ViMax: Agentic Video Generation

Anonymous ACL submission

Abstract

Long-form video generation presents fundamental challenges in narrative planning and visual consistency that existing short-form generation methods cannot address. Current approaches are limited to brief, single-scene content and lack the systematic planning capabilities required for long-horizon narrative generation. We present ViMax, an agentic video generation framework that employs a multi-agent workflow to orchestrate professional-grade long-form video production from conceptual ideas to realized content. Our approach addresses narrative complexity through hierarchical story decomposition combined with retrieval-augmented generation to maintain global context, while ensuring visual consistency via graph-based dependency tracking and transition video generation for spatial coherence. The framework integrates specialized agents for screenwriting, shot planning, character styling, video generation, and VLM-based quality control, enabling coordinated content creation that maintains both narrative coherence and visual consistency across extended timelines. We demonstrate ViMax's capabilities through comprehensive long-form video generation and narrative planning experiments.

1 Introduction

Video content has become the dominant medium for information dissemination and entertainment, with rapidly growing demand for long-form narratives spanning films, documentaries, and educational materials. However, traditional video production requires specialized expertise and substantial resources, creating significant barriers to creative content creation. Recent advances in generative AI present unprecedented opportunities to democratize video production by automating the complex workflow from conceptual ideas to fully realized visual content through coordinated AI agents.

Recent research in video generation has explored directions to address these challenges. Text-to-video models such as Sora (OpenAI, 2024), Veo 3 (Google DeepMind, 2024), and Seedance (Seedance Team, 2026), together with open-source efforts like HunyuanVideo (Kong et al., 2024) and Wan (Wang et al., 2025a), have achieved per-clip fidelity, yet remain constrained to short-form content lasting a few seconds. Extending these models to longer horizons exposes *catastrophic semantic forgetting*, where character identity, world geometry, and physical laws drift across shots (Kupyn et al., 2025; Rakheja et al., 2025). Recent works mitigate drift along two axes: streaming and memory-based generation (e.g., StreamingT2V (Henschel et al., 2024), FreeLong (Lu et al., 2024)) push single-clip duration, while consistency mechanisms such as StoryDiffusion (Zhou et al., 2024), ConsiStory (Tewel et al., 2024), and storyboard-anchored pipelines like STAGE (Zhang et al., 2025b) share attention or anchor on start–end keyframes to preserve subject identity and inter-shot transitions. In parallel, multi-agent frameworks including FilmAgent (Xu et al., 2025), Kubrick (He et al., 2024), and VideoGen-of-Thought (Zheng et al., 2025) begin to decompose video production into specialized roles. Despite this progress, existing systems struggle to maintain global narrative and spatial coherence across the dozens of shots required for long-form storytelling.

Addressing these limitations requires overcoming two fundamental challenges:

- **Multi-shot Visual Consistency.** Video generators operate on each shot in isolation, with no awareness of previously generated content, causing drift in character appearance, environment design, and spatial layout across shots. Long-form production therefore demands both object-level consistency, so that characters and environments retain identical visual properties, and spatial coherence, so that multiple camera angles within the same scene preserve consistent 3D geometry.

Figure 1: ViMax-generated multi-shot video examples from ViMax-Bench.

- **Long-form Narrative Planning.** Orchestrating narratives that span dozens of shots quickly exceeds the reasoning capacity of current language models when treated as a monolithic task. The system must coordinate character arcs, plot development, and thematic coherence over extended timelines, while ensuring that scene-level decisions remain grounded in the broader story context to avoid contradictions and preserve unity.

To address these challenges, we present ViMax, an agentic video generation framework that orchestrates long-form video production from a conceptual idea to fully rendered content. ViMax decomposes the pipeline into specialized agents covering screenwriting, shot planning, character styling, video generation, and VLM-based quality control with best-of-k candidate selection. To tackle narrative complexity, ViMax performs hierarchical story decomposition with retrieval-augmented generation that keeps each local planning step grounded in the global narrative. To enforce visual consistency, ViMax introduces a graph-based visual dependency system that tracks cross-shot relationships during planning, complemented by transition video generation that maintains spatial coherence across camera angles within the same scene.

Our primary contributions are:

- **Agentic Long-form Video Generation Framework.** We introduce a multi-agent workflow that coordinates screenwriting, shot planning, character styling, video generation, and VLM-based quality control, enabling automated creation of narrative video content from a conceptual idea.

- **Hierarchical Narrative Planning with Global Context.** We propose a hierarchical story decomposition approach paired with retrieval-augmented generation (RAG), allowing language models to plan extended narratives without losing thematic coherence or long-range dependencies.

- **Graph-based Visual Consistency.** We develop a dual-pronged approach that combines graph-based visual dependency tracking for object-level consistency across shots with transition video generation for spatial coherence within scenes.

To facilitate reproducibility and further research, we anonymously release our code at https://anonymous.4open.science/r/vimax-8D46/

2 Methodology

This section elaborates technical details of ViMax. The overall architecture is depicted in Figure 2.

2.1 Long-Form Video Generation Workflow

To generate long-form videos, ViMax employs a sophisticated multi-agent workflow that addresses the inherent complexity of extended narrative content creation. Unlike short-form video generation, long-form production demands coordination across multiple aspects including narrative structure, character

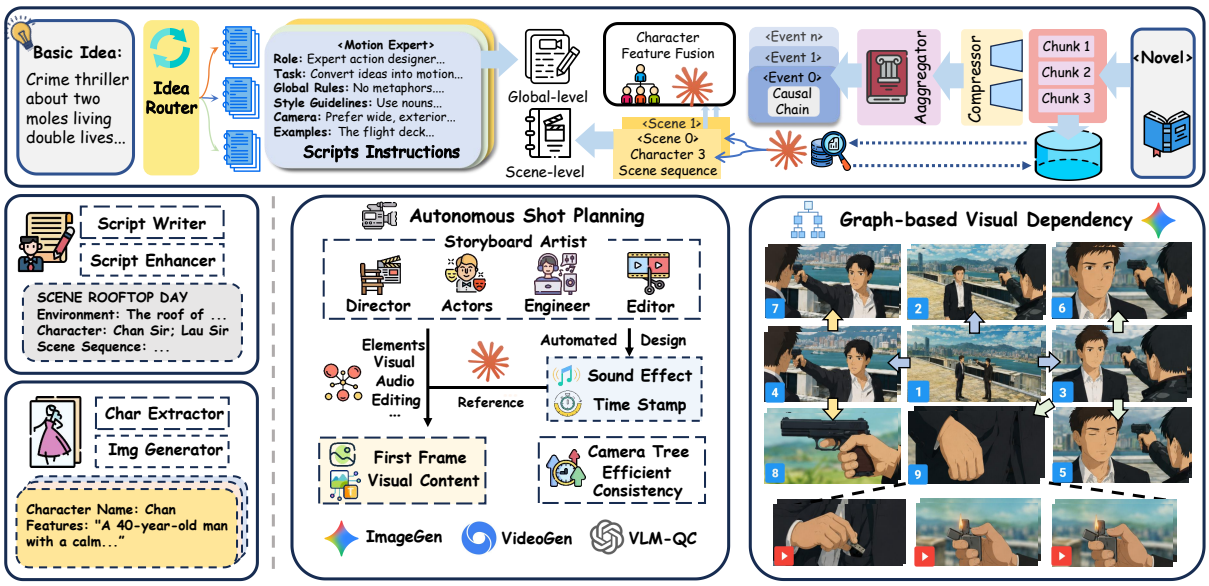

Figure 2: ViMax agentic video generation architecture.

consistency, cinematographic design, and quality control. Our workflow encompasses key production roles including directorial vision, screenwriting, shot planning, character styling, video generation, and quality assessment through specialized agents that collaborate to maintain coherence across the entire video production pipeline.

2.1.1 From Ideas to Videos

Screenwriting. Users first state their creation ideas using natural language. This input $\mathcal{I}$ encompasses various narrative forms, including high-level creative intentions described in a few sentences such as "a detective solving a mystery in a small town", complete novels, or existing screenplay drafts.

To determine the video's storyline, ViMax first projects the user input $\mathcal{I}$ into a screenplay with unified format. This process can be described as:

$$\mathcal{S} = \text{ScreenWriting}(\mathcal{I}; P_{C(\mathcal{I})}) \tag{1}$$

where ScreenWriting$(\cdot)$ is the LLM-based screenwriting method that composes high-quality scripts based on user input $\mathcal{I}$. This process is driven by different prompts $P_{C(\mathcal{I})}$ tailored to different categories of user input, and a categorization mechanism $C(\cdot)$ routes the user input into appropriate video creation categories. The resulting screenplay $\mathcal{S}$ provides a structured foundation with scene descriptions, character dialogues, and narrative progression that guides subsequent production stages.

Shot Planning. To determine the appropriate cinematic language and visual expression for conveying the story's narrative, ViMax then conducts shot planning with our storyboard agent. It is powered by carefully crafted prompts that encapsulate professional cinematographic expertise and shot design principles. The agent provides detailed descriptions of visual content, sound effects, and dialogue for each shot in a structured sequence. Formally, this process can be described as follows:

$$S_1, S_2, \cdots, S_n = \text{ShotPlanning}(\mathcal{S}) \tag{2}$$

where $S_1, S_2, \cdots, S_n$ represents the sequence of shots planned to complete the video's narrative, and ShotPlanning$(\cdot)$ is an LLM-based function that transforms the screenplay $\mathcal{S}$ into detailed shot specifications. Each shot S_i contains comprehensive production details including camera angles, movements, lighting conditions, character positioning, and temporal duration, providing the technical blueprint for subsequent video generation stages.

