# DramaAgent: Agentic Storytelling Video Generation

**Ting Huang**[1*] **Biao Wu**[2*] **Ronghao Chen**[1*] **Zeyu Zhang**[1†]
**Tengfei Cheng**[3] **Qizhen Lan**[4] **Huacan Wang**[5] **Hao Tang**[1‡]
[1]Peking University [2]UTS AAII [3]UOL [4]UTHealth Houston [5]UCAS
[*]Equal contribution. [†]Project lead. [‡]Corresponding author: bjdxtanghao@gmail.com.



Figure 1: **DramaAgent for multi-scene long-horizon storytelling.** Given a narrative prompt, DramaAgent performs hierarchical story planning, maintains persistent character and story states, and coordinates heterogeneous video and audio generators through reflection-guided targeted repair, improving cross-scene coherence, identity consistency, and narrative fidelity.

## Abstract

Recent diffusion and autoregressive video generators have substantially improved short-form synthesis, yet composing multiple independently generated clips into a coherent audio-visual story remains challenging. Character identities may drift across scenes, narrative events can become inconsistent, and dialogue or sound may not align with the evolving visual context. We present **DramaAgent**, a hierarchical, agentic, and model-agnostic framework for *multi-scene long-horizon audiovisual story generation*. Rather than modifying the underlying video backbone, DramaAgent introduces a stateful control layer that combines hierarchical story planning, persistent character conditioning, scene-wise generation, and reflection-guided targeted repair. The framework maintains reusable story and character states across scenes, diagnoses failures along identity, semantic, temporal, and audio-visual dimensions, and maps each diagnosed failure to a corresponding localized repair operation. Experiments on 50 narrative scripts across five video-generation backbones show consistent improvements in semantic alignment, character consistency, narrative fidelity, and scene-level audio-visual compatibility. Controlled same-backbone comparisons further show that these gains cannot be explained solely by stronger generators or additional sampling. Our results suggest that persistent state, failure diagnosis, and targeted repair provide a practical control paradigm for coherent multi-scene audiovisual generation. Code: `https://github.com/AIGeeksGroup/DramaAgent`. Website: `https://aigeeksgroup.github.io/DramaAgent`.

## 1 Introduction

Recent diffusion and autoregressive models have substantially advanced text-to-video generation, especially for short-form visual synthesis (Ho et al., 2022b; Brooks et al., 2024; Ho et al., 2022a; Zheng et al., 2024a; Kong et al., 2024). However, generating coherent *multi-scene long-horizon story videos* by composing multiple short generated clips remains difficult. As generation spans multiple



*Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3609–3624
November 6-10, 2026 ©2026 Association for Computational Linguistics

scenes, existing methods often suffer from narrative drift, unstable character identity, weak temporal continuity, and limited audio-visual consistency (Wei et al., 2024; Zheng et al., 2024a; Zhou et al., 2022; Ho et al., 2022b; Wu et al., 2023). These challenges are particularly severe in story-centric applications such as educational animation, narrative video creation, and marketing content, where coherence must be preserved over long horizons rather than within isolated short clips (Polyak et al., 2024; Wu et al., 2025a; Zhu et al., 2023; Hu et al., 2024; Long et al., 2024).

A straightforward solution is to use stronger video generators. Yet even strong generators and commercial APIs are still primarily optimized for short clips, and their direct use for multi-scene storytelling remains brittle. One-shot prompting tends to accumulate semantic errors over time, while simple clip stitching often fails to preserve consistent characters, scene transitions, and story state. Meanwhile, fully end-to-end training for long-horizon video generation remains costly and difficult to adapt to rapidly evolving video backbone ecosystems. This motivates a complementary perspective: instead of modifying the underlying generator, can multi-scene story generation be improved through a structured control framework built above heterogeneous generators?

We propose **DramaAgent**, a hierarchical, agentic, and model-agnostic framework for multi-scene long-horizon text-to-video-and-audio generation. DramaAgent treats multi-scene storytelling as a structured control problem over heterogeneous video and audio generators, rather than as independent short-clip synthesis. It organizes generation into story planning, persistent character conditioning, scene-wise synthesis, and reflection-guided targeted repair. By transforming multi-scene generation into a staged decision process, DramaAgent maintains reusable story and character states across scenes, improving controllability, character persistence, and cross-scene coherence while remaining compatible with diverse video backbones. As illustrated in Fig. 1, DramaAgent generates coherent animated stories from a single narrative prompt by decomposing the story into controllable scene-level generation targets.

DramaAgent is built on three complementary components. First, **hierarchical story planning** converts an open-ended narrative prompt into ordered scene-level targets, where each target specifies the participating characters, environment, event, emotional state, and transition cue. Second, **persistent character conditioning** constructs reusable character references and carries them across scene generation, candidate selection, and repair so that character identity is maintained as a persistent control signal. Third, **reflection-guided targeted repair** scores generated clips along identity, semantic, temporal, and audio-visual dimensions and regenerates low-quality scenes using constraints derived from the detected failure type. These components allow DramaAgent to coordinate short-form generators under a long-horizon storytelling objective. Instead of treating each clip as an independent generation result, DramaAgent maintains story context across scenes, selects candidates with respect to both local scene requirements and previously accepted content, and repairs local inconsistencies before composing the final audiovisual story. Taken together, these components form a stateful closed-loop process that couples *persistent state, failure diagnosis, and targeted repair* rather than treating planning and regeneration as independent stages.

We evaluate DramaAgent using both automatic and human-centered protocols. Beyond standard visual quality metrics, we examine long-horizon coherence, character consistency, narrative fidelity, scene-level audio-visual consistency, and human preference. Experiments across multiple video backbones show that DramaAgent consistently improves story-level coherence, character consistency, and narrative fidelity over direct generation and strong baselines. Controlled same-backbone and generation-budget-aware comparisons further show that these gains cannot be explained solely by stronger backbones or additional sampling. Together, these results suggest that hierarchical agentic control provides a practical and model-agnostic way to improve controllable multi-scene audiovisual storytelling. Our contributions are as follows:

- We propose **DramaAgent**, a hierarchical and model-agnostic agentic framework for multi-scene long-horizon text-to-video-and-audio generation over heterogeneous backbones.
- We introduce **persistent story-character state** and **reflection-guided targeted repair**, forming a stateful closed-loop process of persistent state, failure diagnosis, and targeted repair to preserve cross-scene identity and narrative context.

- Experiments across multiple video backbones, together with controlled same-backbone and generation-budget-aware comparisons, demonstrate improved long-horizon coherence, character consistency, narrative fidelity, and scene-level audio-visual consistency.

