	2.98	–	–	2.578	–	–	0.268	–
JavisDiT++ (Liu et al., 2025)	1.89	2.56	0.344	0.282	1.763	–	0.034	0.265	0.74
LTX2.3 (HaCohen et al., 2026)	2.01	3.12	0.328	0.393	2.326	–	0.906	0.259	2.22
LTX2.5 (HaCohen et al., 2026)	1.83	2.36	0.409	0.393	3.163	0.108	0.974	0.260	2.57
JoyAI-Echo1.5 (Duan et al., 2026)	2.59	2.85	**0.473**	0.272	2.967	0.126	0.971	0.266	2.44
MiniMax-H3 (MiniMax AI, 2026)	2.35	3.55	0.383	0.328	3.770	**0.915**	0.985	0.271	2.59
CINELONG (Ours)	**4.07**	**4.33**	0.448	**0.486**	**3.948**	0.893	**0.993**	**0.277**	**2.66**

where e_q is a recycled prediction residual of modality q. The generator thus learns to continue from imperfect audio-visual history, making it more robust in long-horizon generation.

5 EXPERIMENTS

We evaluate CINELONG on CINELONG-BENCH in terms of cross-chunk consistency, script adherence, and audio-visual quality. We first compare against existing video and audio-visual generation methods (Section 5.2), then ablate long- and short-term memory, goal-driven memory agent, and audio-visual error recycling (Section 5.3).

5.1 EXPERIMENTAL SETUP

Baselines. We compare against two groups of baselines on our CINELONG-BENCH: video-only multi-shot generators, ShotStream (Luo et al., 2026) and StoryMem (Zhang et al., 2025a); and joint audio-visual generators, JavisDiT++ (Liu et al., 2025), LTX2.3 and LTX2.5 (HaCohen et al., 2026), JoyAI-Echo1.5 (Duan et al., 2026), and MiniMax-H3 (MiniMax AI, 2026). Each baseline is extended to the full horizon with its own continuation mechanism and otherwise uses its native inference settings; Appendix D gives the details. CINELONG generates at 480×832 and 24 fps with 50 denoising steps, using the LTM budget and STM length stated in Appendix B.

Metrics. We group metrics by the claim they test. *Cross-shot consistency*, the target of our memory design, checks whether characters, locations, voices, and audio persist across shots. Subject Recall (SBR) and Background Recall (BGR) are 1–5 scores from a VLM judge (Gemini 3.1 Pro) on characters and locations that return after an absence; Voice Consistency (VC) is the speaker-embedding similarity between shots of the same speaker; and Audio Continuity (AC) is the LAION-CLAP similarity across chunk seams (Wu et al., 2023). *Script adherence* measures whether the output follows the script: S-VLM scores each chunk against its script with VLM, Cut checks whether cuts occur at the scripted times by computing the matching F1 score, Speech Recall (SR) checks spoken dialogue, and CLIP measures frame-text match. *Intra-shot quality* is measured by VBench Imaging Quality (IQ), Subject Consistency (SBC), and Background Consistency (BGC) (Huang et al., 2024), and by SyncNet LSE-C for lip synchronization (Chung & Zisserman, 2016). Please refer to Appendix D for more details.

5.2 MAIN RESULTS

CINELONG leads in long-range recall and audio continuity. CINELONG surpasses all other methods on Background Recall, Subject Recall, and Audio Continuity, and ranks second on Voice Consistency after JoyAI-Echo1.5 (Table 1). Compared with MiniMax-H3, which shares our backbone, CINELONG improves all four cross-shot metrics, most notably Background Recall (2.35 to 4.07) and Audio Continuity (0.328 to 0.486). This displays the effectiveness of our memory mechanism in multishot video generation.



Figure 5: **Qualitative comparison.** Columns align shared narrative time bins across four methods. Matching colors and letter tags pair earlier and later regions within each method; captions beneath the frames name the observed inconsistency.

Consistency does not come at the expense of script adherence or per-shot quality. CINELONG achieves the best S-VLM, Speech Recall, and CLIP Score, and the second-best Cut F1, showing that it follows the requested content, dialogue, and shot boundaries. For intra-shot quality, it achieves the best lip synchronization (LSE-C) (Table 1) and the best subject and background consistency among audio-visual methods, with Imaging Quality close to the best (Table 4 in the appendix). The video-only ShotStream achieves higher VBench scores but has the lowest S-VLM.

User Study. We ran a user study with 20 participants, who together judged about 900 pairwise video comparisons (Figure 4). CINELONG is preferred over LTX2.5, JoyAI-Echo1.5, and StoryMem in 58.8%, 65.0%, and 73.8% of judgments, and loses in only 23.8%, 17.5%, and 13.8%, respectively.

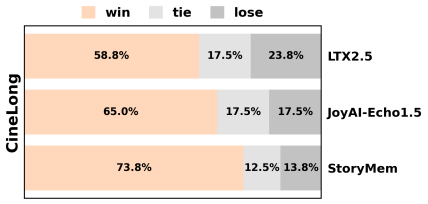


Figure 4: Pairwise human preference.