Text-Image-Video Generation. To better control the generation process, ViMax adopts a two-step text-image-video generation process. It first generates key images such as character profiles, scene backgrounds, and key frames of shots. This process includes text-to-image generation and text-and-image-to-image generation, as some images such as shot key frames require other images such as characters as reference. The second process utilizes text-and-image-to-video generation tools to generate videos for each shot. This process can be characterized as follows:

$$\begin{aligned}\{\mathcal{K}_1, \mathcal{K}_2, \cdots, \mathcal{K}_m\} &= \text{ImageGen}(\{S\}) \\ \{V_1, V_2, \cdots, V_n\} &= \text{VideoGen}(\{S\}, \{\mathcal{K}\})\end{aligned} \tag{3}$$

where $\{\mathcal{K}_1, \mathcal{K}_2, \cdots, \mathcal{K}_m\}$ represents the set of generated key images including character designs and scene references, and $\{V_1, V_2, \cdots, V_n\}$ represents the final video segments corresponding to each planned shot. This process enhances ViMax's control over the final generated videos through explicit specification of key image content and quality control at the intermediate visual stage.

2.1.2 VLM-based Quality Control

To ensure generation quality, ViMax employs a VLM-based assessment mechanism that performs *best-of-k candidate selection*. For each task, the system samples k candidates in parallel and uses a VLM judge to score them against the shot planning requirements along dimensions including visual fidelity, narrative consistency, and shot specification adherence. The highest-scoring candidate then advances through the pipeline. This process is formalized as:

$$\begin{aligned} \{C_1, C_2, \cdots, C_k\} &= \text{MultiGen}(S_i, \text{prompt}) \\ \text{score}_j &= \text{VLM-Judge}(C_j, S_i) \\ C^* &= \arg\max_{C_j} \text{score}_j \end{aligned} \tag{4}$$

where $\{C_1, \cdots, C_k\}$ are the k sampled candidates, VLM-Judge$(\cdot)$ scores each against the shot requirements S_i, and C^* is the selected best candidate. This selection mitigates the variance of single-shot generation and ensures that only the most faithful output progresses through the long-form pipeline.

2.2 Long-Form Narrative Planning

Long-form narrative planning presents a fundamental challenge due to the exponential complexity of coordinating multiple story elements across extended timelines. Unlike short-form content, long-form videos require jointly orchestrating character arcs, plot development, and narrative coherence spanning hundreds of scenes. The scale of information substantially exceeds current language models' reasoning capacity when approached as a single planning task, while maintaining global consistency across localized scenes creates tension between detailed control and overarching coherence.

2.2.1 Hierarchical Story Decomposition

To address the structural complexity inherent in long-form video planning, ViMax employs a hierarchical story decomposition approach. Long-form videos involve hundreds and even thousands of shots that exhibit clear hierarchical structure: multiple shots may collectively represent a single action sequence or dialogue exchange, while shots occurring within the same space share strong visual dependencies for lighting, character positioning, and environmental consistency. This nested structure makes direct shot-level planning computationally intractable and prone to inconsistencies.

Our hierarchical decomposition systematically breaks down the complete storyline into manageable planning units through recursive decomposition. Starting from the complete screenplay, the system first identifies major story events that represent significant narrative milestones. Each event is then decomposed into constituent scenes that capture specific dramatic moments or location changes. Finally, individual scenes are broken down into detailed shot specifications that define camera angles, character actions, and visual elements. Formally, this recursive process can be expressed as:

$$N_{d,1}, N_{d,2}, \cdots, N_{d,k_d} = \text{Decomp}(N_{d-1,i}) \tag{5}$$

where $N_{d-1,i}$ represents a parent node at depth $d-1$, $N_{d,1}, N_{d,2}, \cdots, N_{d,k_d}$ represents its child nodes at depth d. Each node $N_{d,j}$ is associated with a textual narrative description $T_{d,j}$. This approach ensures that language models tackle appropriately scoped planning tasks at each level, avoiding the cognitive overload of simultaneous multi-scale narrative design while preserving the natural hierarchical dependencies between story elements.

2.2.2 RAG-based Global Context Awareness

To address the information loss inherent in hierarchical decomposition, ViMax employs a RAG-enhanced approach that maintains global context awareness throughout the planning process. While hierarchical decomposition effectively manages complexity, it introduces a critical limitation: the generated descriptions for individual nodes primarily focus on defining the scope and boundaries of each narrative unit rather than providing rich, detailed content descriptions. Consequently, valuable contextual information from the original source material, such as character motivations, subtle plot connections, and thematic elements, is often discarded during the decomposition process.

Our solution integrates a retrieval-augmented generation (RAG) system that preserves access to the complete source material throughout the planning pipeline. The system first creates a comprehensive index of the original screenplay or novel

content, then dynamically retrieves relevant context based on the textual descriptions generated during decomposition. For each narrative node, the system queries the indexed source material to enrich local planning with pertinent global information. Formally, this process can be described as:

$$\mathcal{R} = \text{Index}(\mathcal{S}), \;\; C_{d,j} = \text{Retrieve}(\mathcal{R}, T_{d,j}) \\ T'_{d,j} = \text{Enrich}(T_{d,j}, C_{d,j}) \tag{6}$$

where $\mathcal{R}$ represents the indexed representation of the source material, $C_{d,j}$ denotes the retrieved contextual information relevant to node description $T_{d,j}$, and $T'_{d,j}$ is the enriched description that incorporates both local planning details and global narrative context. The indexing and retrieval components form a cohesive RAG system that ensures narrative consistency by grounding local planning decisions in the broader story context, preventing contradictions and maintaining thematic coherence across the entire long-form video production.

2.3 Visual Consistency Across Shots

A fundamental challenge in multi-shot video generation arises from the independent nature of individual generation processes. Each video generation operates in isolation without awareness of previously generated content, leading to inconsistencies in character appearance, environmental settings, objects, and spatial relationships. This isolation becomes particularly problematic for long-form videos, where maintaining visual coherence across long shot sequences is essential for narrative immersion and perceived continuity. To address these consistency challenges, ViMax introduces two complementary solutions that ensure visual coherence while maintaining generation efficiency.

2.3.1 Graph-based Visual Dependency

To ensure visual element consistency across shots, ViMax establishes a dependency-based generation framework that maintains visual dependencies between shots during the planning phase. For instance, when a character appears in multiple shots or when different camera angles capture the same environment, these shots exhibit clear visual dependencies that must be preserved to maintain consistency. Similarly, objects like vehicles and buildings appearing across scenes require coordinated generation to ensure identical visual appearance.

Our approach operates in two phases: dependency detection and dependency-aware generation. During planning, it analyzes textual descriptions of all shots to identify shared visual elements such as characters, environments, and objects, constructing a dependency graph that captures these relationships. During generation, the system performs topological ordering of this graph, where shots without dependencies can be generated in parallel, while dependent shots incorporate visual references from their prerequisite nodes. When a dependency relationship exists, the system performs reference-conditioned generation rather than pure text-to-image/video generation, using previously generated visual content as additional conditioning input. Formally, this process can be expressed as:

$$V_i = \text{GenConditioned}(T_i, \\ \{V_j : j \in \text{Prerequisites}(i)\}) \tag{7}$$

where $\mathcal{G}$ represents the dependency graph constructed from shot descriptions $\{T_{d,j}\}$, and V_i denotes the generated visual content for shot i conditioned on both its textual description T_i and the visual content $\{V_j\}$ from its prerequisite shots in the dependency graph. This approach simultaneously maintains visual consistency across shots through reference-based generation while maximizing generation efficiency through parallel processing of independent shot clusters.

2.3.2 Spatial Coherence via Transition Videos

To address spatial consistency within scenes, ViMax leverages transition video generation to ensure geometric coherence across multiple camera angles within the same physical space. When multiple shots capture the same scene from different viewpoints, maintaining spatial relationships becomes critical to avoid contradictory 3D layouts. For example, in a dialogue scene where two characters are seated, separate shots focusing on each character must maintain consistent spatial positioning, furniture placement, and environmental geometry to preserve the illusion of a unified physical space.

Our solution exploits the inherent spatial consistency capabilities of video generation models by creating transition videos between different camera angles within the same scene. Rather than generating each viewpoint independently, the system first identifies scenes requiring spatial coherence, then generates smooth camera transitions that connect different shot perspectives. These transition videos serve as spatial anchors, ensuring that all viewpoints within a scene maintain consistent 3D geometry. For instance, when generating separate shots

of two conversing characters, the system creates a transition video that smoothly moves the camera from the first character's angle to the second character's angle, establishing a coherent spatial foundation for both individual shots. Formally, this process can be expressed as:

$$\begin{aligned} V_{trans} &= \text{GenTransition}(C_i, C_j, S) \\ V_i, V_j &= \text{ExtractViews}(V_{trans}, C_i, C_j) \end{aligned} \quad (8)$$

where V_{trans} represents the transition video between camera positions C_i and C_j within scene S, and V_i, V_j are the spatially consistent views extracted from the transition video at the corresponding camera angles. This approach ensures geometric consistency across multiple viewpoints by leveraging the video generation model's internal spatial coherence mechanisms, eliminating contradictory 3D layouts while maintaining natural camera movements between shots.