## 2 Related Work

**Long-form and agentic story video generation.** Recent text-to-video research has moved beyond short clips toward story-level and long-form generation, where narrative coherence, character consistency, and temporal continuity across scenes become central challenges. Long-horizon generation has also been explored through hierarchical temporal modeling, such as NUWA-XL (Yin et al., 2023). Methods such as StoryGAN (Li et al., 2019), Make-A-Story (Rahman et al., 2023), Intelligent Grimm (Liu et al., 2024), StoryDiffusion (Zhou et al., 2024), AutoStory (Wang et al., 2025), MovieFactory (Zhu et al., 2023), and VideoStudio (Long et al., 2024), together with benchmarks such as MovieBench (Wu et al., 2025a), highlight the need to preserve characters, events, and transitions over extended horizons. CoIn (Tao et al., 2024a) introduces a lightweight context-aware framework with intra-story feature interchange for story visualization and continuation, while StoryImager (Tao et al., 2024b) unifies story visualization and completion through bidirectional generation and contextual modeling. Recent agentic frameworks such as StoryAgent (Hu et al., 2024), MovieAgent (Wu et al., 2025b), and Mora (Yuan et al., 2024) further improve controllability by decomposing narratives into scripts, storyboards, or scene-level prompts. DramaAgent is related to these works, but focuses on multi-scene long-horizon text-to-video-and-audio generation as a stateful hierarchical control problem over heterogeneous generators. It maintains reusable story and character states across scene generation, candidate selection, and repair, and uses reflection-guided diagnosis to trigger failure-specific regeneration along identity, semantic, temporal, and audio-visual dimensions. Together, these mechanisms couple persistent cross-scene state, failure diagnosis, and targeted repair within a closed-loop generation process.

**Character-consistent video generation.** Maintaining stable character identity is a longstanding challenge in long-form generation. Recent methods address this problem through reference-image conditioning, subject-driven generation, personalized visual prompting, and instance-level visual control (Ye et al., 2023; Gu et al., 2024). For example, Magic-Me (Ma et al., 2024) shows that external visual references can improve subject consistency in generated videos. Related story-generation systems also use storyboard images, keyframes, or visual anchors to stabilize character appearance across scenes. DramaAgent builds on these efforts by treating character references as persistent control signals, propagating fixed references across scene-wise generation, candidate selection, and reflection-guided repair to reduce identity drift over long horizons.

**General video generation backbones.** Recent video generation methods have advanced rapidly with diffusion-based, autoregressive, and hybrid architectures. Diffusion-based models such as VDM (Ho et al., 2022b), Align Your Latents (Blattmann et al., 2023b), SEINE (Chen et al., 2023), I2VGen-XL (Zhang et al., 2023), SVD (Blattmann et al., 2023a), LaVie (Wang et al., 2024), and MotionDirector (Zhao et al., 2024) improve visual realism, motion quality, and text alignment. Large-scale and sequential generation models, including Sora (Brooks et al., 2024), Open-Sora (Zheng et al., 2024b), Hunyuan-Video (Kong et al., 2024), VideoGPT (Yan et al., 2021), VideoPoet (Kondratyuk et al., 2023), and CogVideoX (Yang et al., 2025), further strengthen video synthesis and temporal modeling. While these models provide powerful short-form generation backbones, multi-scene storytelling still requires cross-scene state tracking, character persistence, narrative continuity, and audio-visual coordination. DramaAgent is complementary to this line of work by coordinating heterogeneous video and audio generators under a long-horizon storytelling objective.

## 3 Method

We propose **DramaAgent**, a hierarchical and model-agnostic framework for multi-scene long-horizon text-to-video-and-audio generation. As shown in Fig. 2, DramaAgent coordinates heterogeneous video and audio generators through story decomposition, persistent character references, scene-wise generation, and reflection-guided selection and repair.

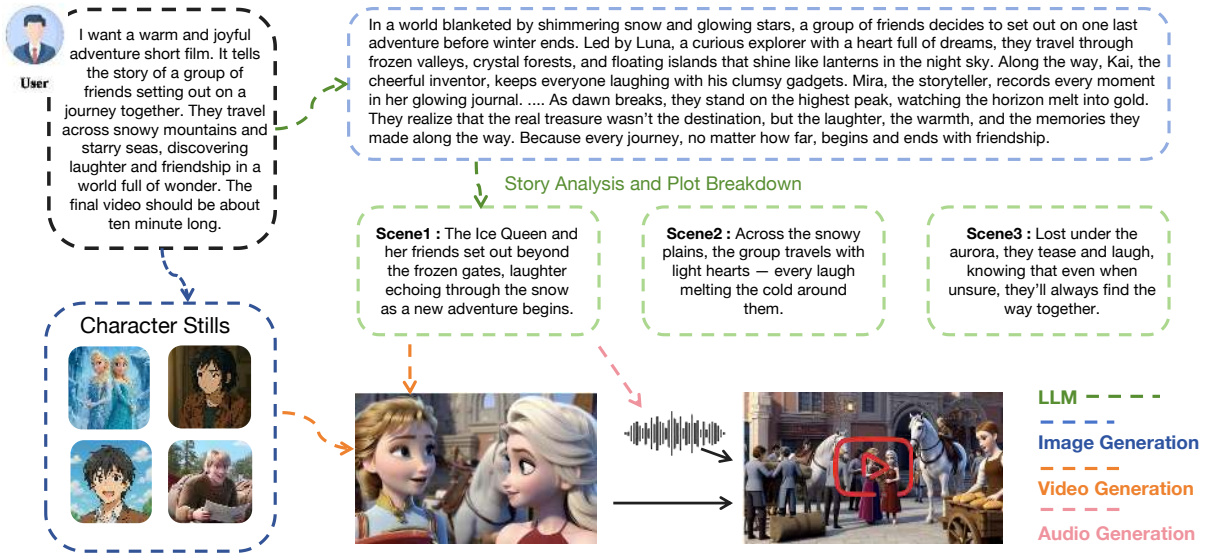



Figure 2: **Overview of DramaAgent.** DramaAgent decomposes a narrative prompt into scene-level targets, constructs persistent character stills, and coordinates image, video, and audio generators through scene-wise generation and reflection-guided repair.

### 3.1 Task Formulation

Given a narrative prompt $P$, our goal is to generate a multi-scene audiovisual story $(\mathcal{V}, \mathcal{A})$ that preserves story coherence, character consistency, cross-scene continuity, and scene-level audio-visual consistency over a long generation horizon. DramaAgent assumes access to a set of heterogeneous generation modules $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$, where each module may correspond to a commercial API or an open-source backbone. From $P$, DramaAgent constructs a temporally ordered scene sequence $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$, and a persistent character reference set $\mathcal{R} = \{R_c \mid c \in \mathcal{C}\}$, where $\mathcal{C}$ denotes the set of story characters. For each scene $S_i$, the framework generates candidate clips, evaluates them with respect to the scene specification and previously accepted context, and selectively repairs low-quality outputs. The accepted video and audio outputs for scene $S_i$ are denoted by $v_i$ and $a_i$, respectively. The final output is obtained by composing all accepted scene-level audiovisual segments: $(\mathcal{V}, \mathcal{A}) = \mathrm{Compose}\left(\{(v_i, a_i)\}_{i=1}^{N}\right)$.

### 3.2 Story Analysis and Plot Breakdown

To support controllable multi-scene generation, DramaAgent first converts the input prompt into a structured story representation. Given a narrative prompt $P$, an instruction-tuned language model expands it into an explicit story draft $\mathcal{N} = \{n_1, n_2, \ldots, n_L\}$, which specifies key narrative elements such as environments, characters, major events, and emotional progression. DramaAgent then decomposes $\mathcal{N}$ into a temporally ordered scene sequence $\mathcal{S}$. Each scene corresponds to a coherent event, transition, or emotional beat. Each scene is represented as

$$S_i = (x_i, \mathcal{C}_i, e_i, m_i, u_i), \tag{1}$$

where $x_i$ denotes the scene description, $\mathcal{C}_i$ the participating characters, $e_i$ the environment, $m_i$ the narrative role or emotional state, and $u_i$ an optional transition cue from the preceding scene. These structured scene targets provide the semantic backbone for both video and audio generation. Representative prompt templates are provided in *App.* D.

