# VIMAX: AGENTIC VIDEO GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long-form video generation requires both systematic narrative planning and visual consistency that current short-clip methods cannot provide. Existing methods generate isolated sequences without narrative structure and lack mechanisms for maintaining character and environmental consistency across scenes. We present ViMax, an agentic video generation framework that addresses video creation through coordinated multi-agent collaboration where specialized components negotiate narrative decisions, visual continuity, and production quality. Our framework employs a hierarchical narrative engine with retrieval-augmented generation for global story coherence and a dependency-aware visual consistency mechanism that tracks character and environmental states across temporal boundaries, while VLM-guided agents continuously monitor and refine both narrative coherence and visual fidelity. The framework enables coordinated agent collaboration to generate extended narrative content from initial concepts to video sequences. This maintains both storytelling integrity and visual coherence across multi-scene timelines.

## 1 INTRODUCTION

Current video generation methods face limitations, producing only short clips without narrative structure while character appearances and environmental elements change unpredictably across scenes. Recent breakthroughs in generative AI offer transformative potential to address these challenges by automating video production pipelines through agentic frameworks, enabling creators to transform conceptual ideas into coherent long-form visual narratives with unprecedented ease. Each shot must advance the story while preserving the characters, environments, and events established earlier.

Research in video generation has explored directions to address these challenges. Text-to-video models such as Sora (OpenAI, 2024), Veo 3.1 (Google DeepMind, 2025), and Seedance (Team Seedance et al., 2026), together with open-source efforts like HunyuanVideo (Kong et al., 2024) and Wan (Team Wan, 2025), have achieved per-clip fidelity. However, these models primarily target short-form content with limited narrative planning. When extended to longer horizons, they suffer from *semantic drift*, where character identity, world geometry, and physical laws drift across shots (Kupyn et al., 2025; Rakheja et al., 2025), breaking narrative immersion. High visual quality within clips therefore does not guarantee coherent progression across an entire story. Maintaining this coherence requires carrying character states, spatial relationships, and narrative context across shot boundaries.

Two approaches have emerged to tackle temporal consistency, yet both face limitations. Streaming methods (StreamingT2V (Henschel et al., 2025), FreeLong (Lu et al., 2024)) extend clip duration but lack explicit long-form narrative planning. Consistency mechanisms such as StoryDiffusion (Zhou et al., 2024), ConsiStory (Tewel et al., 2024), and STAGE (Zhang et al., 2025b) preserve subject identity through shared attention or keyframe anchoring but lack narrative planning. Multi-agent frameworks including FilmAgent (Xu et al., 2025), Kubrick (He et al., 2024), and VideoGen-of-Thought (Zheng et al., 2024) decompose production into specialized roles but fail to coordinate narrative, visual, and temporal constraints. Existing systems struggle to maintain global coherence across dozens of shots required for long-form storytelling. This calls for a unified workflow that coordinates narrative planning, visual dependencies, and quality control across consecutive shots.

Addressing these limitations requires overcoming two fundamental challenges:

- **Multi-shot Visual Consistency.** Current video generators operate on isolated shots without awareness of previous content, causing character appearance, environment design, and spatial



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




Figure 1: Multi-shot video examples generated by ViMax on ViMax-Bench.

layout to drift. Long-form production demands object-level consistency to preserve character and environment properties, and spatial coherence to maintain consistent 3D geometry across angles.

- **Long-form Narrative Planning.** Orchestrating narratives spanning dozens of shots exceeds language models' reasoning capacity when treated monolithically. The system needs to coordinate character arcs, plot development, and coherence while keeping decisions grounded in story context.

To address these challenges, we present ViMax, an agentic framework that orchestrates long-form video production from conceptual ideas to fully rendered content through coordinated agent collaboration. ViMax decomposes this pipeline into specialized agents for screenwriting, shot planning, character styling, video generation, and VLM-based quality control with best-of-$n$ selection.

For narrative complexity, ViMax employs hierarchical story decomposition with retrieval-augmented generation. This keeps local planning grounded in global context. For visual consistency, ViMax introduces a graph-based dependency system that tracks cross-shot relationships during planning. Transition video generation maintains spatial coherence across camera angles within scenes.

Our primary contributions in this work include:

- **Agentic Long-form Video Generation Framework.** We introduce a multi-agent architecture that coordinates screenwriting, shot planning, character styling, video generation, and VLM-based quality control. This framework automates narrative video creation from conceptual ideas.
- **Hierarchical Narrative Planning with Global Context.** We propose hierarchical story decomposition paired with retrieval-augmented generation. This enables language models to maintain thematic coherence and long-range dependencies across extended multi-scene narratives.
- **Graph-based Visual Consistency.** We employ dual-level mechanisms. Graph-based dependency tracking ensures object consistency, while transition video generation maintains spatial coherence.

For the reproducibility of our results, we provide our ViMax code in an anonymous repository 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.