Qualitative example. Figure 5 displays that all the baselines fail when locations and characters return after a long gap. LTX2.5 alters revisited locations and replaces returning characters (A–C). StoryMem renders one location with another's layout and changes identities (D, E). JoyAI-Echo1.5 regenerates revisited locations with new furniture and lighting and changes characters' clothing (F, G). CINELONG keeps returning locations and characters consistent. Videos are in the supplementary material.

5.3 ABLATION STUDIES

Long- and Short-Term Memory. The two memories are complementary, and neither can substitute for the other (Table 2 (a)). Without LTM, the model still sees the preceding chunk through STM, yet Background and Subject Recall collapse from 4.07 and 4.33 to 2.07 and 3.51, close to the memory-free MiniMax-H3 (2.35 and 3.55) in Table 1, and Voice Consistency drops from 0.448 to 0.393: local continuity cannot carry identity across a character's or location's absence. Without STM, the retrieved LTM keeps recall largely intact (3.85 and 4.09), but Audio Continuity falls from 0.486 to 0.427 and Cut from 0.893 to 0.792: sparse keyframes specify what should reappear, not how to continue from the current moment. LTM thus answers *what to remember*, and STM *how to continue*.

Table 2: **Ablations on CINELONG-BENCH.** (a) and (b) reuse the full model's checkpoint and change only inference-time memory; (c) retrains with less recycling. **Bold**: best; underline: second best.Full metrics are reported in Table. 3.

	Setting				Cross-shot Consistency				Script Adherence		LSE-C↑
	LTM chunks	LTM frames	STM	Recycling	BGR↑	SBR↑	VC↑	AC↑	S-VLM↑	Cut↑	
(a) Memory at inference											
w/o STM	GDMA	GDMA	×	A+V	3.85	4.09	0.437	0.427	3.900	0.792	2.67
w/o LTM	–	–	✓	A+V	2.07	3.51	0.393	0.449	**3.967**	0.881	2.59
(b) LTM selection policy											
Recent + Midpoint	Recent	Midpoint	✓	A+V	2.54	3.70	0.441	0.476	3.937	0.879	**2.73**
Random + Midpoint	Random	Midpoint	✓	A+V	3.25	4.07	0.443	0.463	3.856	0.891	**2.73**
Random + HPSv3	Random	HPSv3	✓	A+V	3.19	4.21	0.460	0.482	3.826	0.880	2.72
GDMA + Midpoint	GDMA	Midpoint	✓	A+V	3.77	4.16	0.458	0.464	3.826	0.876	2.62
(c) Error recycling during training											
No Recycling	GDMA	GDMA	✓	None	3.87	4.09	0.455	0.470	3.693	0.910	2.67
Video-only (VER)	GDMA	GDMA	✓	V	3.84	4.27	**0.467**	0.480	3.763	**0.920**	2.72
Full (CINELONG)	GDMA	GDMA	✓	A+V	**4.07**	**4.33**	0.448	**0.486**	3.948	0.893	2.66

Goal-Driven Memory Agent. Goal-driven memory selection. GDMA decides which chunks enter LTM and which frame represents each of their shots; Table 2 (b) varies one decision at a time. Which chunks. A sliding window over the K most recent chunks (Recent + Midpoint) yields the lowest recall (2.54 and 3.70), below even random sampling from the full history (3.25 and 4.07), since returning characters and locations have often left the window. Selecting chunks by prompt relevance (GDMA + Midpoint) raises Background Recall to 3.77. Which frame. Letting GDMA also pick each shot's keyframe raises it to 4.07 and Subject Recall from 4.16 to 4.33, whereas the highest-HPSv3 frame (Random + HPSv3) helps Subject Recall (4.21) but not Background Recall (3.19), as the best-looking frame need not show the required view. Overall, GDMA achieves the best Background Recall, Subject Recall, and S-VLM among all policies (4.07, 4.33, and 3.948), showing that both chunk- and frame-level selection are needed.

Audio-Visual Error Recycling. Recycling closes the gap between the clean history seen in training and the imperfect history generated at inference, and it covers both modalities (Table 2 (c)). Recycling video errors alone, as in SVI, raises Subject Recall from 4.09 to 4.27 and S-VLM from 3.693 to 3.763, but leaves Background Recall unchanged (3.87 to 3.84). Additionally recycling audio errors brings the largest gains, lifting S-VLM to 3.948 and Background Recall to 4.07, even though the added perturbation is acoustic. Audio Continuity also rises at each stage, from 0.470 without recycling to 0.480 with video recycling and 0.486 with AVER. Exposing the generator to its own errors in both streams therefore improves visual recall, script adherence, and cross-chunk audio continuity together, yielding more coherent joint audio-visual generation.

6 DISCUSSION AND CONCLUSION

We presented CINELONG, a long-horizon multi-shot audio-visual generation framework built on complementary memories: LTM recalls characters, voices, and scenes from selected historical chunks, while STM continues ongoing motion and speech. GDMA selects target-relevant chunks and keyframes for LTM at inference, and AVER makes the generator robust to its own imperfect history. We also contribute CINELONG-709K, whose global-to-local annotations make cross-chunk dependencies explicit. On CINELONG-BENCH, CINELONG achieves the strongest subject recall, background recall, and audio continuity without sacrificing script adherence.