### 3.3 Global Plan Validation

Before scene-level generation, DramaAgent performs a lightweight consistency check over the structured scene sequence $\mathcal{S}$. The validator examines character attributes, temporal ordering, causal dependencies, and adjacent-scene transitions, and revises the scene plan when structural conflicts are detected. This stage operates on the global story structure before generation, whereas the reflection mechanism in Sec. 3.5 operates on generated scene outputs and repairs localized identity, semantic, temporal, or audio-visual failures.

### 3.4 Persistent Character Conditioning

After plot breakdown, DramaAgent constructs persistent character references, referred to as **Character Stills**, to stabilize character identity across scenes. From the expanded story, the framework extracts the character set $\mathcal{C}$ and their identity-defining attributes, including appearance, clothing, hairstyle,

age, and style cues. For each character $c \in \mathcal{C}$, DramaAgent forms a character specification $A_c$ and generates a canonical visual reference:

$$R_c = \mathrm{T2I}(A_c), \tag{2}$$

where $\mathrm{T2I}(\cdot)$ denotes a text-to-image generator. These references form a fixed character sheet and are reused throughout the pipeline. For each scene $S_i$, the scene condition is augmented as

$$\tilde{S}_i = (S_i, \{R_c \mid c \in \mathcal{C}_i\}), \tag{3}$$

where $\mathcal{C}_i \subseteq \mathcal{C}$ denotes the characters appearing in scene $i$. When a video generator supports image conditioning, the corresponding character stills are provided as visual inputs. Otherwise, DramaAgent converts them into textual identity constraints. By propagating the same references across generation, candidate selection, and repair, DramaAgent treats character identity as a persistent control signal rather than a one-time prompt attribute.

### 3.5 Reflection-guided Candidate Selection and Repair

After global plan validation and character conditioning, DramaAgent performs scene-wise generation and localized reflection-guided repair. Given the structured scene condition $\tilde{S}_i$, the framework produces a candidate set

$$\mathcal{V}_i = \{v_i^{(1)}, v_i^{(2)}, \ldots, v_i^{(K)}\}, \tag{4}$$

where candidates may come from different backbones, prompt variants, or random seeds under a fixed generation budget. The accepted history before scene $i$ is summarized as

$$h_{i-1} = \mathrm{Summarize}\left(\{S_j, v_j, a_j\}_{j<i}\right), \tag{5}$$

which provides cross-scene context for candidate evaluation and repair.

**Candidate scoring.** Each candidate is evaluated along four dimensions: identity consistency, semantic fidelity, temporal continuity, and audio-visual compatibility. Identity consistency $s_{\mathrm{id}}$ measures whether visible characters match the persistent character references. Semantic fidelity $s_{\mathrm{sem}}$ measures whether the required entities, actions, environments, and events in $\tilde{S}_i$ are present. Temporal continuity $s_{\mathrm{temp}}$ measures whether the candidate is consistent with the accepted history $h_{i-1}$, including scene transition and story progression. Audio-visual compatibility $s_{\mathrm{av}}$ measures scene-level semantic and temporal compatibility between the generated visual content and the dialogue, sound events, and audio plan associated with the scene. The final reflection score is computed as

$$\begin{aligned}\mathrm{Score}(v_i^{(k)}) = \lambda_{\mathrm{id}}\, s_{\mathrm{id}}(v_i^{(k)}, \mathcal{R}) \\ + \lambda_{\mathrm{sem}}\, s_{\mathrm{sem}}(v_i^{(k)}, \tilde{S}_i) \\ + \lambda_{\mathrm{temp}}\, s_{\mathrm{temp}}(v_i^{(k)}, h_{i-1}) \\ + \lambda_{\mathrm{av}}\, s_{\mathrm{av}}(v_i^{(k)}, D_i),\end{aligned} \tag{6}$$

where $D_i$ denotes the scene-level dialogue and audio plan. In the default setting, we use fixed weights $\lambda_{\mathrm{id}} = 0.35$, $\lambda_{\mathrm{sem}} = 0.30$, $\lambda_{\mathrm{temp}} = 0.20$, and $\lambda_{\mathrm{av}} = 0.15$. The detailed scoring prompts and rubric are provided in *App.* B.

**Targeted repair.** The best candidate is selected by

$$v_i^\star = \arg\max_{v_i^{(k)} \in \mathcal{V}_i} \mathrm{Score}(v_i^{(k)}). \tag{7}$$

If $\mathrm{Score}(v_i^\star) \geq \tau$, the scene is accepted. Otherwise, DramaAgent identifies the weakest evaluation dimension:

$$d_i^\star = \arg\min_{d \in \{\mathrm{id,sem,temp,av}\}} s_d(v_i^\star), \tag{8}$$

and rewrites the scene prompt according to the corresponding failure type. For identity failures, the repair prompt strengthens character-reference constraints. For semantic failures, it injects missing entities, actions, or environmental details. For temporal failures, it adds transition cues from the accepted history. For audio-visual failures, it adjusts dialogue timing, sound-event constraints, or visual pacing. The repaired scene is then regenerated and rescored. This process is repeated for at most $R$ rounds or until the score improvement falls below $\epsilon$:

$$\mathrm{Score}(v_{i,t}^\star) - \mathrm{Score}(v_{i,t-1}^\star) < \epsilon. \tag{9}$$

We use $\tau = 0.80$, $\epsilon = 0.02$, and $R = 2$ by default. This mechanism turns refinement into failure-aware correction rather than generic resampling, allowing DramaAgent to repair local inconsistencies before they propagate to later scenes. Unlike global plan validation, which corrects structural inconsistencies before generation, this repair process targets localized failures in generated scene outputs.

### 3.6 Audio Generation and Composition

To complement visual generation, DramaAgent performs scene-aware audio generation for dialogue, ambient sound, and background music.

Table 1: **Main comparison with recent text-to-video generation methods.** Results are reported for script synopsis to keyframe/storyboard generation and multi-scene story video generation. Each method is evaluated as a complete generation system; therefore, this comparison does not isolate backbone or generation-budget effects.

| Method | CLIP↑ | Inception↑ | VBench Metrics /% ↑ | | | | |
|---|---|---|---|---|---|---|---|
| | | | Subject Cons. | Bg Cons. | Motion Smth. | Dyn. Degree | Aesthetic |
| *Script Synopsis to Keyframe/StoryBoard Generation* | | | | | | | |
| StoryDiffusion (Zhou et al., 2024) | 20.46 | 6.24 | - | - | - | - | - |
| AutoStory (Wang et al., 2025) | 20.21 | 6.01 | - | - | - | - | - |
| MovieAgent (Wu et al., 2025b) | 22.12 | 7.23 | - | - | - | - | - |
| **DramaAgent (Ours)** | **23.02** | **7.64** | - | - | - | - | - |
| *Script Synopsis to Movie Generation* | | | | | | | |
| StoryDiffusion (Zhou et al., 2024) + SVD (Blattmann et al., 2023a) | 21.39 | 8.36 | 93.64 | 93.78 | 96.30 | 74.48 | 56.69 |
| StoryDiffusion (Zhou et al., 2024)+ CogVideoX (Yang et al., 2025) | 21.83 | 9.01 | 93.45 | 94.56 | 96.60 | 27.89 | 56.05 |
| AutoStory (Wang et al., 2025) + CogVideoX (Yang et al., 2025) | 20.27 | 7.21 | 91.45 | 93.32 | 95.87 | 70.32 | 52.34 |
| DreamVideo (Wei et al., 2024) | 21.37 | 8.11 | 93.17 | 93.77 | 96.40 | 26.97 | 42.16 |
| Magic-Me (Ma et al., 2024) | 21.72 | 8.34 | 94.01 | 94.68 | 96.41 | 14.86 | 55.89 |
| MovieAgent (Wu et al., 2025b) | 22.25 | 9.39 | 94.72 | 93.52 | 97.84 | **76.27** | 58.63 |
| **DramaAgent (Ours)** | **24.34** | **9.67** | **97.02** | **97.54** | **98.14** | 67.32 | **64.91** |

