ViMax: Agentic Video Generation

Lingxuan Huang[1*], **Sizhe He**[2*], **Hengji Zhou**[2*], **Yan Huang**[2], **Si Wu**[2], **Lianghao Xia**[1†], **Chao Huang**[1†]
[1]The University of Hong Kong, [2]South China University of Technology
xvrhuang@connect.hku.hk, 202320143418@mail.scut.edu.cn, hengjizhou01@gmail.com,
{aihuangy, cswusi}@scut.edu.cn, aka_xia@foxmail.com, chaohuang75@gmail.com

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 extended narratives spanning hundreds of shots. 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 fully 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 productions available at https://github.com/hkuds/vimax.

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 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 AI-driven video generation has explored several directions to address these challenges. Foundational text-to-video models such as Sora (OpenAI, 2024) and Veo 3 (Google DeepMind, 2024) have achieved remarkable fidelity, while open-source frameworks like HunyuanVideo (Kong et al., 2024), CogVideoX (Yang et al., 2024b), and Mora (Yuan et al., 2024) have democratized access to high-resolution synthesis. Alternative architectures like Lumiere (Bar-Tal et al., 2024) explore space-time diffusion to enhance temporal coherence. However, these systems remain fundamentally constrained to short-form content lasting seconds to minutes. They often struggle with "catastrophic semantic forgetting" in extended sequences, failing to maintain the geometric and physical laws required for a consistent world view (Kupyn et al., 2025; Rakheja et al., 2025). New autoregressive approaches like StreamingT2V (Henschel et al., 2024) and training-free extensions like FreeLong (Lu et al., 2024) attempt to mitigate this by employing short-term memory blocks and spectral blending, yet the challenge of semantic drift over minutes of content persists.

Simultaneously, investigations into consistency mechanisms have advanced. Recent works reveal that video models implicitly treat the first frame as a conceptual memory buffer (Chen et al., 2025), allowing robust content customization without architectural changes. While training-free mechanisms like StoryDiffusion (Zhou et al., 2024) and ConsiStory (Tewel et al., 2024) employ shared attention to maintain subject identity, other as FreeNoise (Qiu et al., 2023), utilize noise rescheduling to prevent temporal drift. In the realm of agentic workflows, systems like LAVE (Wang et al., 2024a) introduce LLM-based assistance for editing, while multi-

[1*]Lingxuan Huang, Sizhe He, and Hengji Zhou have equal contribution to this work.

[2†]Lianghao Xia and Chao Huang are the corresponding authors.

agent frameworks like OmniAgent (Wei et al., 2025), FilmAgent (Xu et al., 2025), Kubrick (He et al., 2024), and VideoGen-of-Thought (Zheng et al., 2025) begin to decompose generation tasks. Despite this, methods often operate on footage or struggle to maintain global narrative and spatial coherence across the hundreds of shots required for professional long-form storytelling.

Addressing these limitations requires overcoming two fundamental technical challenges:

- **Long-form Narrative Planning:** The inherent complexity of orchestrating extended narratives spanning hundreds or thousands of shots creates an exponential planning problem that exceeds the reasoning capacity of current language models when approached as a monolithic task. Long-form videos require coordinating multiple story elements including character arcs, plot development, and thematic coherence across extended timelines, while maintaining global narrative consistency throughout localized scene planning. The challenge lies in decomposing complex storylines into manageable planning units without losing the intricate dependencies and contextual relationships that ensure narrative coherence, as local planning decisions must remain grounded in the broader story context to prevent contradictions and preserve thematic unity.

- **Multi-shot Visual Consistency:** The independent nature of video generation processes creates fundamental barriers to maintaining visual coherence across the hundreds of shots required for long-form content. Each generation operates in isolation without awareness of previously generated content, leading to inconsistencies in character appearance, environmental design, object properties, and spatial relationships that undermine the illusion of a unified visual world. This challenge is compounded by the need to ensure both object-level consistency, where characters and environments maintain identical visual properties across different shots, and spatial coherence, where multiple camera angles within the same scene must preserve consistent 3D geometric relationships to avoid contradictory layouts.

To mitigate these challenges, we present ViMax, an agentic video generation framework that employs a sophisticated multi-agent workflow to orchestrate long-form video production from conceptual ideas to fully realized content. Our approach systematically addresses the complexity of video production through specialized agents that coordinate across key production roles including screenwriting, shot planning, character styling, video generation, and quality assessment, with a VLM-based quality control mechanism that ensures professional standards through iterative refinement. For narrative planning complexity, ViMax employs hierarchical story decomposition that systematically breaks down extended storylines into manageable planning units while using retrieval-augmented generation to maintain global narrative context throughout the decomposition process. For visual consistency, ViMax introduces a graph-based visual dependency system that identifies and maintains visual relationships between shots during planning, combined with transition video generation that ensures spatial coherence across multiple camera angles within the same scene. This comprehensive framework enables coordinated content creation that maintains both narrative coherence and visual consistency across hundreds of shots while preserving the professional standards expected in long-form video production.