Limitations and future work. Like other agent-based systems, GDMA relies on an external MLLM, so its selection quality and cost follow the capability of the underlying model; lighter or jointly trained selectors are a natural extension. Our memory is organized at the granularity of chunks and shots, and finer-grained audio-visual memory, for instance at the level of individual speakers or objects, may further benefit dense multi-character scenes. Extending CINELONG to hour-scale narratives and interactive generation is another promising direction.

AI USE STATEMENT

We used generative AI tools for language polishing and dataset annotation. Specifically, Qwen3.6-27B and Qwen3-Omni-30B-A3B generated candidate visual and audio annotations for CINELONG-709K as part of the annotation pipeline (Appendix A). We reviewed all AI-assisted writing and take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe CINELONG in Section 4, and the construction of CINELONG-709K and CINELONG-BENCH in Section 3 and Appendix A. Appendix B provides implementation details of the memory and GDMA, and Section 5.1 details the evaluation protocol. Together, these materials provide the details needed to reproduce the reported results.

REFERENCES

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. URL https://arxiv.org/abs/2005.04208.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *INTERSPEECH*, 2023.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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

Minghong Cai, Qiulin Wang, Zongli Ye, Wenze Liu, Quande Liu, Weicai Ye, Xintao Wang, Pengfei Wan, Kun Gai, and Xiangyu Yue. VideoCanvas: Unified video completion from arbitrary spatiotemporal patches via in-context conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4475–4485, 2026.

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.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

Yixuan Ding, Jiahao Kong, Wei Huang, Ruijie Quan, and Yi Yang. Layerrecall: A state-conditioned memory router for long-horizon consistency in video generation. *arXiv preprint arXiv:2608.28460*, 2026. URL https://arxiv.org/abs/2608.28460.

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

Ridouane Ghermi, Xi Wang, Vicky Kalogeiton, and Ivan Laptev. Long story short: Story-level video understanding from 20k short films. *International Journal of Computer Vision*, 134:396, 2026. doi: 10.1007/s11263-026-02940-x.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. URL https://arxiv.org/abs/2601.03233.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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. URL https://arxiv.org/abs/2606.21661.

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

Chen Li, Peng Zhang, Hanyu Zhou, Jialong Zuo, Fei Wang, Daiguo Zhou, Nong Sang, and Changxin Gao. Tether the subject, release the scene: Query-aware memory routing for long-horizon autoregressive video generation. *arXiv preprint arXiv:2608.26902*, 2026. URL https://arxiv.org/abs/2608.26902.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. OpenHumanVid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7762, 2025a.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable Video Infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025b. URL https://arxiv.org/abs/2510.09212.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, and Tat-Seng Chua. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025. URL https://arxiv.org/abs/2503.23377.

Kaiqi Liu, Yunyao Mao, Ziqi Cai, Zheng Geng, Jing Wang, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, Shuchen Weng, and Boxin Shi. MAVIN: Multi-shot audio-visual generation with customized narrative control. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025. URL https://arxiv.org/abs/2508.03789.

MiniMax AI. Minimax-h3 technical report and model card. https://huggingface.co/MiniMaxAI/MiniMax-H3, 2026. Accessed 2026-08-24.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model. https://qwen.ai/blog?id=qwen3.6-27b, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Christoph Schuhmann. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

Tomas Soucek and Jakub Lokoc. TransNet V2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

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

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.

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. URL https://arxiv.org/abs/2305.10874.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fred Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. *arXiv preprint arXiv:2412.07772*, 2025. URL https://arxiv.org/abs/2412.07772.

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

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2bde8fef08f7ebe42b584266cbcfc909-Abstract-Conference.html.

Project page: `https://cinelong.github.io/project`

A Dataset Construction and Annotation Details

A.1 Dataset Organization and Scale

Corpus scale and video duration. CineLong-709K comprises 708,794 cinematic videos, with individual durations ranging from approximately 17 to 148 seconds and a total duration of approximately 20k hours of high-definition audio-visual footage. The complete video provides the temporal scope for entity registration and cross-chunk annotation; its constituent chunks provide local descriptions within that scope.

Data sourcing and preliminary filtering. We assemble the raw corpus from public datasets, including Condensed Movies (Bain et al., 2020), Short-Films20K (Ghermi et al., 2026), and OpenHumanVid (Li et al., 2025a), as well as publicly available videos from online platforms such as YouTube. We retain source provenance and basic technical metadata throughout processing. Before annotation, a preliminary filter keeps only videos that decode end-to-end, satisfy the target duration and spatial-resolution ranges, and contain a usable audio track. We further exclude samples dominated by severe black borders, persistent subtitles or overlaid text, title cards, openings or end credits, black or near-static content, and gross visual or acoustic corruption. Before VLM inference and chunk-level annotation, we additionally reject near-static candidates using RAFT optical flow (Teed & Deng, 2020) and apply perceptual quality thresholds using the improved aesthetic predictor (Schuhmann, 2022) for visual aesthetics and Audiobox-Aesthetics (Tjandra et al., 2025) for audio.