Given the scene sequence $\mathcal{S}$ and character set $\mathcal{C}$, a language model first produces dialogue scripts $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$, where each $D_i$ is conditioned on the current scene content, previous story context, participating characters, emotional state, and estimated clip duration. Each dialogue line is annotated with speaker identity, emotion, and delivery style for downstream voice assignment and prosody control. DramaAgent then uses a multi-speaker text-to-speech system to synthesize character-specific speech tracks. Ambient sound and background music are generated or retrieved according to the scene environment and narrative role. Finally, the accepted scene clips $\{v_i\}$ and audio tracks $\{a_i\}$ are assembled into the final multi-scene audiovisual output. Scene boundaries are smoothed with visual and auditory transitions, and dialogue timing is aligned with clip duration as closely as possible. The audio module mainly supports scene-level audio-visual consistency by providing dialogue, sound-event, and pacing constraints for generation and reflection. Fine-grained lip synchronization and precise sound-effect timing still depend on the underlying video and audio generators, and are discussed as limitations.

# 4 Experiments

We evaluate DramaAgent from four perspectives: overall generation quality, robustness across heterogeneous video backbones, the contribution of each control component, and human-perceived cross-scene coherence. Beyond visual quality, our evaluation focuses on whether the generated story preserves character identity, narrative progression, cross-scene continuity, and scene-level audio-visual consistency.

## 4.1 Experimental Setup

**Evaluation data.** We evaluate DramaAgent on narrative-to-video and script-to-storyboard generation tasks. The evaluation set contains 50 narrative scripts spanning multiple genres, including fantasy stories, conversational scenes, emotional monologues, and action sequences. Each script contains 10–20 sentences and corresponds to approximately 30–60 seconds of video content after scene decomposition and composition.

**Generation backbones.** For the video generation stage, we use heterogeneous backbones including Wan2.2-I2V-Flash, Keling, Jimeng, Hailuo 3.0, and Seedance. These backbones are accessed through their corresponding generation interfaces, and DramaAgent adapts scene prompts and conditioning inputs to each interface. Unless otherwise stated, scene-level clips are generated with a duration of 5 seconds at 1280×720 resolution. DramaAgent uses $K = 3$ candidates per scene and performs at most $R = 2$ rounds of reflection-guided repair by default. Controlled comparisons that vary the sampling or repair budget explicitly report their generation calls per scene. Detailed generation settings, backbone versions, API access dates, and reflection hyperparameters are provided in *App.* A and *App.* B.

**Metrics.** Following the VBench protocol (Huang et al., 2024), we report *Subject Consistency*, *Background Consistency*, *Motion Smoothness*, *Dynamic Degree*, and *Aesthetic Quality*. For the visual-fidelity analysis, we additionally report *Imaging Quality*. We further report *CLIP* and *Inception* scores to assess semantic alignment and visual quality. CLIP is computed using the OpenAI CLIP ViT-B/32 model by averaging frame-level cosine similarities between uniformly sampled video frames and the corresponding text prompt. Inception

Score is computed using an ImageNet-pretrained Inception-v3 model over the same sampled video frames. Since automatic visual metrics cannot fully capture multi-scene storytelling quality, we also conduct blind human evaluation on temporal coherence, character consistency, narrative fidelity, and scene-level audio-visual consistency. The human-evaluation protocol and participant statistics are provided in *App.* C. Runtime and approximate API cost statistics are reported in *App.* A.1 and Tab. 6.

### 4.2 Main Results

Tab. 1 compares DramaAgent with recent text-to-video and story video generation methods under two settings: *Script Synopsis to Keyframe/Storyboard Generation* and *Script Synopsis to Multi-scene Story Video Generation*. DramaAgent achieves strong overall performance in both settings.

**Script synopsis to keyframe/storyboard generation.** In the storyboard generation setting, DramaAgent obtains a CLIP score of 23.02 and an Inception score of 7.64, outperforming StoryDiffusion, AutoStory, and MovieAgent. These results indicate that structured story planning and character-aware generation produce more semantically grounded intermediate visual targets from high-level narrative synopses. Since storyboard quality directly affects downstream video generation, this setting evaluates whether DramaAgent can provide a stronger visual foundation for multi-scene story synthesis.

**Script synopsis to multi-scene story video generation.** In the multi-scene story video generation setting, DramaAgent reaches a CLIP score of 24.34 and an Inception score of 9.67. It also obtains the highest scores in Subject Consistency, Background Consistency, Motion Smoothness, and Aesthetic Quality among the compared methods. Compared with StoryDiffusion+SVD, StoryDiffusion+CogVideoX, AutoStory+CogVideoX, DreamVideo, Magic-Me, and MovieAgent, DramaAgent produces stronger semantic alignment and more stable character consistency. These results suggest that structured control is effective for organizing scene-level generations into a coherent multi-scene story.

### 4.3 Controlled Analysis

To isolate the contribution of DramaAgent from backbone choice and additional generation attempts, we conduct fixed-backbone and generation-budget-aware analyses.

**Fixed-backbone analysis.** Tab. 2 reports a progressive ablation using the same Seedance video backbone. Adding story planning, character stills, and reflection-guided repair progressively improves performance, showing that the gains arise from the proposed control components rather than the backbone choice alone.

Table 2: **Component ablation on a fixed backbone.** Progressively adding story planning, character stills, and reflective repair on Seedance.

| Method | CLIP↑ | Inception↑ | VBench Metrics / % ↑ | | | | |
|---|---|---|---|---|---|---|---|
| | | | Subject Cons. | Bg Cons. | Motion Smth. | Dyn. Degree | Aesthetic |
| Direct Generation (Seedance) | 22.90 | 8.65 | 92.10 | 95.60 | 97.10 | 64.20 | 62.30 |
| + Story Planning | 23.55 | 8.95 | 93.80 | 96.30 | 97.45 | 65.10 | 62.90 |
| + Character Stills | 24.12 | 9.41 | 96.85 | 97.33 | 97.96 | 66.84 | 64.37 |
| + Reflective Repair (Full) | 24.34 | 9.67 | 97.02 | 97.54 | 98.14 | 67.32 | 64.91 |

**Backbone- and generation-budget-controlled comparison.** To further distinguish targeted repair from generic additional sampling, we compare DramaAgent with direct generation, open-loop story planning, best-of-$K$ sampling, and generic iterative repair using the same Seedance backbone. Tab. 3 explicitly reports the generation budget for each variant. DramaAgent achieves the strongest semantic alignment, subject consistency, and dynamic degree while requiring only a modest increase in generation calls over generic repair. These results indicate that its gains cannot be explained solely by backbone strength or additional sampling, but also arise from persistent cross-scene state and failure-specific targeted repair.