The primary contributions of this work include:

- **Agentic Long-form Video Generation Framework:** We introduce a comprehensive multi-agent workflow that systematically addresses the complexity of long-form video production through specialized agents coordinating across screenwriting, shot planning, character styling, video generation, and VLM-based quality control, enabling automated creation of narrative content from conceptual ideas to videos.

- **Hierarchical Narrative Planning with Global Context Awareness:** We propose a novel hierarchical story decomposition approach that breaks down complex storylines into manageable planning units while employing retrieval-augmented generation (RAG) to maintain global narrative context, enabling language models to handle extended narratives without losing thematic coherence or contextual dependencies.

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

scenes, ensuring professional-grade visual continuity throughout long-form video production.

2 Methodology

This section elaborates technical details of ViMax. The overall architecture is depicted in Figure 1.

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 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 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 the quality of generated content, ViMax employs a VLM-based quality assessment mechanism that evaluates and selects the best results from multiple generation candidates. For each generation task, the system generates multiple candidates and uses a VLM judge to evaluate which candidate best matches the requirements specified in the shot planning. The VLM judge assesses various quality dimensions including visual fidelity, narrative consistency, and adherence to shot specifications. If all candidates fail to meet quality standards, the

Figure 1: ViMax agentic video generation architecture.

Figure 2: VLM-based quality control workflow.

system enters an iterative refinement loop where prompts are automatically modified and refined based on the VLM feedback to improve generation quality. This process can be 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 $\{C_1, C_2, \cdots, C_k\}$ represents multiple generation candidates, VLM-Judge$(\cdot)$ evaluates each candidate against shot requirements, C^* is the selected best candidate. This quality-driven approach ensures that only high-quality content progresses through the production pipeline, maintaining consistency and professional standards across the entire long-form video.

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 orchestrating character arcs, plot development, and narrative coherence spanning hundreds of scenes. The scale of information 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:

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

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 hundreds of shots is essential for creating cohesive viewing experiences. 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:

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

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 gen-

eration 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 Implementation Details

We implement ViMax using a modular multi-agent framework designed for efficient parallelization, utilizing CPU-based orchestration.

3.1 Model Specifications

Our framework leverages state-of-the-art foundation models to power specific agent roles.

- **Reasoning & Planning:** The Director, Screenwriter, and Shot Planning agents utilize `GPT-5` to handle complex narrative reasoning and long-context dependency tracking.
- **Visual Analysis:** The Quality Control Agent employs `Gemini-2.5-Pro` for visual consistency scoring and candidate judging.
- **Generation:** Keyframes are synthesized using `Gemini-2.5-Flash-Image` for speed and adherence to prompts, while video motion synthesis is executed via `Veo3.1-Fast`.

3.2 Agent Configurations and Prompting

To bridge the gap between general-purpose LLMs and cinematographic requirements, we implement prompting strategies and constraints for each agent.

Screenwriting Strategies. We employ a *Repetition for Precision* prompting strategy within the Script Enhancer. Since the downstream video generation models synthesize shots independently, the agent is instructed to redundantly embed global visual traits (e.g., "wearing a tattered red scarf") into every individual shot description. This ensures that visual constraints are not lost due to the limited context window of the video generation model.

Character Portrait Design Logic. The Character Extractor Agent analyzes the smallest narrative units to identify characters. To ensure global consistency across the narrative, the agent recursively integrates character features in a bottom-up manner. Features that remain the same for a character across different scenes are merged into **Static Features** (the core identity), while varying attributes are treated as **Dynamic Features** (e.g., attire). When designing a character's visual portrait, the base appearance is first constructed from the Static Features. Dynamic Features are then layered on top, according to scene-specific needs, to create the final look for any given narrative moment.

Quality Control Hyperparameters. To mitigate the stochasticity of generative models, we employ a Best-of-N selection strategy. For every required keyframe, the Image Generation Agent synthesizes a batch of $K = 3$ candidates. The VLM-Judge evaluates the candidate $\hat{I}$ that maximizes the joint

Figure 3: Case study: ViMax-generated keyframes.

Figure 4: Case study: Seko-generated keyframes.

probability of semantic alignment and visual consistency with the parent anchor node.