Chunk-level annotation within long videos. We formulate the dataset as sequential *chunk chains* aligned with the native temporal input budget of the target generator. We first detect shot boundaries on the source timeline with TransNetV2 (Soucek & Lokoc, 2020), then randomly partition the video into consecutive, non-overlapping temporal intervals that preserve temporal order and cover the video without gaps. Each chunk is a temporally contextualized generation unit that may contain multiple camera shots. We record the shots overlapping each chunk and their temporal extents within that chunk. A chunk boundary may occur inside a continuous shot, action, or utterance; these cases are retained as supervision and represented through explicit continuity fields. Within each video sequence, chunks span 8.0 to 25.0 seconds and average 16.8 seconds and 4.6 detected camera shots. These statistics describe the local annotation units, not the duration or shot count of the complete video. The ordered chunk annotations retain their links to the full sequence through shared entity references and cross-chunk continuity fields.

A.2 Structured Multimodal Annotations

The structured annotation in Figure 2 combines local audio-visual descriptions with persistent entity references. Each chunk record contains the fields shown in Figure 2: the chunk index and number of shots, subject definitions, the scene setting, a shot-by-shot integrated multimodal description, the overall soundscape, and non-diegetic music.

A.3 Five-Stage Automatic Annotation Pipeline

Reliable chunk-level annotation requires global entity coverage, local visual grounding, and sequence-level revision. We decompose automatic annotation into five constrained stages with structured inputs and outputs. Qwen3.6-27B (Qwen Team, 2026) handles the visually centered passes, including entity registration, chunk-level entity presence detection, and sequence-level revision. Qwen3-Omni-30B-A3B (Xu et al., 2025) drafts chunk captions from the native video and audio streams, conditioned on the entity registry and presence table, and identifies speech, speaker cues, sound events, ambience, music, and audio-visual correspondences.

Stage 1: Sparse global entity registration. Processing every frame of a long video with a VLM incurs prohibitive visual-token cost. We sample a small set of temporally distributed representative frames over the complete sequence,

$$\mathcal{K}^{\mathrm{reg}} = S_{\mathrm{sparse}}(V) = \{I_{\tau_1}, \ldots, I_{\tau_M}\}, \qquad M \ll T, \tag{5}$$

where T is the total number of source frames. The sampler preserves coverage of the complete timeline and retains representative views from visually distinct shots after redundancy reduction. Frames are provided to the VLM in chronological order. For sequences that exceed the VLM context limit, sparse frames are processed in temporal blocks and the resulting text records are merged. The VLM outputs a text-only registry

$$\mathcal{R} = \{\mathcal{R}_{\text{person}}, \mathcal{R}_{\text{object}}, \mathcal{R}_{\text{scene}}\}, \tag{6}$$

whose entries contain canonical IDs, entity types, and concise identity attributes. This stage creates registry records and does not generate frame-level or chunk-level captions.

Stage 2: Chunk-level entity presence detection. For each chunk Y_t, a VLM matches visible content against the global registry and produces an entity-presence table

$$\mathcal{P}_t = \{\mathcal{P}_t^{\text{person}}, \mathcal{P}_t^{\text{object}}, \mathcal{P}_t^{\text{scene}}\}. \tag{7}$$

The table contains only registered entities supported by the chunk, together with their current state, active scene or subregion, and observable relations. Persistent IDs connect reappearing entities across distant chunks, while a hierarchical scene ID links different camera views of the same physical environment.

Stage 3: Grounded chunk-caption generation. The captioning VLM receives the current audio-visual chunk, its presence table, and the global registry, and produces the structured fields in Appendix A.2. Subject definitions describe the target state visible in the current chunk. Shot descriptions retain actions and event progression required for video generation.

Stage 4: Cross-chunk consistency revision. A sequence-level VLM reviews the ordered chunk captions together with the registry and presence tables. It resolves contradictory identity attributes, normalizes scene IDs across viewpoint changes, and distinguishes evidence-supported state transitions from annotation drift. It also assigns a cross-chunk continuity label by checking whether camera motion, background ambience, actions, or entity states carry across each chunk boundary.

Stage 5: Speech and audio-boundary alignment. WhisperX (Bain et al., 2023) provides automatic speech recognition and word-level timestamps through voice-activity detection and forced alignment. Qwen3-Omni-30B-A3B combines these timestamps with audio-visual evidence to associate each utterance with its active entity and chunk interval. Utterances crossing a chunk boundary receive explicit `continues_from_previous` and `continues_to_next` markers, preserving syntactic continuity without duplicated text. The stage also records non-speech sound events, background ambience, and music descriptors aligned to the corresponding chunk.

Benchmark source isolation. All source sequences used in CINELONG-BENCH (Section 3.2) are excluded before training-data construction. Source-level exclusion does not by itself establish entity-level isolation, nor does it rule out overlap with backbone pretraining.

B MEMORY AND AGENT IMPLEMENTATION

Memory budget. We use $K = 2$ historical chunks for LTM during training and inference, with the same budget across all internal comparisons. When fewer than two historical chunks are available, all are retained. STM is attached separately and does not consume an LTM slot. Both memories are empty at the first step unless an initial reference is supplied. This is a fixed implementation setting, not a requirement of the memory formulation.

Backbone and fine-tuning. All experiments build on the pretrained MiniMax-H3 (33B) joint audio-visual backbone (MiniMax AI, 2026), which we fine-tune with LoRA (Hu et al., 2022).

