ONE SENTENCE, ONE DRAMA: PERSONALIZED SHORT-FORM DRAMA GENERATION VIA MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing approaches for digital short-drama production typically rely on one-shot LLM generated scripts and loosely coupled pipelines, which fail to satisfy three key requirements of short-drama generation: (1) narrative pacing, resulting in weak hooks, insufficient escalation, and unattractive endings; (2) spatial consistency, leading to drifting scene layouts and inconsistent character positions across clips; and (3) production-level quality control, requiring extensive manual review and correction across script and visual stages. We present *One Sentence, One Drama*, a hierarchical multi-agent framework that transforms a user's single-sentence idea into a fully produced short drama through structured intermediate modules and iterative refinement. Our approach is built upon three key components: (1) a multi-agent debate-based story generation module that enforces short-drama pacing and narrative coherence; (2) a 3D-grounded first-frame generation mechanism that establishes a shared spatial reference for consistent character positioning and scene layout across clips; and (3) multi-stage reviewer loops that perform comprehensive error detection and targeted revision across script, visual, and video generation stages. We also introduce scene-level BGM matching and scene transition planning to improve the audience's immersive experience. To systematically evaluate this task, we introduce *Short-Drama-Bench*, a benchmark that extends standard video quality metrics with short-drama-specific criteria. Experimental results demonstrate that our method significantly outperforms existing pipelines in narrative quality, cross-clip consistency, and overall viewing experience.

1 INTRODUCTION

Recent advances in video foundation models have substantially improved automated short-clip generation. Models such as Sora (Brooks et al., 2024), Seedance (Seedance et al., 2026), Kling (Kuaishou Technology, 2024), and Veo (Google DeepMind, 2025) have demonstrated strong capabilities in visual fidelity, motion realism, and prompt following. These models provide a powerful basis for generating high-quality video clips from textual or visual conditions. Recent long-form generation pipelines have explored combining large language model planning with video synthesis. Systems such as MovieAgent (Wu et al., 2025), StoryMem (Zhang et al., 2025), and ScriptAgent (Mu et al., 2026) decompose long-video creation into multiple stages, representing an important step toward automated long-form video production. Nevertheless, these methods are primarily designed for organizing clips into longer videos and do not explicitly model the distinctive narrative dynamics of short dramas, which demand dense dramatic hooks—characterized by rapid conflict onset, high-frequency escalation and reversals, and fast-paced payoff within a highly compressed duration (Cao et al., 2026).

More recently, Toonflow (HBAI Ltd, 2026) and Xiaoyunque (ByteDance, 2026) have adapted generative models to short-drama production workflows. However, they still face three major limitations. First, they often rely on a ready-made story input, which shifts the burden of short-drama writing to the user (Cao et al., 2026). When only a brief idea is provided, they simply use one-shot LLM expansion, leading to weak dramatic hooks and unsatisfactory storylines. Second, they usually create clips using loosely connected generation units (Guo et al., 2025b; Elmoghany et al., 2026), causing cross-clip spatial inconsistencies such as drifting scene layouts, abrupt character position changes, and unresolved prop states. Third, their outputs typically require substantial manual

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.

inspection and correction across script, keyframe, and video stages before reaching production-level quality, due to diverse errors in pacing, character consistency, dialogue accuracy, spatial layout, prop states, and action continuity (Guo et al., 2025b; Elmoghany et al., 2026; Meng et al., 2026).

To address these challenges, we present *One Sentence, One Drama*, a hierarchical multi-agent framework for generating an entire short drama from a single-sentence idea. Our framework decomposes the generation process into a hierarchy of structured and revisable intermediate modules. Specifically, our framework consists of three core components. First, we introduce a **multi-agent debate-based story generation** module that improves short-drama pacing and narrative coherence by explicitly modeling opening hooks, conflict escalation, ending suspense, and storyline consistency through synergistic debate and revision. To counter the over-correction of iterative revision, which tends to delete memorable hooks and reversals, an **Idea Bank** records the removed elements and selectively revives them in a final round. Second, we propose **3D-grounded first-frame generation** to address cross-clip spatial drift. By constructing a scene-level 3D world model and aligning frames within a shared spatial coordinate system, the method enables consistent character positioning and scene layout across clips, even under severe viewpoint changes or scene re-establishment. We design **geometry-adaptive character registration** to place the single-frame character mesh into the scene-level 3D world at the correct size and distance. Third, we design **multi-stage reviewer loops** across script, prompt, keyframe, and video generation to enforce constraints on pacing, spatial relations, prop states, physical plausibility, and action continuity. In addition, we incorporate scene-level BGM matching and transition planning to further enhance the immersive viewing experience.

To verify our framework, we introduce *Short-Drama-Bench*, a novel and challenging benchmark that augments standard video-quality metrics with short-drama-specific criteria, including narrative engagement, spatial continuity, and full-production viewing experience. It consists of 50 diverse story prompts spanning 7 popular categories (e.g., Time Travel & Rebirth, Suspense & Thriller, Sweet Romance) and 17 fine-grained subcategories. Each subcategory contains 2–3 representative samples, covering a broad range of commonly observed short-drama patterns and narrative structures.

To further reflect the practical complexity of this task, we generate full short-drama outputs for all benchmark prompts, resulting in a total of approximately 239 minutes of video content. The generated results include a mixture of long-, medium-, and short-duration dramas, consisting of 2 long-form dramas (≈ 30 minutes each), 5 medium-length dramas (≈ 10 minutes each), and 43 short dramas (≈ 3 minutes each). This large-scale generation setup highlights the long-horizon consistency challenges of the task, as models must maintain narrative coherence, character consistency, and spatial continuity across hundreds of sequential clips. These characteristics make Short-Drama-Bench significantly more demanding than conventional short video benchmarks that focus on isolated clip generation. Experimental results demonstrate that our agentic framework consistently outperforms existing generation pipelines in narrative quality, cross-clip consistency, and overall viewing experience.

In summary, our main contributions in this work are as follows:

- We formulate single-sentence short-drama generation as a structured generation problem that requires jointly modeling narrative pacing, spatial consistency, and production-level coherence. We propose *One Sentence, One Drama*, a hierarchical multi-agent framework that transforms one-shot generation into a controllable and self-refining process.
- We introduce two key technical innovations to address the core challenges of this task: (i) a multi-agent debate-based story generation module, with an Idea Bank with final-round revival, to improve short-drama pacing and narrative coherence, and (ii) 3D-grounded first-frame generation, with geometry-adaptive character registration, to enforce cross-clip spatial consistency via a shared spatial coordinate system.
- We present *Short-Drama-Bench*, a diverse and challenging benchmark with 50 prompts across 7 categories and 17 subcategories, along with short-drama-specific evaluation metrics. Our benchmark enables systematic evaluation of narrative quality, spatial continuity, and full-production viewing experience.

2 Personalized Short-Form Drama Generation

Fig. 1 shows our hierarchical sentence-to-video pipeline. A single-sentence input is transformed into structured story plans and scene/clip-level scripts (Fig. 1.A), scene-level visual assets and paired

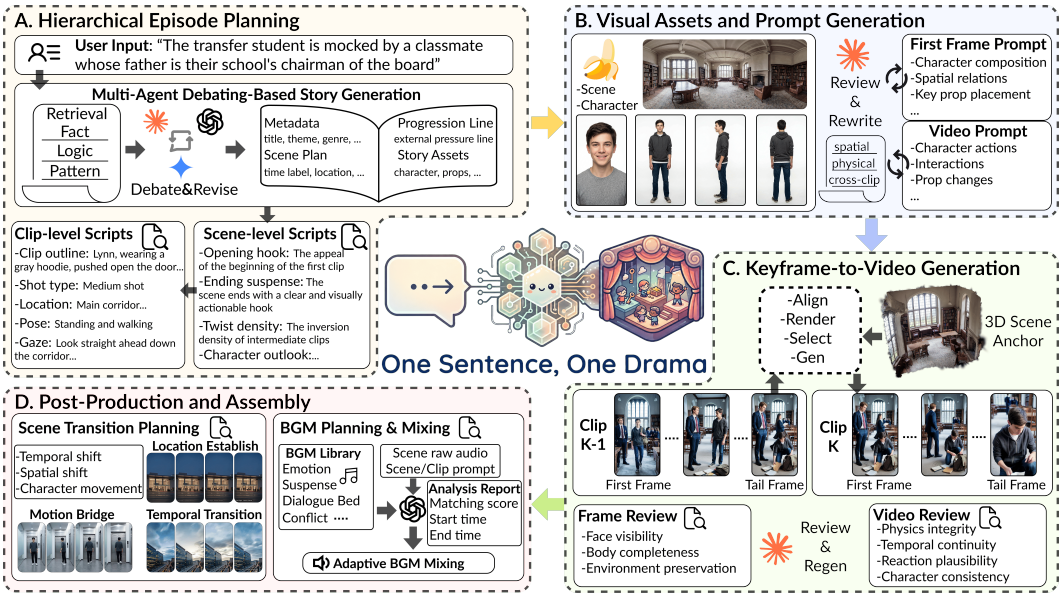


Figure 1: Overview of our personalized short-form drama generation with four stages. Given a single-sentence input, the system generates structured story and clip scripts through retrieval and multi-agent debate (A), expands into visual assets and paired first-frame/video prompts with prompt review (B), produces 3D-anchored keyframe-to-video clips with frame and video review loops (C), and assembles the final drama with scene transition planning and adaptive BGM mixing (D).

prompts (Fig. 1.B), 3D-anchored keyframe-to-video generation (Fig. 1.C), and post-production with scene transitions and BGM (Fig. 1.D). Reviewer loops are inserted across stages for quality control and cross-stage consistency. Section 2.1 describes episode planning with atom corpus construction, retrieval, and multi-agent debate-based story generation. Section 2.2 presents visual assets and prompt generation. Section 2.3 introduces our strategy for 3D-grounded next-keyframe and next-clip generation. Section 2.4 illustrates the cross-clip transition planning and drama BGM mixing.

2.1 Hierarchical Episode Planning

Atom Script Corpus Construction. Directly expanding a short-drama script from a single prompt often causes weak pacing and unstable local logic. To address this, we build an atomized corpus from 1,102 high-performing short-drama scripts and construct two retrieval banks. Each script is distilled into a structured script card, and together, these cards contain 16,311 beat-level units, encoding cues such as opening action, conflict function, and closing hook visual. These embedded beats form the *Pattern Bank*, which provides reusable pacing and dramatic packaging priors. In parallel, we split scripts into overlapping local chunks to form the *Logic Bank*, preserving causal context such as motivations, evidence activation, consequence transitions, and scene continuity. Thus, the corpus is transformed into transferable patterns and logic atoms rather than copied directly. Details are provided in Appendix P.

Multi-Agent Debating-Based Story Generation. Given a user's sentence as a logline, we first expand it into a seed text containing a preliminary story skeleton. Based on these, an LLM produces a problem-driven retrieval plan with three routes: fact, logic, and pattern. Fact retrieval invokes web search for externally constrained content, such as law, medicine, and history. Logic retrieval queries the *Logic Bank* for local causal support, while pattern retrieval queries the *Pattern Bank* for relevant short-drama structures. The retrieved references are summarized into fact, logic, and pattern atoms, providing factual, causal, and pacing priors for story drafting. Combining all these, the pipeline generates a structured story core, containing story-level metadata and the scene plan.

Next, we introduce scene-level script review and rewrite through a multi-agent debating loop. The draft story, story core, and retrieval atoms are reviewed by three independent LLM judges. When these judges provide conflicting revision suggestions, we send these suggestions to GPT-5.4 Pro as the final decider. The issues are passed to a reviser for local rewriting rather than full regeneration.

Idea Bank with Final-Round Revival. Iterative revision has a systematic bias: each round removes risk, deleting memorable elements such as hooks, reversals, and foreshadowing along the way, so the script becomes safer but flatter. To remedy this over-correction, we record every element deleted or weakened during revision, together with its reason, in an Idea Bank. After the revision loop terminates, a final-round revival step restores those elements that can return without harming narrative logic or visual executability. More details are given in Appendix C.4 and Fig. 4.

Scene-level and Clip-level Script Synthesis. After obtaining the story core containing the rewrite scene plan, we synthesize clip-level scripts for visual generation. Each scene is then decomposed into temporally ordered clip-level scripts, where each clip specifies its local narrative description, shot type, characters, key props, dialogue or audio cues, actions, interactions, and so on. We also extract each clip's initial and ending states before visual asset generation.

Finally, a clip-level review and rewrite loop is designed for short-drama pacing. The reviewer evaluates the opening hook, ending suspense, and twist density. Based on the evaluation, we perform partitioned rewriting: the opening-hook review revises only the first clip to ensure opening attraction; the ending-suspense review revises only the last clip to ensure a clear and visually actionable hook that motivates continued viewing; and the twist-density review revises only the middle clips to increase reversals, escalations, or information reveals. This strategy strengthens short-drama rhythm while preserving the story structure and core.

2.2 Visual Assets and Prompt Generation

Scene-level Visual Assets. We expand the structured script into scene-level visual assets for subsequent keyframe generation and video rendering. Specifically, for each scene, we generate a 360° panorama from the scene description, spatial anchors, and the initial character–prop layout. This panorama serves as an environment reference for viewpoint selection and for maintaining cross-clip spatial consistency. We construct scene-level character assets, which are used to produce multi-view character references based on the wardrobe descriptions. The optional personalization operation from user-provided portraits is detailed in Appendix Q. These spatial and character assets are jointly used later in the first-frame prompting, keyframe review, and video generation.

Keyframe-Video Prompt for Clip Generation. Given the clip-level script and the scene-level visual assets, we construct a paired keyframe-video prompt for each clip. The keyframe prompt specifies the static first frame, including character composition, spatial relations, key prop placement, and camera viewpoint. The video prompt describes the temporal development from that starting frame, including character actions, interactions, prop changes, and local narrative progression.

To improve prompt executability before rendering, we introduce a prompt-level reviewer loop. The reviewer checks spatial consistency, physical plausibility, and cross-clip continuity, and further verifies prop continuity across adjacent clips. When violations are detected, the system first outputs the issue list, root-cause analysis, and targeted revision suggestions, and then rewrites the corresponding keyframe or video prompt. In this way, many spatial, physical, and continuity errors can be corrected at the text level before the first frame and video generation.

2.3 Keyframe-to-Video Generation with 3D Priors

Current clip-based video generation pipelines (Wu et al., 2025) often synthesize each clip as an independent storyboard shot, or reuse the previous clip's tail frame as the next initial frame. This easily leads to scene drift and struggles to adapt to moving views. To address these issues, we introduce consistent first-frame synthesis via 3D scene grounding.

Scene Anchor Initialization. For each scene, we first generate a person-free 360° panorama P and reconstruct a scene-level 3D world $\mathcal{W}$ using Marble (World Labs, 2026). Since P covers the complete scene, we can sample multiple candidate views from the canonical space. Given candidate view parameters $v_1, \ldots, v_K$, we obtain empty background candidates $B_k = \Pi(P; v_k), \quad k = 1, \ldots, K$, where Π denotes panorama-to-perspective projection. For each background B_k, an image generation model (Google DeepMind, 2026) synthesizes a character-conditioned first-frame candidate I_k using the background and the scene-level character references. A vision-language model (Qwen Team, 2026a) then selects the view that best supports character placement while preserving the scene layout. The selected pair is denoted as $(B^\star, I^\star)$.

