MAViS2: A Multi-Agent Framework for Interactive and Adaptive Long-Sequence Video Storytelling

Anonymous ACL submission

Figure 1: MAViS2 allows users to intervene in and refine the generated content in real time, thereby facilitating flexible exploration of diverse visual narratives and creative directions.

Abstract

Existing long-sequence video generation frameworks often overlook scriptwriting, rely on a fixed text-to-image followed by image-to-video paradigm, lack post-production support, and offer limited user interaction, resulting in poor viewing quality and limited user control. To address these limitations, we propose **MAViS2**, a multi-agent framework for interactive and adaptive long-sequence video storytelling. MAViS2 decomposes the video creation process into three coordinated stages: scriptwriting, video clip generation, and post-production, each handled by multiple specialized agents. In the scriptwriting stage, we propose a **Scriptwriting Workflow** that progressively improves the expressiveness of the scripts. In the video generation stage, MAViS2 uses **Adaptive Generation Planning** to select an appropriate generation strategy for each shot and dynamically adjusts it based on the global memory and the evaluation of generated results, thereby significantly increasing visual diversity while reducing constraints on scriptwriting. In the post-production stage, MAViS2 integrates basic video editing, voice-over synthesis, background music matching, and subtitle composition to improve the completeness of the final output. Moreover, MAViS2 supports **Fine-grained Human-in-the-loop Control**, allowing users to intervene and make fine-grained adjustments at any stage, thereby flexibly exploring diverse visual narratives and creative directions. Experimental results demonstrate that MAViS2 outperforms existing methods in terms of visual quality, narrative coherence, and overall viewing experience. MAViS2 offers a novel solution for long-sequence video storytelling, supporting a wide range of storytelling tasks, including end-to-end generation, video understanding, Wikipedia-based external knowledge retrieval, prequel and sequel creation, and video remaking.

1 Introduction

In recent years, with the rapid development of video generation models (OpenAI, 2024; MiniMax, 2024; GenmoTeam, 2024), multi-modal large models (Brown et al., 2020; Anil et al., 2023), and agent technologies (Yang et al., 2023a; Yin et al., 2023; Li et al., 2023), several multi-agent frameworks (Wang et al., 2024a; Xie et al., 2024) have

Methods	Complete Script	Gen. Planning	Audio	Video Editing	Refinement	Automated	HITL Control
Video Model							
MovieDreamer (He et al., 2024)	×	-	×	×	×	✓	×
VGoT (Zheng et al., 2025)	×	-	×	×	×	✓	×
LCT (Guo et al., 2025)	×	-	×	×	×	✓	✓
Image Model							
StoryGen (Liu et al., 2024)	×	-	×	×	×	✓	×
SEED-Story (Yang et al., 2024b)	×	-	×	×	×	✓	×
DreamStory (Zhao et al., 2024)	×	-	×	×	×	✓	×
Agent-based Tool Orchestration							
Mora (Yuan et al., 2024)	×	×	×	×	×	✓	×
AesopAgent (Wang et al., 2024a)	×	×	×	×	×	✓	×
UniVA (Liang et al., 2025)	×	✓	✓	✓	×	✓	✓
Multi-agent Framework							
DreamFactory (Xie et al., 2024)	×	×	×	×	×	×	×
StoryAgent (Hu et al., 2024)	×	×	×	×	✓	✓	×
MovieAgent (Wu et al., 2025b)	×	×	✓	×	×	×	×
MAViS (Wang et al., 2025)	×	×	✓	×	✓	✓	×
MAViS2 (ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: **Properties of MAViS2 vs Other Storytelling Works**: We summarize seven metrics of long-sequence storytelling works. These works span video models, image models, agent-based tool orchestration, and multi-agent frameworks. "Complete Script": ability in generating a complete, rich, and filmable script rather than a simple storyboard; "Gen. Planning": supports generation strategy selection; "Audio": the output includes synchronized audio; "Video Editing": supports video editing; "Refinement": supports evaluating and refining generated results instead of treating one-step outputs as final; "Automated": enables end-to-end automated generation without human intervention; "HITL Control": allows users to intervene in and refine the generated content.

emerged to automate the creation of long-sequence videos. These frameworks have made significant strides in extending video length and complexity.