Backbone tokenization. The causal visual VAE maps video to normalized latents $z_t^v \in \mathbb{R}^{24\times F\times H/16\times W/16}$. A $1 \times 2 \times 2$ patchification produces tokens $x_t^v \in \mathbb{R}^{L_v\times 96}$ with $L_v = F(H/32)(W/32)$; the audio encoder produces aligned tokens $x_t^a \in \mathbb{R}^{L_a\times d_a}$. Both modalities condition the joint flow-matching backbone.

STM alignment. The backbone temporal grid restricts visual STM to $L^v_{\mathrm{STM}} = 5k + 2$ latent frames, corresponding to $17k + 5$ pixel frames. The fixed setting $k = 1$ gives seven latent frames, 22 pixel frames ($22/24 \approx 0.92$ seconds), and 37 aligned audio latent frames. This is a packing constraint of the backbone. Both LTM and STM retain original temporal positions through multimodal rotary embeddings. When the predecessor is also selected for LTM, its audio is split into an LTM prefix and the STM suffix, so each audio token appears once at its original position; this also prevents clean LTM audio from duplicating an STM suffix perturbed by AVER.

Replay buffers and strength. Detached prediction residuals are stored in bounded FIFO buffers matched by latent resolution and noise grid. Residual capture reuses the available velocity predictions. The strength multiplier is $\gamma_q \sim \mathcal{U}(1 - f, 1 + f)$, where f controls the modulation range. The full configuration uses $p_v = p_a = 0.6$, $\alpha_v = 0.5$, and $\alpha_a = 0.2$. No Recycling sets $p_v = p_a = 0$; Video-Only Error Recycling sets $p_v = 0.6$ and $p_a = 0$, keeping the visual settings identical to the full configuration. All variants train against the same clean target.

Candidate reduction. Frames are sampled at stride two (12 fps at 24 fps output). Within each shot, CLIP-L embeddings retain a frame only when its cosine distance to the nearest retained embedding exceeds τ. The threshold increases from 0.10 to 0.60 in steps of 0.05 when the shot budget is exceeded. HPSv3 then retains candidates within 1.5 of the shot's maximum and protects its highest-scoring frame. Chunk-level pruning removes lower-scoring candidates only from shots with multiple representatives. The configured budget is $\min(12, S_j + \lfloor T_j/4 \rfloor)$ for duration T_j in seconds.

Agent execution. GDMA supports Gemini 3.1 Pro or a locally deployed Qwen3.6-27B model (Qwen Team, 2026). The response includes a usefulness flag, a relevance score in $[0, 1]$, and per-shot frame choices. Chunks are ranked by relevance to fill $\min(K, t)$ slots, and chunk indices and frame choices are validated before the selected observations are encoded. Communication failure after two retries or an incomplete required frame assignment stops generation, rather than silently replacing GDMA with another policy.

C Full Ablation Results

Table 3 reports all twelve metrics of Table 1 for every row of Table 2, whose setting columns define each row.

Table 3: **Full ablation results on CineLong-Bench.** Rows and blocks follow Table 2; metric groups follow Table 1. The shaded row is the full model.

	Cross-chunk Consistency				Script Adherence				Per-shot Quality			
Setting	BGR↑	SBR↑	VC↑	AC↑	S-VLM↑	Cut↑	SR↑	CLIP↑	IQ↑	SC↑	BGC↑	LSE-C↑
Full (CineLong)	4.07	4.33	0.448	0.486	3.948	0.893	0.993	0.277	0.702	0.940	0.949	2.66
(a) Memory at inference												
w/o STM	3.85	4.09	0.437	0.427	3.900	0.792	0.992	0.277	0.705	0.943	0.950	2.67
w/o LTM	2.07	3.51	0.393	0.449	3.967	0.881	0.980	0.277	0.711	0.940	0.947	2.59
(b) LTM selection policy												
Recent + Midpoint	2.54	3.70	0.441	0.476	3.937	0.879	0.989	0.278	0.701	0.941	0.949	2.73
Random + Midpoint	3.25	4.07	0.443	0.463	3.856	0.891	0.992	0.277	0.699	0.941	0.950	2.73
Random + HPSv3	3.19	4.21	0.460	0.482	3.826	0.880	0.995	0.277	0.702	0.940	0.949	2.72
GDMA + Midpoint	3.77	4.16	0.458	0.464	3.826	0.876	0.991	0.277	0.701	0.941	0.949	2.62
(c) Error recycling during training												
No Recycling	3.87	4.09	0.455	0.470	3.693	0.910	0.994	0.276	0.706	0.941	0.949	2.67
Video-only (VER)	3.84	4.27	0.467	0.480	3.763	0.920	0.994	0.276	0.703	0.940	0.949	2.72

D Evaluation Details

This appendix complements Section 5.1 with the baseline settings, metric implementations, prompt formats, ablation configurations, and supplementary cases. All methods are evaluated on the com-

Table 4: **Main comparison on CINELONG-BENCH. Bold**: best. underline: second best. †: video-only. –: not applicable.