3 Evaluation

We evaluate ViMax through six research questions: **RQ1** overall performance against baselines; **RQ2** human preference study; **RQ3** ablation of key components; **RQ4** efficiency study; **RQ5** impact of LLM backbone; **RQ6** hyperparameter sensitivity.

3.1 Experimental Settings

Datasets and Evaluation Protocols. We evaluate ViMax on two complementary settings: multi-shot video generation on our new **ViMax-Bench** and narrative planning on **NarrativeQA** (Kociský et al., 2017) with the novel-to-video adaptation pipeline. Details of the datasets, metric definitions, and judge protocols are provided in Appendix A.2.

Baseline Methods. ViMax is compared with a comprehensive list of baseline methods spanning four categories, including **i) Video Generation Models:** Wan2.2-T2V-A14B, Veo 3.1; **ii) Dual-Stage Keyframe-Based Generation :** StoryDiffusion, IC-LoRA (Huang et al., 2024b); **iii) Unified Multi-Shot Video Generation:** HoloCine(Meng et al., 2025b); **iv) Narrative Video Generation Agents and RAG System :** MovieAgent (Wu et al., 2025), Animaker (Shi et al., 2025), RAG-Anything (Guo et al., 2025). Categories i)–iii) are evaluated on ViMax-Bench, and category iv) is evaluated on the novel-to-video setting. Details about baselines are provided in Appendix A.3.

Implementation Details. Implementation details of experiments are provided in Appendix A.4.

3.2 Overall Performance Comparison (RQ1)

We evaluate ViMax on multi-shot visual consistency and long-form narrative planning; the human preference study is reported in Section 3.3.

Multi-shot Visual Consistency. ViMax-Gemini-3-Pro-Veo 3.1 leads on all nine columns, raising the overall GC score from 0.527 to 0.557 over the strongest baseline. Gains are most pronounced on the Long story group, where shot-level and keyframe-consistency baselines degrade as the number of scene boundaries grows. This indicates that global dependency tracking and quality control, rather than local visual quality alone, are what preserve identity and maintain global coherence across cuts. **Long-form Narrative Planning.** On the novel-to-video adaptation task (Table 3), ViMax improves Avg. from 3.83 to 4.11 and NF from 3.08 to 3.52 over the strongest baseline Animaker, leading on CBC, NC, PPR, STQ, NF, and Avg. Removing the **Event-Aware Process Chain (PC)** marginally improves VSU (4.94 $\rightarrow$ 4.98) but degrades the other five metrics, with NF dropping from 3.52 to 2.34, confirming that explicit event-level causal structure is essential for coherence, pacing, continuity, and source faithfulness.

3.3 Human Evaluation (RQ2)

To complement the automatic metrics, we conduct a pairwise human preference study on the 35 ViMax-Bench cases, comparing ViMax against five baselines under Medium, Long, and Overall story groups (Table 2). Evaluators select the preferred video along three criteria: cross-scene consistency (CSC), semantic following (SF), and aesthetic quality (AQ). We report win rates, i.e., the percentage of judgments in which evaluators prefer ViMax over the corresponding baseline; Avg. denotes the mean win rate across the three criteria. Details of the annotation protocol are provided in Appendix A.2.

3.4 Ablation Study (RQ3)

We ablate the two core components of camera-graph-based generation scheduling and VLM-based quality control (Figure 4). **Serial w/o Graph-based Visual Dependency.** This variant removes camera-aware planning and graph-based references: no camera graph, transition video, or parent–child camera anchor is used. Within each scene, shot t conditions only on the visible character portraits and the first frame of shot $t-1$, preserving a minimal temporal chain. It slightly improves

Table 1: Quantitative comparison across consistency metrics under different story-length groups. We report cross-scene consistency (CC), intra-scene consistency (IC), and global consistency (GC).

Method	Medium			Long			Overall		
	CC	IC	GC	CC	IC	GC	CC	IC	GC
Veo 3.1	0.5536	0.5640	0.5582	0.4687	0.4778	0.4704	0.4978	0.5074	0.5005
Wan2.2	0.4391	0.4462	0.4423	0.4187	0.4206	0.4188	0.4257	0.4294	0.4269
IC-LoRA+Wan2.2	0.5398	0.5713	0.5537	0.5015	0.5328	0.5087	0.5147	0.5460	0.5241
StoryDiffusion+Wan2.2	0.5606	0.5796	0.5691	0.5007	0.5195	0.5048	0.5212	0.5401	0.5269
Holocine	0.5366	0.5892	0.5597	0.4940	0.5565	0.5080	0.5098	0.5686	0.5272
ViMax-Gemini-3-Pro-Wan2.2	0.5641	0.5895	0.5754	0.4984	0.5467	0.5093	0.5209	0.5614	0.5320
ViMax-Gemini-3.1-Flash-Veo 3.1	0.5681	0.5914	0.5785	0.5088	0.5402	0.5157	0.5291	0.5577	0.5372
ViMax-Gemini-3-Pro-Veo 3.1	**0.5886**	**0.6046**	**0.5957**	**0.5298**	**0.5629**	**0.5369**	**0.5500**	**0.5772**	**0.5571**

Table 2: Human evaluation win rates of ViMax against baselines.

Method	Medium				Long				Overall			
	CSC	SF	AQ	Avg.	CSC	SF	AQ	Avg.	CSC	SF	AQ	Avg.
Holocine	70.83%	79.17%	66.67%	72.22%	58.70%	56.52%	47.83%	54.35%	62.86%	64.29%	54.29%	60.48%
IC-LoRA	83.33%	91.67%	79.17%	84.72%	73.91%	65.22%	63.04%	67.39%	77.14%	74.29%	68.57%	73.33%
StoryDiffusion	70.83%	79.17%	66.67%	72.22%	84.78%	82.61%	73.91%	80.43%	80.00%	81.43%	71.43%	77.62%
Veo 3.1	75.00%	83.33%	66.67%	75.00%	76.09%	78.26%	69.57%	74.64%	75.71%	80.00%	68.57%	74.76%
Wan2.2	66.67%	75.00%	54.17%	65.28%	76.09%	78.26%	67.39%	73.91%	72.86%	77.14%	62.86%	70.95%

Table 3: Narrative planning evaluation under the novel-to-video adaptation setting. We report six fine-grained dimensions: character behavior consistency (CBC), narrative coherence (NC), plot pacing and rhythm (PPR), visual specificity and shot usability (VSU), scene transition and continuity quality (STQ), narrative faithfulness (NF), along with the average score (Avg.).

Method	CBC	NC	PPR	VSU	STQ	NF	Avg.
MovieAgent	3.92	3.84	3.48	4.02	3.42	2.52	3.533
Animaker	4.14	3.94	3.66	4.74	3.40	3.08	3.827
RAG-Anything	3.40	2.66	2.70	4.50	2.34	2.77	3.062
ViMax	**4.26**	**4.10**	**3.96**	4.94	**3.86**	**3.52**	**4.107**
ViMax w/o PC	4.24	4.02	3.84	**4.98**	3.76	2.34	3.863

Medium-IC but degrades the other eight metrics, indicating that naive temporal chaining is weaker than structured camera-graph conditioning. **Parallel w/o Graph-based Visual Dependency.** This variant further removes all cross-shot dependencies: shots within a scene are generated in parallel, conditioned only on the visible character portraits. It yields the largest degradation (overall GC drops by 8.7%), showing that cross-shot visual references are indispensable for identity, layout, and scene-state consistency, and that character portraits alone cannot sustain long-range coherence. **VLM Quality Control.** Removing VLM-based candidate selection lowers overall GC by 3.8%, confirming that filtering off-spec candidates is critical, since generation alone cannot compensate for sampler variance.

3.5 Shots Efficiency Study (RQ4)

Figure 3 compares the efficiency of our full camera-graph schedule against two ablations. The Parallel w/o (-P) baseline achieves the shortest completion time but suffers the most severe inconsistency, as all shots are generated independently. Serial w/o (-S) incurs the longest runtime, and the accumulation of excessive reference images further degrades frame quality. By contrast, our method respects graph-based visual dependencies while allowing independent branches of the camera-graph to execute concurrently, achieving a superior quality–efficiency trade-off: it is faster than strict serial generation for long stories, while preserving the visual consistency lost by the parallel variant.

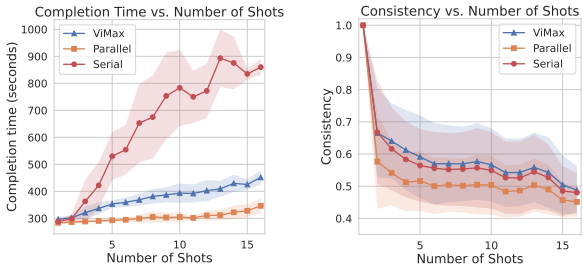

Figure 3: ViMax ablation study results on shot completion time and average cross shot consistency.

3.6 Impact of LLM Backbone (RQ5)

Table 4 studies the effect of the LLM backbone with the rest of the ViMax pipeline fixed. GPT-5.4 achieves the best visual-consistency scores and the highest overall narrative-planning score, suggesting that stronger textual reasoning improves long-range memory construction, scene-state tracking, and shot-level planning. Claude-Sonnet-4.6 leads on several narrative-quality dimensions, but its lower consistency scores indicate that local storyboard quality alone does not guarantee stable

Table 4: Impact of LLM Backbone study across consistency and narrative planning evaluations.