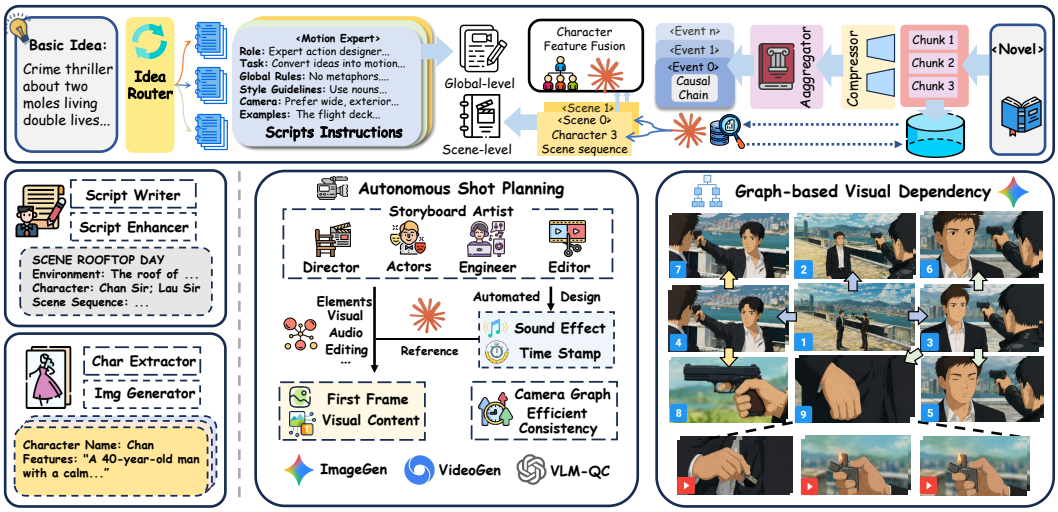


Figure 2: The ViMax agentic video generation architecture.

### 2.1 Long-Form Video Generation Workflow

To generate long-form videos, ViMax employs a multi-agent workflow that addresses the complexity of narrative content creation. Unlike short-form video generation, long-form production demands coordination across aspects including narrative structure, character consistency, cinematographic design, and quality control. Our workflow encompasses 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:

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

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

**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. The two generation stages are formally expressed as follows:

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

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 $\{V_1, V_2, \cdots, V_L\}$ 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-$n$ candidate selection*. For each task, the system samples $n$ 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 for subsequent processing. This process is formalized as:

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

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

### 2.2 LONG-FORM NARRATIVE PLANNING

Long-form narrative planning requires coordinating character arcs, plot development, and narrative coherence across scenes. As stories grow, planning all details at once becomes difficult, motivating a hierarchy that balances local control with global context throughout the full production workflow.

#### 2.2.1 HIERARCHICAL STORY DECOMPOSITION

To organize multi-shot narratives, ViMax uses hierarchical story decomposition. Shots form action sequences or dialogue exchanges, while shots within a scene share dependencies in lighting, character positioning, and environment. This structure motivates planning events and scenes before shots.

Our hierarchical decomposition systematically breaks down the complete storyline into manageable planning units through recursive decomposition. Starting from the complete screenplay, the system identifies major story events. Each event contains a summary and an *Event-Aware Process Chain (PC)*: an ordered list of intermediate actions, decisions, and causal steps describing how the event unfolds. This event representation guides decomposition into scenes capturing specific dramatic moments or location changes. Scenes are then decomposed into shots specifying camera angles, character actions, and visual elements while retaining context from the parent event and scene:

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

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 hierarchy scopes each planning task while preserving 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, may be omitted during compression and decomposition.

Our RAG system grounds scene generation in the original novel $\mathcal{I}$. We first index the novel to preserve access to its detailed context. For each event, the steps in its process chain serve as retrieval queries. Retrieved passages are reranked and combined with the event representation to guide scene generation. Formally,

$$\begin{aligned} \mathcal{R} = \text{Index}(\mathcal{I}), \;\; C_j = \text{Retrieve}(\mathcal{R}, \mathcal{P}_j) \\ T'_j = \text{Enrich}(T_j, C_j) \end{aligned} \tag{6}$$

where $N_j$ denotes the $j$-th event, $\mathcal{R}$ indexes passages from the original novel, $\mathcal{P}_j$ contains the process-chain steps of event $N_j$, and $C_j$ collects the retrieved and reranked context. The enriched representation $T'_j$ combines the event description with these passages through Enrich, providing source-grounded context for scene planning. This event-level retrieval connects the planned sequence of actions to relevant source details, including character motivations and plot connections, supporting scene and shot planning throughout the adaptation process.

### 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 consistency across shots, ViMax establishes a dependency-based generation framework that maintains shot-level visual dependencies through a camera-level dependency graph. For instance, shots involving recurring characters or capturing the same environment from different camera angles exhibit visual dependencies that must be preserved over time. Similarly, objects such as vehicles and buildings appearing across multiple shots within the same scene require coordinated generation to maintain a consistent appearance despite changes in viewing perspective.