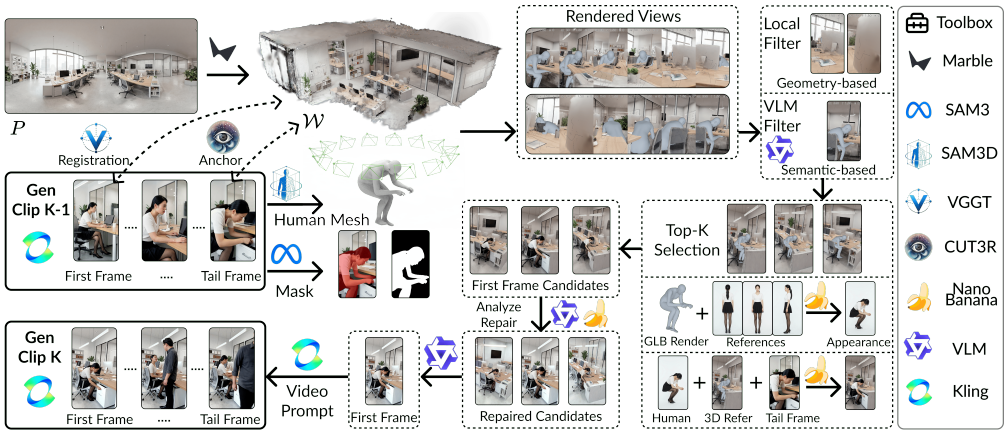


Figure 2: Consistent first-frame synthesis via 3D scene grounding. We reconstruct a scene-level 3D world from a panorama, register generated clips and the human mesh into the shared coordinate system, and synthesize the next first-frame through geometry-semantic-aware camera selection, character-conditioned generation, and frame quality analysis and repair.

First-Frame Registration, Video Trajectory Anchoring, and Human Alignment. Although $I^\star$ is generated from the selected background $B^\star$, character insertion may cause small viewpoint or focal-length shifts. We therefore register $I^\star$ back to the 3D world $\mathcal{W}$. Since $B^\star$ is cropped from the panorama, its pose $T_{B^\star}$ and intrinsics are known. After masking the character region, we use VGGT (Wang et al., 2025a) to estimate the relative transform $\Delta T_{I^\star \to B^\star}$ and initialize $T_{I^\star} = T_{B^\star} \Delta T_{I^\star \to B^\star}$. We resolve the scale ambiguity by aligning the VGGT depth of $B^\star$ with the depth rendered from $\mathcal{W}$, and further refine rotation, translation, and focal length on the background.

After generating the clip from $I^\star$, we anchor its camera trajectory to the same world. We sample video frames and use CUT3R (Wang et al., 2025b) to recover a local trajectory, depth, and intrinsics. Each frame is expressed relative to the first frame and anchored by $T_t = T_{I^\star} \Delta T_t$, where ΔT_t is the scale-calibrated relative pose. We further refine the tail-frame pose T_{tail} by aligning background-only regions in a local temporal window using color, edge, and depth consistency. Next, the character is aligned to the shared coordinate system. From the tail frame, SAM 3D Body (Yang et al., 2026b) reconstructs a human mesh with corresponding 2D/3D body keypoints. We register the generated mesh to the tail frame based on the keypoints and the person mask from SAM3 (Carion et al., 2026).

Geometry-Adaptive 3D Character Registration. Registering the reconstructed character mesh into the scene-level 3D world $\mathcal{W}$ is challenged by an inherent scale–depth ambiguity: the human mesh is recovered from a single frame and thus has an unknown coordinate transformation. The character mask and 2D keypoints constrain its image-space position and orientation, but not its distance to the camera—a person twice as large and twice as far away yields the same mask. We therefore select the registration strategy according to the geometric cues available in the shot. (i) **Occlusion matching**: when the character is occluded by a scene object, moving the body along the viewing ray changes the visible silhouette; we localize it at the depth whose rendered silhouette best matches the person mask. (ii) **Ground-depth reading**: when the character is unoccluded but its feet are visible and in contact with the ground, the ground depth beneath the feet, read from $\mathcal{W}$, directly determines the character depth, without requiring a standing pose. (iii) **Floor contact**: for standing characters, we place the body's lowest point on the floor of $\mathcal{W}$ to determine its position. This registers the human, tail-frame camera, and 3D scene in a common coordinate system for next-clip planning.

Next-Shot Consistent First-Frame Generation. The pipeline for next-first-frame and next-clip generation is shown in Fig. 2. Given the tail-frame pose and aligned human model, we sample geometrically feasible cameras for the next clip on a local spherical shell by varying azimuth, radius multiplier, and elevation. For each camera, we render the background from $\mathcal{W}$, the human mesh, and their rough composite. The candidates are filtered in two stages. The geometric filter (local filter) removes views that are too close to scene surfaces, heavily occluded in face visibility, or lack sufficient valid background. The semantic filter (VLM filter) uses a VLM to verify whether scene anchors from the next-clip prompt are visible in the rendered background.

For each of the top-K selected cameras, we generate a view-conditioned appearance image. The mesh render provides pose, silhouette, and viewpoint constraints, while multi-view character references preserve identity, clothing, and appearance. We then synthesize the next clip's first frame from the rendered reference image with human mesh, generated appearance image, and previous tail frame, so that the layout follows the 3D geometry while identity and cross-shot continuity are preserved.

Finally, a VLM checks the generated first frame for background blur, warped boundaries, missing details, and brightness or color-temperature mismatch. An image generation model repairs the background while preserving the human, camera view, and scene layout, followed by conservative color correction. The resulting first frame is then passed to the frame and video review loop. For clips with multiple characters, we additionally use the nearest previous frame in which other characters are visible to reconstruct their 3D models, and use the center of all involved characters as the camera target; details are provided in Appendix E.

2.4 Post-Production and Assembly

Diverse Scene Transitions. Unlike pipelines (ByteDance, 2026; Wu et al., 2025; Mu et al., 2026; Zhang et al., 2025) that simply concatenate independently generated scenes, ours explicitly designs transitions between adjacent scenes. As shown on the left of Fig. 1.D, the transition type is selected according to temporal shift, spatial shift, and character movement. If two scenes are continuous in both time and space, we use a direct cut to preserve action immediacy. If the location is unchanged but time advances, we generate a temporal transition with a short text overlay. If the story moves to a substantially different location, we use a location-establishing shot to clarify the upcoming time and place. If the transition involves local spatial movement with narrative meaning, we generate a motion-bridge transition. This space-time-aware planning improves scene-to-scene continuity and viewing smoothness without adding unnecessary narrative burden.

BGM Planning and Mixing. Since raw audio from video generators may contain artifacts, mismatched music, or inconsistent ambience, we introduce scene-level BGM for emotional continuity. As shown on the right of Fig. 1.D, we organize a short-drama BGM library of $8{,}122$ tracks into 8 functional categories, such as dialogue bed, suspense, and conflict escalation, and 40 fine-grained subcategory buckets using provider metadata including genre, instrument, and speed. For each scene, an LLM selects primary and backup buckets from the scene overview, clip descriptions, clip-level BGM moods, and bucket descriptions. GPT-Audio (OpenAI, 2026) then scores candidate segments by emotional, narrative, rhythm, and transition fit, and selects the best segment as the scene BGM. We mix the selected BGM with generated scene audio using adaptive volume control, including dialogue-aware base volume adjustment, LUFS-based loudness calibration, and speech-preserving dynamic compression. This maintains scene-level musical coherence while preserving dialogue clarity. More details are provided in Appendix D.2.

3 Experiments

3.1 Script Corpus and BGM Library Setup

To strengthen narrative planning, we build a structured short-drama database from $1{,}102$ original short-drama scripts, which are distilled into $16{,}311$ beat cards and segmented into $4{,}750$ logic chunks. We also build a BGM library with $8{,}122$ tracks, covering 8 high-level categories and 40 fine-grained subcategories. More details about the corpus and library can be seen in Appendix N.

3.2 Experimental Settings

Short-Drama-Bench. Our proposed Short-Drama-Bench consists of 50 story prompts spanning 7 popular categories, such as Time Travel & Rebirth, Suspense & Thriller, and Sweet Romance, which include 17 fine-grained subcategories. Each subcategory contains 2–3 representative samples, covering a broad range of short-drama patterns. More details can be seen in Appendix M.

Evaluation Benchmarks and Metrics. To comprehensively evaluate the generated short drama, we utilize three groups of metrics across different benchmarks. First, standard video-generation metrics from VBench (Huang et al., 2023) are adopted to measure low-level video quality. We also evaluate

story visualization ability based on ViStoryBench (Zhuang et al., 2026), but adapt its protocol from image-based assessment to multi-frame video evaluation, better fitting our task setting.

Second, we introduce short-drama-specific metrics for the requirements of our task. For *Narrative Hook*, *Opening Hook* is scored on the first 10 seconds of the drama and *End Hook* on the ending of each scene. For *Narrative Flow*, *Escalation Effect* is scored on the middle portion of each scene to measure conflict escalation, and *Narrative Coherence* on the complete drama to measure logic clarity.

For *Continuity*, we evaluate adjacent clips by sampling three frames from the previous clip's tail and another three from the next clip's beginning. For *Audio & Transition*, we evaluate BGM Emotion Alignment at the scene level and transition naturalness. For model-based evaluation, we choose Gemini 3.1 Pro (Google DeepMind, 2026), Qwen3.5-Omni-Plus (Qwen Team, 2026b), and Seed 2.0 Pro (ByteDance Seed Team, 2026).

Third, to complement model-based evaluation, we conduct a human study with 20 annotators. For each method, we sample the same evaluation units as used in the model-based metrics. All samples are anonymized and presented in randomized order to reduce method-specific bias. Annotators rate each sample on a 5-point Likert scale according to the short-drama criterion. We report the score averaged over all annotators. More details can be seen in Appendix F.

Baselines. We compare our method with two groups of baselines. The first group includes five story-visualization and long-form video generation pipelines: ScriptAgent (Mu et al., 2026), MovieAgent (Wu et al., 2025), StoryMem (Zhang et al., 2025), LongLive (Yang et al., 2026a), and HoloCine (Meng et al., 2026). Since these methods are not originally designed to take a single-sentence short-drama idea as input, we use Claude Opus 4.6 (Anthropic, 2026) to expand each prompt into the required format of each baseline for a fair comparison. The second group includes two short-drama products, Toonflow (HBAI Ltd, 2026) and Xiao Yun Que (ByteDance, 2026). We use their built-in LLM-based script expansion interface to convert each single-sentence prompt into an executable production script. All baselines are evaluated on the same 50 prompts in *Short-Drama-Bench*. Closed-source commercial systems are marked with † in the tables. More details can be found in Appendix G.

3.3 Qualitative Analysis

As shown at the top of Fig. 3, we compare representative outputs in terms of spatial continuity and short-drama pace. The left part compares the tail frame of clip $N-1$ with the first frame of clip N. Baselines show noticeable cross-clip drift in both character positions and background layouts. For example, in Xiao Yun Que (ByteDance, 2026), the boss and employee interact across an office partition in the previous tail frame, but move to a corridor-like space in the next first frame. Our method better preserves the scene layout and character blocking by our 3D-grounded first-frame generation mechanism significantly.

The right part of Fig. 3 compares generated scripts. Baselines often produce weak openings or scene endings without creating sufficient curiosity. In contrast, our multi-agent debate module strengthens opening conflicts and ending hooks that motivate continued viewing. The examples also show the role of our multi-stage reviewer loops: while commercial platforms such as Toonflow (HBAI Ltd, 2026) and Xiao Yun Que (ByteDance, 2026) often require manual involvement, our script, image, and video reviewers automatically analyze such errors and trigger targeted revisions. These results illustrate how our framework jointly improves spatial consistency, narrative pacing, and production-level quality control. Moreover, we show a drama's complete script, assets, and clip structure in the bottom part of Fig. 3 for better illustration.

3.4 Quantitative Analysis

As shown in Table 1, our method achieves strong performance across standard video metrics, story-visualization metrics, and the proposed short-drama-specific metrics. On VBench (Huang et al., 2023) and ViStoryBench (Zhuang et al., 2026), our method improves both general visual quality and story-level consistency, indicating that the proposed framework does not trade off low-level video quality for higher-level narrative control.

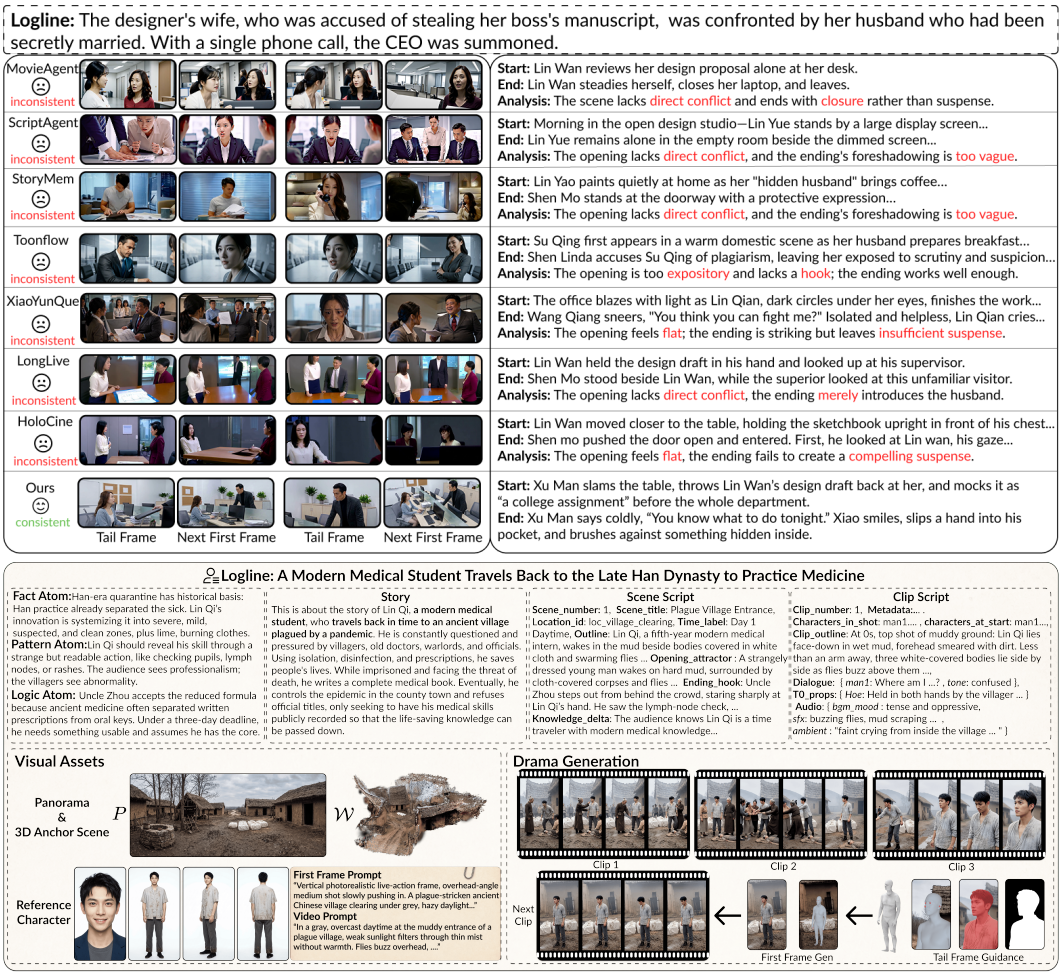


Figure 3: Qualitative examples. Top: comparison between our generated results and baselines on cross-clip visual continuity and drama pacing. Bottom: visualization of our complete drama generation process, including atom retrieval, story structurization, scene/clip script synthesis, visual asset generation, prompt synthesis, and cross-clip 3D-consistent generation.