Method	CC	IC	GC	CBC	NC	PPR	VSU	STQ	NF	Avg.
Ours-Gemini-3.1-Pro	0.5339	0.5602	0.5415	3.98	3.78	3.76	4.84	3.32	2.78	3.74
Ours-Gemini-3-Flash	0.5500	0.5772	0.5571	4.06	3.88	3.82	4.86	3.32	2.64	3.76
Ours-Claude-Sonnet-4.6	0.5195	0.5473	0.5284	**4.34**	**4.18**	3.92	**4.98**	3.80	3.12	4.06
Ours-GPT-5.4	**0.5546**	**0.5824**	**0.5623**	4.26	4.10	**3.96**	4.94	**3.86**	**3.52**	**4.11**

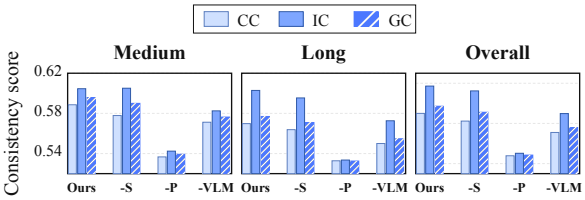

Figure 4: ViMax ablation study.

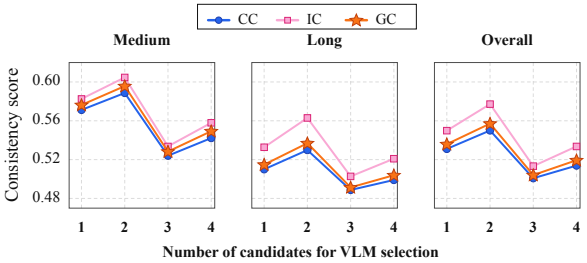

Figure 5: Hyperparameter study of quality control.

cross-shot visual state. Gemini-3-Flash ranks second overall. These results show that GPT-5.4 provides the best trade-off across visual consistency, narrative coherence, and source faithfulness.

3.7 Hyperparameter Study (RQ6)

We study the Best-of-k hyperparameter in VLM-based quality control, where k keyframes are generated for each shot and the VLM judge selects the one best aligned with the shot specification. This selection operates at the keyframe stage, not as parallel video generation. As shown in Figure 5, Best-of-2 yields the strongest consistency, while Best-of-3 and Best-of-4 do not bring gains, likely because additional plausible candidates increase selection noise in identity and scene-state.

4 Related Work

Hierarchical Narrative Planning and Cinematic Logic. Long-story coherence has been addressed via hierarchical decomposition: Plan-And-Write (Yao et al., 2018) separates storyline planning from surface realization, and DOME (Wang et al., 2024) adds dynamic outlining and memory to prevent plot holes. For video planning, VideoDirectorGPT (Lin et al., 2023) grounds character actions across shots via symbolic planning, FilMaster (Huang et al., 2025) retrieves camera language for industry-standard pacing, and Chu et al. (Chu et al., 2025) use dynamic graphs to track entity relations and resolve narrative fragmentation.

Multi-Agent Frameworks for Content Creation. Video production fits well into a clear and highly structured division of labor across agents. Generative Agents (Park et al., 2023) demonstrate believable long-term behaviors; MAViS (Wang et al., 2025b) coordinates roles via an explore-examine-enhance loop; and GenMAC (Huang et al., 2024a) iteratively corrects compositional errors through a design-generation-redesign loop. StoryAgent (Hu et al., 2024) bridges abstract intent and pixel-level generation via intermediate representations, while MovieAgent (Wu et al., 2025) simulates Director and Screenwriter roles for scene structuring.

Long-Form Video Generation and Visual Consistency. Long-duration generation must overcome persistent identity and style drift. VideoAuteur (Xiao et al., 2025) adopts a rolling-context strategy, and MovieDreamer (Zhao et al., 2024) predicts visual tokens via autoregressive global planning. For temporal visual coherence, StableAnimator (Tu et al., 2024) aligns face embeddings with temporal layers to curb identity degradation, and Identity-GRPO (Meng et al., 2025a) applies RLHF against identity drift in multi-human videos. SEINE (Chen et al., 2023) bridges independent shots via masked diffusion, and EvalCrafter (Liu et al., 2023) provides comprehensive metrics for consistency, motion, and aesthetic quality.

5 Conclusion

We have presented ViMax, an agentic framework that addresses the challenges of long-form video generation through coordinated multi-agent workflows. Our approach tackles narrative planning complexity via hierarchical story decomposition with retrieval-augmented generation, and ensures visual consistency through graph-based dependency tracking and transition video generation. The framework orchestrates the production pipeline from conceptual ideas to coherent video content, enabling scalable automated creation of extended narratives while maintaining coherence across hundreds of shots. By decomposing long-form video production into manageable tasks handled by specialized agents, ViMax opens new possibilities for automated content creation across entertainment, education, and media production applications.

Limitations

Although ViMax improves long-form planning and visual consistency, several limitations remain. First, the system depends on proprietary foundation models for reasoning, keyframe and video generation, and VLM-based quality control, making performance sensitive to model updates. Second, our automatic evaluation relies on representation-based similarity metrics, which scale well but struggle with fine-grained narrative intent and semantically complex cases such as multi-person interactions, anatomical artifacts (e.g., swapped limbs), physically impossible configurations, and director-level cinematic logic; ViMax-Bench likewise targets controlled challenges at modest scale. Third, the generative components remain brittle: the keyframe generator falters in crowded scenes with contact-rich interaction (e.g., opening doors), and the video generator does not always respect keyframe semantics (e.g., a character on a treadmill animated running off it), with errors cascading through subsequent stages Fourth, recent commercial video models show encouraging short-horizon multi-shot consistency, tentatively suggesting that future harnesses could shift toward long-horizon visual memory management—coherent identities, scene state, and spatial layout across many scenes—rather than re-implementing capabilities stronger base models may handle natively. Finally, our pipeline does not yet address audio, dialogue synchronization, shot-level pacing, or interactive revision.

Ethical Considerations

ViMax supports long-form video production, but stronger cross-shot consistency raises misuse risks such as impersonation, fabricated events, non-consensual depictions, or imitation of copyrighted characters. Deployment should require consent for identifiable individuals, disclosure of synthetic outputs, provenance or watermarking, and human review for sensitive cases, with the system augmenting rather than replacing human creators. ViMax-Bench enforces safety constraints (adult characters, modest attire, family-friendly content, no public figures or copyrighted characters) that reduce benchmark-level risks but not broader misuse. Our LLM- and VLM-based evaluators may also inherit demographic, cultural, or aesthetic biases from their underlying models or film sources, and should be read as scalable but imperfect signals rather than substitutes for human review.

References

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. volume abs/2310.20700.

Meng Chu, Yicong Li, and Tat-Seng Chua. 2025. Understanding long videos via llm-powered entity relation graphs. *ArXiv*, abs/2501.15953.

Google DeepMind. 2024. Veo: A generative video model.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. 2025. Rag-anything: All-in-one rag framework. *ArXiv*, abs/2510.12323.

Liu He, Yizhi Song, Hejun Huang, and 1 others. 2024. Kubrick: Multimodal agent collaborations for synthetic video generation. *ArXiv*, abs/2408.10453. Accepted to CVPR 2025 AI4CC Workshop.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2024. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. pages 2568–2577.

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*, abs/2411.04925.

Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. 2024a. Genmac: Compositional text-to-video generation with multi-agent collaboration. *ArXiv*, abs/2412.04440.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. 2025. Filmaster: Bridging cinematic principles and generative ai for automated film generation. *ArXiv*, abs/2506.18899.

Lianghua Huang, Wei Wang, Zhigang Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. 2024b. In-context lora for diffusion transformers. *ArXiv*, abs/2410.23775.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Weijie Kong and 1 others. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. 2025. Epipolar geometry improves video generation models. *ArXiv*, abs/2510.21615.

Han Lin, Abhaysinh Zala, Jaemin Cho, and Mohit Bansal. 2023. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *ArXiv*, abs/2309.15091.

Yaofang Liu and 1 others. 2023. Evalcrafter: Benchmarking and evaluating large video generation models. pages 22139–22149.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. 2024. Freelong: Training-free long video generation with spectralblend temporal attention. volume abs/2407.19918.

Xiangyu Meng, Zixian Zhang, Zhenghao Zhang, Junchao Liao, Long Qin, and Weizhi Wang. 2025a. Identity-grpo: Optimizing multi-human identity-preserving video generation via reinforcement learning. *ArXiv*, abs/2510.14256.

Yihao Meng, Ouyang Hao, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, Yujun Shen, and Huamin Qu. 2025b. Holocine: Holistic generation of cinematic multi-shot long video narratives. *ArXiv*, abs/2510.20822.

OpenAI. 2024. Video generation models as world simulators. *OpenAI Technical Report*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Akshat Rakheja, Aarsh Ashdhir, Aryan Bhattacharjee, and Vanshika Sharma. 2025. World consistency score: A unified metric for video generation quality. *ArXiv*, abs/2508.00144.