Our approach operates in two phases: dependency detection and dependency-aware generation. During planning, it analyzes the textual descriptions of all shots within a scene, assigns the shots to cameras, and constructs a camera dependency graph that captures their visual relationships. During generation, the system follows a topological ordering of this graph. Cameras whose dependencies have been satisfied can be generated in parallel, while dependent cameras incorporate visual anchors propagated from their prerequisite nodes to preserve cross-shot visual consistency. Formally,

$$\begin{aligned} \mathcal{G} &= \text{BuildCameraGraph}\left(\{S_i\}_{i=1}^{\ell}\right), \\ \mathcal{K}_i &= \text{GenConditioned}\left(S_i, A_i\right), \end{aligned} \tag{7}$$

where $\ell$ counts shots in the current scene and $i$ indexes them. $\mathcal{G}$ denotes the camera dependency graph constructed from the shot specifications by $\text{BuildCameraGraph}(\cdot)$, with its nodes and edges representing cameras and their visual dependencies, respectively. $S_i$ and $\mathcal{K}_i$ denote the specification and generated key frame of shot $i$, while $A_i$ denotes the visual anchor associated with the shot's assigned camera and propagated according to $\mathcal{G}$. $\text{GenConditioned}(\cdot)$ denotes reference-conditioned key-frame generation using the shot specification and its corresponding visual anchor. This approach maintains visual consistency across shots while maximizing generation efficiency through parallel processing of independent camera branches during inference across complex cinematic sequences.

#### 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 spatial consistency capabilities of video generation models by creating transition videos between camera angles within the same scene. Rather than generating each viewpoint independently, the system 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 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 shots. Formally, this process can be expressed as:

$$\begin{aligned} V_{trans} &= \text{GenTransition}(C_i, C_j, S) \\ U_i, U_j &= \text{ExtractViews}(V_{trans}, C_i, C_j) \end{aligned} \tag{8}$$

where $V_{trans}$ represents the transition video between camera positions $C_i$ and $C_j$ within scene $S$, and $U_i, U_j$ denote extracted reference views at the corresponding camera angles, distinct from the final shot videos $V_i, V_j$ in the generation workflow. These views provide shared spatial references for subsequent shots, leveraging the video model's internal spatial coherence to reduce incompatible layouts across camera angles within a scene as the camera moves between planned viewpoints.

## 3 EVALUATION

We evaluate ViMax to assess its generation capabilities 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., 2018) with the novel-to-video adaptation pipeline to assess consistency and coherence. Details of the datasets, metric definitions, and judge protocols are provided in Appendix A.2 for reproducibility.

**Baseline Methods**. For ViMax-Bench, we compare ViMax with four categories of methods: **i) shot-level video generators:** Wan2.2-T2V-A14B and Veo 3.1; **ii) dual-stage keyframe-based pipelines:** StoryDiffusion and IC-LoRA (Huang et al., 2024b); **iii) multi-shot generation:** HoloCine (Meng et al., 2025b); and **iv) agentic video-generation systems:** MovieAgent (Wu et al., 2025) and AniMaker (Shi et al., 2025). For novel-to-video planning, we compare against MovieAgent, AniMaker, and RAG-Anything (Guo et al., 2025). Details and adaptation protocols are in Appendix A.3.

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

### 3.2 OVERALL PERFORMANCE COMPARISON (RQ1)

**Multi-shot Visual Consistency.** ViMax-Gemini-3-Pro-Veo 3.1 leads nine settings, improving Overall GC from 0.5272 (HoloCine) to 0.5571 and outperforming the baselines MovieAgent (0.4716) and AniMaker (0.5167). Its gains on Long stories support dependency tracking across scene boundaries. The gains across story lengths demonstrate robustness beyond short-range continuity.

**Long-form Narrative Planning.** In Table 3, ViMax improves Avg. from 3.827 to 4.107 and NF from 3.08 to 3.52 over AniMaker, leading non-ablated baselines. Section 3.4 analyzes effects.

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 |
| MovieAgent-Gemini-3-Pro-Veo 3.1 | 0.4761 | 0.5340 | 0.5016 | 0.4410 | 0.5086 | 0.4560 | 0.4530 | 0.5174 | 0.4716 |
| AniMaker-Gemini-3-Pro-Veo 3.1 | 0.5523 | 0.6014 | 0.5737 | 0.4771 | 0.5224 | 0.4875 | 0.5030 | 0.5486 | 0.5167 |
| 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: Novel-to-video narrative planning results. See Appendix A.2 for metric definitions. Best and second-best values are bolded and underlined, respectively.

| 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 |
| ViMax w/o RAG | **4.36** | **4.28** | **3.98** | 4.96 | **3.94** | 3.14 | **4.110** |
| ViMax w/o StoryboardArtist | 4.04 | 4.00 | 3.72 | 3.10 | 3.60 | 3.49 | 3.658 |

### 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). Win rates represent the percentage of judgments favoring ViMax over the corresponding baseline for each criterion, and Avg. denotes their mean across the three criteria. Details of the annotation protocol are provided in Appendix A.2.