However, as shown in Table 1, despite the progress these works have offered valuable insights into long video generation, they still exhibit notable shortcomings:

• Underdeveloped Scriptwriting: Existing frameworks either do not support scriptwriting (Yuan et al., 2024), generate a simple storyboard without considering the capabilities of the generation tools (Liang et al., 2025), or overly constrain the scriptwriting to be compatible with generative tools (Wang et al., 2025). As a result, the scripts are often rigid and lack expressive depth.

• Fixed T2I+I2V Paradigm: Many existing frameworks (Xie et al., 2024; Hu et al., 2024; Wu et al., 2025b) generate long-sequence videos by combining text-to-image (T2I) and image-to-video (I2V) generations. However, real-world AI movie creation requires coordinated use of multiple models and tools to achieve greater creative freedom and diversity. The T2I+I2V paradigm inherently limits the expressiveness of long-sequence videos.

• Limited Post-production: Post-production tasks, such as video editing, voice-over synthesis, background music (BGM) matching, and subtitle composition, can significantly enhance the viewing experience and immersion. However, many existing frameworks lack support for post-production.

• Lack of User Interaction: The quality of a story video often depends on the viewer's experience, and real-time content adjustment based on user feedback is particularly important. Yet, many existing frameworks focus solely on end-to-end generation, overlooking user controllability.

To address these issues, we propose **MAViS2**, a multi-agent framework for interactive and adaptive long-sequence video storytelling. MAViS2 decomposes the long-sequence video storytelling process into three independent but collaborative stages: scriptwriting, video clip generation, and post-production.

• Scriptwriting: We propose a **Scriptwriting Workflow**. This process enhances and enriches the script step by step through a progressive flow of "*initial generation* → *structure enhancement* → *content augmentation* → *narration optimization* → *narrative closure*", improving its expressiveness.

• Video Clip Generation: We propose **Adaptive Generation Planning**, in which MAViS2 selects an appropriate generation strategy for each shot and dynamically adjusts it based on the global memory and the evaluation of generated results, thereby significantly increasing visual diversity while reducing constraints on scriptwriting.

• Post-production: MAViS2 integrates basic video editing, voice-over synthesis, BGM matching, and subtitle composition, comprehensively enhancing the overall viewing experience and completeness of the final video.

• **Fine-grained Human-in-the-loop Control**: MAViS2 enables interactive human-in-the-loop (HITL) control. As illustrated in Figure 1, MAViS2

allows users to intervene and make fine-grained adjustments at any stage of the generation process, thereby flexibly exploring diverse visual narratives and creative directions.

To ensure orderly collaboration, MAViS2 adopts a pyramid-structured organization consisting of a User Proxy, Group Managers, and specialized agents. As shown in Figure 2, the User Proxy receives user instructions and dispatches group-level tasks to the Group Managers. Each Group Manager decomposes the group-level tasks, assigns them to specialized agents, and provides feedback to ensure generation quality across stages. Moreover, leveraging video captioning and retrieval-augmented generation (RAG), MAViS2 can generate long-sequence video storytelling with targeted styles and content tailored to users' multimodal inputs.

MAViS2 offers a novel solution for long-sequence video storytelling, supporting a wide range of tasks such as end-to-end generation, video understanding, Wikipedia-based external knowledge retrieval, prequel and sequel creation, and video remaking. Experiments demonstrate that MAViS2 achieves state-of-the-art performance and garners the highest user preference in terms of expressiveness and visual quality.

2 Related Works

2.1 Video Generation

Recent advances in video generation (He et al., 2022; Wang et al., 2024c) have significantly improved text-to-video (T2V) (ModelScopeTeam, 2023; Wang et al., 2024b) and image-to-video (I2V) (Xing et al., 2025; Zhang et al., 2023) paradigms, driven by large-scale foundation models (Google, 2025; OpenAI, 2025; Research, 2025). Meanwhile, auto-regressive approaches (Weng et al., 2024; Henschel et al., 2024; Yang et al., 2025; Huang et al., 2025) and video extension methods (KlingAI, 2024; Runway, 2024) have attracted growing attention by extending the duration of generated videos. However, these methods still suffer from limited length and are unable to solely complete full AI movie generation. MAViS2 addresses long-video generation from a system-level perspective, expanding the capabilities of existing models and enabling minute-scale video synthesis.