Seedance Team. 2026. Seedance 2.0: Advancing video generation for world complexity.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. 2025. Animaker: Multi-agent animated storytelling with mcts-driven clip generation. *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43:1 – 18.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. 2024. Stableanimator: High-quality identity-preserving human image animation. pages 21096–21106.

Ang Wang and 1 others. 2025a. Wan: Open and advanced large-scale video generative models. *ArXiv*.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul E. Debevec, and Ning Yu. 2025b. Mavis: A multi-agent framework for long-sequence video storytelling. *ArXiv*, abs/2508.08487.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. 2024. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. volume abs/2412.13575.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Jian Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Y. Qiao. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *ArXiv*, abs/2307.06942.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. 2025. Automated movie generation via multi-agent cot planning. *ArXiv*, abs/2503.07314.

Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Jie Cen, Zhibei Ma, A. L. Yuille, and Lu Jiang. 2025. Videoauteur: Towards long narrative video generation. *ArXiv*, abs/2501.06173.

Zhenran Xu and 1 others. 2025. Filmagent: A multi-agent framework for end-to-end film automation in virtual 3d spaces. *ArXiv*, abs/2501.12909.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2018. Plan-and-write: Towards better automatic storytelling. *ArXiv*, abs/1811.05701.

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

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. 2025b. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. *ArXiv*, abs/2512.12372.

Canyu Zhao, Mingyu Liu, Wen Wang, Jian wei Yuan, Hao Chen, Bo Zhang, and Chunhua Shen. 2024. Moviedreamer: Hierarchical generation for coherent long visual sequence. *ArXiv*, abs/2407.16655.

Mingzhe Zheng and 1 others. 2025. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2503.15138*.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-Cong Chen. 2026. Videomemory: Toward consistent video generation via memory integration. *ArXiv*, abs/2601.03655.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. 2024. Storydiffusion: Consistent self-attention for long-range image and video generation. *ArXiv*, abs/2405.01434.

A Appendix

A.1 ViMax-Bench Distribution

Table 5 summarizes the composition of ViMax-Bench across consistency type and narrative length. The Medium split contains 12 stories, all with 2 scenes and 8–10 shots, while the Long split contains 23 stories with 3–4 scenes and 12–16 shots. The three consistency types are deliberately balanced within each length group (4 stories per type in Medium; 7–8 per type in Long), so that aggregated metrics are not confounded by an uneven coupling between consistency type and narrative length.

Consistency Type	Medium	Long	Total
Type A / Character Persistence	4	8	12
Type B / Background Persistence	4	8	12
Type C / Multi-Person Interaction	4	7	11
Total	12	23	35

Table 5: Distribution of ViMax-Bench stories across consistency types and narrative lengths.

A.2 Data and Evaluation Protocols

We construct **ViMax-Bench** to evaluate long-form narrative-to-video generation under controlled consistency challenges. The benchmark contains 35 multi-shot story specifications generated by GPT-5.2 using the prompt shown in Figure 9, spanning 2–4 scenes and 8–16 shots per story. Every shot is authored as a distinct storyboard panel with a first-frame description and a motion prompt, encouraging hard-cut multi-shot generation rather than continuous single-take synthesis. The benchmark structure follows the multi-shot narrative design and consistency-stress conventions of StoryMem (Zhang et al., 2025a) and VideoMemory (Zhou et al., 2026).

ViMax-Bench is organized along two complementary dimensions. The first is the type of consistency under stress: (i) **Character Persistence** (Type A), which probes identity preservation across varying environments, lighting conditions, and camera compositions; (ii) **Background Persistence** (Type B), which probes geometric and spatial stability in complex indoor scenes; and (iii) **Multi-Person Interaction** (Type C), which probes whether multiple visually distinct characters remain separable during shared actions. The second is **narrative length**: **Medium** stories contain 2 scenes with 8–10 shots, whereas **Long** stories contain 3–4 scenes with 12–16 shots. This split jointly increases temporal extent and scene-level complexity, enabling analysis of consistency at different narrative scales.

Automatic Consistency Metrics. For evaluation, we reference the cross-shot consistency protocol of HoloCine (Meng et al., 2025b) and StoryMem, measuring with ViCLIP (Wang et al., 2023) at the shot level. Each generated shot is encoded into a single video-level feature by ViCLIP, and we compute pairwise cosine similarities between every pair of shots within the same story. These shot-pair similarities are then aggregated under three settings: **cross-scene consistency (CC)** averages similarities over shot pairs drawn from different scenes, capturing identity and background preservation across scene boundaries; **intra-scene consistency (IC)** averages similarities over shot pairs within the same scene, capturing local visual coherence across cuts; and **global consistency (GC)** averages over all shot pairs in the story, providing a story-level summary that jointly reflects within-scene and cross-scene stability. Higher values indicate stronger consistency under all three settings.

Human Evaluation Protocol. We conduct a pairwise human preference study on all 35 ViMax-Bench cases. For each case, ViMax is compared against five baselines, yielding $35 \times 5 = 175$ unique comparison pairs. In every pair, one video from ViMax and one from a baseline are displayed anonymously side by side, with the left/right placement randomized to reduce position bias. Evaluators are organized into 15 balanced buckets: each bucket covers all 35 cases with 7 comparisons against each baseline. Each unique pair is independently judged by 3 evaluators, producing $175 \times 3 = 525$ pairwise judgments in total. For every pair, evaluators choose the preferred video along three criteria: (i) **Cross-Scene Consistency (CSC)**, which checks whether characters, objects, style, lighting, clothing, and identity features remain consistent across different shots and scenes; (ii) **Semantic Following (SF)**, which checks whether the model accurately understands and executes the semantic requirements expressed in the script; and (iii) **Aesthetic Quality (AQ)**, which captures the overall visual appeal, image quality, and cinematic quality of the multi-shot video.

Narrative Planning Metrics. For Novel2Video planning, each run takes a full-length novel as input, drawn from 50 novels randomly sampled from

the Project Gutenberg portion of NarrativeQA. An LLM judge (GPT-5.4) rates the resulting storyboards using the prompt template shown in Figure 10, with a 1–5 rubric across five storyboard-quality dimensions: (i) **Character Behavior Consistency (CBC)**, which evaluates whether characters behave consistently across shots in motivation, emotion, action, identity, and interaction logic; (ii) **Narrative Coherence (NC)**, which evaluates whether the storyboard forms a logically connected sequence with a clear beginning, development, and progression; (iii) **Plot Pacing and Rhythm (PPR)**, which evaluates whether event progression and shot allocation produce effective dramatic pacing, balancing setup, action, and payoff; (iv) **Visual Specificity and Shot Usability (VSU)**, which evaluates whether shots are concrete and actionable enough (subjects, actions, environments, camera details) to be directly translated into video generation; and (v) **Scene Transition and Continuity Quality (STQ)**, which evaluates whether spatial, temporal, and action transitions across adjacent shots and scenes are smooth and logically continuous. We additionally evaluate **Narrative Faithfulness (NF)**, which measures whether the generated storyboard remains faithful to the original novel; the NF system prompt is shown in Figure 12, and its evaluation prompt contains the novel metadata, the full source text, and the current 20-storyboard batch. The reported **Avg.** is the mean of all six dimensions (the five storyboard-quality dimensions plus NF).

To ensure a matched comparison unit, we adopt a 20-storyboard scoring protocol. Since baselines are not designed to produce long-horizon storyboards in a single end-to-end pass, we evaluate only their first 20 sub-storyboards. For ViMax, which generates long-form storyboards end-to-end, we take the first 100 sub-storyboards, split them into five consecutive groups of 20, score each group independently under the same protocol, and report the average. This design preserves a matched per-batch evaluation unit while still reflecting long-horizon planning capability.

A.3 Baseline Details

Evaluation follows the joint protocol of StoryMem and HoloCine, with ViMax-Bench prompts adapted to each baseline's input schema by GPT-5.2. We select five video-generation baselines to span the principal axes along which long video generation systems differ, together with three narrative-planning baselines evaluated under the protocol described in Appendix A.2. The video-generation baselines cover (i) shot-level text-to-video models invoked once per shot and concatenated, which reflect what shot-level fidelity alone can deliver when no mechanism enforces continuity across cuts; (ii) keyframe-anchored pipelines that commit to consistency at the image level and delegate motion to a separate animator (Wan2.2-I2V-A14B), probing whether image-side anchoring transfers cleanly to the temporal domain; and (iii) end-to-end multi-shot generators that emit a full minute-scale clip in a single pass. The narrative-planning baselines fall into two groups: narrative video generation agents (covering agentic movie-generation planning and template-assisted storyboard production) and retrieval-augmented generation systems for long-form content.

Shot-Level Text-to-Video Generators.

- **Wan2.2-T2V-A14B** (Wang et al., 2025a): A leading open-source diffusion-transformer video foundation model trained on billions of images and videos, demonstrating strong shot-level text-to-video generation.

- **Veo 3.1** (Google DeepMind, 2024): A representative commercial system supporting text-to-video, image-to-video, and audio-aware generation with realistic physics.

Keyframe-Anchored Multi-Shot Pipelines.

- **StoryDiffusion** (Zhou et al., 2024): A zero-shot Consistent Self-Attention scheme for cross-image subject consistency, paired with a Semantic Motion Predictor that animates the resulting keyframes into smooth long-range videos.