### 3.4 Ablation Study (RQ3)

We isolate narrative-planning and visual-generation components in Table 3, Figure 4, and the qualitative comparison in Appendix Figure 7 to assess their individual contributions. The comparisons also reveal whether gains in narrative quality coincide with stronger cross-shot visual consistency.

**Narrative-Planning Components.** Replacing StoryboardArtist with deterministic screenplay segmentation preserves story order and shot count but removes agent-based visual articulation; VSU falls from 4.94 to 3.10 and Avg. from 4.107 to 3.658. The w/o PC variant retains ordered event summaries but omits their intermediate actions and causal steps (Section 2.2.1); VSU rises from 4.94 to 4.98, while NF drops from 3.52 to 2.34. Removing RAG increases Avg. from 4.107 to 4.110 but lowers NF from 3.52 to 3.14, isolating retrieval's role in retaining source-specific details. The three modules serve complementary purposes: visual articulation, global event structure, and source grounding. Aggregate planning scores should thus be interpreted alongside source faithfulness.

**Graph-Based Visual Dependency.** The serial variant (Ser.) removes camera-aware planning, graph-based references, and transition videos. Each shot conditions on character portraits and the preceding shot's first frame, retaining only a temporal chain. Medium-IC slightly improves, but the other eight

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** |

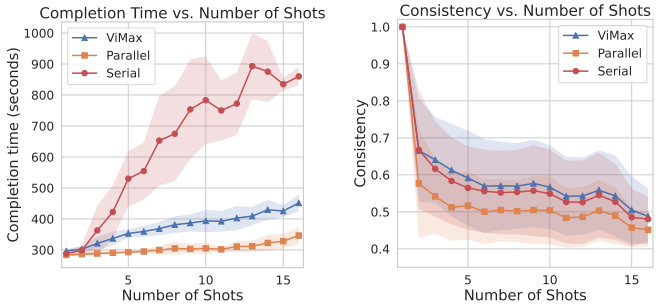


Figure 3: ViMax ablation study results on shot completion time and average cross shot consistency.

metrics decline. The parallel variant (Par.) removes all cross-shot dependencies; Overall GC drops by 8.7%, indicating that portraits alone do not preserve identity, layout, and scene state. **VLM Quality Control.** Removing VLM candidate selection (NoQC) lowers Overall GC by 3.8%. This supports filtering off-spec keyframes before they become references for later shots, reducing error propagation through the visual dependency graph rather than judging each generated shot only in isolation.

### 3.5 Shots Efficiency Study (RQ4)

Figures 3 and 6 compare scheduling efficiency over 35 ViMax-Bench stories. Parallel execution is fastest (310s) but yields the lowest Overall GC (0.5085); serial execution reaches 0.5511 but requires 835s. By executing independent graph branches concurrently, the ViMax achieves the best GC (0.5571) in 405s. Costs include API usage only, and latency is measured end to end. ViMax offers the best consistency–latency balance, with parallel execution suitable for latency-constrained settings. This balance retains cross-shot references while allowing independent branches to run concurrently.

### 3.6 Impact of LLM Backbone (RQ5)

With other components fixed, GPT-5.4 obtains the highest visual-consistency and mean narrative-planning scores in Table 4. Its narrative advantage over Claude is small (4.11 vs. 4.06). GPT-5.4 also judges narratives, so potential self-preference limits conclusions about general planning ability. Claude leads on several narrative dimensions despite lower visual-consistency scores; Gemini-3-Flash ranks second on Overall GC. These results should be interpreted jointly across the evaluation metrics.

### 3.7 Hyperparameter Study (RQ6)

We study the Best-of-$n$ hyperparameter in VLM-based quality control, where $n$ 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 and Appendix Figure 6, Best-of-2 yields Overall GC of 0.5571 at an average cost of $17.500 and 405s per task. Best-of-3 and Best-of-4 increase cost and latency ($17.770/415s and $18.040/425s) while reducing GC to 0.5041 and 0.5194, respectively. This supports $n = 2$ as the default operating point and shows that additional candidates can increase selection noise in identity and scene state. Candidate count should thus balance consistency, per-task generation cost, and end-to-end latency.

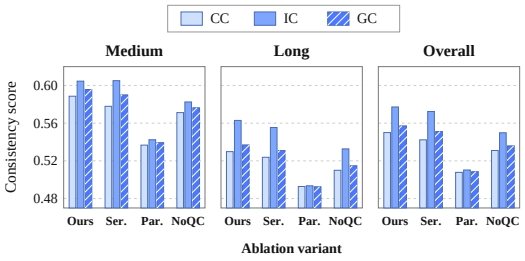


Figure 4: ViMax ablation study.