On the short-drama-specific metrics, long-form story-visualization baselines such as MovieAgent (Wu et al., 2025), ScriptAgent (Mu et al., 2026), LongLive (Yang et al., 2026a), HoloCine (Meng et al., 2026) and StoryMem (Zhang et al., 2025) generally perform worse on *Opening Hook* and *End Hook*, since they are not explicitly optimized for the compressed pacing and frequent suspense points required by short dramas. Their continuity scores also expose different failure modes. MovieAgent and HoloCine generate multiple clips largely from independent textual descriptions, making cross-clip spatial relations difficult to preserve. ScriptAgent reuses the previous tail frame as the next clip's first frame, which helps preserve local appearance continuity but limits viewpoint changes. StoryMem introduces memory mechanisms to preserve character and scene information across clips, while LongLive uses KV caching and frame-level attention sinks to improve visual consistency in long videos. However, these methods primarily focus on visual consistency and continuous generation, without explicitly addressing short-drama pacing, opening hooks, or end-of-scene suspense. Moreover, the videos generated by StoryMem, LongLive, and HoloCine are silent, leading to a relatively low *BGM Emotion Alignment* score. Short-drama production platforms such as Toonflow (HBAI Ltd, 2026) and Xiao Yun Que (ByteDance, 2026) achieve relatively better narrative and viewing-experience scores, but their visual generation is typically conditioned on a limited set of reference images or visual prompts, leading to missing viewpoints, inconsistent scene layouts, and unstable cross-clip character blocking. In addition, their BGM mainly relies on the native audio generation capability of the underlying video model, e.g., Seedance 2.0 (Seedance et al., 2026), without explicit scene-level music planning or transition optimization.

Table 1: Quantitative evaluations. **Top-left:** comparison on standard video and story-visualization benchmarks. **Bottom-left:** comparison on our proposed Short-Drama-Bench metrics, covering narrative hooks, narrative flow, cross-clip continuity, and audio-transition quality. **Right:** human rating on the same short-drama criteria, averaged over 20 annotators across the benchmark. The radar plot normalizes each axis by the score of our method to show superior performance over baselines.

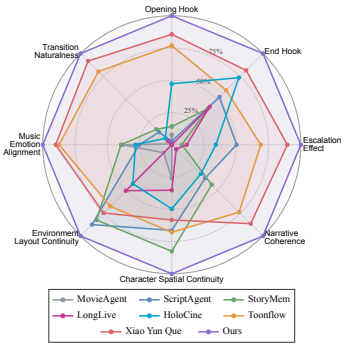
General Video and Story Metrics

Method	VBench			ViStoryBench			
	Subject Consist. ↑	Background Consist. ↑	Motion Smooth. ↑	Scene Score ↑	Cross Char. Consist. ↑	Global Char. Action ↑	Single Char. Action ↑
MovieAgent (Wu et al., 2025)	0.8672	0.8876	0.9084	2.5751	0.1543	2.7883	1.6812
ScriptAgent (Mu et al., 2026)	0.9031	0.9092	0.9235	2.6422	0.1926	2.3243	1.6789
StoryMem (Zhang et al., 2025)	0.8002	0.8719	0.9085	3.0333	0.1753	2.4167	1.6847
LongLive (Yang et al., 2026a)	0.9624	0.9358	0.9711	2.9421	0.1884	2.0922	1.9788
HoloCine (Meng et al., 2026)	0.9086	0.9195	0.9741	3.4033	0.2072	2.5367	2.6478
Toonflow (HBAI Ltd, 2026)	0.9197	0.9303	0.9462	3.1451	0.3405	2.2180	2.6258
Xiao Yun Que (ByteDance, 2026)†	0.9216	0.8697	0.9455	3.2527	0.4539	2.4204	2.8928
Ours	**0.9723**	**0.9465**	**0.9830**	**3.5334**	**0.4606**	**3.3376**	**3.0833**

Short-Drama-Bench

Method	Narrative Hook		Narrative Flow		Continuity		Audio & Transition	
	Opening Hook ↑	End Hook ↑	Escalation Effect. ↑	Narrative Coherence ↑	Character Spatial Cont. ↑	Environment Layout Cont. ↑	BGM Emotion Alignment ↑	Transition Naturalness ↑
MovieAgent (Wu et al., 2025)	2.34	1.62	1.89	1.26	2.15	1.98	2.42	1.94
ScriptAgent (Mu et al., 2026)	1.80	2.90	2.98	2.90	2.89	3.80	2.09	2.26
StoryMem (Zhang et al., 2025)	2.28	2.72	2.06	3.08	3.14	3.54	1.46	2.23
LongLive (Yang et al., 2026a)	1.94	2.67	2.10	2.07	2.39	2.90	1.06	1.83
HoloCine (Meng et al., 2026)	3.03	3.28	2.64	2.80	2.64	2.72	1.70	2.00
Toonflow (HBAI Ltd, 2026)	3.73	3.01	3.41	3.86	2.92	3.21	3.55	3.44
Xiao Yun Que (ByteDance, 2026) †	3.85	3.68	3.84	4.22	2.75	3.39	3.58	3.66
Ours	**4.26**	**3.75**	**4.06**	**4.62**	**3.52**	**4.05**	**3.86**	**3.85**

Human Rating



3.5 Ablation Study

As shown in Table 2, we conduct ablations by removing four key components from our full framework. Removing Story Gen mainly degrades narrative-related metrics, including opening hooks, ending hooks, and escalation, showing that multi-agent debate is critical for short-drama pacing and suspense construction. Removing 3D First-Frame causes the largest drop in continuity metrics, while leaving narrative scores relatively stable, confirming that 3D grounding primarily addresses cross-clip spatial drift. Removing Multi-Stage Review leads to consistent degradation across all metrics, indicating that iterative feedback is necessary to correct accumulated errors across script, prompt, keyframe, and video stages. Removing Transition & BGM mainly hurts BGM Emotion Alignment and transition naturalness, while clip-level continuity remains nearly unchanged because it is evaluated from adjacent generated clip frames before post-production assembly. These results suggest that each component targets a distinct failure mode and jointly contributes to the full system's overall performance. Isolation ablations of the Idea Bank and the geometry-adaptive registration are reported in Appendix L.

Table 2: Ablation study on Short-Drama-Bench. We remove four components from the full system: story generation, 3D first-frame synthesis, multi-stage review, and transition/BGM planning.

Variant	Narrative Hook		Narrative Flow		Continuity		Audio & Transition	
	Opening Hook ↑	End Hook ↑	Escalation Effect. ↑	Narrative Coherence ↑	Character Spatial Cont. ↑	Environment Layout Cont. ↑	BGM Emotion Alignment ↑	Transition Naturalness ↑
w/o Story Gen	3.48	3.02	3.31	3.82	3.34	3.84	3.70	3.67
w/o 3D First-Frame	4.26	3.60	3.92	4.51	2.81	3.02	3.85	3.77
w/o Multi-Stage Review	3.86	3.32	3.66	4.28	3.04	3.48	3.47	3.42
w/o Transition & BGM	4.22	3.72	4.02	4.49	3.52	4.05	3.15	3.03
Full (Ours)	**4.26**	**3.75**	**4.06**	**4.62**	**3.52**	**4.05**	**3.86**	**3.85**

4 Conclusion

We presented *One Sentence, One Drama*, a hierarchical multi-agent framework for generating complete personalized short dramas from a single-sentence idea. Our framework addresses three central challenges in this setting: short-drama pacing, cross-clip spatial consistency, and production-level quality control. It combines retrieval-augmented multi-agent story generation, 3D-grounded first-frame synthesis, multi-stage reviewer loops, and scene-level transition and BGM planning. We further introduced *Short-Drama-Bench*, a benchmark with short-drama-specific evaluation criteria. Experiments across automatic metrics, adapted story-visualization evaluation, and human ratings show that our method improves narrative engagement, visual continuity, and overall viewing experience over existing pipelines. These results suggest that structured agentic generation is a promising direction for controllable long-horizon video creation.

AI USE STATEMENT

We used generative AI tools to assist in implementing parts of the pipeline code and to polish the English writing (grammar, spelling, and word choice). We did not use generative AI tools to develop the method, design experiments, or create figures. All AI-assisted code was verified and tested by the authors, all edited text was reviewed by the authors, and we take full responsibility for the content of this work.

ETHICS STATEMENT

The script corpus used for retrieval is released under the MIT License. The BGM library contains only royalty-free or commercially usable tracks (Appendix I). As with any video generation system, our framework could be misused to produce harmful or misleading content; the safeguards implemented in the released system are described in Appendix A.

REPRODUCIBILITY STATEMENT

The full implementation, prompts, reviewer schemas, and evaluation scripts are provided in the supplementary material and at `https://anonymous.4open.science/r/one_sentence_one_drama_ICLR/`. Benchmark prompts, model versions, baseline settings, and hardware are documented in Appendix M, Appendix G, and Appendix J.

REFERENCES

Anthropic. Claude Opus 4.6 System Card. `https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf`, February 2026. System card.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

ByteDance. Xiao yun que ai agent. `https://xyq.jianying.com`, 2026. Closed-source commercial product built on Seedance 2.0. Accessed: 2026-04-22.

ByteDance Seed Team. Seed2.0 Model Card: Towards Intelligence Frontier for Real-World Complexity. `https://seed.bytedance.com/en/seed2`, 2026. Model card.

Gengchen Cao, Tianke He, Yixuan Liu, and RAY LC. Audience in the loop: Viewer feedback-driven content creation in micro-drama production on social media. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026. URL `https://arxiv.org/abs/2511.16719`.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model, 2025. URL `https://arxiv.org/abs/2504.13074`.

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

Dramaland. Dramaland short drama creator service platform. https://www.dramaland.com/, 2026. Accessed: 2026-05-06. Public platform quotation for Hongguo short-drama production tiers: A-level 2000 CNY/min, S-level 3000 CNY/min, and S+-level 5000 CNY/min.

Mohamed Elmoghany, Liangbing Zhao, Xiaoqian Shen, Subhojyoti Mukherjee, Yang Zhou, Gang Wu, Viet Dac Lai, Seunghyun Yoon, Ryan Rossi, Abdullah Rashwan, Puneet Mathur, Varun Manjunatha, Daksh Dangi, Chien Nguyen, Nedim Lipka, Trung Bui, Krishna Kumar Singh, Ruiyi Zhang, Xiaolei Huang, Jaemin Cho, Yu Wang, Namyong Park, Zhengzhong Tu, Hongjie Chen, Hoda Eldardiry, Nesreen Ahmed, Thien Nguyen, Dinesh Manocha, Mohamed Elhoseiny, and Franck Dernoncourt. Infinitystory: Unlimited video generation with world consistency and character-aware shot transitions, 2026. URL https://arxiv.org/abs/2603.03646.

Google AI for Developers. Gemini 3 pro image preview. https://ai.google.dev/gemini-api/docs/models/gemini-3-pro-image-preview, 2026. Accessed: 2026-04-24.

Google DeepMind. Veo 3 technical report. https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf, 2025. Technical report.

Google DeepMind. Gemini 3.1 pro preview. https://ai.google.dev/gemini-api/docs/models, 2026. Model gemini-3.1-pro-preview, accessed March 2026.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling, 2025a. URL https://arxiv.org/abs/2512.15702.

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

HBAI Ltd. Toonflow. https://github.com/HBAI-Ltd/Toonflow-app, 2026. Open-source project under AGPL-3.0 license. Accessed: 2026-04-22.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

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

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

Kuaishou Technology. Kling ai. https://klingai.com, 2024. Accessed: 2026-04-22.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion, 2026. URL https://arxiv.org/abs/2602.07775.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023.

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

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

Chenyu Mu, Xin He, Qu Yang, Wanshun Chen, Jiadi Yao, Huang Liu, Zihao Yi, Bo Zhao, Xingyu Chen, Ruotian Ma, Fanghua Ye, Erkun Yang, Cheng Deng, Zhaopeng Tu, Xiaolong Li, and Linus. The script is all you need: An agentic framework for long-horizon dialogue-to-cinematic video generation, 2026. URL https://arxiv.org/abs/2601.17737.

OpenAI. GPT-Audio API Documentation, 2026. URL https://platform.openai.com/docs/models/gpt-audio. Accessed: 2026-04-30.

Shiya Pi. Intensifying competition in the short-drama market poses challenges for long-form video platforms. Sina Finance, March 2025. URL https://finance.sina.com.cn/roll/2025-03-06/doc-inensrzt1029804.shtml. Accessed: 2026-05-06. The article reports that common short-drama production costs are about 10,000 CNY per minute.

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

Qwen Team. Qwen3.5-omni technical report, 2026b. URL https://arxiv.org/abs/2604.15804.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You

Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025b.

World Labs. Marble: A multimodal world model. https://www.worldlabs.ai/blog/marble-world-model, 2026. Accessed: 2026-04-24.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, YINGCONG CHEN, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 90037–90059, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/91a1610c6ed9e02d33f826b46f472b92-Paper-Conference.pdf.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery, 2026b. URL https://arxiv.org/abs/2602.15989.

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, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 83048–83077, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ce31378e9f41d8907e97dab172b6c559-Paper-Conference.pdf.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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*, 2025.

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

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention, 2025. URL https://arxiv.org/abs/2412.02259.

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

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

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, Xuanyang Zhang, Xianfang Zeng, Zhewei Huang, Gang Yu, and Chi Zhang. Vistorybench: Comprehensive benchmark suite for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9455–9467, June 2026.

APPENDIX OVERVIEW

This appendix provides additional details for the related work, generation pipeline, benchmark construction, evaluation protocol, implementation settings, prompts, and responsible-use discussion.