	Cross-shot Consistency				Script Adherence				Intra-shot Quality			
Method	BGR↑	SBR↑	VC↑	AC↑	S-VLM↑	Cut↑	SR↑	CLIP↑	IQ↑	SBC↑	BGC↑	LSE-C↑
ShotStream† (Luo et al., 2026)	2.83	2.86	–	–	1.737	–	–	0.247	**0.711**	**0.975**	**0.968**	–
StoryMem† (Zhang et al., 2025a)	3.60	2.98	–	–	2.578	–	–	0.268	0.683	0.909	0.935	–
JavisDiT++ (Liu et al., 2025)	1.89	2.56	0.344	0.282	1.763	–	0.034	0.265	0.666	0.814	0.885	0.74
LTX2.3 (HaCohen et al., 2026)	2.01	3.12	0.328	0.393	2.326	–	0.906	0.259	0.560	0.846	0.916	2.22
LTX2.5 (HaCohen et al., 2026)	1.83	2.36	0.409	0.393	3.163	0.108	0.974	0.260	0.673	0.840	0.918	2.57
JoyAI-Echo1.5 (Duan et al., 2026)	2.59	2.85	**0.473**	0.272	2.967	0.126	0.971	0.266	0.655	0.921	0.939	2.44
MiniMax-H3 (MiniMax AI, 2026)	2.35	3.55	0.383	0.328	3.770	**0.915**	0.985	0.271	0.710	0.918	0.944	2.59
CINELONG (Ours)	**4.07**	**4.33**	0.448	**0.486**	**3.948**	0.893	**0.993**	**0.277**	0.702	0.940	0.949	**2.66**

plete chains of CINELONG-BENCH (30 scripts × 10 chunks; Section 3.2). Evaluator models, judge prompts, and thresholds are identical across methods.

D.1 BASELINE DETAILS

Shared protocol. All methods receive the same CINELONG-BENCH scripts, converted deterministically into each model's native prompt format: shot-level models receive one prompt per scripted shot, chunk-level models receive the whole chunk with its cut timestamps, and the video-only models receive dialogue as speaking actions without the spoken text. Shot-level models are limited to about 5 s per call, so shots longer than this limit are clipped and are only let to generated the 5 s segments; all evaluation aligns outputs to the script through the generation timeline, so this does not affect how shots are matched.

ShotStream (Luo et al., 2026). We use the released checkpoint at its native 480×832 and 16 fps with 4 distilled sampling steps. Each scripted shot (up to 81 frames) starts a new stream. Following the official implementation, the model is conditioned on 6 context frames sampled uniformly over all previous shots of the story, together with their captions. It does not generate audio.

StoryMem (Zhang et al., 2025a). We follow the official pipeline: the first shot is generated by Wan2.2-T2V-A14B, and every later shot by Wan2.2-I2V-A14B with the MI2V LoRA, conditioned on a memory bank of up to 10 keyframes (the first 3 kept fixed). Keyframes are extracted from the model's own outputs with the official CLIP- and HPSv3-based selection. We use 480×832, 16 fps, up to 81 frames per shot, and 40 sampling steps. It does not generate audio.

JavisDiT++ (Liu et al., 2025) JavisDiT++ has no continuation or memory mechanism. We generate every scripted shot independently with the official text-to-audio-video model (480×864, 16 fps, up to 81 frames, 50 sampling steps) and concatenate the clips along the timeline.

LTX2.3 and LTX2.5. (HaCohen et al., 2026) Both use the official two-stage pipeline at 1024×1536 and 24 fps with 30 sampling steps. They are chained by prefix extension: every call after the first re-encodes the last 57 frames (8 latent frames) of the previous clip and injects them as clean latents at the start of the sequence; the prefix is removed after decoding. The prefix length follows the official extension example; we also condition on the prefix audio so that the soundtrack continues across calls. LTX2.3 is called once per shot (up to 121 new frames). LTX2.5 is called once per chunk (345 new frames) with the whole multi-shot chunk script, including cut timestamps.

JoyAI-Echo1.5. (Duan et al., 2026) We use the official model at 736×1280 and 25 fps with 8 sampling steps, one request per chunk (361 frames). Memory across chunks is built by the official VLM Director: after each chunk, a VLM (Gemini 3.1 Pro) reviews candidate frames and selects a reference frame for each character at its first appearance and a representative frame of the previous shot, each paired with the audio of its source chunk. These form the memory slots of the next request.

MiniMax-H3. (MiniMax AI, 2026) We use the official FL2VA mode of Minimax-H3 with its default settings (768×1344, 24 fps, CFG-distilled, 49 sampling steps), one request per chunk (345

frames). Chunk 0 is generated from text only; every later chunk is conditioned on the last decoded frame of the previous chunk as its first frame. No audio or other context crosses chunk boundaries.

D.2 Metric Implementation

All metrics are computed per case (one ten-chunk chain) and averaged over the 30 cases. Audio metrics are not applicable to the video-only baselines, and Cut is not applicable to shot-level generators, whose cuts are placed by the scheduler.

Shot units. VC, IQ, SBC, BGC, and LSE-C are computed on shot units. For shot-level generators, each generation call is one unit. For chunk-level generators, hard cuts are detected with the PySceneDetect AdaptiveDetector (threshold 3.0, minimum 12 frames; F1 = 0.94 at 0.5 s tolerance on videos with known cut times); cuts within 0.5 s of a chunk boundary are ignored, and each segment is assigned to the scripted shot it overlaps most. Segments shorter than 0.5 s are dropped, and multi-shot chunks without any detected cut are excluded.