Table 3: **Backbone- and generation-budget-controlled comparison on Seedance.** All variants use the same video backbone and evaluation setting.

| Method | Calls/Scene↓ | CLIP↑ | Subject Cons.↑ | Dyn. Degree↑ |
|---|---|---|---|---|
| Direct Generation | 1.0 | 22.90 | 92.10 | 64.20 |
| Open-loop Story Planning | 1.0 | 23.55 | 93.80 | 65.10 |
| Best-of-$K$ | 3.0 | 23.73 | 94.65 | 65.74 |
| Generic Iterative Repair | 3.8 | 23.87 | 95.41 | 66.21 |
| **DramaAgent** | 4.1 | **24.34** | **97.02** | **67.32** |

A complete wall-clock time and cost comparison under the same Seedance setting is reported in Tab. 7 in *App.* A.2, further contextualizing the quality–efficiency trade-off. We further analyze the sensitivity to the acceptance threshold $\tau$ in Tab. 9 and *App.* B.1.

### 4.4 Ablation Study

**Effect of backbone choice.** Tab. 4 compares DramaAgent with different video generation backbones. Although performance varies across generators, DramaAgent remains consistently competitive, suggesting that the framework is not tied to a specific backbone. Seedance achieves the best balance between semantic alignment and aesthetic quality, while Jimeng and Hailuo 3.0 perform strongly on motion-related metrics.

**Component ablation on a fixed backbone.** Tab. 2 reports a progressive ablation on the same back-




Figure 3: **Qualitative results across three narrative scenes.** DramaAgent maintains recurring characters and coherent scene progression across different story contexts. The examples illustrate the role of structured planning, persistent conditioning, and reflection-guided repair in multi-scene storytelling.

Table 4: **Effect of backbone choice.** DramaAgent is evaluated with five video generation backbones under the same control framework.

| Method | CLIP↑ | Inception↑ | VBench Metrics / % ↑ | | | | |
|---|---|---|---|---|---|---|---|
| | | | Subject Cons. | Bg Cons. | Motion Smth. | Dyn. Degree | Aesthetic |
| DramaAgent (Wan2.2-I2V-Flash) | 23.47 | **9.56** | 94.12 | 93.95 | 96.92 | 70.15 | 54.32 |
| DramaAgent (Keling) | 23.78 | 9.35 | 95.64 | 95.82 | 97.43 | 66.92 | 58.14 |
| DramaAgent (Jimeng) | 23.96 | 9.38 | 96.42 | 96.71 | 98.02 | 80.45 | 61.27 |
| DramaAgent (Hailuo 3.0) | 22.91 | 9.12 | 96.34 | 97.27 | **98.61** | **84.31** | 61.32 |
| DramaAgent (Seedance) | **24.12** | 9.41 | **96.85** | **97.33** | 97.96 | 66.84 | **64.37** |

bone. Story planning improves semantic alignment, character stills strengthen identity preservation, and reflective repair adds further gains, validating the complementary roles of the three components.

### 4.5 Audio-Visual and Human Evaluation

**Visual fidelity.** To examine whether improved cross-scene consistency comes at the expense of perceptual quality, we compare DramaAgent with native Seedance under identical generation settings. DramaAgent preserves comparable VBench Imaging Quality (64.72 vs. 64.18) while receiving higher blind human preference for sharpness (53.1% vs. 46.9%), texture detail (53.1% vs. 46.9%), fewer artifacts (56.2% vs. 43.8%), and overall visual fidelity (54.9% vs. 45.1%). These results indicate that the consistency gains are not obtained at the expense of perceptual quality.

**Automatic audio-visual evaluation.** We further evaluate scene-level audio-video semantic alignment, temporal synchronization, and audio–text relevance. DramaAgent achieves scores of 70.6, 72.4, and 69.8, respectively, outperforming the variant without audio planning (61.8/64.5/58.9) and the variant without audio-visual reflection (65.9/67.8/64.7). These results show that audio planning and audio-visual reflection both contribute to scene-level semantic and temporal compatibility.

**Human evaluation.** Automatic metrics alone cannot fully capture multi-scene audiovisual storytelling quality. We therefore conduct a blind pairwise human evaluation with 15 evaluators on temporal coherence, character consistency, narrative fidelity, and scene-level audio-visual consistency, with each sample independently assessed by at least three annotators. DramaAgent is preferred in temporal coherence (66.7%), character consistency (71.3%), narrative fidelity (63.3%), and audio-visual alignment (68.0%). Inter-rater agreement reaches Fleiss' $\kappa = 0.42$. These results indicate stronger scene continuity, character persistence, narrative faithfulness, and semantic compatibility between audio and visual content. Detailed evaluation criteria, participant information, and annotation protocols are provided in Tab. 10 and *App.* C. This evaluation focuses on scene-level semantic and temporal audio-visual compatibility rather than fine-grained lip synchronization, which remains partly constrained by the underlying generators.

### 4.6 Qualitative Analysis

**Cross-scene storytelling.** Fig. 3 shows generated results across three narrative scenes. DramaAgent

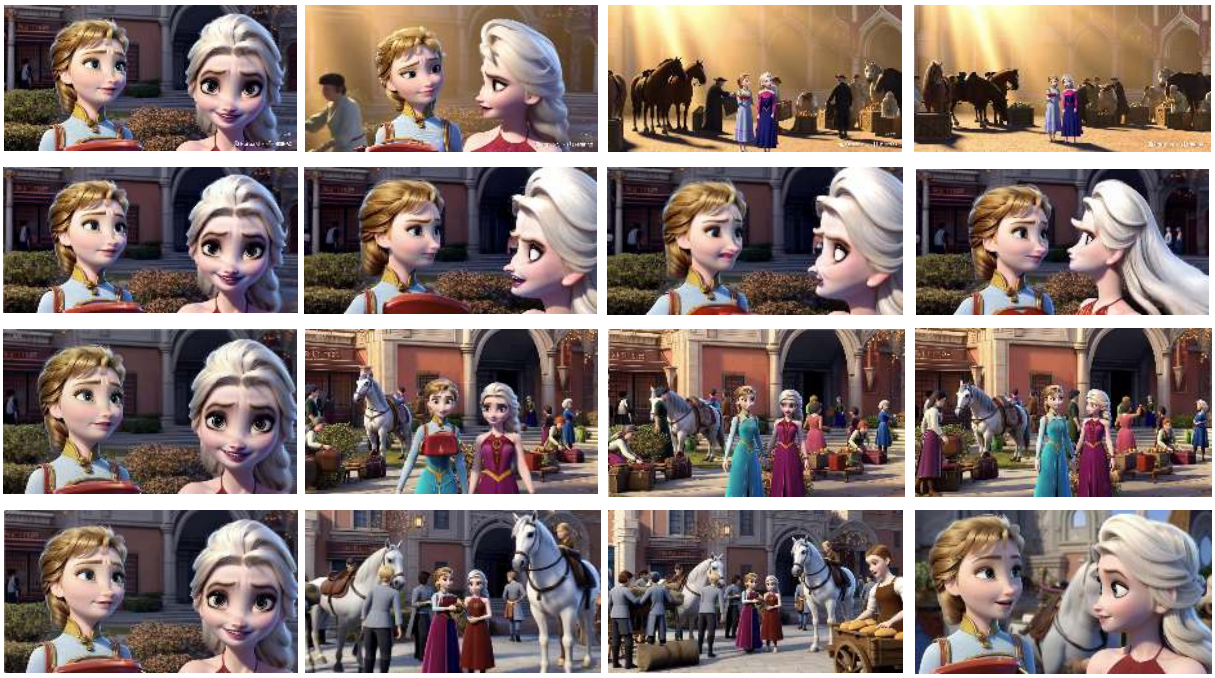


Figure 4: **Qualitative comparison across four video generation backbones.** We compare Hailuo, Keling, Jimeng, and SeedDance under the same narrative prompt and character references, highlighting their differences in identity consistency, temporal stability, and scene realism.