- **Appendix A: Broader Impacts.** Potential positive impacts on creative access and production cost, as well as copyright and licensing concerns.
- **Appendix B: Related Work.** Extended discussion of video generation, story visualization, and short-drama generation.
- **Appendix C: Multi-Agent Debating-Based Story Generation.** Details of atom script corpus construction, problem-driven retrieval, story drafting, and debate-based polishing.
- **Appendix D: Diverse Transition Clips and BGM Planning.** Details of scene transition design, BGM subcategory buckets selection, audio scoring, and adaptive mixing.
- **Appendix E: Multi-Character 3D-Consistent First-Frame Generation.** Extension of the 3D-grounded first-frame pipeline to multi-character clips.
- **Appendix F: Human Rating Detail.** Human rating protocol, anonymization, randomization, and score aggregation.
- **Appendix G: Detailed Experimental Settings.** Hardware settings, baseline execution environment, retry policy, and 3D candidate selection.
- **Appendix H: Time and Cost Analysis.** Wall-clock runtime estimates and API cost analysis.
- **Appendix I: Limitations.** Practical limitations including cost, human-in-the-loop interaction, and audio licensing.
- **Appendix J: Multi-Stage Review Metrics and Judge Models.** Internal reviewer metrics, judge models, decision rules, and external benchmark evaluation settings.
- **Appendix K: Reliability and Sensitivity Analysis.** Human–LLM evaluation agreement, robustness to LLM judge choice, statistical confidence of performance improvements, and human-evaluation reliability.
- **Appendix L: Isolation Ablations of the Idea Bank and Geometry-Adaptive Registration.**
- **Appendix M: Short-Drama-Bench Prompts and Generated Videos.** Benchmark categories, prompt topics, and representative generated video examples.
- **Appendix N: Script Library and BGM Library Details.** Statistics and construction details of the script retrieval corpus and BGM library.
- **Appendix O: Prompt Templates for Each Stage.** Prompt templates for evaluation, text review, image review, and video review.
- **Appendix P: Structures and Examples of Narrative Memory Units.** Exact structures and representative examples of BeatCards, Pattern Atoms, Logic Chunks, and Logic Atoms.
- **Appendix Q: Personalization from user-provided portraits.**
- **Supplementary Material Files.** The submitted supplementary package contains the code and a demo short-drama video showcase, including one Chinese-dubbed and one English-dubbed generated video.

A Broader Impacts

Our framework may broaden access to short-drama creation by reducing the gap between a human creative idea and a complete audio-visual production. By turning a single-sentence concept into scripts, visual assets, coherent video clips, transitions, and BGM, the system can lower production cost and technical barriers for independent creators, educators, small studios, and users without professional filmmaking resources. It may also support faster prototyping of narrative ideas, multilingual short-drama production, and more diverse forms of personalized storytelling.

Safety measures in the released system. The released implementation already contains three safeguards, all of which are verifiable in the code.

(i) *Script-level content constraints.* The script generation prompt (`step1_story_init_full_script.py`) explicitly restricts sensitive visual content: when death, injury, violence, blood, remains, or burning are involved, only their narrative function is kept and expressed in a restrained, non-gory manner suitable for image and video generation; elaboration into detailed horror imagery is prohibited.

(ii) *Dialogue-level term replacement.* During dialogue polishing (`step1_1_clip_dialogue_role_polish.py`), a built-in list of about 30 high-risk terms (e.g., corpse, gore, bloody, rotting, and their Chinese counterparts) triggers a local replacement patch by the polishing LLM. Only the flagged words are rewritten while the plot is locked—dialogue structure, events, conflict intensity, and prop positions remain unchanged—and the patched text is required to contain no remaining risk terms; for example, "a rotting corpse" becomes "the outline of the deceased under a white sheet".

(iii) *Generator-side filtering.* All images and videos are produced by commercial generators with built-in content filtering (Gemini 3.1 Pro Image Preview (Google AI for Developers, 2026) and Kling (Kuaishou Technology, 2024)). These measures complement the multi-stage reviewer loops, which can be extended with explicit safety criteria without changing the architecture.

B Related Work

B.1 Video Generation

Recent video foundation models have rapidly advanced text- and image-to-video generation in visual fidelity, motion realism, and prompt adherence. Representative works include closed-source systems such as Sora (Brooks et al., 2024), Veo (Google DeepMind, 2025), Kling (Kuaishou Technology, 2024), and Seedance (Seedance et al., 2026), as well as open-source counterparts such as CogVideoX (Yang et al., 2025), HunyuanVideo (Kong et al., 2025), and Wan (Wan et al., 2025). However, these models are typically limited to 5–15 seconds per clip. To go beyond this limit, autoregressive long-video methods (Yin et al., 2025; Huang et al., 2025; Cui et al., 2025; Zhu et al., 2026; Guo et al., 2025a; Zhao et al., 2026; Li et al., 2026; Yang et al., 2026a; Chen et al., 2025) roll out clips sequentially with teacher- or self-forcing, scaling generation to the 1-minute level. Nevertheless, both families do not model multi-shot pacing, cross-scene spatial layout, or production-level coherence, and even 1-minute level outputs fall short of the multi-minute, multi-scene requirement of short dramas. They thus serve as per-shot rendering backends in our framework, but cannot by themselves ensure long-horizon planning or cross-clip consistency.

B.2 Story Visualization

To extend single-clip models towards narrative video, prior work explores LLM-guided planning, memory-conditioned generation, and multi-agent collaboration. Early efforts such as VideoDirectorGPT (Lin et al., 2023), VideoStudio (Long et al., 2024), and StoryDiffusion (Zhou et al., 2024) use layout planning or shared self-attention to improve cross-scene consistency, while VideoGen-of-Thought (Zheng et al., 2025), StoryAgent (Hu et al., 2024), MovieAgent (Wu et al., 2025), and ScriptAgent (Mu et al., 2026) adopt multi-agent or chain-of-thought decomposition to organize scripts, storyboards, and shots. StoryMem (Zhang et al., 2025) further reformulates multi-shot generation as iterative synthesis conditioned on a visual memory bank. Despite these advances, these systems are primarily designed for general storytelling or movie-style narratives rather than short dramas. Most of them require carefully curated inputs. For example, MovieAgent (Wu et al., 2025) needs a full

script and a character bank with reference portraits (Wu et al., 2025), and StoryMem (Zhang et al., 2025) expects detailed per-shot prompts (Zhang et al., 2025). Many also rely on local fine-tuning of the image or video modules (Wu et al., 2025; Zhang et al., 2025; Hu et al., 2024), which shifts much of the creative burden to the user. Their shots are typically produced as loosely coupled five-second clips (Wu et al., 2025; Mu et al., 2026; Hu et al., 2024), with audio either absent (Zhang et al., 2025) or added post-hoc, leading to visible disconnection between adjacent shots. Moreover, these systems do not explicitly model dense hooks, frequent reversals, and compressed payoff structures, and thus tend to produce unsatisfactory pacing.

B.3 Short-Drama Generation

A few recent systems specifically target short-drama production. Toonflow (HBAI Ltd, 2026) is an open-source workflow that converts a full novel into a short drama through sequential character extraction, script generation, storyboard drawing, and video synthesis, while Xiaoyunque (ByteDance, 2026) is a closed-source commercial product built on Seedance 2.0 (Seedance et al., 2026). Despite their popularity, both systems share several limitations. On the script side, Toonflow (HBAI Ltd, 2026) requires a complete novel as input, while Xiaoyunque (ByteDance, 2026) appears to rely on one-shot LLM expansion, leading to weak hooks and brittle narrative logic. On the visual side, keyframes are generated independently from a few reference images, causing spatial drift and inconsistent character placement across clips. They also depend on manual inspection for quality control, and neither models scene-level audio or transitions, typically reusing the video model's built-in audio and concatenating scenes via hard cuts. In contrast, our framework addresses these issues through retrieval-augmented multi-agent story generation, 3D-grounded synthesis, multi-stage reviewer loops, and scene-level BGM matching with space-time-aware transition planning.

C Details of the Multi-Agent Debating-based Story Generation Framework

C.1 Atom Script Corpus Construction

As shown in Fig. 4 (Retrieval Bank Construction), directly expanding a full short-drama script from a single logline often suffers from two complementary failure modes: weak short-drama pacing, such as unconvincing openings and underpowered ending hooks, and unstable local causal coherence, where character actions are under-motivated, evidence becomes effective at unclear moments, and scene-to-scene consequences fail to connect smoothly. To address these issues, we construct two complementary retrieval banks from a corpus of 1,102 high-performing short-drama scripts released under the MIT License. First, we distill each source script into a structured script card containing script-level metadata, plot summaries, and further decompose these cards into 16,311 reusable beat-level units. Each beat unit encodes key structural cues such as the opening action, beat summary, and closing hook visual, and is mapped into an embedding space to support retrieval of transferable short-drama patterns. This forms our Pattern Bank, which provides reusable pacing and packaging priors for new stories. Second, we segment the original scripts into overlapping local text chunks to preserve short-range causal context across boundaries. These chunks constitute a Logic Bank that supports retrieval of local narrative evidence, including motivation chains, evidence activation conditions, and consequence transitions.

C.2 Problem-Driven Retrieval

As shown in Fig. 4, directly expanding a complete short-drama script from a single logline often lacks the external support needed for coherent and compelling story planning. To address this, we first use an LLM to expand the input logline into a seed text containing a preliminary narrative skeleton and key conflict cues. Based on this seed text, the LLM further analyzes what kinds of support are missing and generates a structured retrieval plan with three complementary retrieval routes. First, for externally grounded content such as professional knowledge, historical details, legal constraints, and institutional procedures, we invoke web search to retrieve factual evidence. Second, for local narrative validity, including character motivation, evidence activation conditions, scene-to-scene consequence chaining, and knowledge-state transitions, we retrieve the top-k relevant local chunks from the Logic Bank as causal support. Third, for short-drama-specific dramatic packaging, including

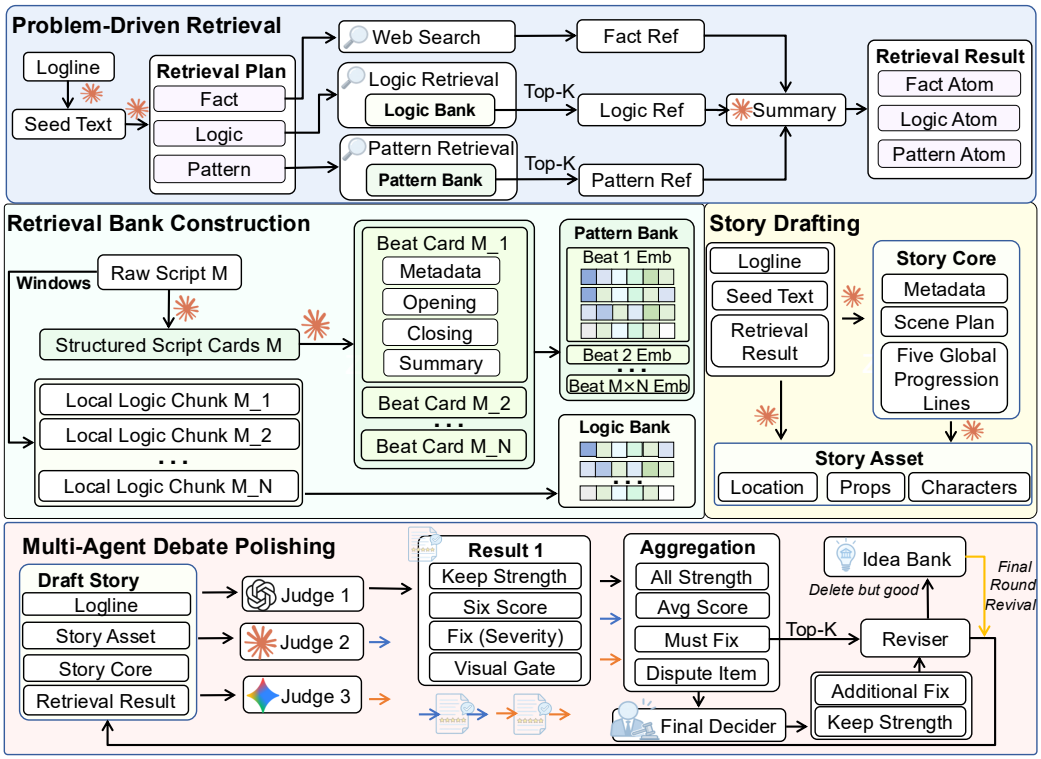


Figure 4: The Multi-Agent Debating-based Story Generation Framework.

opening design, conflict presentation, reversal pacing, and ending hooks, we retrieve the top-k most relevant beat-level cards from the Pattern Bank by computing similarity over multiple beat views and aggregating them with weighted ranking. Finally, we feed all retrieved evidence into a summarization module, which compresses the raw retrieval outputs into reusable structured units, namely Fact Atoms, Logic Atoms, and Pattern Atoms. This process provides complementary factual, causal, and pacing priors for downstream story planning, while also avoiding direct copying of source scripts by transforming retrieved content into abstract, transferable units.

C.3 Story Drafting

After obtaining the logline, the seed text, and the retrieved factual, logical, and pattern-level support, we perform story drafting in two stages. We first construct a story core, which specifies both story-level metadata and a structured scene plan for the entire drama. Concretely, the story core defines the title, theme, genre, and overall narrative framing, and, for each scene, predicts the scene title, spatiotemporal boundary, outline, opening attractor, key progression steps, scene goal, escalation beats, and ending hook. To maintain global consistency beyond isolated scene planning, we introduce five cross-scene progression lines: the external pressure line, the protagonist response line, the resolution mechanism line, the emotional progression line, and the knowledge-state line. The first four lines organize the escalation of external constraints, the evolution of the protagonist's strategy and resources, the gradual setup of the eventual resolution, and the trajectory of emotional tension, respectively. The knowledge-state line records, after each scene, what the audience and the in-story characters know, what remains hidden, and which new evidence or state changes have been introduced, thereby improving information control and scene-to-scene coherence. We then derive story assets from the resulting story core, including character assets, location assets, and prop assets. These assets provide stable identity and appearance descriptions for major characters, reusable spatial descriptions and visual attributes for core locations, and functional as well as symbolic descriptions for key props.

C.4 Multi-Agent Debate Polishing

In the Multi-Agent Debate Polishing stage, we submit the drafted story to three independent frontier LLM judges for parallel review. Given the same input, each judge returns a structured evaluation that includes keep strengths, six rubric scores, must-fix issues with severity levels, and a visual executability gate. The six scoring dimensions are logical integrity, opening strength, hook continuity, narrative clarity, reversal pacing, and payoff resolution. For each must-fix issue, the judge further specifies the supporting evidence, the recommended fix direction, and the target object that should be revised. The visual executability gate provides an additional non-score signal indicating whether key turning points can be reliably grounded in downstream scene expansion and video generation.

We then perform deterministic aggregation over the three reviews to merge, deduplicate, and summarize the outputs into a unified set of retained strengths, average rubric scores, candidate must-fix issues, and disputed items. Disputed items arise when judges exhibit large score discrepancies on the same dimension, assign substantially different severities to the same issue, propose conflicting fix directions, or expose high-risk signals such as failed visual-executability checks or critically low logical-integrity scores. These disputed items are further routed to a Final Decider, implemented with GPT-5.4 Pro, which selectively determines whether a disputed issue should be fixed, what minimal-change principle should be followed, and which strengths must be explicitly protected. We then pass the top-k aggregated must-fix issues together with the decider's rulings to a Reviser. Rather than regenerating the entire draft, the Reviser performs patch-based local revision: it outputs structured patches that replace only the targeted scene plans, or, when necessary, the global cross-scene progression lines. For each revised scene, the patch rewrites the scene outline together with its dependent fields, including the opening attractor, key progression steps, scene goal, escalation beats, ending hook, and knowledge-state update, thereby enforcing minimal but coherent modifications while preserving unaffected parts of the story.