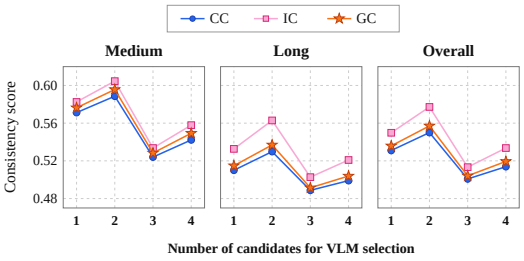


Figure 5: Quality-control hyperparameters.

## 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, and FilMaster (Huang et al., 2025) retrieves camera language for pacing. For video understanding, Chu et al. (Chu et al., 2025) use entity-relation graphs to track visual context. These studies motivate linking high-level story structure to concrete shot decisions. Our framework couples hierarchical decomposition with retrieval of relevant source context throughout the planning process.

**Multi-Agent Frameworks for Content Creation.** Video production fits into a division of labor across agents. Generative Agents (Park et al., 2023) demonstrate long-term behaviors; MAViS (Wang et al., 2025) coordinates roles via an explore-examine-enhance loop; and GenMAC (Huang et al., 2024a) 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. Role specialization requires shared context to keep agents' revisions compatible with downstream decisions. Our workflow connects these roles through intermediate artifacts and shot-level visual feedback.

**Long-Form Video Generation and Visual Consistency.** Long-duration generation must overcome 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 coherence, StableAnimator (Tu et al., 2025) 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., 2024) provides metrics for consistency, motion, and aesthetic quality. Appearance persistence and spatial coherence require coordination. Our framework combines cross-shot dependencies, reference conditioning, and quality control to preserve visual context.

## 5 Conclusion

We have presented ViMax, an agentic framework for long-form video generation via coordinated multi-agent workflows. It tackles narrative planning via hierarchical story decomposition with retrieval-augmented generation, and ensures visual consistency through graph-based dependency tracking and transition video generation, with the dependency graph also exposing shot-level parallelism for efficient large-scale generation. We introduce ViMax-Bench, a benchmark targeting controlled cross-shot consistency challenges. By decomposing production into specialized agents, ViMax enables scalable creation of coherent and extended narratives, opening up new possibilities for automated content creation across entertainment, education, and media production. Looking forward, we view ViMax as a step toward fully agentic filmmaking, where richer multimodal generation, long-horizon memory, and interactive co-creation can be layered on top of increasingly capable foundation models to bring automated cinematic creation closer to practical industry deployment.

## Ethics Statement

Mitigating impersonation, misinformation, and copyright misuse requires consent, provenance, and human review. Benchmark safeguards are limited, and evaluators may inherit source biases.

## AI Use Statement

AI assisted language polishing, coding, and LaTeX formatting, and powered narrative/visual generation, benchmark specification, and LLM/VLM evaluation (see the main text and Appendix A.2).

## References


Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. SEINE: Short-to-long video diffusion model for generative transition and prediction. *ArXiv*, abs/2310.20700, 2023. URL https://api.semanticscholar.org/CorpusID:264820183.

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

Google DeepMind. Veo 3.1. *Google DeepMind*, 2025. URL https://deepmind.google/models/veo/.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. RAG-Anything: All-in-one RAG framework. *ArXiv*, abs/2510.12323, 2025. URL https://api.semanticscholar.org/CorpusID:282064544.

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

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2568–2577, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Henschel_StreamingT2V_Consistent_Dynamic_and_Extendable_Long_Video_Generation_from_Text_CVPR_2025_paper.html.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. StoryAgent: Customized storytelling video generation via multi-agent collaboration. *ArXiv*, abs/2411.04925, 2024. URL https://api.semanticscholar.org/CorpusID:273878062.

Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. GenMAC: Compositional text-to-video generation with multi-agent collaboration. *ArXiv*, abs/2412.04440, 2024a. URL https://api.semanticscholar.org/CorpusID:274514370.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. FilMaster: Bridging cinematic principles and generative AI for automated film generation. *ArXiv*, abs/2506.18899, 2025. URL https://api.semanticscholar.org/CorpusID:279999866.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context LoRA for diffusion transformers. *ArXiv*, abs/2410.23775, 2024b. URL https://arxiv.org/abs/2410.23775.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

Weijie Kong et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Orest Kupyn, Théo Uscidda, Marta Tintore Gazulla, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *ArXiv*, abs/2510.21615, 2025. URL https://arxiv.org/abs/2510.21615.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via LLM-Guided planning. *ArXiv*, abs/2309.15091, 2023. URL https://arxiv.org/abs/2309.15091.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond H. Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EvalCrafter_Benchmarking_and_Evaluating_Large_Video_Generation_Models_CVPR_2024_paper.html.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. FreeLong: Training-free long video generation with SpectralBlend temporal attention. *ArXiv*, abs/2407.19918, 2024. URL https://api.semanticscholar.org/CorpusID:271533485.