4 Case Study

This section provides visual evidence of ViMax's capabilities in synthesizing coherent, long-form narratives. We showcase several case studies of video key frames automatically generated by our multi-agent framework. A representative segment demonstrating long-term character and scene consistency is shown in Figure 3 and Figure 5.

4.1 Experimental Setup

For a fair comparison, both ViMax and the baseline were provided with identical high-level user ideas (e.g., "A short, characterful nature vignette following a young river otter...").

- **ViMax Settings:** The user idea was processed through our multi-agent system to generate a structured script, followed by hierarchical storyboard planning, and finally video synthesis.
- **Baseline Settings:** The same user idea was fed directly into SenseTime Seko (SenseTime, 2025) for storyboards generation.

Figure 5: Case study: ViMax-generated keyframes.

Figure 6: Case study: Seko-generated keyframes.

We generated six distinct long-form videos for each method. Figures 3 and 5 illustrate keyframes from the ViMax outputs, while Figures 4 and 6 display the corresponding baseline results.

4.2 Visual Consistency and Identity Preservation

A critical failure mode in long-form video generation is spatiotemporal inconsistency, where character identities and environmental details fail to remain stable across extended durations.

Character Identity. As shown in the comparison between Figure 5 (ViMax) and Figure 6 (Baseline), the baseline model struggles to maintain protagonist identity across non-continuous shots. In the baseline generated sequence, the protagonist's clothing attributes (e.g., color and style) and facial features fluctuate significantly between scenes. In contrast, ViMax successfully anchors visual identity. By utilizing the Character Extractor Agent to decouple static features and the Graph-based Dependency Tracking to enforce reference-conditioned generation, ViMax ensures the protagonist remains recognizable with consistent attire and facial structure throughout the narrative arc.

Spatial Coherence. Figure 3 and Figure 5 demonstrate ViMax's ability to handle complex spatial relationships. In multi-angle scenes (e.g., dialogue sequences), ViMax utilizes transition video

Figure 7: Case study: VLM-guided quality control.

VLM Judger System Prompt:
"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, 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:
{
"best_image_index": 2,
"reason": "All three candidate images closely match the references and description, but Candidate Image 2 best captures the specified actions.\n\n-**Candidate Image 0:**\n - Character Consistency: 9/10. The boy and dog closely match the references.\n - Spatial Consistency: 10/10. The backyard layout with the boy on the left and dog on the right is correct.\n - Description Accuracy: 8/10. The dog is standing rather than 'crouched in anticipation.'\n\n-**Candidate Image 1... This makes it the best overall choice."
}

Figure 8: VLM-guided image selection example

generation to act as a spatial anchor. This ensures relative positions of characters and background objects remain geometrically consistent when camera angles switch. The baseline Figure 4 and Figure 6, lacking this spatial reference awareness, frequently hallucinates contradictory layouts, such as often contradictory layout illusions, such as chaotic environments, confused reference objects, or characters switching positions between camera cuts.

4.3 Narrative Logic and Pacing

Beyond visual consistency, the logical progression of the generated content is crucial. The baseline videos often resemble a montage of loosely related clips, lacking a clear causal chain. Conversely, ViMax leverages Hierarchical Narrative Planning to decompose the user idea into a coherent screenplay. The generated shots exhibit professional narrative and spatial pacing, ensuring that character actions logically lead to subsequent plot points.

4.4 VLM-Guided Quality Control

To validate our Quality Control Agent, we visualize the candidate selection process in Figure 7. For every keyframe storyboard, the Image Generation Agent synthesizes $K = 3$ candidates. The VLM-Judge critically evaluates these candidates using a prompting strategy (Figure 8) that weighs character consistency, spatial logic, and semantic alignment. As demonstrated in the case study, the VLM successfully distinguishes between visually high-quality candidates and those that accurately reflect the script—for instance, rejecting a candidate where a dog is standing in favor of one where it is "crouched in anticipation." This discriminative filtering significantly reduces the rate of semantic hallucinations compared to single-pass generation.

5 Related Works

Recent advancements in generative media have shifted focus from short, independent clips to coherent, long-form narratives. This section reviews the foundational literature across three key dimensions: hierarchical narrative planning, multi-agent orchestration, and visual consistency mechanisms.