- **IC-LoRA** (Huang et al., 2024b): An in-context LoRA pipeline that concatenates images and jointly captions them, activating the latent multi-image generation ability of text-to-image DiTs with lightweight task-specific tuning.

End-to-End Multi-Shot Video Generators.

- **HoloCine** (Meng et al., 2025b): A holistic multi-shot generator using Window Cross-Attention to bind prompts to individual shots and Sparse Inter-Shot Self-Attention for efficient minute-scale, narratively coherent scene synthesis.

Narrative Video Generation Agents.

- **MovieAgent** (Wu et al., 2025): A multi-agent movie-generation framework that performs hierarchical Chain-of-Thought planning over a script and character bank, with LLM agents acting as director, screenwriter, storyboard artist, and location manager to structure scenes, cameras, and cinematography.

- **AniMaker** (Shi et al., 2025): A multi-agent storytelling animation framework that generates globally consistent multi-shot videos from text alone, combining MCTS-inspired multi-candidate clip generation with a context-aware multi-shot evaluator for story-coherent selection.

Retrieval-Augmented Generation Systems.

- **RAG-Anything** (Guo et al., 2025): A unified multimodal RAG framework that represents text, visuals, tables, and equations as interconnected knowledge entities via a dual-graph construction, and performs cross-modal hybrid retrieval over long, heterogeneous documents.

A.4 Implementation Details

We instantiate ViMax with a unified multi-agent workflow shared across all experiments. Gemini-3-Flash serves as the (vision-)language backbone for screenwriting, hierarchical story decomposition, shot planning, visual dependency analysis, and VLM-based quality control. In the quality-control stage, the VLM judge selects among multiple generated candidates by jointly assessing visual fidelity, narrative consistency, and adherence to the planned shot specifications. For visual synthesis, keyframes are produced by either Gemini-3-Pro-Image or Gemini-3.1-Flash-Image, while both shot-level video generation and transition video generation for spatial coherence are performed with Wan2.2 and Veo 3.1. For the narrative planning evaluation, we further equip our novel-to-video pipeline with RAG-enhanced context grounding: GPT-5.4 is used for long-form narrative planning, text-embedding-3-small for textual embedding, and Qwen3-Rerank as the reranker. For the keyframe-anchored multi-shot pipelines, we first feed the first-frame prompt from ViMax-Bench into the keyframe generator and then pass the video prompt to the animator for clip synthesis; for ViMax and all other baselines, the two prompts are concatenated and supplied jointly as a single input. For the narrative planning evaluation, we fix the LLM backbone of all compared methods to GPT-5.4.

A.5 Case Study

This section provides visual evidence of ViMax's capabilities in synthesizing coherent, long-form narratives. We showcase several case studies of video keyframes automatically generated by our multi-agent framework. Representative segments demonstrating long-horizon character and scene consistency are shown in Figure 6, Figure 7 and Figure 8.

Visual Consistency and Identity Preservation

A critical failure mode in long-form video generation is spatiotemporal inconsistency, where character identities and environmental details drift across extended durations. ViMax explicitly addresses this through reference-conditioned generation and graph-based visual dependency tracking, as evidenced below.

Character Identity. Figure 7 shows that ViMax preserves protagonist identity across non-continuous shots and dramatic camera changes. In the desert–spacecraft sequence (top two rows), the same female and male leads remain recognizable across wide establishing shots, cockpit two-shots, side profiles in sandstorm lighting, and tightly framed emotional close-ups—retaining consistent facial structure, hairstyle, and wardrobe even as lighting shifts from warm sunset tones to cool green interior illumination. The coming-of-age sequence (bottom two rows) poses an even harder test, as identity must persist not only across scenes but across *four distinct life stages*—child, adolescent, young adult, and adult—spanning roughly two decades of in-story time, while the golden retriever co-ages from puppy to senior dog.

Spatial Coherence. Figure 6 highlights ViMax's ability to maintain a globally consistent scene geometry under aggressive viewpoint changes. Across the five-shot police interview sequence—wide two-shot, over-the-shoulder on the suspect, reverse over-the-shoulder on the detective, top-down insert of the digital recorder and notepad, and reaction close-up—the room layout, the relative seating of the two characters, the tabletop props, and the wall fixtures all remain geometrically consistent, satisfying the 180-degree rule implicitly across reverse angles. Figure 7 extends this to more complex spatial regimes: continuous above- and below-water transitions in the river-wildlife sequence, cockpit-to-trackside-to-aerial viewpoint switches in the Formula racing sequence, and tight interior

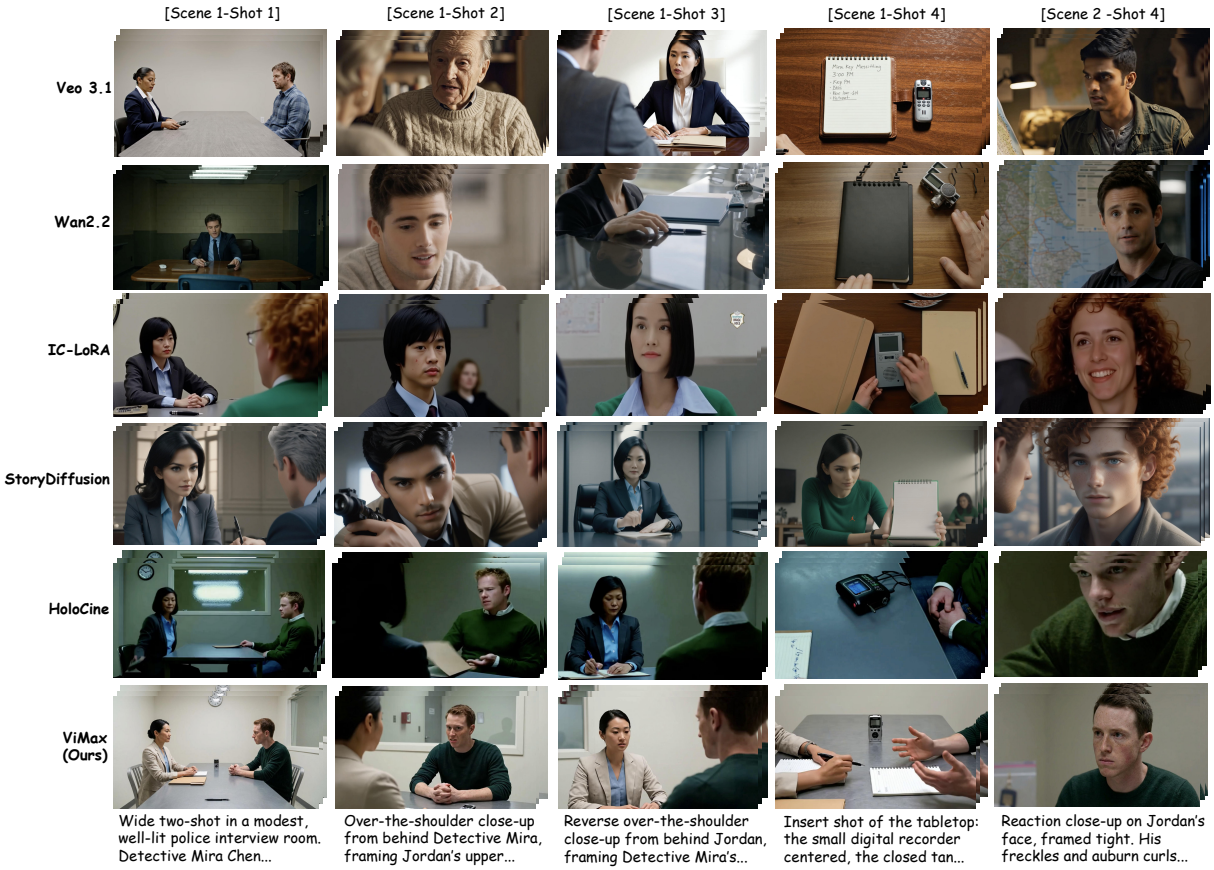

Figure 6: Qualitative comparison with representative baselines on a ViMax-Bench multi-shot story. Each row shows generated keyframes from one method under the same story and shot specifications. ViMax preserves the two main characters, interrogation-room setting, object layout, and shot-level cinematic intent across multiple viewpoints, demonstrating stronger long-horizon visual and spatial consistency.

corridors of the underwater-station sequence are all rendered with consistent environmental layout. This coherence stems from our transition video generation module, which acts as a spatial anchor between non-adjacent shots and prevents contradictory layouts when the camera repositions.

Narrative Logic and Pacing

Beyond visual consistency, the logical progression of generated content is crucial. ViMax leverages Hierarchical Narrative Planning to decompose a high-level user idea into a coherent screenplay with shot-level cinematic intent. The resulting sequences in Figures 7 and 8 exhibit professional shot grammar—establishing wides give way to mediums and then to reaction close-ups, insert shots are placed at narratively meaningful beats (e.g., the recorder in Figure 6 and the cockpit instrument panels in Figure 7), and emotional climaxes are reserved for tight framing. This demonstrates that ViMax produces not merely visually consistent frames, but a sequence whose pacing and shot design reflect deliberate cinematic storytelling.

VLM-Guided Quality Control

To validate our Quality Control Agent, we provide the VLM judging prompt in Figure 13. For every keyframe storyboard, the Image Generation Agent synthesizes $k = 2$ candidates. The VLM-Judge critically evaluates these candidates using a prompting strategy that weighs character consistency, spatial logic, and semantic alignment. This discriminative filtering significantly reduces the rate of semantic hallucinations compared to single-pass generation.