2.2 AI Agent

Empowered by large language models (Touvron et al., 2023; Brown et al., 2020), AI agents (Park et al., 2023; Li et al., 2023; Hong et al., 2023; Park et al., 2023) have demonstrated broad applicability across diverse tasks. Multimodal models (Achiam et al., 2023; Anil et al., 2023) further extend agent capabilities across modalities (Wang et al., 2024d; Hang et al., 2024; Yang et al., 2023b; Zheng et al., 2024), while Chain-of-Thought (CoT) prompting (Wei et al., 2023; Wang et al., 2023) improves agents' multi-step reasoning performance (Schick et al., 2023; Gao et al., 2023). In addition, some works (Liu et al., 2025; Madaan et al., 2023) benefit from HITL interaction (Natarajan et al., 2024). In this work, multiple agents employ CoT-based multi-step reasoning while supporting HITL interaction throughout the generation process.

2.3 Video Agents

Recent advances have spawned video agents for long-sequence generation. Some works (Yuan et al., 2024; Liang et al., 2025) focus on multi-model system integration to better serve creators by orchestrating multiple models and tools, while others (Yang et al., 2024a; Wang et al., 2024a) emphasize agent-driven story planning but lack end-to-end execution and controllability. Existing frameworks (He et al., 2024; Wu et al., 2025b) further over-rely on T2I+I2V paradigm, limiting shot-level flexibility. In contrast, MAViS2 integrates Scriptwriting Workflow, Adaptive Generation Planning, post-production, and HITL interaction for flexible, high-quality video storytelling.

3 MAViS2

We propose **MAViS2**, a multi-agent framework for interactive and adaptive long-sequence video storytelling. As illustrated in Figure 2, given a user prompt P, MAViS2 generates a long-sequence video storytelling that comprises N shots:

$$\mathcal{F} : P \to \hat{\mathcal{V}}, \quad \hat{\mathcal{V}} = \{\{V_i,\ A_i\} \mid i = 1, 2, \ldots, N\}, \quad (1)$$

where V_i is the video clip and A_i is the audio track.

MAViS2 decomposes the long-sequence video storytelling process into three independent yet collaborative stages: scriptwriting, video clip generation, and post-production. In the scriptwriting stage, we propose a **Scriptwriting Workflow** to progressively enhance and enrich the script. In the video clip generation stage, we introduce **Adaptive Generation Planning**, in which MAViS2 selects an appropriate generation strategy for each shot. In the post-production stage, MAViS2 integrates