Xiangyu Meng, Zixian Zhang, Zhenghao Zhang, Junchao Liao, Long Qin, and Weizhi Wang. Identity-GRPO: Optimizing multi-human identity-preserving video generation via reinforcement learning. *ArXiv*, abs/2510.14256, 2025a. URL https://api.semanticscholar.org/CorpusID:282138850.

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

OpenAI. Video generation models as world simulators. *OpenAI Technical Report*, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. URL https://api.semanticscholar.org/CorpusID:258040990.

Akshat Rakheja, Aarsh Ashdhir, Aryan Bhattacharjee, and Vanshika Sharma. World consistency score: A unified metric for video generation quality. *ArXiv*, abs/2508.00144, 2025. URL https://api.semanticscholar.org/CorpusID:280417376.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. AniMaker: Multi-agent animated storytelling with MCTS-driven clip generation. *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025. URL https://api.semanticscholar.org/CorpusID:279318504.

Team Seedance et al. Seedance 2.0: Advancing video generation for world complexity. *ArXiv*, abs/2604.14148, 2026. URL https://arxiv.org/abs/2604.14148.

Team Wan. Wan: Open and advanced large-scale video generative models. *ArXiv*, abs/2503.20314, 2025. URL https://arxiv.org/abs/2503.20314.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43:1 – 18, 2024. URL https://api.semanticscholar.org/CorpusID:267412997.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. StableAnimator: High-quality identity-preserving human image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tu_StableAnimator_High-Quality_Identity-Preserving_Human_Image_Animation_CVPR_2025_paper.html.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul E. Debevec, and Ning Yu. MAViS: A multi-agent framework for long-sequence video storytelling. *ArXiv*, abs/2508.08487, 2025. URL https://api.semanticscholar.org/CorpusID:280635223.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. *ArXiv*, abs/2412.13575, 2024. URL https://api.semanticscholar.org/CorpusID:274822787.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A large-scale video-text dataset for multimodal understanding and generation. *ArXiv*, abs/2307.06942, 2023. URL https://arxiv.org/abs/2307.06942.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent CoT planning. *ArXiv*, abs/2503.07314, 2025. URL https://api.semanticscholar.org/CorpusID:276929150.

Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Jiepeng Cen, Zhibei Ma, Alan Yuille, and Lu Jiang. VideoAuteur: Towards long narrative video generation. *ArXiv*, abs/2501.06173, 2025. URL https://arxiv.org/abs/2501.06173.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. FilmAgent: A multi-agent framework for end-to-end film automation in virtual 3D spaces. *ArXiv*, abs/2501.12909, 2025. URL https://api.semanticscholar.org/CorpusID:275788131.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-And-Write: Towards better automatic storytelling. *ArXiv*, abs/1811.05701, 2018. URL https://api.semanticscholar.org/CorpusID:53306064.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. StoryMem: Multi-shot long video storytelling with memory. *ArXiv*, abs/2512.19539, 2025a. URL https://api.semanticscholar.org/CorpusID:284077920.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. STAGE: Storyboard-anchored generation for cinematic multi-shot narrative. *ArXiv*, abs/2512.12372, 2025b. URL https://api.semanticscholar.org/CorpusID:283896552.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. MovieDreamer: Hierarchical generation for coherent long visual sequence. *ArXiv*, abs/2407.16655, 2024. URL https://arxiv.org/abs/2407.16655.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention. *ArXiv*, abs/2412.02259, 2024. URL https://arxiv.org/abs/2412.02259.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-Cong Chen. VideoMemory: Toward consistent video generation via memory integration. *ArXiv*, abs/2601.03655, 2026. URL https://api.semanticscholar.org/CorpusID:284532367.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. *ArXiv*, abs/2405.01434, 2024. URL https://api.semanticscholar.org/CorpusID:269502120.

# 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.

| 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 11, 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 design references the multi-shot narrative settings and consistency challenges studied in StoryMem (Zhang et al., 2025a) and VideoMemory (Zhou et al., 2026). ViMax-Bench targets hard-cut, multi-shot narrative generation grounded in cinematic shot design. Its specifications incorporate cinematographic techniques such as over-the-shoulder and reverse-angle shots, insert shots, and varied camera compositions, while requiring recurring characters across scene changes and spatial consistency under substantial viewpoint shifts. This design provides a controlled setting for studying the cross-shot narrative and visual-consistency challenges addressed by ViMax. 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.** We follow the cross-shot consistency protocols of HoloCine (Meng et al., 2025b) and StoryMem and measure shot-level similarity with ViCLIP (Wang et al., 2023). Each generated shot is encoded into one video-level feature, and cosine similarity is computed for every shot pair within the same story. IC averages pairs within the same scene, CC pairs across different scenes, and GC all pairs. The scenario labels Character Persistence, Background Persistence, and Multi-Person Interaction define the challenge covered by each story; they are not manually assigned scores and require no expert annotation. All three reported metrics are automatic, and higher values indicate greater feature stability.