During revision, any strong idea, hook, or memorable dramatic design that is removed or softened in order to improve logic, clarity, or continuity is explicitly recorded in an Idea Bank. After each revision round, the updated draft re-enters the same multi-judge review and aggregation loop. The process terminates once the draft satisfies predefined quality thresholds, or after at most 3 rounds. Finally, we perform a final-round revival step, which revisits the Idea Bank and selectively restores a small number of previously removed ideas when they can be reintroduced without harming the current logical integrity or executability. This final step mitigates over-correction during iterative polishing and helps recover strong hooks, payoffs, and memorable dramatic moments.

D Details of Diverse Transition Clips and BGM Planning & Mixing

D.1 Diverse Transition through Scenes

Unlike conventional long-video generation pipelines that simply concatenate independently generated scenes, we explicitly model the transitional relationship between adjacent scenes. Hard cuts are easy to implement, but they often introduce two problems in long-form narratives: visually abrupt pacing, which weakens the viewing experience, and ambiguous temporal or spatial context, which makes it difficult for viewers to infer when and where the next scene takes place. As shown in Fig. 5, we introduce diverse transition clips that are selected according to the temporal shift, spatial shift, and character movement between consecutive scenes. When two scenes are continuous in both time and space, we use a direct cut to preserve the immediacy of the action. When the location remains largely unchanged but time advances substantially, we generate a temporal transition, such as a time-lapse exterior shot of the office building, with a short text overlay indicating the elapsed time. When the story moves to a substantially different location, we generate a location-establishing transition, using an exterior or entrance shot of the next location together with a text overlay that clarifies the upcoming time and place. When the transition involves only a local spatial change and the character movement itself carries narrative information, we generate a motion-bridge transition, such as a character walking through a corridor or moving toward an elevator, to visually connect the two scenes. This space-time-aware transition planning improves scene-to-scene continuity, interpretability, and viewing smoothness without adding unnecessary narrative burden.

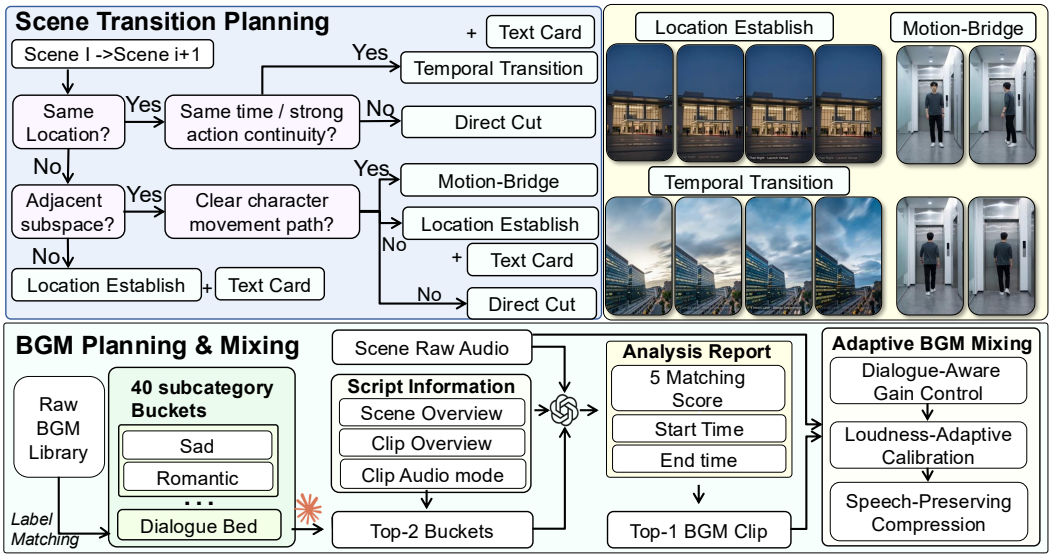


Figure 5: Our Diverse Transition Clips and BGM Planning & Mixing

D.2 BGM Planning & Mixing

Built-in audio from video generation models often contains artifacts, mismatched music, or inconsistent background sound across clips. Since each scene in our pipeline consists of multiple generated clips, we add a scene-level BGM track to improve emotional consistency and reduce perceptual discontinuities. As shown in Fig. 5, we first construct a short-drama-oriented BGM library organized into 8 functional categories and 40 fine-grained subcategories. Candidate tracks are assigned to these subcategories using provider-side metadata, including genre, vartag, instrument, and speed. For each scene, we use the scene overview, clip descriptions, clip-level BGM moods, and subcategory bucket descriptions to let an LLM select the most suitable primary and backup BGM buckets. We then call GPT-Audio to evaluate full candidate tracks from the selected subcategories buckets. Given the scene's original audio and each candidate BGM, GPT-Audio predicts a scene-length segment and scores it by emotional fit, narrative fit, rhythm fit, and transition fit. The highest-scoring track segment is selected as the BGM for the entire scene.

Finally, we mix the selected BGM with the generated scene audio using adaptive volume control. We first lower the BGM base volume for dialogue-dense scenes, then calibrate the BGM level using the LUFS gap between the scene audio and BGM segment. We further apply speech-preserving dynamic compression so that BGM is reduced during dialogue-heavy regions and remains stronger in non-dialogue regions. This produces coherent scene-level music while preserving dialogue clarity.

E Multi-Character 3D-Consistent First-Frame Generation

The main 3D-grounded first-frame generation pipeline in Section 2.3 describes the single-character case. For multi-character clips, most steps remain unchanged: we still use the scene-level 3D world, video trajectory anchoring, geometry-aware camera sampling, character-conditioned first-frame generation, and frame-level review. The main difference is that we must first place all required characters into the same 3D world before sampling the next first-frame camera.

Multi-Character Registration. Given two adjacent clips, we identify the set of characters that appear in the current clip and are still required in the next clip. The tail frame usually contains the primary character, which is registered to the 3D world using the procedure described in the main text. For other characters that are not clearly visible in the tail frame, we scan the current clip backward until finding a frame where the character is sufficiently visible and separable from the background. For each selected frame, SAM 3D Body (Yang et al., 2026b) reconstructs a character mesh and body keypoints, while SAM3 (Carion et al., 2026) provides the corresponding person mask.

To localize this character in the scene-level 3D world, we recover the camera pose of the selected frame using CUT3R (Wang et al., 2025b) on the video prefix ending at that frame, and anchor it to the same world coordinate system as in the single-character case. We then initialize the character transform from the reconstructed body keypoints and refine its depth along the camera–character ray. Specifically, we translate the mesh forward and backward along this ray and render it from the selected-frame camera. The final position is chosen such that the visible rendered silhouette best matches the SAM3 person mask, while preserving the 2D keypoint alignment. This refinement reduces depth ambiguity and gives a more reliable estimate of each character's position in the shared 3D scene. After this step, all involved characters are represented as 3D meshes registered in the same world coordinate system.

Multi-Character Camera Sampling. Once all required characters are placed in the 3D world, we modify the next-shot camera sampling strategy. Instead of centering the spherical sampling region around a single character, we compute the center of all involved characters and use it as the camera target. Candidate cameras are then sampled with different radii, azimuths, and elevations around this multi-character center. This encourages the next first frame to keep the required characters inside the same view while preserving their relative spatial arrangement.

For each candidate camera, we render the scene background and all registered character meshes. The local geometric filter rejects views where any required character falls outside the image, becomes too small, is severely occluded, or has insufficient visible body/face area. It also removes cameras that are too close to scene surfaces or contain too little valid background. The remaining candidates are passed to the semantic VLM filter, which checks whether the rendered view supports the next-clip prompt, including scene anchors, character interaction direction, and relative blocking. The top-ranked views are then used for character-conditioned first-frame synthesis.

First-Frame Synthesis and Review. For each selected view, the rendered multi-character meshes provide pose, scale, and spatial-layout constraints, while multi-view character references preserve identity and clothing. The image generation model synthesizes the next first frame conditioned on the rendered scene, the character references, and the previous clip context. Finally, the frame reviewer checks whether all required characters appear with correct identities, whether their relative positions remain consistent with the previous clip, and whether the background agrees with the scene-level 3D world. Frames that fail these checks are repaired or resampled. This extension allows the system to handle multi-character interactions while maintaining spatial continuity across adjacent clips.

F Human Rating Details

We conduct a human rating study to complement model-based evaluation. We recruit 20 annotators. All participants receive monetary compensation at a rate consistent with local wage standards. Before rating, participants are given written instructions describing the evaluation criteria, the 5-point Likert scale, and the meaning of each score, where 1 indicates very poor quality and 5 indicates excellent quality with respect to the target criterion. All evaluated samples are anonymized: method names are removed, and participants are not informed which system generated each sample. For each metric, we use the same evaluation units as in the model-based protocol, including the opening segment of the drama for opening hook, scene-ending segments for end hook, scene-level segments for escalation effect, the complete drama for narrative coherence, adjacent clip boundaries for spatial and layout continuity, scene-level audio for BGM Emotion Alignment, and scene-boundary segments for transition naturalness. Samples from different methods are randomly shuffled for each participant to reduce ordering bias and method-specific bias. Each participant rates the assigned samples independently according to the corresponding short-drama criterion. For aggregation, we first average each participant's scores over all evaluation units belonging to the same method and metric, such as multiple clips, scene boundaries, or scene-level segments. We then average these per-participant scores across the 20 participants to obtain the final human rating for each method and metric. The final human rating results are reported in Table 1.

G Detailed Experiment Settings

All experiments are conducted with API-based image, video, language, and audio generation modules, together with local 3D and vision inference modules. For our method, the local components can be

run on a single NVIDIA RTX A6000 GPU with 48GB memory. The main GPU memory requirement comes from CUT3R, which is used for video trajectory estimation and frame pose anchoring. When a 48GB GPU is unavailable, these local 3D modules can also be executed on CPU, but with substantially slower runtime. VGGT, SAM 3D Body, SAM3, and rendering are also executed locally.

For baseline evaluation, most baselines can be run on the same 48GB A6000 setup. LongLive, HoloCine, and StoryMem are run on an H200 GPU due to their higher memory and runtime requirements. All baselines are evaluated on the same 50 prompts from Short-Drama-Bench using the evaluation protocol described in the main paper.

We use the same retry policy across generation stages. For text review, first-frame review, and generated video review, each failed item can be revised or regenerated at most three times. If the sample still fails after the maximum retry count, we keep the best available candidate according to the corresponding reviewer score. For 3D-consistent first-frame generation, we keep the top-8 candidate camera views after geometric and semantic filtering, and select the final first frame from these candidates using the reviewer score described in Appendix J.

Baseline selection and adoption. We compare against the most widely adopted public short-drama and story-visualization systems. ToonFlow is the most popular open-source short-drama workflow with 15.3k GitHub stars, followed by StoryMem (767), ScriptAgent (357), MovieAgent (358), LongLive (2.6k), and HoloCine (703); Xiao Yun Que is a leading commercial short-drama product with approximately 1.22M monthly visits. GitHub statistics are as of September 2026, and monthly visits are as of December 2025.

ToonFlow (HBAI Ltd, 2026): official release v1.1.6, default configuration. Xiao Yun Que (ByteDance, 2026): official public interface with the recommended configuration, accessed in March 2026. LongLive (Yang et al., 2026a), HoloCine (Meng et al., 2026), and StoryMem (Zhang et al., 2025): Claude Opus 4.6 (Anthropic, 2026) converts the single sentence into its required structured input, giving them more information than our method receives. MovieAgent (Wu et al., 2025) and ScriptAgent (Mu et al., 2026): internal expansion LLMs replaced by Claude Opus 4.6.

H Time and Cost Analysis

Time Analysis. We report approximate wall-clock generation time for producing one complete 10 min short drama under our evaluation setting. The runtime depends on API latency, queueing time, video duration, and the number of reviewer-triggered retries, so the numbers should be interpreted as practical estimates rather than fixed constants.

For our method, story generation and multi-agent script refinement take about 10–15 minutes. Image generation is performed through external APIs and can be parallelized across scenes and clips, so the first-frame, panorama, and visual-asset generation stage takes about 2–4 minutes in practice. Scene-level 3D world construction takes about 2–4 minutes per world. The dominant cost in wall-clock time is video generation, which takes nearly one hour for a typical short drama under our API setting. Overall, our pipeline usually takes about 74–90 minutes to produce a complete short drama.

We also compare the practical runtime of representative baselines under the same benchmark prompts. Xiao Yun Que (ByteDance, 2026) typically takes about 1.5–2 hours per drama, while Toonflow (HBAI Ltd, 2026) takes about 2–3 hours. ScriptAgent (Mu et al., 2026) relies on API-based video generation but has a longer sequential pipeline, taking about 4 hours. HoloCine (Meng et al., 2026) requires about 9 hours per drama. LongLive takes about 1 hour per drama. MovieAgent (Wu et al., 2025) and StoryMem (Zhang et al., 2025) require heavier local generation and take about 35 hours in our evaluation setting. These comparisons show that our framework improves generation quality while maintaining practical runtime, mainly because image and scene-level asset generation can be parallelized and local 3D inference is only a small fraction of the full pipeline time.

Cost Analysis. We estimate the API cost for generating a one-minute short drama under the 1080P setting. For our method, the main API costs come from video generation, image generation, text/model review, and 3D world construction. Using Kling v3 Pro image-to-video at \$0.168/s, 60 seconds of video generation costs \$10.08. Image generation uses about 30 generated images, including first frames, panoramas, backgrounds, and repair images, costing about \$4.02 at \$0.134 per image. Text generation and review cost about \$2.0 per minute, and one World Labs Marble world

costs about $1.2. Without reviewer-triggered regeneration, the base cost is therefore about $17.3/min. In practice, around half of the text, image, and video generation budget is spent on reviewer-triggered retries or repair, leading to an average cost of about $25–$27/min.

Compared with existing short-drama platforms, this cost is slightly higher but remains in a similar range. In our evaluation setting, Xiao Yun Que costs about $24.36/min, while Toonflow costs about $21.53/min, including approximately $15.51 for video generation, $4.02 for image generation, and $2.0 for text generation. The additional cost of our method mainly comes from multi-stage review, regeneration after failed review, and 3D world construction. These costs improve quality and cross-clip consistency, but they remain an important target for future optimization.

The cost remains much lower than professional short-drama production. The public Dramaland quotation for Hongguo short dramas corresponds to about $293/min for A-level productions, $439/min for S-level productions, and $732/min for S+ productions (Dramaland, 2026). For live-action productions with human actors, industry reports indicate a typical production cost of about $1,464/min (Pi, 2025). In comparison, our estimated API cost of about $25–$27/min is higher than existing automated short-drama platforms, but remains substantially lower than professional short-drama production while providing stronger controllability, reviewer-based refinement, and cross-clip spatial consistency.