maintains recurring character appearances and coherent narrative progression across different story contexts. These examples illustrate how structured planning, persistent character conditioning, and reflection-guided repair jointly support cross-scene consistency in multi-scene storytelling. Additional case studies are provided in *App.* E.

**Cross-backbone comparison.** Fig. 4 compares different video generation backbones under the same narrative prompt and character references. These examples illustrate complementary strengths across backbones while showing that the same DramaAgent control framework can be applied to heterogeneous generators. For example, some backbones produce stronger motion continuity, while others better preserve character appearance or visual style.

## 5 Conclusion

We propose **DramaAgent**, a hierarchical and model-agnostic framework for multi-scene long-horizon text-to-video-and-audio generation. DramaAgent formulates multi-scene storytelling as a structured control problem over heterogeneous video and audio generators, combining story planning, persistent character conditioning, scene-wise generation, reflection-guided targeted repair, and audio composition. Experiments across multiple backbones and controlled same-backbone comparisons show that DramaAgent improves cross-scene coherence, character consistency, narrative fidelity, and scene-level audio-visual consistency over direct generation and strong baselines. These results suggest that stateful hierarchical control with persistent context and failure-specific repair provides a practical direction for controllable multi-scene audiovisual storytelling.

## Limitations

DramaAgent has several limitations. First, our evaluation focuses on 30–60-second multi-scene stories and therefore does not establish scalability to substantially longer, movie-length generation. Maintaining a globally coherent narrative state, character identity, and audio-visual consistency over much longer horizons remains an important direction for future work. Second, DramaAgent relies on third-party video and audio generation APIs, so outputs, latency, pricing, and conditioning interfaces may change over time. Third, as an orchestration framework rather than a unified world-consistent generative model, it may still struggle with complex interactions, physical contact, large pose/camera changes, and residual identity drift. Finally, generated videos may inherit artifacts, biases, and safety risks from the underlying models and should be clearly disclosed as synthetic content in real-world use.

**Acknowledgements.** This work was supported by the Fundamental Research Funds for the Central Universities, Peking University.

## References


Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, and 1 others. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

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

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, and 1 others. 2024. Video generation models as world simulators. *OpenAI Blog*, 1(8):1.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*.

Yuchao Gu, Yipin Zhou, Yunfan Ye, Yixin Nie, Licheng Yu, Pingchuan Ma, Kevin Qinghong Lin, and Mike Zheng Shou. 2024. Roictrl: Boosting instance control for visual generation. *arXiv preprint arXiv:2411.17949*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, and 1 others. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

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

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818. IEEE.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, and 1 others. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, and 1 others. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6329–6338.

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

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

Ze Ma, Daquan Zhou, Xue-She Wang, Chun-Hsiao Yeh, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. 2024. Magic-me: Identity-specific video customized diffusion. In *European Conference on Computer Vision*, pages 19–37. Springer.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, and 1 others. 2024. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2493–2502. IEEE.

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. 2024a. Coin: A lightweight and effective framework for story visualization and continuation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10659–10668.

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. 2024b. Storyimager: A unified and efficient framework for coherent story visualization and completion. In *European Conference on Computer Vision*, pages 479–495. Springer.

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

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, and 1 others. 2024. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. 2024. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7589–7599. IEEE.

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

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

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, and 1 others. 2025. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pages 83048–83077.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, and 1 others. 2023. Nuwa-xl: Diffusion over diffusion for extremely long video generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1309–1320.

Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, and 1 others. 2024. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. 2024. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and 1 others. 2024a. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2412.02259*.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024b. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.

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

Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. 2023. Moviefactory: Automatic movie creation from text using large generative models for language and images. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9313–9319.

## A Detailed Generation Settings

Table 5 summarizes the generation settings used in our experiments for reproducibility. Unless otherwise stated, all methods are evaluated on the same set of 50 narrative scripts, where each script contains 10–20 sentences and corresponds to approximately 30–60 seconds of video content after scene decomposition and composition.

Across different backbones, the upper-level DramaAgent framework is kept unchanged, including structured story planning, persistent character conditioning, and reflective repair. Only the underlying video generator is varied. When a backbone supports direct image conditioning, character stills are provided as visual inputs; otherwise, the same references are converted into textual identity constraints and injected into the scene prompt.

**Reproducibility note.** Because some commercial APIs may evolve over time, exact outputs can vary across access dates. To improve reproducibility, we record the API variant and access date for each experiment, while keeping the scene-level budget, resolution, repair rounds, and upper-level DramaAgent logic fixed within each comparison group.

### A.1 Runtime and Cost Statistics

Table 6 reports practical runtime and approximate API usage statistics for the default setting. Under the default configuration ($K = 3$ candidates per scene, at most $R = 2$ repair rounds, 5-second clips, and 1280×720 resolution), DramaAgent requires on average 75 seconds for initial scene generation and 28 seconds for each repaired scene. Across the full pipeline, the average end-to-end generation time for a 30–60 second script is approximately 18–32 minutes, depending on the number of scenes, the activated repair rounds, and the selected backbone APIs. On average, DramaAgent makes 4.1 API calls per scene, including candidate generation and optional regeneration calls. The approximate generation cost is about $7.8 per minute of final video under the default setup. These values should be interpreted as practical reference numbers rather than fixed constants, since commercial API latency and pricing may vary across providers and access dates.

### A.2 Comparative Efficiency Analysis

To contextualize the overhead of selective repair, we compare DramaAgent with direct generation, best-of-$K$ sampling, and generic iterative repair under the same Seedance setting.

Under comparable generation budgets, DramaAgent requires only 0.3 additional calls per scene over generic iterative repair, while providing stronger generation quality as reported in Tab. 3.

## B Reflection Settings and Repair Policy

**Repair policy.** Table 8 summarizes the default reflection hyperparameters. For each scene, DramaAgent first generates $K = 3$ candidate clips and computes the reflection score in Eq. 6. If the best candidate score is below $\tau = 0.80$, the framework triggers targeted repair. The repair route is chosen according to the weakest sub-score among $s_{\text{id}}$, $s_{\text{sem}}$, $s_{\text{temp}}$, and $s_{\text{av}}$: low identity consistency strengthens character-reference conditioning, low semantic fidelity rewrites the prompt with missing entities or actions, low temporal continuity injects stronger transition cues from the previous scene, and low audio-visual compatibility adjusts dialogue timing or regenerates the clip under tighter temporal constraints. We perform at most $R = 2$ repair rounds per scene, and stop early when the score improvement is below $\epsilon = 0.02$.