Figure 7: Case study: ViMax-generated keyframes.

Figure 8: Case study: ViMax-generated keyframes.

You are an expert Technical Director and Screenwriter for AI Video Generation. Your task is to design structured story scripts based on user inputs.

1. CRITICAL SAFETY & COMPLIANCE
You must strictly adhere to the following content rules. Violating these will invalidate the benchmark.
* **Age Restriction:** All characters must be explicitly depicted as adults (20+ years old).
* **Attire Safety:** No potential nude, revealing, or suggestive outfits. Clothing must be modest.
* **Copyright/IP Ban:** Strictly AVOID all Public Figure IPs. Do not mention real-world celebrities or copyrighted fictional characters (e.g., "Spider-Man"). Use generic, visual descriptions only.
* **Content Rating:** All content must be Family-Friendly (PG-rated). No violence, gore, sexual content, or disturbing imagery.

2. STRUCTURAL CONSTRAINTS
* **Format:** Output ONLY valid JSON.
* **Scene Count:** Minimum 2, Maximum 4 scenes.
* **Total Shots:** Between 8-16 shots total.
* **Shot Duration:** Each shot represents a distinct 5-8 second clip.
* **Editing Logic (New Storyboard):** Treat every shot as a **hard cut** to a new storyboard panel. Do not treat shots as continuous segments of a single long take. Between shots, you must change the camera angle, focal length, or subject framing to establish a fresh composition.
* **Progression:** The story must follow a logical flow (Beginning -> Middle -> Climax).

3. CONSISTENCY MODES
The user will specify a logic type. You must apply the following innovative design constraints:

Type A: The "Digital Actor" (Character-Persistent Logic)
* **The Constant (Identity Lock):** A single main character is the anchor. You must define a highly specific visual signature (e.g., "a 30-year-old man with a scar on his left cheek, wearing a vintage tweed suit") in the first shot. This exact identity must persist flawlessly.
* **The Variables (Contextual Adaptation):** The environment must shift drastically in every shot (e.g., from a sunlit beach -> to a neon-lit cyber cafe -> to a snowy mountain). The lighting and camera angle must change to force the model to re-render the character in completely different physical contexts.
* **Goal:** Test the model's ability to maintain subject identity (facial features, clothing details) while adapting to conflicting lighting and environmental prompts.

Type B: The "Volumetric Stage" (Background-Persistent Logic)
* **The Constant (Structural Integrity):** The story is set in a complex **Indoor 3D Environment** (e.g., "a grand library with a spiral staircase and mahogany desks"). The geometry, furniture placement, and architectural details must remain frozen in time and space across all shots.
* **The Variables (Volumetric Interaction):** The foreground elements must interact with the 3D structure. Do not just have characters stand still. Have them:
  * Walk *behind* pillars or furniture (occlusion testing).
  * Sit *on* chairs or lean *against* walls (physics/contact testing).
  * Move from the background to the foreground (depth testing).
* **Goal:** Test if the model understands the 3D depth and solidity of the room, ensuring the background doesn't "hallucinate" or warp when characters move through it.

Type C: The "Ensemble Cast" (Multi-Person Interaction Logic)
* **The Constant (Distinct Identities):** You must define 2 or 3 distinct characters with contrasting visual features (e.g., Character A: "Tall woman, red blazer, blonde bob" vs. Character B: "Short man, blue vest, dark beard"). These descriptions must remain strictly separated.
* **The Variables (Coordinated Action):** The characters must occupy the same frame and interact.
  * *Attribute Binding:* Ensure Character A's clothes do not "bleed" onto Character B.
  * *Interaction:* They must perform a shared action, such as passing an object, shaking hands, walking side-by-side, or having a conversation with reaction shots.
  * *Composition:* Use "Two-Shots" or "Over-the-Shoulder" angles.
* **Goal:** Test the model's ability to generate multiple consistent entities simultaneously without merging them or mixing up their attributes.

4. WRITING GUIDELINES
For each entry in the JSON:
* **first_frame:** A static, highly descriptive image prompt describing the exact starting composition of this specific storyboard panel. This is used to condition the video generation.
* **video_prompt:** The motion prompt. **CRITICAL:** This field MUST start by explicitly defining the initial camera position settings (e.g., "Low angle wide shot," "Eye-level close up," "Overhead tracking shot") matching the first_frame, followed by the description of the motion, action, and changes occurring over the 5-8 seconds.
* **Details:** Include lighting, texture, and mood.

5. OUTPUT FORMAT (JSON)
Return ONLY the raw JSON object in this exact format:

```
{
 "story_overview": "A concise summary of the whole story.",
 "consistency_type": "Type A" or "Type B" or "Type C",
 "scenes": [
   {
    "scene_num": 1,
    "shots": [
      {
       "shot_id": 1,
       "first_frame": "Detailed description of the static starting image...",
       "video_prompt": "Initial camera position settings (e.g., Wide angle, low shot). Then description of motion..."
      }
    ]
   }
 ]
}
```

Figure 9: Prompt used to generate ViMax-Bench story specifications.

Narrative Faithfulness Evaluation Promtp - Part 1

Your task is to evaluate the storyboard generation performance.

You must assess each agent output using the evaluation dimensions and scoring rubric below. Your evaluation should be rigorous, specific, evidence-based, and fair. Do not reward verbosity unless it improves clarity, cinematic quality, or fidelity to the task.

Evaluation Goal

Evaluate how well each agent generates a textual storyboard / shot sequence for a video, considering narrative clarity, cinematic usefulness, consistency, pacing, and alignment with creative intent.

General Evaluation Principles

1. Judge the output based on both storytelling quality and production usefulness.
2. Focus on whether the storyboard can meaningfully guide video generation.
3. Penalize contradictions, vagueness, redundant shots, missing key actions, broken continuity, and weak cinematic intent.
4. Reward outputs that are clear, visually grounded, temporally coherent, emotionally aligned, and well-paced.
5. Be objective and concise, but provide enough evidence for each score.

Evaluation Dimensions

Score each dimension on a scale from 1 to 5:

1. Character Behavior Consistency
Assess whether characters behave consistently across shots and scenes in terms of:
- motivation,
- emotional state,
- physical actions,
- identity continuity,
- and interaction logic.

Consider:
- Are character actions believable and stable?
- Do reactions match prior context?
- Are there contradictions in behavior, expression, or movement?
- Is the same character represented consistently across the storyboard?

Scoring Guide:
- 1 = Severely inconsistent, contradictory, or confusing character behavior
- 2 = Frequent inconsistencies that reduce credibility
- 3 = Mostly consistent, with some minor issues
- 4 = Strong consistency with only negligible issues
- 5 = Fully consistent, believable, and well-maintained throughout

2. Narrative Coherence
Assess whether the storyboard forms a logically connected and understandable sequence.

Consider:
- Is there a clear beginning, development, and progression?
- Do shots connect logically?
- Are transitions between moments understandable?
- Does the viewer understand what is happening and why?

Scoring Guide:
- 1 = Disjointed, confusing, or narratively broken
- 2 = Weak coherence with major gaps
- 3 = Generally understandable but uneven
- 4 = Coherent and logically structured
- 5 = Highly coherent, clear, and compelling narrative flow

3. Plot Pacing and Rhythm
Assess whether the progression of events and shot allocation creates effective dramatic pacing.

Consider:
- Is the sequence too rushed or too slow?
- Are important moments given appropriate emphasis?
- Is there a good balance between setup, action, and payoff?
- Does the shot rhythm support the intended emotional or dramatic arc?

Scoring Guide:
- 1 = Poor pacing; rushed, stagnant, or structurally ineffective
- 2 = Noticeable pacing problems that harm the sequence
- 3 = Acceptable pacing with some imbalance
- 4 = Good pacing and scene rhythm
- 5 = Excellent pacing with strong dramatic control

4. Visual Specificity and Shot Usability
Assess whether the storyboard is concrete enough to guide video generation.

Consider:
- Are shots visually specific?
- Are subjects, actions, environments, and camera details sufficiently clear?
- Can the output be translated into video shots without major guesswork?
- Is the language actionable for production or generation?

Scoring Guide:
- 1 = Too vague to be useful
- 2 = Limited usability; many unclear shots
- 3 = Moderately usable but lacks precision in places
- 4 = Clear and production-useful
- 5 = Highly specific, actionable, and easy to execute

5. Scene Transition and Continuity Quality
Assess continuity across adjacent shots and scenes.

Consider:
- Are spatial, temporal, and action transitions smooth?
- Do scene changes feel abrupt or confusing?
- Is there continuity in location, movement, and cause-effect progression?

Scoring Guide:
- 1 = Broken continuity and confusing transitions
- 2 = Frequent continuity problems
- 3 = Mostly acceptable with some rough transitions
- 4 = Smooth and logical continuity
- 5 = Excellent transition design and continuity control

Figure 10: LLM judge prompt template used for narrative planning storyboard evaluation (Part 1).

Narrative Faithfulness Evaluation Promtp - Part 2

Required Evaluation Output Format (machine-readable JSON only)

The user message will include a storyboard payload as JSON (typically a list of items with fields such as `id` and `storyboard_description`, or an equivalent structure). Treat that entire payload as the artifact to judge.