I LIMITATIONS

Our framework has several limitations. First, the improved controllability and production quality come with higher generation cost. Because our system includes multi-stage generation, 3D world construction, automatic review, and retry mechanisms, its estimated API cost is higher than some existing short-drama platforms. For a one-minute generated drama, our average API cost is about $25-$27 per minute, while Xiao Yun Que costs about $24.36 per minute under our evaluation setting. Although our system achieves better generation quality and Xiao Yun Que is a closed-source commercial platform whose internal pipeline and true operating cost are not fully observable, reducing cost remains important for large-scale deployment.

Second, the current framework emphasizes automatic generation and has limited human-in-the-loop interaction. Future systems could expose reviewer scores and diagnostic feedback to users through an interactive production interface. For example, clips with low reviewer scores could be automatically regenerated, clips with high scores could be accepted directly, and borderline cases could be routed to human creators for selecting whether and how to revise them. Such a hybrid workflow may reduce unnecessary retries while preserving production-level quality control.

Third, audio licensing is an important practical constraint for short-drama production. To reduce copyright risk, our current BGM library mainly contains royalty-free or commercially usable music, which limits the diversity of available styles and emotional expressions. A future system could integrate a larger licensed music library and provide users with explicit purchase or licensing options when a matched track is selected, improving audio quality while satisfying commercial publishing requirements.

J MULTI-STAGE REVIEW METRICS AND JUDGE MODELS

Our system uses reviewer models in two settings. Internal reviewers are used during generation to trigger rewriting, regeneration, or candidate selection. External judges are used only after generation for benchmark evaluation. We summarize the metrics, judge models, and decision rules below.

J.1 TEXT-LEVEL REVIEW

Table 3 summarizes the text-level reviewers used before visual generation. These reviewers check story quality, prompt executability, prop continuity, and whether the next clip requires additional scene information for 3D-consistent first-frame generation.

Table 3: Text-level reviewer configurations.

Stage	Configurations
Story bible / planning review	**Role:** Evaluates the story core before clip expansion, focusing on narrative logic, short-drama pacing, hook design, and visual executability. **Judge models:** gpt-5.4-2026-03-05, claude-sonnet-4-6, gemini-3.1-pro-preview; Final Decider: gpt-5.4-pro-2026-03-05. **Metrics:** logic_integrity, open_grab, hook_continuity, clarity, reversal_pacing, payoff_resolution, each scored 0–10. **Gate:** visual_executability_gate in {pass, borderline, fail}. **Rule:** Low key scores, strong judge disagreement, or visual failure triggers arbiter-guided revision. Key thresholds: logic, open_grab, and clarity ≥ 7.
Clip spatial text audit	**Role:** Checks whether the clip-level text prompt is spatially executable before image or video generation. **Judge model:** claude-opus-4-6. **Metrics:** spatial_consistency, physics_plausibility, cross_clip_continuity, overall, each scored 0–10. **Issue types:** POSITION, GAZE, ENTRY_EXIT, PROP, CONTINUITY, OVERCROWDING. **Rule:** If pass=false, the system rewrites the corresponding prompt, state, or prop description.
Prop continuity audit	**Role:** Verifies whether key props have valid sources, movements, ownership changes, and end states across adjacent clips. **Judge model:** claude-opus-4-6. **Metrics:** prop_source_continuity, prop_motion_plausibility, overall, each scored 0–10. **Rule:** If pass=false, the system applies a minimal prompt rewrite before generation.
Next-clip scene-information audit	**Role:** Determines from the next clip's textual description whether the next first frame requires additional scene information beyond the previous tail frame. **Judge model:** claude-opus-4-6. **Outputs:** needs_extra_scene_information, has_large_character_or_camera_reposition, required_scene_anchors. **Rule:** If needs_extra_scene_information=true, the system does not simply reuse the previous tail frame and instead invokes the 3D-consistent first-frame generation path.

J.2 First-Frame and Tail-Frame Image Review

Table 4 summarizes the image-level reviewers used for first-frame selection and tail-frame routing. These reviewers select 3D-consistent first-frame candidates and determine whether a previous tail frame contains sufficient visual context for direct reuse.

J.3 Video-Level Review

Table 5 summarizes the video-level reviewers used after clip generation. They evaluate visual physics, temporal continuity, reaction plausibility, and whether character entrances, exits, and presence states match the clip script.

Table 4: First-frame and tail-frame image reviewer configurations.

Stage	Configurations
3D-consistent first-frame candidate selection	**Role:** Selects the best image candidate among 3D-consistent rendered views based on image quality, character integrity, and spatial continuity. **Judge model:** Qwen/Qwen3.5-397B-A17B. **Metrics:** temporal_continuity, layout_consistency, background_quality, person_scene_interaction, character_integrity, color_continuity. **Score range:** each metric is scored 0–5; total score is 0–30. **Rule:** Any metric below 3 rejects the candidate; among accepted candidates, the highest total score is selected.
Tail-frame close-up detector	**Role:** Checks whether the previous clip's tail frame is too local or context-poor to directly serve as the next first frame. **Judge model:** Qwen/Qwen3.5-397B-A17B. **Outputs:** is_local_closeup, shot_scale, visible_environment, confidence in 0–1. **Rule:** If the tail frame is a local close-up or contains insufficient visible environment, it is not directly reused as the next first frame.

J.4 Audio and BGM Review

Table 6 summarizes the audio reviewers used for scene-level BGM planning and selection. The system first selects suitable BGM subcategory buckets from textual scene context, then scores candidate audio segments to choose the final soundtrack.

J.5 External Benchmark Evaluation

Table 7 summarizes the judge models and metrics used only for final evaluation. These benchmark judges are separate from the internal reviewer loops and are applied to the generated videos from the same 50 Short-Drama-Bench prompts.

Table 5: Generated video reviewer configurations.

Stage	Configurations
Generated video spatial audit	**Role:** Checks whether the generated video is physically plausible and visually continuous across time, using sampled video frames rather than audio. **Judge model:** Qwen/Qwen3.5-397B-A17B, used as a visual/video reviewer over sampled frames. **Metrics:** physics_integrity, temporal_continuity, reaction_plausibility, each scored 0–10. **Overall:** local overall is the mean of the above three metrics. **Rule:** Pass iff overall ≥ 5 and character presence consistency passes; otherwise the system rewrites the prompt or regenerates the video.
Character presence consistency	**Role:** Verifies whether character presence, entrance, exit, and on-screen status follow the clip script. This metric does not evaluate identity similarity or action quality. **Judge model:** Qwen/Qwen3.5-397B-A17B, used as a visual/video reviewer over sampled frames. **Inputs:** characters_at_start, characters_at_end, and ordered character_events from the clip script. **Checks:** whether the characters visible at the video beginning match characters_at_start; whether the characters visible at the video ending match characters_at_end; whether scripted entrances and exits occur in the correct order; whether any character appears without narrative support; whether a character remains visible after a scripted exit; and whether an unintended character dominates the frame. **Metric:** character_presence_consistency, binary score 0 or 10. **Rule:** The score must be 10; otherwise the video fails and is sent to prompt rewriting or regeneration.

Table 6: BGM selection and audio reviewer configurations.

Stage	Configurations
BGM bucket selection	**Role:** Selects the most suitable functional BGM categories for each scene according to its narrative content and emotional trajectory. **Judge model:** claude-opus-4-6. **Inputs:** scene overview, clip descriptions, clip-level BGM moods, and subcategory bucket descriptions. **Outputs:** primary and backup BGM subcategory buckets. **Rule:** Candidate tracks are retrieved from the selected subcategory buckets before audio-level scoring.
BGM segment scoring	**Role:** Scores candidate BGM segments and selects the best segment/window for the scene-level soundtrack. **Judge model:** gpt-audio-2025-08-28. **Metrics:** emotion_fit, narrative_fit, rhythm_fit, transition_fit, each scored 0–10. **Rule:** the highest-scoring segment is selected.

Table 7: External benchmark evaluation settings.

Benchmark	Configurations
Short-Drama-Bench	**Role:** Evaluates short-drama-specific quality on the 50 benchmark prompts and their generated videos. **Judge models:** gemini-3.1-pro-preview, qwen3.5-omni-plus-2026-03-15, seed-2-0-pro-260328. **Metrics:** opening_hook, end_hook, escalation_effect, narrative_coherence, character_spatial_continuity, environment_layout_continuity, bgm_emotion_alignment, transition_naturalness. **Score range:** each metric is scored 1–5. **Aggregation:** scores are averaged across judges, prompts, and evaluation units.
VBench	**Role:** Evaluates general video quality on the generated videos for the same 50 benchmark prompts. **Evaluator:** VBench automatic metric modules. **Metrics:** subject_consistency, background_consistency, motion_smoothness. **Score range:** scores follow the native VBench normalization.
ViStoryBench adaptation	**Role:** Evaluates story-visualization quality on the generated videos for the same 50 benchmark prompts. **Judge model:** gpt-5.4-2026-03-05. **Metrics:** scene alignment, cross-character consistency, global character action, and single-character action. **Protocol:** we sample video frames at 1 FPS and apply the image-based ViStoryBench story-visualization criteria to the sampled frames. **Aggregation:** scores are averaged across prompts, scenes, and sampled frames.

K Reliability and Sensitivity Analysis

This section evaluates the reliability of our evaluation protocol and examines whether the main conclusions are sensitive to evaluator choice. We analyze four aspects: agreement between human and LLM-based evaluation, robustness to the choice of LLM judge, confidence intervals for the improvement over the strongest baseline, and uncertainty and inter-rater reliability in the human study.

K.1 Human–LLM Evaluation Agreement

We first examine whether the LLM-based evaluation reflects human judgments. For each of the 50 prompts and each of the eight Short-Drama-Bench metrics, we average the ratings from 20 raters and the reference scores from three LLM judges separately for each of the eight methods. We then compute Pearson and Spearman correlations across the eight paired method scores within each prompt.

As shown in Table 8, the median prompt-level Pearson correlations range from 0.878 to 0.960, while the median Spearman correlations range from 0.848 to 0.949 across the eight metrics. Across all 3,200 method–prompt–metric pairs, the LLM-judge reference scores are approximately 0.38 points higher than the human ratings on average.

Table 8: Prompt-level correlations between human ratings and LLM-judge reference scores across the eight Short-Drama-Bench metrics. For each prompt and metric, correlations are computed across eight methods.

Evaluation metric	**Pearson** r	**Spearman** ρ
Opening Hook	0.960	0.929
End Hook	0.909	0.892
Escalation Effect	0.950	0.931
Narrative Coherence	0.955	0.949
Character Spatial Continuity	0.878	0.848
Environment Layout Continuity	0.921	0.884
BGM Emotion Alignment	0.936	0.881
Transition Naturalness	0.958	0.883

K.2 Robustness to LLM Judge Choice

We next examine whether the benchmark conclusion depends on the choice of a particular LLM judge. Using the same generated dramas, we score all eight methods with three judge models. As shown in Table 9, the judges assign different absolute scores but agree on the top three methods: Ours, Xiao Yun Que, ToonFlow.

Among the three judge models used in the main evaluation—Gemini, Qwen, and Seed—the median pairwise item-level Spearman correlation is 0.823, with a range of 0.800–0.839. Krippendorff's α among these three judges is 0.804. These results indicate that the ranking of the leading methods is robust to the choice of LLM judge, despite differences in score calibration.

Table 9: Average Short-Drama-Bench scores assigned by three LLM judges. All three judges rank ours first, followed by Xiao Yun Que and ToonFlow.

Judge model	**MovieAgent**	**ScriptAgent**	**StoryMem**	**LongLive**	**HoloCine**	**ToonFlow**	**Xiao Yun Que**	**Ours**
Gemini 3.1 Pro Preview	1.82	2.57	2.42	2.12	2.65	3.25	3.48	**3.84**
Qwen3.5-Omni-Plus	2.10	2.85	2.73	1.92	2.44	3.56	3.78	**4.18**
Seed 2.0 Pro	1.93	2.68	2.54	2.36	2.72	3.36	3.60	**3.97**

K.3 Statistical Confidence of Performance Improvements

We quantify the statistical uncertainty of the improvement over the strongest baseline using paired resampling. For the LLM-based evaluation, we perform paired bootstrap resampling over the 50 Short-Drama-Bench prompts with $B = 10{,}000$ bootstrap samples. For the human evaluation, the confidence interval accounts for variation across both prompts and evaluators. Xiao Yun Que is the strongest baseline in both evaluation settings.

As reported in Table 10, our method improves the average LLM-judge score by 0.375, with a 95% confidence interval of $[0.282, 0.470]$. In the independent human evaluation, our method improves the average score by 0.349, with a 95% confidence interval of $[0.191, 0.515]$. Both confidence intervals remain entirely above zero, showing that the improvements are robust to prompt and evaluator variation.

Table 10: Score improvements of our method over Xiao Yun Que, the strongest baseline in both LLM-based and human evaluations. The 95% confidence intervals are computed using paired resampling, and positive differences favor our method.

Evaluation setting	**Score difference**	**95% confidence interval**
LLM-based evaluation	**0.375**	$[0.282,\ 0.470]$
Human evaluation	**0.349**	$[0.191,\ 0.515]$

K.4 Uncertainty and Inter-Rater Reliability in Human Evaluation

The human study includes 20 evaluators and 50 prompts, resulting in 1,000 ratings for each evaluated method. The scores in Table 11 are averaged over Opening Hook, End Hook, Escalation Effect, Narrative Coherence, Character Spatial Continuity, Environment Layout Continuity, BGM Emotion Alignment, and Transition Naturalness. The confidence intervals account for variation across both evaluators and prompts.

We additionally report $\mathrm{ICC}(2, 20)$, which measures the reliability of the aggregated ratings from the 20 evaluators. The $\mathrm{ICC}(2, 20)$ values range from 0.79 to 0.86, indicating consistent aggregated human judgments. Our method achieves the highest human-evaluation mean, obtaining 3.52.

Table 11: Human-evaluation means, standard deviations, 95% confidence intervals, and inter-rater reliability for the eight evaluated methods. Each method receives 1,000 ratings from 20 evaluators over 50 prompts.

Method	**Mean**	**SD**	**95% CI**	**ICC**$(2, 20)$
MovieAgent	1.67	0.70	$[1.53, 1.81]$	0.83
ScriptAgent	2.28	0.77	$[2.12, 2.45]$	0.79
StoryMem	2.27	0.78	$[2.11, 2.45]$	0.82
LongLive	1.79	0.72	$[1.68, 1.91]$	0.85
HoloCine	2.24	0.82	$[2.11, 2.39]$	0.86
ToonFlow	2.97	0.85	$[2.79, 3.16]$	0.82
Xiao Yun Que	3.17	0.86	$[2.99, 3.35]$	0.85
Ours	**3.52**	0.79	$[3.35, 3.70]$	0.81

L Isolation Ablations of the Idea Bank and Geometry-Adaptive Registration

Here we isolate the two algorithm-level designs, keeping all other settings identical.

Idea Bank with final-round revival. We generate scripts with and without the Idea Bank under identical settings and score them with the story-bible reviewer of Table 3 on its six dimensions (0–10).