Cross-shot consistency. SBR and BGR are scored by Gemini 3.1 Pro (temperature 0, JSON output) on a 1–5 scale. Targets are selected with a fixed seed independently of the method, so every method is scored on the same targets. Inputs are muted and resized to a 480-pixel short side, and prompts never contain method names. BGR targets are chunks that return to a previously visited location after the preceding chunk was set elsewhere. The VLM receives the returning chunk and the chunk of the last visit, and scores only the location (layout, props, lighting, and palette), ignoring camera changes and changes stated in the script. SBR targets are pairs of shots showing the same character in two chunks separated by at least one chunk where the character is absent. Each side is represented by the four frames with the largest detected faces. The VLM first localizes the character, scoring 1 if it is not found, and then compares stable identity cues (face, hair, outfit, and body), ignoring expression, viewpoint, and scripted costume changes.

For VC, the dialogue stem is separated with TIGER-DnR and transcribed with Whisper-large-v3. Transcript segments are aligned to the scripted lines by monotonic dynamic programming and inherit the speaker of a line when their character-level similarity is at least 0.3. Aligned segments of at least 0.6 s are embedded with CAM++; embeddings are averaged per speaker and shot, and VC is the mean cosine similarity over all shot pairs of the same speaker. Cases without aligned speech are excluded.

AC is measured at every chunk seam of the stitched output. We take 2 s of audio on each side of the seam, embed both windows with LAION-CLAP, and use their cosine similarity; AC averages this over the seams of a case.

Script adherence. S-VLM is scored by Gemini 3.1 Pro (temperature 0) for every chunk except the first. The judge receives the chunk re-encoded at 6 fps and 384-pixel width with its audio, together with the chunk script (subjects, shots, dialogue, and sound requirements), and gives a 1–5 score on high-level script semantics: characters, actions, relations, scenes, props, and shot order. It is instructed not to judge identity, voice, exact cut timing, verbatim dialogue, or audio-visual synchrony, which are covered by other metrics. S-VLM averages the chunk scores of a case.

For Cut, scripted within-chunk cut times are matched one-to-one with detected cuts by Hungarian matching; matched pairs within 0.5 s are true positives. True positives, false positives, and false negatives are summed over the chunks of a case before computing F1.

For SR, the audio of each chunk with scripted dialogue is transcribed with Whisper-large-v3 and compared with the concatenated scripted lines of that chunk. Chinese, Cantonese, Japanese, and Thai are tokenized by character, and other languages by lowercased words without punctuation. After alignment, reference tokens that are neither deleted nor substituted count as hits, and SR is the number of hits divided by the number of reference tokens, pooled over the chunks of a case.

CLIP is computed on every scripted shot over its scheduled interval, so it does not depend on cut detection. Eight uniformly sampled frames are encoded with CLIP ViT-L/14 and averaged. The text is the shot description with dialogue and sound descriptions removed and subject tags replaced by nouns; each sentence is truncated to 77 tokens and the sentence embeddings are averaged. The score is the cosine similarity between the two.

Intra-shot quality. IQ, SBC, and BGC use the official VBench implementation on shot units. IQ scores frames with MUSIQ after resizing the longer side to 512 pixels, which partly reduces resolution differences across methods. SBC and BGC compare each frame with the previous frame and the first frame using DINO ViT-B/16 and CLIP ViT-B/32 features, respectively.

LSE-C is computed with SyncNet on shot units whose scripted shot contains dialogue. Frames are resampled to 25 fps, faces are detected with S3FD and tracked, and LSE-C is the median minus the minimum of the average audio-visual distance over offsets of ± 15 frames, averaged over all face tracks of a case. Because most shots are shorter than 4 s, we require tracks of at least 25 frames and faces of at least 64 pixels instead of the official 100 frames and 100 pixels, so absolute values are lower than those reported in prior work.

D.3 Prompt Formats for Evaluation

All methods are driven by the same structured chunk annotations (Figure 6). Baseline prompts are derived from the annotated fields by deterministic concatenation, without re-paraphrasing, and are kept within each model's text-encoder budget. Depending on the native input granularity of each generator, we use three prompt formats.

Structured chunk prompts. Ours and MiniMax-H3 consume the structured record verbatim. This includes the subject definitions, the scene setting, and the shot-by-shot integrated multimodal description, where each shot is marked by `[Shot` k`]` and its cut timestamp. It also includes speaker-tagged dialogue enclosed in `<d>` tags, the overall soundscape, and non-diegetic music. Each chunk is generated in a single call. MiniMax-H3 generates the first chunk from text alone. For every later chunk, it only conditions on the last frame of the preceding chunk, with the official first-frame reference line prepended to the prompt. Ours additionally reads keyframes retrieved from its memory.

Flattened chunk prompts. LTX-2.5 and JoyAI-Echo1.5 also generate one chunk per call, but they accept only free-form text. We therefore flatten the record into a single paragraph. The scene setting, subject definitions, and visual style come first, and entity tags such as `<Subject 1>` become plain "Subject 1". Each shot boundary is verbalized with its timestamp (e.g., "At 00:05.500, the scene cuts to ..."). Dialogue is rewritten as quoted speech with its speaker and language, and the soundscape and music descriptions are appended once at the end. The same text is fed to all three models. They differ only in how cross-chunk context is supplied: LTX-2.5 extends a video prefix, while the JoyAI-Echo use memory reference slots.