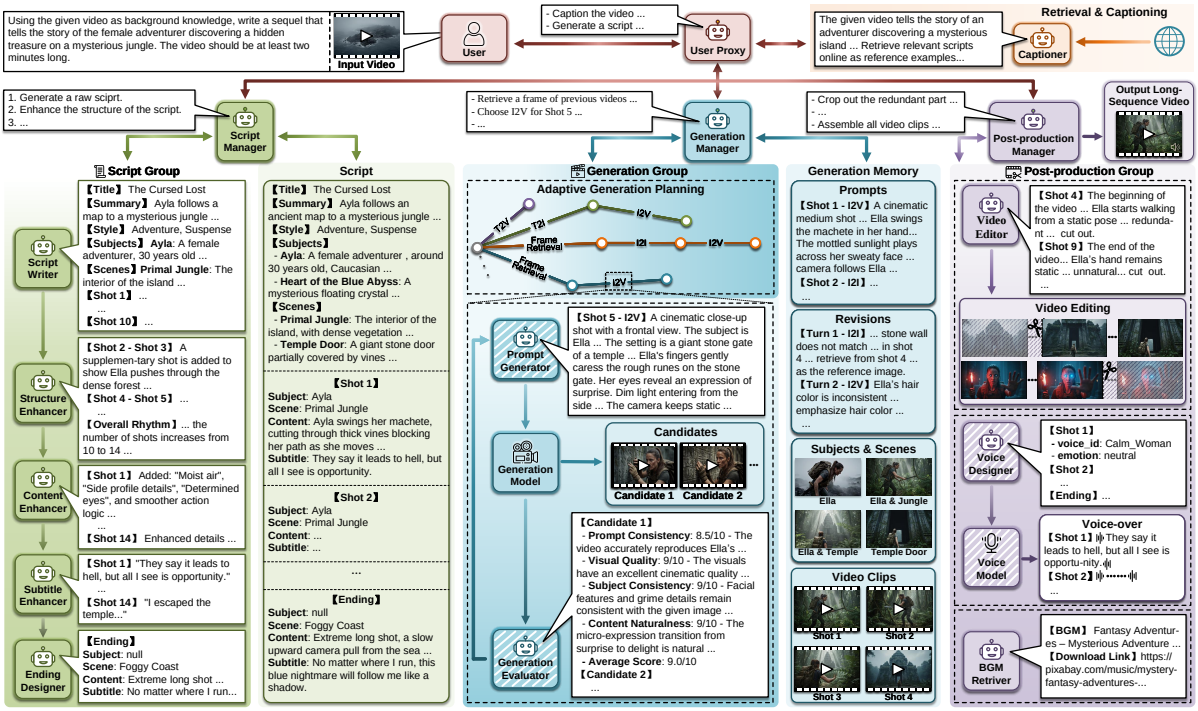

Figure 2: Illustration of MAViS2 framework. ▨ indicates component inherited from MAViS (Wang et al., 2025).

video editing and audio generation to enhance the overall viewing experience and completeness of the final long-sequence video storytelling. Furthermore, through **Fine-grained Human-in-the-loop Control**, MAViS2 allows users to intervene and make fine-grained adjustments at any stage of the generation process, thereby flexibly exploring diverse visual narratives and creative directions.

Section 3.1 outlines the Scriptwriting Workflow, Section 3.2 details Adaptive Generation Planning, Section 3.3 introduces the post-production stage, and Section 3.4 discusses the Fine-grained Human-in-the-loop Control.

3.1 Scriptwriting Workflow

The scripts generated by LLMs in a single turn are often underdeveloped, with abrupt transitions between shots and poor content consistency. This leads to severe fragmentation in the generated long-sequence videos, where adjacent shots lack proper transitions, buildup, and atmospheric reinforcement, ultimately degrading the overall viewing experience. On the other hand, naively injecting a large number of scriptwriting guidelines into the system message tends to disperse the agent's attention, making it difficult to ensure that all rules are consistently satisfied.

To obtain a more complete and expressive script, we propose a **Scriptwriting Workflow** that enhances and enriches the script step by step through a progressive flow of "*initial generation* → *structure enhancement* → *content augmentation* → *narration optimization* → *narrative closure*".

Initial Generation Upon receiving a scriptwriting task, the Scriptwriter G_{sw} generates a structured draft script $S_1 = G_{\text{sw}}(P)$.

Structure Enhancement The Structure Enhancer G_{sse} performs structural refinement on the draft script: $S_2 = G_{\text{sse}}(S_1)$. Without altering the narrative trajectory, it adds supplementary shots—such as multi-angle shots, reverse shots, long takes, and bridging shots—to make proper transitions, buildup, and atmospheric reinforcement.

Content Augmentation The Content Enhancer G_{sce} refines the script, focusing on resolving inconsistencies, smoothing conflicts, and enriching necessary details such as environmental cues and lighting descriptions across shots: $S_3 = G_{\text{sce}}(S_2)$. By addressing these issues at the script level, MAViS2 mitigates visual inconsistency in the resulting long-sequence videos.

Narration Optimization Narration generated directly by LLMs is often overly theatrical, unnatural, or unsuitable for grounded storytelling. Therefore, the Narration Enhancer G_{sne} rewrites the narration to better align with the narrator's perspective and tone, ensuring that it reads as authentic storytelling rather than Shakespearean prose: $S_n = G_{\text{sne}}(S_3)$.

Narrative Closure The Ending Designer G_{sed} crafts a thematically meaningful ending shot as a narrative closure, consolidating the narrative arc and leaving room for audience interpretation and emotional resonance: $S_e = G_{\text{sed}}(S_3)$.