As shown in Table 12, the gains concentrate on the memorability-related dimensions (payoff $+0.2$, reversal pacing $+0.2$, opening and hook continuity $+0.1$), matching the design goal of restoring strong elements without harming logic. The over-correction it addresses is not specific to short dramas and applies to any LLM-based iterative narrative creation, such as story writing or plot planning for games.

Table 12: Isolation ablation of the Idea Bank (story-bible reviewer scores, 0–10).

Setting	Logic	Opening	Hook cont.	Clarity	Reversal	Payoff
w/o Idea Bank	8.17	8.98	8.87	8.32	8.59	8.61
w/ Idea Bank (Ours)	**8.17**	**9.07**	**8.97**	**8.37**	**8.77**	**8.82**

Geometry-adaptive character registration. The ablation baseline projects the reconstructed mesh onto the registration frame and adjusts its placement by grid search so that the projected silhouette and keypoints match the person in the image; since enlarging the mesh while moving it proportionally farther away leaves the projection unchanged, this alignment cannot determine the character's distance. Our method shares every other step and additionally solves the distance from scene geometry. We measure placement accuracy with three metrics: *scene penetration rate*, the fraction of the 70 body keypoints that fall inside scene solids such as the floor or tabletops; *foot-to-floor distance*, the distance from the foot keypoints to the floor plane, reported as a fraction of the character's height, which grows whether the character sinks or floats; and *visible-region IoU*, the overlap between the visible part of the character rendered inside the 3D world and the person region visible in the source frame, which drops when incorrect placement changes how furniture occludes the character. As shown in Table 13, our registration nearly eliminates penetration and floating and substantially improves occlusion consistency. The same procedure applies to placing a single-image human reconstruction into any virtual scene, e.g., for world models or VR/AR content.

Table 13: Isolation ablation of the geometry-adaptive registration.

Metric	Projection-only baseline	Ours
Scene penetration rate ↓	21.4%	**0.7%**
Foot-to-floor distance (% of human height) ↓	63.2%	**3.9%**
Visible-region IoU ↑	0.27	**0.67**

M Short-Drama-Bench Prompts and Generated Videos

Table 14 lists the prompt topics used in Short-Drama-Bench. The benchmark covers seven high-level short-drama genres, including underdog comeback, social realism, ancient court intrigue, suspense and thriller, time travel and rebirth, sweet romance, and corporate/business war. Each genre is further divided into fine-grained subcategories, with each subcategory containing representative story prompts. These prompts are designed to evaluate whether the generation pipeline can handle diverse narrative settings, character relationships, conflict structures, and genre-specific storytelling patterns.

Category	Subcategory	Specific Topics
Underdog Comeback	Marriage Comeback	1. My Unremarkable Husband Turns Out to Be the Company Chairman 2. After Being Abandoned at the Wedding, She Returned as an Investor Capable of Buying the Groom's Empire 3. The Day the Divorce Papers Were Signed, His Ex-Wife's Company Went Public 4. The Daughter-in-Law Kicked Out by Her Mother-in-Law Became the New Owner of Her Company Three Years Later

Category	Subcategory	Specific Topics
	Hidden Identity	1. The Humiliated Stable Boy Turns Out to Be the Long-Lost Heir to the Kingdom 2. The Security Guard Everyone in the Company Looks Down On Has Five World Leaders' Private Numbers in His Phone 3. The Transfer Student Mocked by Classmates Whose Father Is Their School's Chairman of the Board
	Career Comeback	1. The designer's wife, who was accused of stealing her boss's manuscript, was confronted by her husband who had been secretly married. With a single phone call, the CEO was summoned. 2. The Intern Publicly Humiliated by the Director Was Sitting in the Director's Chair a Year Later 3. The Designer Reduced to Tears by a Client Went On to Win an International Design Award
Social Realism	Workplace Injustice	1. The Woman Fired for Being Pregnant Returned as the Company's Biggest Client 2. The Middle Manager 'Optimized Out' at 35 Built a Startup Team from an Unemployment Group Chat 3. The Engineer Forced to Sign a Non-Compete Discovered the Boss Had Already Violated His Own
	Medical & Survival	1. The Night the Hospital Refused Her Surgery, She Livestreamed Everything 2. Her Father's Life-Saving Pill Costs 700 Yuan Each, So the Daughter Went to India to Find the Manufacturer Herself 3. In the Three Months She Was Misdiagnosed with Cancer, She Saw Everyone Around Her for Who They Really Are
	Family Ethics	1. When the Mother Who Favored Sons Over Daughters Fell Ill, Only the Neglected Daughter Came 2. The Parents Gave the House to Their Son but Left the Debt to Their Daughter 3. The Whole Family Pooled Money for the Brother to Study Abroad, but the Sister Got Into a Better School on Her Own
Ancient Court Intrigue	Harem Power Struggle	1. The Abandoned Consort in the Cold Palace Is Determined to Put the Crown Prince on the Throne 2. Sentenced to Death on Her First Day in the Palace, She Traded a Bowl of Poison for the Empress's Secret 3. She Pretended to Be Out of Favor for Three Years While Secretly Building a Shadow Guard That Answers Only to Her
	Court Conspiracy	1. The Poisoned Princess Married the Enemy Prince Only to Burn the Empire from Within 2. Everyone Believed the Chancellor Was Loyal — Only the Crown Prince Knew He Killed the Late Emperor

Category	Subcategory	Specific Topics
		3. The Exiled General's Daughter Returns with Her Father's Former Army
	Women Breaking the Rules	1. She Disguised Herself as a Man to Top the Imperial Exam, Only to Be Exposed in the Golden Hall 2. The Princess Who Knew No Martial Arts Talked Down a Hundred Thousand Rebels with Words Alone 3. The Merchant's Daughter Who Married into the General's Household Brought Down Military Corruption with Her Ledger
Suspense & Thriller	Digital-Age Thriller	1. The Serial Killer Had Been Lurking in Their Group Chat All Along 2. The Missing Intern Sent a Video from the CEO's Private Basement 3. The Social Media Post She Deleted Became the Only Lead in the Case
	Closed-Space Mystery	1. In a Snowed-In Mountain Lodge, One of the Eight Guests Is a Fugitive from Ten Years Ago 2. An Elevator Malfunction Traps Six People — One of Them Has a Bloody Knife in Their Bag 3. Halfway Through a Murder Mystery Game, Someone Realizes the Script Is Based on a Real Person in the Room
	Twist Thriller	1. She Called the Police to Report Her Husband Missing, but They Found a Clue in Her Own Car Trunk 2. Three Women Went Missing in a Row — the Person Who Filed the Report Turned Out to Be the Suspect 3. The Therapist Discovered That Her Most Dangerous Patient Is Actually Her Own Husband
Time Travel & Rebirth	Professional Time Travel	1. A Modern Medical Student Travels Back to the Late Han Dynasty to Practice Medicine 2. A Chemistry PhD Travels to an Era of Witches, Kings, and Poisons 3. A Modern Forensic Scientist Travels to Ancient Times and Overturns a Wrongful Conviction Through Autopsy
	Rebirth & Revenge	1. A Female Lawyer Is Reborn as a Queen Accused of Treason 2. She Is Reborn to the Day Before Her Murder — This Time She Gathers All the Evidence First 3. After Rebirth She Didn't Rush to Seek Revenge — She First Got Admitted to Law School
Sweet Romance	Status Gap / Contract Romance	1. She Fake-Married Her Roommate for a Visa — Then It Stopped Being Fake 2. The CEO's Blind Date Turns Out to Be the Girl He Anonymously Sponsored for Ten Years
	Reconciliation Romance	1. Five Years After Breaking Up, They Reunite in the ER — She's His Attending Physician 2. He Finally Found His First Love, but She No Longer Remembers Him

Category	Subcategory	Specific Topics
Corporate & Business War	Business Showdown	1. On Her First Day at Work, She Discovered the Company's Biggest Corporate Spy Is Her Own Mentor 2. Two Interns Both Fell for the Same Proposal — One Chose to Plagiarize, the Other Chose to Innovate 3. The Founder Kicked Out by His Partners Started Over — Taking the Core Technology with Him

Table 14: Short-Drama-Bench prompt topics.

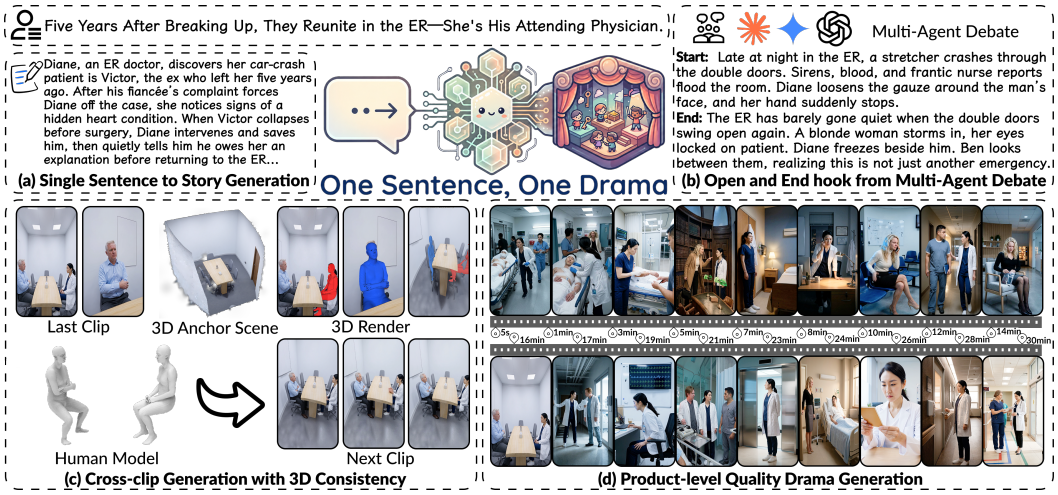


Figure 6: Representative short dramas generated by our framework from single-sentence ideas, covering different genres, durations, and character settings.

Fig. 6, Fig. 7 and Fig. 8 present representative generated video examples from Short-Drama-Bench, illustrating the visual results produced under different prompt topics and narrative settings.

N Script Library and BGM Library Details

To strengthen narrative planning, we use a short-drama database that includes 1,102 short-drama scripts released under the MIT License, which are distilled into 16,311 beat cards and 4,750 logic chunks. This structured database provides retrieval-based references for plot rhythm, conflict escalation, and genre-specific storytelling patterns, giving the generation pipeline stronger narrative logic and short-drama style priors.

As shown in Fig. 9, we build a BGM library with 8,122 tracks, covering 8 high-level categories and 40 fine-grained subcategories; each category is paired with a textual description that guides subsequent audio matching according to the scene rhythm and emotional intent.

O Prompt Templates for Each Stage

O.1 Evaluation Prompt

Fig. 10 shows the unified prompt template used for model-based Short-Drama-Bench evaluation. The same template is instantiated with different metrics, scopes, rubrics, and reference contexts to evaluate narrative quality, continuity, audio alignment, and transition naturalness.



Figure 7: Representative generated videos (Gallery 1/2).

O.2 Text Review

Fig. 11 shows the clip-level script review prompt used before visual generation. It separately evaluates the first-clip hook, last-clip ending hook, and middle-clip twist density, enabling targeted rewriting without changing unrelated parts of the scene.

O.3 Image Review

Figs. 12 and 13 show the prompt and scoring criteria used for selecting 3D-consistent first-frame candidates. The reviewer evaluates temporal continuity, coarse layout consistency, background quality, character integrity, color continuity, and person-scene interaction before video generation.



Figure 8: Representative generated videos (Gallery 2/2).

O.4 Video Review

Fig. 14 shows the prompt template used for generated video review. The reviewer checks physical realism, temporal continuity, reaction plausibility, and character presence consistency, and failed clips are sent to prompt revision or regeneration.

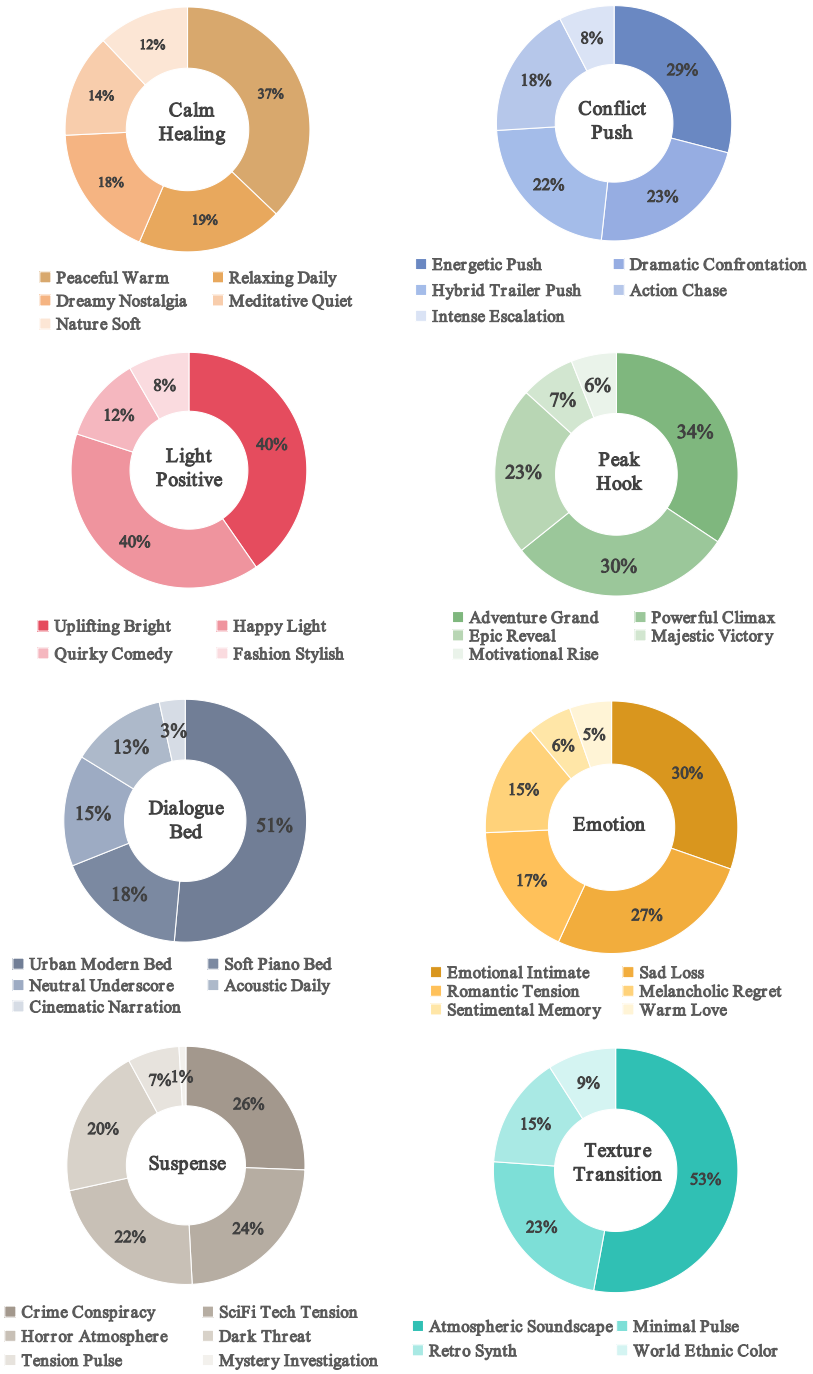


Figure 9: Our BGM Datasets

Short-Drama-Bench Evaluation Prompt Template

You are a rigorous short-drama video benchmark judge.