5.1 Hierarchical Narrative Planning and Cinematic Logic

The challenge of maintaining logical coherence in long stories has been addressed through hierarchical decomposition. Work by Plan-And-Write (Yao et al., 2018) demonstrated that separating storyline planning from surface realization yields significantly more coherent narratives. This paradigm has been further advanced by DOME (Wang et al., 2024b), which introduces dynamic hierarchical outlining with memory enhancement to prevent plot holes in long-form generation. To address the hallucination issues in long-form generation, RapID (Gu et al., 2025) proposes a retrieval-augmented framework that generates a preliminary outline before synthesis, ensuring structural integrity. Storyteller (Li et al., 2025) treats narrative generation as a state-tracking problem, utilizing Subject-Verb-Object (SVO) triplets to maintain logical flow. VideoDirectorGPT (Lin et al., 2023) and StoryVerse (Wang et al., 2024c) employ symbolic narrative planning to ground character actions. Collaborative authorship frameworks like CollabStory (Venkatraman et al., 2024) and Agents' Room (Huot et al., 2025)

demonstrate that assigning specialized roles to agents enhances narrative depth compared to single-model generation. FilMaster (Huang et al., 2025) retrieves camera language from corpuses to ensure industry-standard pacing, while Chu et al. (Chu et al., 2025) introduce dynamic graph-based modeling to track complex entity relations and gradually resolve narrative fragmentation over time.

5.2 Multi-Agent Frameworks for Content Creation

The complexity of video production necessitates a division of labor, effectively modeled by multi-agent systems. Generative Agents (Park et al., 2023) proved the viability of simulating believable long-term behaviors. In video production, MAViS (Wang et al., 2025b) coordinates roles via an explore-examine-enhance loop, while DreamFactory (Xie et al., 2024) uses Chain-of-Thought prompting to align styles across scenes. Recent agentic frameworks have introduced more sophisticated control mechanisms: GenMAC (Huang et al., 2024) employs a design-generation-redesign loop to iteratively correct compositional errors, and AniME (Zhang et al., 2025) utilizes a director-oriented "Model Context Protocol" to centralize task scheduling. Specific attention has also been paid to subject consistency and specialized feedback loops in agentic workflows. StoryAgent (Hu et al., 2024) integrates intermediate representations to bridge the gap between abstract user intent and pixel-level generation. PersonaVlog (Hou et al., 2025) introduces a feedback and rollback mechanism (FRM) to ensure character consistency in vlogs by treating generation as a state-traversal problem. Similarly, SciTalk (Park et al., 2025) employs a user-simulator agent to critique scientific videos for accuracy before final rendering. MovieAgent (Wu et al., 2025) further formalizes the production pipeline by simulating roles like Director and Screenwriter to automate scene structuring. Additionally, LAVE (Wang et al., 2024a) explores human-AI co-creation in post-production, employing agents to assist with footage retrieval and narrative brainstorming.

5.3 Long-Form Video Generation and Visual Consistency

Generating long-duration videos requires overcoming the drift in character identity and visual style. VideoAuteur (Xiao et al., 2025) addresses this with a rolling-context strategy, while MovieDreamer (Zhao et al., 2024) uses autoregressive global planning to predict visual tokens. Foundational improvements in model architecture, such as the scaling laws demonstrated by Wan (Wang et al., 2025a), provide robust backbones. To ensure visual coherence, recent methods have moved beyond simple conditioning. StableAnimator (Tu et al., 2024) introduces an identity-preservation adapter that aligns face embeddings with temporal layers to prevent identity degradation during motion. Identity-GRPO (Meng et al., 2025) applies reinforcement learning with human feedback (RLHF) to specifically penalize identity drift in multi-human videos, while ReMix (Zhou et al., 2025) aligns the noise space of reference and target images to enable consistent character editing. For spatial logic and transitions, TaleCrafter (Gong et al., 2023) utilizes explicit layout generation to ground character positions before synthesis. SEINE (Chen et al., 2023) and VidSplice (Xie et al., 2025) focus on generative transitions, using masked diffusion and video inpainting to create smooth bridges between independent shots. Epipolar Geometry (Kupyn et al., 2025) constraints have been introduced as preference signals to enforce 3D consistency, and Direct-a-Video (Yang et al., 2024a) allows for precise decoupled control of camera movement and object motion. To rigorously assess these capabilities, new benchmarks like VBench (Huang et al., 2023) and EvalCrafter (Liu et al., 2023) have established comprehensive metrics for consistency, motion smoothness, and aesthetic quality.

6 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 video content, enabling 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.

References

Omer Bar-Tal et al. 2024. Lumiere: A space-time diffusion model for video generation. *SIGGRAPH Asia 2024 Conference Papers*.

Jingxi Chen, Zongxia Li, Zhichao Liu, Guangyao Shi, Xiyang Wu, Fuxiao Liu, Cornelia Fermuller, Brandon Y. Feng, and Yiannis Aloimonos. 2025. First frame is the place to go for video content customization.