**Human Evaluation Protocol.** The evaluation covers all 35 ViMax-Bench stories and 438 generated shots. For each story, ViMax is compared against five baselines, yielding $35 \times 5 = 175$ unique comparison pairs. In each pair, a ViMax video and a baseline video are displayed anonymously side by side, with left/right placement randomized to reduce position bias. Fifteen evaluators with AI-related backgrounds participated. We construct 15 balanced annotation buckets, each containing all 35 stories with 7 comparisons against each baseline, and assign one evaluator to each bucket. Every unique pair appears in three buckets and is therefore judged independently by three evaluators, producing $175 \times 3 = 525$ pairwise judgments in total.

Evaluators select the preferred complete video along three criteria: (i) **Cross-Scene Consistency (CSC)**, covering recurring characters, objects, clothing, identity features, visual style, and action or behavioral characteristics across shots and scenes; (ii) **Semantic Following (SF)**, covering the depicted subjects, actions, interactions, environments, and narrative progression; and (iii) **Aesthetic Quality (AQ)**, covering visual quality, composition, visual appeal, and cinematic presentation. The human study complements CC/IC/GC by directly assessing semantic execution and overall presentation in addition to visual continuity.

**Narrative Planning Metrics.** Table 3 reports six narrative-planning metrics and their mean. 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 12, 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 14, 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 native input schema by GPT-5.2. We use seven video-generation baselines spanning (i) shot-level text-to-video models invoked independently and concatenated, (ii) keyframe-anchored pipelines that delegate motion to Wan2.2-I2V-A14B, (iii) unified multi-shot generation, and (iv) agentic video-generation frameworks. All methods receive the same 35 underlying stories and shot requirements. Each baseline follows its standard generation procedure without ViMax-specific reference propagation, transition generation, or post-processing. We separately evaluate three narrative-planning baselines under the matched protocol in Appendix A.2.

**Shot-Level Text-to-Video Generators.**

- **Wan2.2-T2V-A14B** (Team Wan, 2025): 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, 2025): 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.

**Agentic Video-Generation Frameworks.**

- **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.

For the ViMax-Bench comparison in Table 1, we retain the native planning and coordination of MovieAgent and AniMaker while using Gemini-3-Pro-Image and Veo 3.1 as their image and video backbones. Reporting the backbone in each method name makes this adaptation explicit.

**Novel-to-Video Planning Baselines.** MovieAgent and AniMaker are also evaluated as planning systems on the 50-novel task, with video synthesis disabled so that their structured storyboards can be scored under the same judge protocol as ViMax. We additionally compare against the following retrieval system:

- **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 Additional Component and Cost Analyses

Figure 6 presents a Pareto analysis of configurable operating points within ViMax. The default uses Gemini-3-Flash for language reasoning, Gemini-3-Pro-Image for keyframes, and Veo 3.1 for video. Figure 7 complements the aggregate metrics with a shared four-shot example, making identity, composition, object-state, and scene-layout changes directly inspectable across component variants.

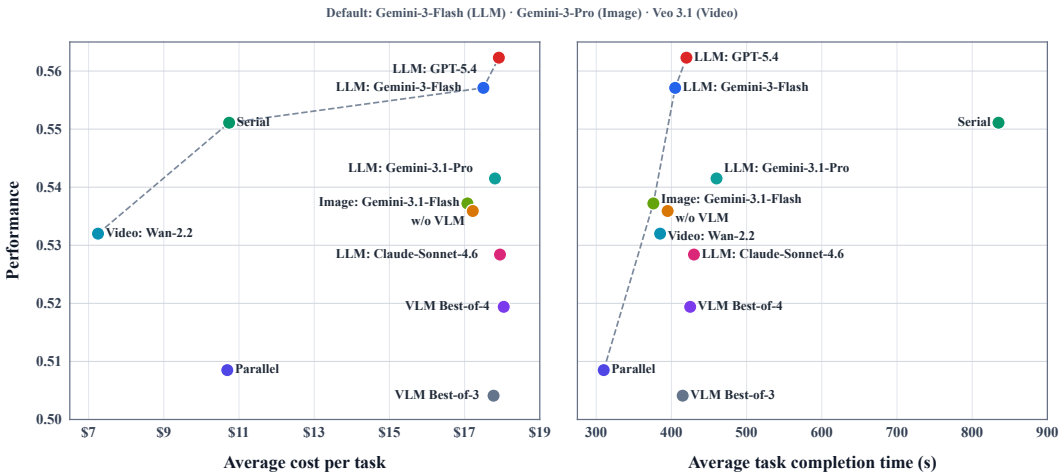


Figure 6: Pareto frontier of cost, quality, and latency over 35 ViMax-Bench stories. Performance is GC, and dashed lines connect Pareto-efficient configurations. Costs include API usage only, and latency is measured end to end.




Figure 7: Qualitative component ablation under the same four-shot specification. Rows compare removal of VLM quality control, serial and parallel alternatives to graph-based scheduling, and full ViMax. The examples expose complementary failure modes in recurring-character identity, viewpoint and composition, and spatial layout.