Metric: `{metric}`
Scope: `{scope}`
Rubric: `{rubric}`

Use a 1–5 integer score:
1 = weak / mostly ineffective
2 = partially effective but with clear issues
3 = mostly effective with minor issues
4 = excellent and clearly satisfies the metric
5 = exceptional; reserve this only for outstanding cases with strong evidence, almost no visible weakness, and especially strong short-drama impact

Reference context:
`{context}`

Return strict JSON only:
`{"score": 1, "analysis": "brief evidence-based explanation"}`

Metric-specific fields:
`opening_hook` / drama: Evaluate whether the first 10 seconds quickly attract viewers, establish stakes, and create immediate curiosity.
`end_hook` / scene: Evaluate whether the scene ending creates sufficient motivation to continue watching.
`escalation_effect` / scene: Evaluate whether the middle portion contains effective conflict escalation, rising pressure, or power reversal.
`narrative_coherence` / drama: Evaluate whether the story logic remains clear, non-confusing, and internally consistent.
`character_spatial_continuity` / clip_pair: Evaluate adjacent clip continuity, including character identity, screen position, body orientation, and spatial relationship continuity.
`environment_layout_continuity` / clip_pair: Evaluate adjacent clip continuity, including background layout, room geometry, object placement, and camera-direction consistency.
`bgm_emotion_alignment` / scene: Evaluate whether the background music emotion aligns with scene mood, tension, pacing, and dramatic intent.
`transition_naturalness` / scene_pair: Evaluate whether the previous scene tail, transition clip, and next scene head connect naturally in action, camera, mood, and story logic.

Figure 10: Prompt template for model-based Short-Drama-Bench evaluation.

Scene-Level Script Review Prompt Template

You are a reviewer for viral short-drama scripts designed for AI video production.

You will receive the full story information and all clips of a single scene in structured JSON format. Your task is to output scores, reasons, and improvement suggestions for three metrics:
1) `hook`: evaluate only the opening attraction of the first clip.
2) `scene_end`: evaluate only the ending hook of the last clip.
3) `twist`: evaluate only the reversal density of the middle clips, excluding the first and last clips.

Boundary rules:
- When scoring `hook`, do not use the last clip or `ending_hook` as supporting evidence.
- When scoring `scene_end`, do not use the opening attraction of the first clip as supporting evidence.
- When scoring `twist`, do not use the first or last clip as the main evidence.
- If the scene has fewer than 3 clips, set `twist.score` to 8 and `twist.improvements` to an empty array.

Number of improvement suggestions:
- Score 1–3: exactly 3 improvement suggestions.
- Score 4–6: exactly 2 improvement suggestions.
- Score 7: exactly 1 improvement suggestion.
- Score 8–10: exactly 0 improvement suggestions.

Scoring rubric:
`hook`: 1–3 = flat opening, weak conflict, cannot attract viewers; 4–6 = conflict exists but intensity or pacing is insufficient; 7 = acceptable and can retain viewers; 8–10 = highly attractive, with clear conflict and emotional momentum.
`scene_end`: 1–3 = no escalation or suspense; 4–6 = ending intention exists but the hook is weak; 7 = clear ending hook; 8–10 = strong escalation and clear anticipation for the next scene.
`twist`: 1–3 = little or no middle reversal, or mostly repetitive progression; 4–6 = some turns exist but density or quality is moderate; 7 = acceptable middle-clip turning points; 8–10 = high reversal density and strong reversal quality.

Additional constraints:
- Review only; do not rewrite the script.
- Improvement suggestions must be actionable and specific.
- Avoid text-dependent visual solutions, wet-body/water-stain descriptions, and ellipses in dialogue.
- All strings must be single-line strings; if a line break is needed, use `\n`.
- Return strict JSON only, without Markdown or explanation.

Output format:

```
{
  "hook": {"score": 1-10, "reason": "...", "improvements": ["..."]},
  "scene_end": {"score": 1-10, "reason": "...", "improvements": ["..."]},
  "twist": {"score": 1-10, "reason": "...", "improvements": ["..."]}
}
```

User prompt template:

```
Please output the hook/scene_end/twist review JSON for this scene:
{payload_json}
```

Figure 11: Prompt template for scene-level script review.

3D-Consistent First-Frame Candidate Selection Prompt

You are selecting the best first frame for the next video clip.

You will receive exactly three images:
Image 1 is the previous tail frame with the person.
Image 2 is `candidate_rank{rank}`'s geometry-only environment render reference.
Use Image 2 only for coarse spatial layout and large object relationships: walls, desks, cabinets, glass, doors, floor, and whether the person appears in a plausible spatial position.
Ignore Image 2's blur, noise, color, texture, material quality, lighting, and small details.
Do not evaluate or compare overall camera direction, camera angle, or view direction against Image 2.
Image 3 is `candidate_rank{rank}`, the locally refined candidate first frame to evaluate.

Evaluate only Image 3. Do not compare it with other candidates.
Use Image 1 to judge temporal continuity and character consistency.
Use Image 2 only to judge coarse background layout and large object spatial consistency.
Do not use Image 2 for color judgment. Judge color continuity only by comparing Image 3 against Image 1.

Scoring rules:
Score each criterion from 0 to 5 using integers only.
There are six criteria, and the maximum total score is 30.
Any score below 3 rejects the candidate, even if the total score is high.
Be strict, not generous. Do not give 5 unless there is nearly no visible problem. Do not give 4 if the flaw is clearly visible.
For every criterion, provide a concrete explanation in `score_explanations`. Do not leave any explanation empty.

Figure 12: Prompt template for 3D-consistent first-frame candidate selection.

3D-Consistent First-Frame Candidate Selection Criteria

General scoring scale:
5 = almost no issue; the criterion is essentially flawless.
4 = acceptable with only tiny flaws; usable without obvious disruption.
3 = noticeable flaw but barely usable; keep it only if other candidates are worse.
2 = unacceptable and must reject.
1 = severe failure and must reject.
0 = completely wrong, missing, or impossible to judge; must reject.

Criteria:
`temporal_continuity`: Judge whether Image 3 continues naturally from Image 1 in character status, pose tendency, spatial relation, and scene continuity.
`layout_consistency`: Judge whether Image 3 preserves the coarse room layout and large object relationships indicated by Image 2.
`background_quality`: Judge whether the background in Image 3 is visually usable, without severe blur, broken geometry, missing regions, or distorted large objects.
`person_scene_interaction`: Judge whether the person is physically well integrated with the scene, without penetration, floating, clipping, or impossible occlusion.
`character_integrity`: Judge whether the visible person in Image 3 has complete and plausible body structure, clothing, and silhouette, without severe identity or body artifacts.
`color_continuity`: Judge whether Image 3 is visually compatible with Image 1 in brightness, contrast, color temperature, and overall tone.

Detailed rubric for `person_scene_interaction`:
Inspect person-scene interaction part by part before scoring: hands, fingers, feet, shoes, legs, knees, back, hips, clothing edges, hair, and the body silhouette.
For each body part, check whether it penetrates desks, drawers, chairs, cabinets, walls, glass panels, floor, monitors, shelves, or other objects.
Also check impossible occlusion order, floating limbs, clipped shoes, hands merging into drawers, back or hips sinking into furniture, and dark or transparent edge artifacts around the person.
5 = all contacts and occlusions are natural; no penetration, floating, clipping, black edge, or transparent edge artifact.
4 = mostly natural, with only tiny edge or contact imperfections.
3 = noticeable but not fatal contact issue, such as slight hand-drawer merging, odd foot contact, or local black edge.
2 = visible penetration or impossible occlusion, such as hand through drawer, shoe into floor, or back/hips sinking into furniture. Reject.
1 = multiple severe penetration/contact failures. Reject.
0 = person-environment geometry is unusable. Reject.

Return strict JSON only:

```
{
  "scores": {
    "temporal_continuity": 0,
    "layout_consistency": 0,
    "background_quality": 0,
    "person_scene_interaction": 0,
    "character_integrity": 0,
    "color_continuity": 0
  },
  "total_score": 0,
  "rejected": true,
  "score_explanations": {
    "temporal_continuity": "...",
    "layout_consistency": "...",
    "background_quality": "...",
    "person_scene_interaction": "...",
    "character_integrity": "...",
    "color_continuity": "..."
  }
}
```

Figure 13: Scoring criteria and output format for 3D-consistent first-frame candidate selection.

Generated Video Review Prompt Template

Please score the generated video according to the following four complementary dimensions.

1) `physics_integrity`: physical realism.
Check whether the video contains floating objects, gravity violations, object penetration, abnormal human body structure, duplicate entities, or other physically implausible artifacts.

2) `temporal_continuity`: temporal continuity.
Check whether objects or characters flicker, disappear, teleport, abruptly change state, or miss necessary transition actions. Also check whether an object has already been handed over or moved but later reappears at its original position without explanation.

3) `reaction_plausibility`: reaction plausibility.
Check whether character facial expressions and body actions match the scene stimulus, emotional context, and narrative situation.

4) `character_presence_consistency`: character presence consistency.
Only check whether the entrance, exit, and on-screen presence of characters match the script requirements. This score must be either 0 or 10: assign 0 if any inconsistency exists, and assign 10 only if it is fully consistent.

General scoring scale for the first three dimensions:
0–2 = severe errors; clearly unusable, with major physical violations or severe temporal jumps.
3–4 = many visible issues; the video feels clearly inconsistent and requires rewriting key segments.
5–6 = basically watchable but with obvious flaws; targeted repair is needed.
7–8 = mostly stable, with only minor visible issues.
9–10 = stable and natural, with almost no visible logical or physical problems.

Additional rules for `character_presence_consistency`:
- Characters visible at the beginning of the video should match `characters_at_start`.
- If `character_events` is not empty, the video must show these character-set changes in the specified order.
- Characters visible at the end of the video should match `characters_at_end`.
- If the script says a character exits but the character remains continuously visible, `character_presence_consistency` must be 0.
- If a character suddenly appears without a scripted entrance or narrative support, `character_presence_consistency` must be 0.

Return strict JSON only:

```
{
  "physics_integrity": 0,
  "temporal_continuity": 0,
  "reaction_plausibility": 0,
  "character_presence_consistency": 0,
  "analysis": "brief evidence-based explanation"
}
```

Figure 14: Prompt template for generated video review.

P STRUCTURES AND EXAMPLES OF NARRATIVE MEMORY UNITS

This appendix provides the exact structures and representative examples of the narrative memory units introduced in Section 2.1. Our retrieval module contains two complementary offline memories. The *Pattern Bank* stores structured BeatCards that describe reusable short-drama pacing and staging patterns, whereas the *Logic Bank* stores overlapping Logic Chunks that preserve local causal context. After retrieval, the selected records are further abstracted into Pattern Atoms and Logic Atoms before being passed to the story writer.

The examples below are translated, lightly condensed, and anonymized to illustrate the data structures without reproducing source-specific dialogue or character names.

P.1 PATTERN BANK: FROM BEATCARD TO PATTERN ATOM

Each normalized script is decomposed into an ordered sequence of BeatCards. A BeatCard contains the fields `beat_id`, `beat_type`, `hook_type`, `opening_action`, `closing_hook_visual`, `beat_summary`, `characters`, `location`, and `keywords`. The cards are embedded and retrieved according to the structural needs of the seed story:

$$\text{script} \rightarrow \text{BeatCards} \rightarrow \text{retrieval} \rightarrow \text{Pattern Atom}.$$

Representative BeatCard and Its Pattern Atom

Source category: Wedding reversal

`beat_id`: 1
`beat_type`: Opening conflict
`hook_type`: Threat
`opening_action`: The groom arrives to collect the bride and finds forty cups of liquor arranged at the entrance.
`closing_hook_visual`: The bride's mother blocks the doorway with both arms and announces that no one may pass until every cup is emptied.
`beat_summary`: A ceremonial entrance is turned into a public coercion test, escalating an interpersonal obstruction into social humiliation.
`characters`: [`groom`, `bride`, `bride's mother`, `onlookers`]
`location`: Villa entrance
`keywords`: [`bride collection`, `forced drinking`, `public humiliation`]

Retrieved abstraction—Pattern Atom
Stage the obstruction as a physical blockade at a threshold that the protagonist must cross, and escalate the demand in public so that refusal carries a social cost.

The Pattern Atom removes source-specific characters, dialogue, and setting details while preserving the transferable dramatic mechanism: a physical threshold externalizes the obstruction, and public escalation increases the cost of refusal.

P.2 LOGIC BANK: FROM LOGIC CHUNK TO LOGIC ATOM

Each normalized script is segmented into source-text windows of approximately 900 characters with an overlap of 120 characters. Each Logic Chunk stores `doc_id`, `source_name`, and `text`. Unlike BeatCards, these chunks are not organized around predefined dramatic beats. They retain local narrative context for resolving specific causal questions concerning motivation, evidence activation, consequences, relationship changes, timing, and closure:

$$\text{script} \rightarrow \text{overlapping Logic Chunks} \rightarrow \text{retrieval} \rightarrow \text{Logic Atom}.$$

Representative Logic Chunk and Its Logic Atom

doc_id: [anonymized]
source_name: [anonymized source script]
text (translated and condensed): A consort lies motionless after making a life-saving sacrifice. Only at the moment of her death does the beneficiary learn that she once saved him, leaving no opportunity for gratitude or repayment.

Retrieval query: What makes a past sacrifice a credible motive for revenge?

Retrieved abstraction—Logic Atom
A past sacrifice becomes a credible motive for revenge when it is revealed too late to repay: the beneficiary learns of it at the moment of loss, so the debt can never be settled.

The Logic Atom converts the retrieved local context into a causal constraint for story drafting. In this example, the motivation does not arise from the sacrifice alone; it depends on the delayed revelation, the irreversible loss, and the impossibility of repayment. Thus, BeatCards and Pattern Atoms provide reusable pacing and staging priors, while Logic Chunks and Logic Atoms provide local causal support. Only the abstracted atoms, rather than the retrieved source records, are passed to the downstream story-generation stage.

Q Personalization from User-Provided Portraits

Our framework is personalized at two levels: the drama is generated from the user's own single-sentence idea, and the user may cast themselves or anyone they choose as the protagonists by uploading a portrait.

Concretely, the scene-level character assets in Section 2.2 accept either a generated seed portrait or a user-uploaded photo. From an uploaded portrait, we generate multi-view character references under the wardrobe descriptions from the story core, and these references are then used throughout the pipeline: conditioning first-frame synthesis, providing the appearance transferred onto the rendered mesh in the 3D-grounded branch, and serving as the character-reference elements passed to the video generator. No fine-tuning or per-user training is required, and uploaded and generated identities are treated identically, so a user can supply portraits for only some characters and let the framework generate the rest.