### B.1 Sensitivity to Acceptance Threshold

We study the sensitivity to the acceptance threshold $\tau$ while keeping the remaining reflection settings fixed.

As shown in Tab. 9, performance remains stable across a broad threshold range. Increasing $\tau$ beyond 0.80 substantially increases repair activation and API calls while yielding only marginal or no quality improvement. We therefore use $\tau = 0.80$ as the default quality–efficiency operating point.

## C Human Evaluation Protocol

This section describes the human-evaluation protocol used to complement automatic metrics for multi-scene story generation.

**Evaluation dimensions.** We evaluate generated outputs on four human-centered dimensions: *temporal coherence*, *character consistency*, *narrative fidelity*, and *audio-visual alignment*. These dimensions capture cross-scene properties that are not fully reflected by standard automatic visual metrics.

**Participants.** We recruit 15 human evaluators with prior experience in video content assessment, multimedia generation, or related visual-language tasks.

Table 5: **Reproducibility settings for video generation backbones.** We report the backbone/API variant, access date, supported conditioning interface, and scene-level generation budget used in our experiments. All DramaAgent variants share the same upper-level planning, persistent character conditioning, and reflective repair framework; only the underlying video generator changes across rows.

| Backbone | Version / API Variant | Access Date | Input Type | Candidates / Scene | Repair Rounds | Clip Duration | Resolution |
|---|---|---|---|---|---|---|---|
| Wan2.2-I2V-Flash | Wan2.2-I2V-Flash | 2025-11 | image + text | 3 | 2 | 5s | 1280×720 |
| Keling | Keling | 2025-11 | image + text | 3 | 2 | 5s | 1280×720 |
| Jimeng | Jimeng | 2025-11 | text-only / multimodal | 3 | 2 | 5s | 1280×720 |
| Hailuo 3.0 | Hailuo 3.0 | 2025-11 | image + text | 3 | 2 | 5s | 1280×720 |
| Seeddance 3.0 | Seeddance 3.0 | 2025-11 | image + text | 3 | 2 | 5s | 1280×720 |

Table 6: **Runtime and cost statistics under the default setting.** Reported values are averages over the evaluation set and may vary with backbone API latency and access date.

| Statistic | Value |
|---|---|
| Average initial generation time / scene | 75 s |
| Average repaired generation time / scene | 28 s |
| Average end-to-end time / 30–60s script | 18–32 min |
| Average API calls / scene | 4.1 |
| Approximate cost / minute of final video | $7.8 |

Table 7: **Comparative generation efficiency on Seedance.**

| Method | Calls/Scene↓ | Time/Story↓ | Cost/Final Min.↓ |
|---|---|---|---|
| Direct Generation | 1.0 | 5–8 min | $1.9 |
| Best-of-$K$ Sampling | 3.0 | 13–23 min | $5.7 |
| Generic Iterative Repair | 3.8 | 17–29 min | $7.2 |
| **DramaAgent** | 4.1 | 18–32 min | $7.8 |

Table 8: **Default reflection settings.** We use the following default hyperparameters for candidate scoring and selective reflective repair.

| Parameter | Value |
|---|---|
| $\lambda_{\text{id}}$ | 0.35 |
| $\lambda_{\text{sem}}$ | 0.30 |
| $\lambda_{\text{temp}}$ | 0.20 |
| $\lambda_{\text{av}}$ | 0.15 |
| Acceptance threshold $\tau$ | 0.80 |
| Early-stop threshold $\epsilon$ | 0.02 |
| Max repair rounds $R$ | 2 |
| Candidates per scene $K$ | 3 |

Table 9: **Sensitivity to the acceptance threshold $\tau$.**

| $\tau$ | CLIP↑ | Subject Cons.↑ | Calls/Scene↓ | Repair Activation (%) |
|---|---|---|---|---|
| 0.70 | 23.96 | 96.18 | 3.45 | 31.2 |
| 0.75 | 24.18 | 96.63 | 3.78 | 45.6 |
| 0.80 | 24.34 | 97.02 | 4.10 | 58.7 |
| 0.85 | 24.39 | 97.11 | 4.38 | 71.5 |
| 0.90 | 24.36 | 97.08 | 4.65 | 83.4 |

All participants are blind to the method identities during evaluation.

**Evaluation protocol.** We use blind pairwise comparison for generated samples. For each comparison, evaluators are shown outputs from two methods under the same narrative prompt and are asked to indicate the preferred result along the four evaluation dimensions. Each sample is evaluated independently by at least 3 annotators.

**Evaluation criteria.** *Temporal coherence* measures whether scene transitions, motion progression, and event continuity remain stable across time. *Character consistency* measures whether the same characters preserve their visual identity across scenes. *Narrative fidelity* measures whether the generated output remains faithful to the story content, event sequence, and emotional progression specified in the input prompt. *Audio-visual alignment* measures whether audio and visual content remain semantically compatible and temporally coherent at the scene level; it does not evaluate fine-grained lip synchronization.

**Statistics.** For each evaluation dimension, we report the average preference rate of DramaAgent across all evaluated samples. Inter-rater agreement is measured by Fleiss' kappa. The detailed human-evaluation results, including preference rates for temporal coherence, character consistency, narrative fidelity, and audio-visual alignment, are reported in Table 10.

# D Prompt Templates and Reflection Settings

This section provides representative prompt templates and key settings used in DramaAgent, including story analysis and plot breakdown, character still generation, reflective rewriting, and repair control parameters.

## D.1 Story Analysis and Plot Breakdown Prompt

To convert a narrative prompt into a structured story representation, we use an instruction-tuned language model to expand the input into a longer story draft and decompose it into ordered scene-level units. A representative prompt template is shown below.

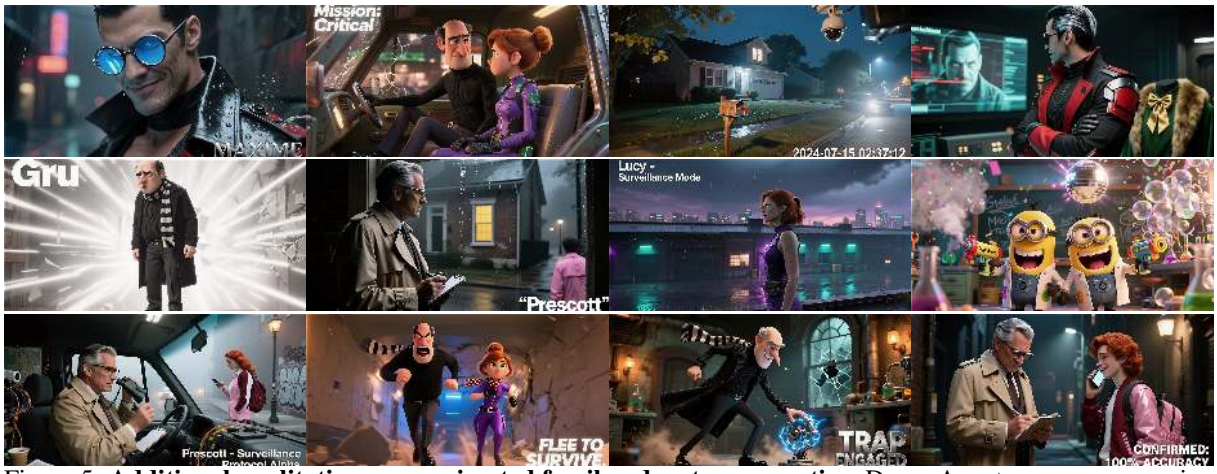



Figure 5: **Additional qualitative case: animated family-adventure narrative.** DramaAgent preserves recurring character identity, expressive interaction, and multi-scene story continuity in a stylized family-oriented animated story.