### A.6 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 8, Figure 9 and Figure 10.

#### 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.

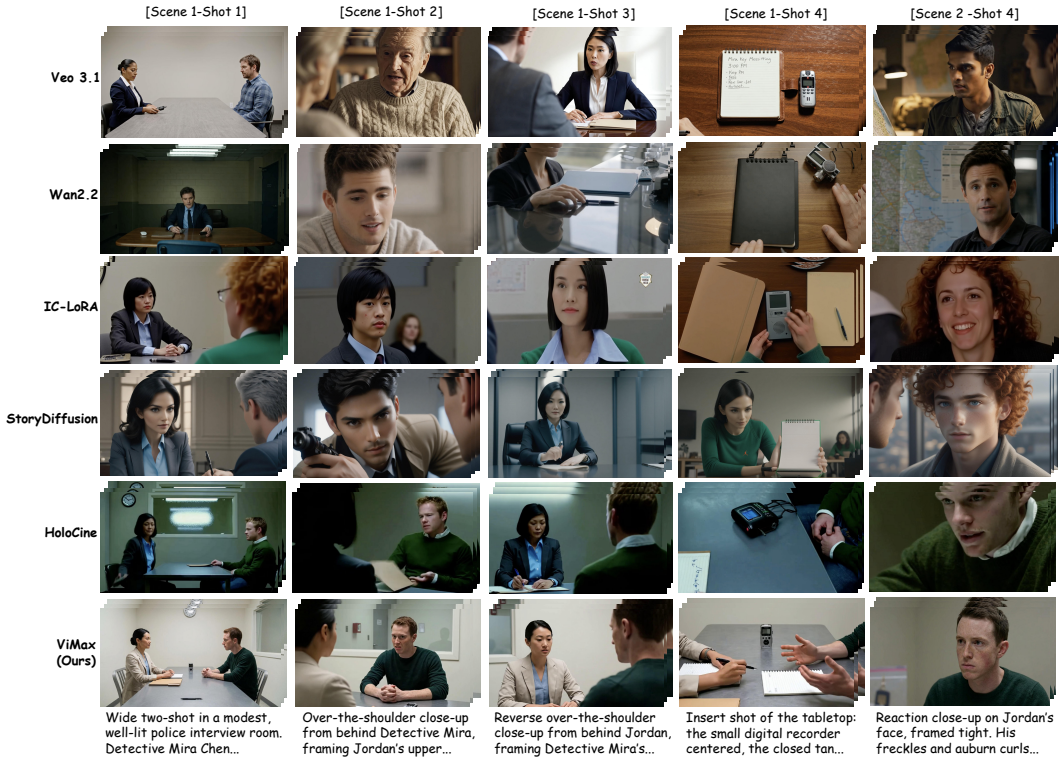


Figure 8: Qualitative comparison with representative baselines on a ViMax-Bench multi-shot story. Each row shows generated keyframes from one method under the same story and shot specifications. ViMax preserves the two main characters, interrogation-room setting, object layout, and shot-level cinematic intent across multiple viewpoints, demonstrating stronger long-horizon visual and spatial consistency.

**Character Identity.** Figure 10 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 8 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 9 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 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.



Figure 9: Case study: ViMax-generated keyframes.



Figure 10: Case study: ViMax-generated keyframes.

### 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 9 and 10 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 8 and the cockpit instrument panels in Figure 9), 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 15. For every keyframe storyboard, we use best-of-$n$ selection with $n = 2$ candidates synthesized by the Image Generation Agent. The VLM-Judge critically evaluates these candidates using a prompting strategy that weighs character consistency, spatial logic, and semantic alignment. This filtering is designed to reduce semantic mismatches with the shot specification before selected keyframes are reused downstream.

## A.7 LIMITATIONS

ViMax inherits base-model errors that can propagate across shots. As generators improve, orchestration may shift toward long-horizon visual memory. Current feedback refines individual outputs but does not accumulate experience to improve planning and coordination strategies. Learning from failures to improve how agents plan, coordinate, and revise their decisions remains an open direction.

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 11: Prompt used to generate ViMax-Bench story specifications.

**Narrative Planning Evaluation Prompt - 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 12: LLM judge prompt template used for narrative planning storyboard evaluation (Part 1).

**Narrative Planning Evaluation Prompt - 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.

Figure 13: 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 14: 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 description (the desired final image, not an edit instruction).

**### VLM Response (Illustrative Best-of-2 Example)**
Candidate images are indexed 0 and 1 in this example.
The JSON below has a single-line reason string; visual wrapping is for display.


```
{
  "best_image_index": 1,
  "reason": "Candidate Image 1 best matches the references and target description. Both
    candidates preserve character identity and the backyard layout, but Image 0 shows the
    dog standing, whereas Image 1 depicts it crouched in anticipation as requested."
}
```


Figure 15: VLM judge prompt used for quality-control image selection.