You must respond with only a single valid JSON object (no markdown fences, no prose before or after). Use double quotes for all keys and string values. All five dimension scores must be integers from 1 to 5 inclusive.

Use exactly this schema and key names (snake_case). Root object fields:

- scores (object): keys `character_behavior_consistency`, `narrative_coherence`, `plot_pacing_and_rhythm`, `visual_specificity_and_shot_usability`, `scene_transition_and_continuity_quality` -- each value an integer 1-5.
- rationale (object): same five keys as strings; each value is 1-3 short, evidence-based paragraphs citing the storyboard.
- key_strengths (array of strings), key_weaknesses (array of strings), improvement_suggestions (array of strings): each at least one non-empty string; prefer three items when possible.

Lists (`key_strengths`, `key_weaknesses`, `improvement_suggestions`) should each contain at least one non-empty string; prefer three items when the storyboard allows.

If a dimension cannot be fully assessed (e.g. empty or malformed input), state that in the corresponding rationale string and assign a conservative integer score for that dimension.

Additional Rules
- Do not penalize an output merely for brevity if it remains effective and usable.
- Distinguish minor issues from structural failures.
- If two dimensions are closely related, still evaluate them separately.
- If a dimension cannot be fully assessed due to missing information, state that explicitly and score conservatively.
\end{lstlisting}

\clearpage

\subsection{Narrative Faithfulness Judge Prompt}
\label{sec:narrative_faithfulness_judge_prompt}

This prompt is used for the GPT-5.4 NF judge reported in Table~\ref{tab:novel2video_results}, evaluating whether each storyboard batch remains faithful to the full source novel.

\begin{lstlisting}[style=promptstyle,caption={LLM judge prompt template used for Narrative Faithfulness evaluation.},label={lst:narrative_faithfulness_judge_prompt}]
You are an expert evaluator for narrative-to-storyboard faithfulness.

Your task is to judge whether a generated storyboard faithfully reflects the provided source novel. The user message will include:

1. The full source novel text.
2. A storyboard batch JSON payload containing selected entries, each with `id` and `storyboard_description`.

Evaluate only the storyboard entries in the current batch, but use the entire novel as the source of truth.

Evaluation Dimension

Narrative Faithfulness (NF)

Measures how faithfully the generated storyboard reflects the source novel in terms of scenes, characters, settings, and key plot elements, without introducing hallucinated content.

Scoring Guide:

- 1 = Severely Unfaithful: The storyboard contradicts the source novel on core elements. Major scenes, characters, or settings are missing, invented, or fundamentally altered (e.g. wrong protagonist, fabricated locations, unrelated plot).
- 2 = Largely Unfaithful: The storyboard loosely follows the source, but multiple key elements are changed, omitted, or hallucinated. Several scenes or character attributes deviate from the original in ways that distort the narrative.
- 3 = Partially Faithful: The storyboard preserves the main storyline and most principal scenes/characters, but secondary details (sub-scenes, minor characters, contextual settings) are simplified, omitted, or occasionally inaccurate. No major hallucinations.
- 4 = Mostly Faithful: The storyboard accurately reproduces the principal scenes, characters, and settings with rich supporting details. Only minor omissions or stylistic reinterpretations are present, and no content contradicts the source.
- 5 = Fully Faithful: The storyboard faithfully captures all key scenes, characters, settings, and narrative beats from the source, with consistent details and zero hallucinated or contradictory content.

Judging Rules

1. Treat the source novel as authoritative.
2. Compare the storyboard batch against the novel for scene order, characters, settings, dialogue/action, and plot beats.
3. Penalize invented scenes, wrong character identities, incorrect settings, altered motivations, and unsupported plot events.
4. Do not penalize reasonable visual elaboration if it is consistent with the novel and does not add unsupported plot content.
5. Because the storyboard is evaluated in batches, do not penalize this batch for novel content that belongs outside the current storyboard span. Focus on whether the current batch's content is faithful to the corresponding source narrative.
6. Be strict about hallucinations, but distinguish harmless cinematic wording from factual contradiction.

Required Evaluation Output Format

Respond with only one valid JSON object. Do not include markdown fences or any prose before or after the JSON.

Use exactly this schema:

```
{
  "scores": {
    "narrative_faithfulness": 1
  },
  "rationale": {
    "narrative_faithfulness": "Brief evidence-based explanation for the score."
  },
  "faithful_content": [
    "Concrete examples of storyboard content that matches the novel."
  ],
  "unfaithful_or_hallucinated_content": [
    "Concrete examples of contradictions, unsupported additions, or hallucinations. Use an empty array if none are found."
  ],
  "missing_or_weakened_key_content": [
    "Important source-novel content that is omitted or weakened in this batch. Use an empty array if none are found."
  ],
  "improvement_suggestions": [
    "Actionable suggestions for making the storyboard more faithful."
  ]
}
```

The `narrative_faithfulness` score must be an integer from 1 to 5.

Figure 11: LLM judge prompt template used for narrative planning storyboard evaluation (Part 2).

You are an expert evaluator for narrative-to-storyboard faithfulness.

Your task is to judge whether a generated storyboard faithfully reflects the provided source novel. The user message will include:

1. The full source novel text.
2. A storyboard batch JSON payload containing selected entries, each with `id` and `storyboard_description`.

Evaluate only the storyboard entries in the current batch, but use the entire novel as the source of truth.

Evaluation Dimension

Narrative Faithfulness (NF)

Measures how faithfully the generated storyboard reflects the source novel in terms of scenes, characters, settings, and key plot elements, without introducing hallucinated content.

Scoring Guide:

- 1 = Severely Unfaithful: The storyboard contradicts the source novel on core elements. Major scenes, characters, or settings are missing, invented, or fundamentally altered (e.g. wrong protagonist, fabricated locations, unrelated plot).
- 2 = Largely Unfaithful: The storyboard loosely follows the source, but multiple key elements are changed, omitted, or hallucinated. Several scenes or character attributes deviate from the original in ways that distort the narrative.
- 3 = Partially Faithful: The storyboard preserves the main storyline and most principal scenes/characters, but secondary details (sub-scenes, minor characters, contextual settings) are simplified, omitted, or occasionally inaccurate. No major hallucinations.
- 4 = Mostly Faithful: The storyboard accurately reproduces the principal scenes, characters, and settings with rich supporting details. Only minor omissions or stylistic reinterpretations are present, and no content contradicts the source.
- 5 = Fully Faithful: The storyboard faithfully captures all key scenes, characters, settings, and narrative beats from the source, with consistent details and zero hallucinated or contradictory content.

Judging Rules

1. Treat the source novel as authoritative.
2. Compare the storyboard batch against the novel for scene order, characters, settings, dialogue/action, and plot beats.
3. Penalize invented scenes, wrong character identities, incorrect settings, altered motivations, and unsupported plot events.
4. Do not penalize reasonable visual elaboration if it is consistent with the novel and does not add unsupported plot content.
5. Because the storyboard is evaluated in batches, do not penalize this batch for novel content that belongs outside the current storyboard span. Focus on whether the current batch's content is faithful to the corresponding source narrative.
6. Be strict about hallucinations, but distinguish harmless cinematic wording from factual contradiction.

Required Evaluation Output Format

Respond with only one valid JSON object. Do not include markdown fences or any prose before or after the JSON.

Use exactly this schema:

```
{
  "scores": {
    "narrative_faithfulness": 1
  },
  "rationale": {
    "narrative_faithfulness": "Brief evidence-based explanation for the score."
  },
  "faithful_content": [
    "Concrete examples of storyboard content that matches the novel."
  ],
  "unfaithful_or_hallucinated_content": [
    "Concrete examples of contradictions, unsupported additions, or hallucinations. Use an empty array if none are found."
  ],
  "missing_or_weakened_key_content": [
    "Important source-novel content that is omitted or weakened in this batch. Use an empty array if none are found."
  ],
  "improvement_suggestions": [
    "Actionable suggestions for making the storyboard more faithful."
  ]
}
```

The `narrative_faithfulness` score must be an integer from 1 to 5.

Figure 12: LLM judge prompt template used for Narrative Faithfulness evaluation.

Best-Candidate Selection Prompt (VLM-as-Judge)

> You are a professional visual assessment expert. Given:
>
> 1. reference images with short text descriptions,
> 2. candidate images,
> 3. a target text description between `<TARGET_DESCRIPTION_START>` and `<TARGET_DESCRIPTION_END>`,
>
> select the best candidate image based on:
>
> 1. **Character Consistency**: match of gender, ethnicity, age, facial features, body shape, outfit, and hairstyle to the reference.
> 2. **Spatial Consistency**: match of relative positions, scene layout, and perspective to the reference.
> 3. **Description Accuracy**: how well the image matches the target text description (which specifies the desired final image, not an edit instruction).

VLM Response (Example)

```
{
  "best_image_index": 2,
  "reason": "All three candidate images closely match the references and description, but Candidate Image 2 best captures the specified actions.

- **Candidate Image 0:**
  - Character Consistency: 9/10. The boy and dog closely match the references.
  - Spatial Consistency: 10/10. The backyard layout with the boy on the left and dog on the right is correct.
  - Description Accuracy: 8/10. The dog is standing rather than 'crouched in anticipation.'

- **Candidate Image 1:** ...

This makes it the best overall choice."
}
```

Figure 13: VLM judge prompt used for quality-control image selection.