Figure 6: **Additional qualitative case: action-crime ensemble narrative.** The example illustrates multi-character coordination, dynamic scene progression, and stable identity preservation in a high-motion cinematic setting.

Table 10: **Human evaluation results.** DramaAgent is preferred over baseline pipelines across all four evaluation dimensions.

| Metric | DramaAgent Preference |
|---|---|
| Temporal Coherence | 66.7% |
| Character Consistency | 71.3% |
| Narrative Fidelity | 63.3% |
| Audio-Visual Alignment | 68.0% |
| Inter-rater Agreement (Fleiss' Kappa) | 0.42 |

> **Input:** A short narrative prompt describing the story theme, characters, and emotional tone.
>
> **Instruction:** Expand the prompt into a coherent story with clear event progression. Then decompose the story into a sequence of scene-level descriptions. Each scene should correspond to a semantically coherent event or emotional beat. For each scene, provide: (1) a concise scene description, (2) participating characters, (3) environment, (4) emotional state or narrative role, and (5) a transition cue if needed.

The output of this stage is used to construct the scene sequence $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$ in the main pipeline.

### D.2 Character Stills Prompt

To build persistent character references, DramaAgent first extracts the main characters and their visual attributes from the expanded story. It then generates a canonical visual reference (*Character Still*) for each character. A representative prompt template is shown below.

> **Input:** Character name and visual attributes (e.g., hairstyle, clothing, facial appearance, age, style cues).
>
> **Instruction:** Generate a high-quality portrait of the character that preserves the specified identity-defining traits. The portrait should serve as a canonical reference image for subsequent scene generation. Maintain consistent facial structure, hairstyle, costume, and overall artistic style.

These generated character stills form the persistent character reference set $\mathcal{R} = \{R_c\}$ used




Figure 7: **Additional qualitative case: fantasy mystery narrative.** DramaAgent supports atmosphere-heavy story generation with coherent clue progression, recurring character consistency, and visually stable magical world-building.




Figure 8: **Additional qualitative case: mythic fantasy narrative.** The generated story maintains stylized visual design and coherent narrative development across ritual, conflict, resolve, and farewell scenes.

throughout scene generation and reflective repair.

### D.3 Reflective Rewrite Prompt

When the reflective repair module detects a low-quality scene candidate, DramaAgent applies a failure-specific prompt rewriting strategy. A representative reflective rewrite prompt is shown below.

> **Input:** Current scene description, detected failure type, previously accepted scene summary, and character references.
>
> **Instruction:** Rewrite the scene prompt to correct the detected failure while preserving the original story intent. If the failure is identity inconsistency, strengthen appearance and character-reference constraints. If the failure is semantic omission, explicitly inject missing entities, actions, or environment details. If the failure is temporal discontinuity, add transition information from the previous scene. If the failure is audio-visual mismatch, adjust dialogue timing or visual pacing constraints.

This rewritten prompt is then used for targeted re-generation rather than full-sequence re-generation.

### D.4 Reflection Thresholds and Repair Settings

In our experiments, DramaAgent generates 3 candidate clips per scene and performs at most 2 rounds of reflective repair. The reflective module evaluates candidate clips with four criteria: identity consistency, semantic fidelity, temporal continuity, and audio-visual compatibility. These criteria are aggregated into the reflection score described in the main text.

The repair process is triggered when the best candidate score falls below the acceptance threshold $\tau$. Repair terminates when either the maximum number of rounds is reached or the score improvement between two successive rounds falls below $\epsilon$. In practice, these settings provide a good balance between output quality and computational cost.

## E Additional Qualitative Results

This section provides additional qualitative examples beyond the representative cases shown in the main paper. We include diverse story generations to further illustrate the genre generalization, multi-scene coherence, and character consistency of DramaAgent under long-form narrative generation.

Figures 5–8 present additional qualitative case studies spanning multiple narrative genres, including animated family adventure, action-crime ensemble, fantasy mystery, and mythic fantasy stories. These examples complement the main-paper visualizations by showing that DramaAgent is not restricted to a single visual style or story template. Instead, the framework supports long-form generation across substantially different genres while preserving scene-level progression and recurring character identity.

Across these examples, several consistent patterns can be observed. First, DramaAgent maintains recognizable character appearance across multiple scene transitions, even when the story changes location, lighting, or emotional tone. Second, the generated stories exhibit coherent narrative development rather than isolated clip-level plausibility: early frames establish characters and setting, middle frames develop conflict, and later frames move toward confrontation or resolution. Third, the framework remains effective across diverse visual regimes, including cinematic live-action prompts, stylized animation, fantasy environments, and mythology-inspired compositions.

Figure 5 shows an animated family-oriented adventure with recurring characters and light comedic interactions, highlighting the ability of DramaAgent to preserve expressive identity and warm cross-scene continuity. Figure 6 presents an action-crime narrative with multi-character coordination, tactical progression, and dynamic urban scenes, illustrating that the framework can support visually complex stories with sustained identity consistency. Figure 7 demonstrates a fantasy mystery narrative centered on clue discovery, mentor guidance, and magical exploration, showing that DramaAgent can maintain atmospheric consistency and structured narrative progression in dialogue- and world-building-heavy stories. Figure 8 further presents a mythology-inspired story with ritual restoration, large-scale conflict, personal resolve, and emotional farewell, suggesting that the framework can preserve both stylized world design and long-horizon emotional continuity.

Overall, these qualitative cases support the central claim of this work: by combining structured planning, persistent character conditioning, and reflective refinement, DramaAgent improves long-form story coherence across diverse prompts and visual styles.

## F Discussion

DramaAgent takes a control-oriented view of long-form text-to-video-and-audio generation. Rather than introducing a new backbone, it studies how structured planning, persistent character conditioning, and reflective repair can improve story-level coherence over extended generation horizons. Our results suggest that this agentic control layer is a practical way to make long-form generation more controllable and more robust across heterogeneous video generators.

At the same time, evaluating long-form audio-visual generation remains an important open problem. Current automatic metrics mainly capture clip-level visual properties, such as consistency, smoothness, or aesthetic quality, but are still limited in measuring story-level coherence across scenes, long-horizon character persistence, and alignment between generated audio and visual events. As a result, even when automatic scores improve, they may not fully reflect whether a generated story is perceived as globally coherent by human viewers.

We see two important directions for future work. First, the field would benefit from richer long-form multimodal benchmarks that jointly evaluate narrative progression, identity consistency, temporal continuity, and audio-visual alignment over extended sequences. Second, more reliable automatic evaluators are needed for story-level assessment, especially evaluators that can reason jointly over text, video, and audio rather than scoring each modality in isolation. Such advances would make it easier to measure progress on coherent long-form generation and to build stronger feedback loops for automatic repair and refinement.