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.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. Interactive story visualization with multiple characters. *SIGGRAPH Asia 2023 Conference Papers*.

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

Hongchao Gu, Dexun Li, Kuicai Dong, Hao Zhang, Hang Lv, Hao Wang, Defu Lian, Yong Liu, and Enhong Chen. 2025. Rapid: Efficient retrieval-augmented long text generation with writing planning and information discovery. In *Annual Meeting of the Association for Computational Linguistics*.

Liu He, Yizhi Song, Hejun Huang, et al. 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.

Xiaolu Hou, Bing Ma, Jiaxiang Cheng, Xuhua Ren, Kai Yu, Wenyue Li, Tianxiang Zheng, and Qinglin Lu. 2025. Personavlog: Personalized multimodal vlog generation with multi-agent collaboration and iterative self-correction. *ArXiv*, abs/2508.13602.

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. 2024. 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.

Ziqi Huang et al. 2023. Vbench: Comprehensive benchmark suite for video generative models. pages 21807–21818.

Fantine Huot, Joshua Maynez, Shashi Narayan, et al. 2025. Agents' room: Collaborative narrative generation. In *International Conference on Learning Representations (ICLR)*.

Weijie Kong et al. 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.

Jiaming Li, Yukun Chen, Ziqiang Liu, Minghuan Tan, Lei Zhang, Yunshui Li, Run Luo, Longze Chen, Jing Luo, Ahmadreza Argha, Hamid Alinejad-Rokny, Wei Zhou, and Min Yang. 2025. Storyteller: An enhanced plot-planning framework for coherent and cohesive story generation. In *Annual Meeting of the Association for Computational Linguistics*.

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 et al. 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. 2025. Identity-grpo: Optimizing multi-human identity-preserving video generation via reinforcement learning. *ArXiv*, abs/2510.14256.

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

Jong Inn Park, Maanas Taneja, Qianwen Wang, and Dongyeop Kang. 2025. Stealing creator's workflow: A creator-inspired agentic framework with iterative feedback loop for improved scientific short-form generation. *ArXiv*, abs/2504.18805.

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*.

Haonan Qiu, Menghan Xia, Yong Zhang, Yin-Yin He, Xintao Wang, Ying Shan, and Ziwei Liu. 2023. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *ArXiv*, abs/2310.15169.

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

SenseTime. 2025. Sensetime seko: Ai short video creation platform. Accessed: 2025-11-28.

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.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2024. Collabstory: Multi-llm collaborative story generation and authorship analysis. volume abs/2406.12665.

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

Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024a. Lave: Llm-powered agent assistance and language augmentation for video editing. *Proceedings of the 29th International Conference on Intelligent User Interfaces*.

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. 2024b. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. volume abs/2412.13575.

Yi Wang, Qian Zhou, and David Ledo. 2024c. Storyverse: Towards co-authoring dynamic plot with llm-based character simulation via narrative planning. *Proceedings of the 19th International Conference on the Foundations of Digital Games*.

Zheng Wei et al. 2025. Hollywood town: Long-video generation via cross-modal multi-agent orchestration. *ArXiv*, abs/2510.22431.

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.

Ming Xie, Junqiu Yu, Qiaole Dong, et al. 2025. Vidsplice: Towards coherent video inpainting via explicit spaced frame guidance. *ArXiv*, abs/2510.21461.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F. Bissyand, and Saad Ezzini. 2024. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *ArXiv*, abs/2408.11788.

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

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. 2024a. Direct-a-video: Customized video generation with user-directed camera movement and object motion. *ACM SIGGRAPH 2024 Conference Papers*.

Zhuoyi Yang et al. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *ArXiv*, abs/2408.06072.

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.

Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Weixiang Sun, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. 2024. Mora: Enabling generalist video generation via a multi-agent framework. *ArXiv*, abs/2403.13248.

Lisai Zhang, Baohan Xu, Siqian Yang, Mingyu Yin, Jing Liu, Chao Xu, Siqi Wang, Yidi Wu, Yuxin Hong, Zihao Zhang, Yanzhang Liang, and Yudong Jiang. 2025. Anime: Adaptive multi-agent planning for long animation generation. *ArXiv*, abs/2508.18781.

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 et al. 2025. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2503.15138*.

Benjia Zhou, Bin Fu, Pei Cheng, Yanru Wang, Jiayuan Fan, and Tao Chen. 2025. Remix: Towards a unified view of consistent character generation and editing. *ArXiv*, abs/2510.10156.

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.