Finally, the Script Manager M_{s} aggregates the refined script, polished narration, and ending shot into the final camera-ready script, which serves as the input for the video clip generation stage:

$$S = M_{\text{s}}(S_3, S_n, S_e). \tag{2}$$

Details of the script structure, supplementary shots, and content augmentation are provided in Appendix B.

3.2 Adaptive Generation Planning

Different shots impose different requirements on visual grounding, subject consistency, and background consistency, and therefore demand different generation strategies. A generation strategy may consist of a single or multiple generation steps, such as T2V, video extension, image retrieval + I2V, T2I + I2V, and image retrieval + image-to-image generation (I2I) + I2V, etc. To address this, we propose **Adaptive Generation Planning**, which enables the agent framework to adaptively select the most appropriate video generation strategy for each shot.

First, for each shot $\hat{s} \in S$, the Generation Manager M_{g} selects a generation strategy $\mathcal{R}$ by jointly considering the shot content, the global script context, and the generation memory $\mathcal{M}$:

$$\mathcal{R} = M_{\text{g}}(\hat{s}, S, \mathcal{M}). \tag{3}$$

The generation memory maintains a record of all previously generated shots, including their prompts, revision histories, subject images, background images, and resulting video clips.

Next, the Prompt Generator G_{gp} composes a structured input prompt corresponding to the generation tool $G_{\text{gt}} \in \mathcal{R}$ of the current step in generation strategy to produce a set of candidate images or video clips $\mathcal{C}_0$ using N_g random seeds:

$$\mathcal{C}_0 = \{G_{\text{gt}}(p_0, k),\ k = 1, 2, \ldots, N_g\},\ p_0 = G_{\text{gp}}(\hat{s}). \tag{4}$$

Subsequently, the Generation Evaluator E_{g} scores these candidates according to prompt consistency, visual quality, content naturalness, and subject consistency, selects the best candidate c_0, and provides the evaluation result e_0:

$$[c_0, e_0] = E_{\text{g}}(\mathcal{C}_0) \tag{5}$$

Then, the Prompt Generator analyzes the identified deficiencies of the selected candidate and initiates the next iteration of prompt generation:

$$p_{j+1} = G_{\text{gp}}(\hat{s}, e_j, p_j) \tag{6}$$

This internal loop of generation–evaluation–refinement provides the Generation Manager a provisional best candidate. It then analyzes the remaining issues and decides whether to further adjust the shot content, proceed to the next step in the generation strategy, or reselect a different generation strategy. Once the video clip for a shot is finalized, the Generation Manager stores the resulting clip together with its associated subject and background images in the generation memory, providing visual and contextual references for subsequent shots.

Through this adaptive and feedback-driven process, MAViS2 ultimately generates a sequence of video clips $\{V_i \mid i = 1, 2, \ldots, N\}$. Details of generation planning, input prompt structure, and evaluation are provided in Appendix C.

3.3 Post-production Stage

With the generated video clips, MAViS2 performs video editing, voice design and synthesis, BGM matching, and final video assembly in the post-production stage.

First, the Video Editor G_{pe} trims each video clip according to the script[1], removing redundant segments or content that is misaligned with the script:

$$\mathcal{V} = \{G_{\text{pe}}(V_i, S),\ i = 1, 2, \ldots, N\}. \tag{7}$$

Next, the Voice Designer G_{pv} assigns a voice identity (voice ID) and appropriate emotional tone to each shot, and synthesizes the corresponding voice-over using a text-to-speech (TTS) model G_{pt}.

$$\mathcal{A} = \{G_{\text{pt}}(\hat{s}, S_v),\ \forall \hat{s} \in S\}, \quad S_v = G_{\text{pv}}(S, \mathcal{V}). \tag{8}$$

Finally, the BGM Retriever G_{pb} selects and downloads a BGM, and the Post-production Manager M_{p} synchronizes and assembles the edited video clips, voice-overs, subtitles, and BGM into the final long-sequence video storytelling:

$$\hat{\mathcal{V}} = M_{\text{p}}(\mathcal{V}, \mathcal{A}, S, \text{BGM}), \quad \text{BGM} = G_{\text{pb}}(S). \tag{9}$$

3.4 Fine-grained Human-in-the-loop Control

MAViS2 supports **Fine-grained Human-in-the-loop Control**, enabling interactive user intervention and fine-grained adjustments at any stage of

[1] In addition to video trimming, the Video Editor also supports invoking multiple video editing models.