Per-shot prompts. JavisDiT++, LTX-2.3, StoryMem, and ShotStream generate one shot per call, so each chunk is split into one prompt per shot. Every per-shot prompt repeats the scene setting, subject definitions, and visual style, followed by the description of that shot alone. Cut timestamps are dropped, since shot boundaries are placed by the scheduler rather than generated by the model. For the audio-video models, the prompt also carries the shot's own dialogue and the chunk soundscape; JavisDiT++ generates each shot independently, whereas LTX-2.3 extends the previous shot through a video prefix. For the video-only models, dialogue and sound descriptions are removed and only the speech action is kept (e.g., "Subject 1 shouts."). StoryMem conditions later shots on keyframes from its memory bank, and ShotStream further prepends a condensed caption of the whole chunk, followed by a `shot{`k`}:` tag, to each per-shot prompt.

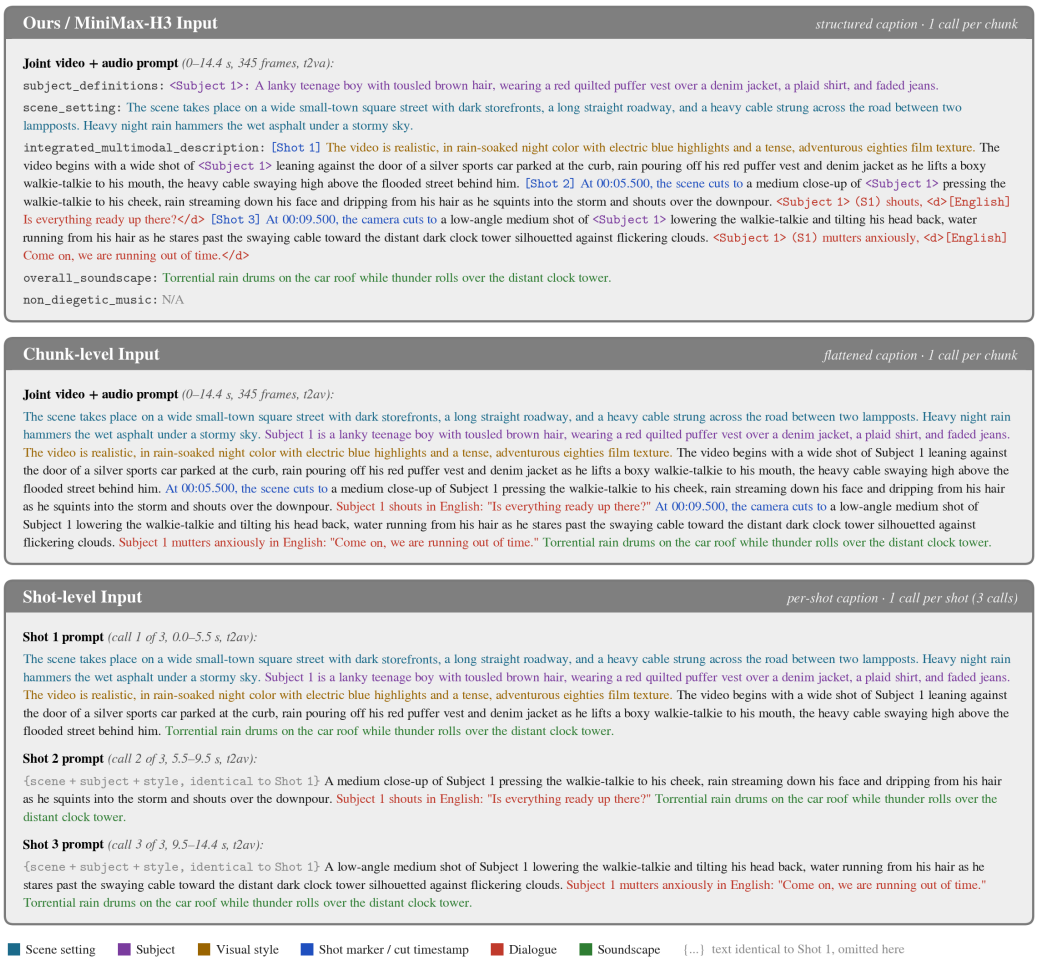


Figure 6: **Input prompts for different generation granularities.** We show the first chunk of one evaluation case (14.4 s, three shots, two dialogue lines). **Top:** Ours and MiniMax-H3 take the structured record verbatim, covering subject definitions, scene setting, the shot-by-shot description with `[Shot k]` markers and cut timestamps, speaker-tagged dialogue, soundscape, and music. The whole chunk is generated in a single call. **Middle:** Chunk-level baselines (LTX-2.5 and JoyAI-Echo1.5) receive the same content flattened into one paragraph, still in a single call. Cut timestamps are kept as text, but field boundaries and speaker tags are lost. **Bottom:** Shot-level baselines (Javis-DiT++ shown; LTX-2.3, StoryMem, and ShotStream follow the same scheme) are called once per shot. Each call repeats the shared context and describes only its own shot, without cut timestamps or information about other shots. The repeated prefix of Shots 2 and 3 is collapsed into a gray `{...}` for display only; each call receives the full text. Colors mark the scene setting, subjects, visual style, shot markers and cut timestamps, dialogue, and soundscape.