Method	Keyframe Generation		Video Clip Generation					
	CLIP ↑	Inception ↑	T. Flick. ↑	M. Smooth. ↑	Sub. Cons. ↑	Bg. Cons. ↑	Aesthetic ↑	I. Quality ↑
VGoT (Zheng et al., 2025)	22.37	8.53	99.00	99.31	**98.94**	**98.74**	**80.11**	64.59
Mora (Yuan et al., 2024)	33.98	12.68	97.32	99.23	95.23	94.88	64.36	71.48
MovieAgent (Wu et al., 2025b)	31.71	10.72	97.07	99.20	93.29	94.51	63.39	71.26
Univa (Liang et al., 2025)	34.32	12.84	99.08	99.33	95.63	96.08	63.14	72.95
MAViS (Wang et al., 2025)	34.22	12.81	**99.09**	99.53	95.72	96.12	63.17	72.91
MAViS2 (ours)	**34.45**	**12.92**	99.02	**99.56**	95.81	96.27	64.25	**73.08**

Table 2: **Evaluation of Automatic Metrics for Keyframe Generation and Video Clip Generation.** "T. Flick.", "M. Smooth.", "Sub. Cons.", "Bg. Cons.", "Aesthetic", and "I. Quality" refer to "Temporal Flickering", "Motion Smoothness", "Subject Consistency", "Background Consistency", 'Aesthetic Quality', and "Imaging Quality" from metrics in VBench, respectively. ↑ indicates that a higher value is more desirable. **Bold** indicates the best results.

Method	Narrative ↑	Visual ↑	User Align. ↑	Sub. Cons. ↑	Sub. Natural. ↑	Bg. Cons. ↑	Bg. Real. ↑
VGoT (Zheng et al., 2025)	0.56	1.41	1.40	2.25	1.12	1.69	2.25
Mora (Yuan et al., 2024)	3.09	4.51	3.93	4.49	3.65	5.65	3.10
MovieAgent (Wu et al., 2025b)	4.21	4.23	4.49	4.49	4.21	3.95	3.10
Univa (Liang et al., 2025)	15.45	18.03	16.57	17.13	17.13	16.67	16.34
MAViS (Wang et al., 2025)	14.89	13.52	18.26	14.89	16.85	15.82	15.21
MAViS2 (ours)	**61.80**	**58.31**	**55.34**	**56.74**	**57.02**	**56.21**	**60.00**

Table 3: **User Study on the Performance of Long-Sequence Video Storytelling** (Voting Results). "Narrative", "Visual", "User Align.", "Sub. Cons.", "Sub. Natural.", "Bg. Cons.", and "Bg. Real." refer to "Narrative Expressiveness", "Visual Quality", "User Prompt Alignment", "Subject Consistency", "Character Naturalness", "Background Consistency", and "Background Realism", respectively. ↑ indicates that a higher value is more desirable. **Bold** indicates the best results.

the generation process. The interactions are jointly handled by the User Proxy and Group Managers.

Given that the overall generation pipeline involves multiple agents, various tasks, and diverse working guidelines, relying solely on the User Proxy for intent understanding, fine-grained task decomposition, and task execution would be unreliable and unstable. To ensure orderly collaboration, MAViS2 adopts a hierarchical pyramid-structured organization consisting of the User Proxy, Group Managers, and specialized agents. Specifically:

• The User Proxy receives user instructions, decomposes them into group-level tasks, and dispatches these tasks to the corresponding Group Managers.

• Each Group Manager further decomposes a group-level task into a sequence of individual tasks and selects suitable specialized agents.

• After completing their assigned tasks, the specialized agents return their outputs to the Group Manager, which aggregates the results and submits a consolidated response back to the User Proxy.

• Guided by the 3E (Explore–Examine–Enhance) Principle (Wang et al., 2025), the Group Manager reviews and refines specialized agent outputs to ensure generation quality at each stage and progressive optimization throughout the pipeline.

Given multimodal inputs, MAViS2 can generate long-sequence video storytelling with targeted styles and content tailored to user needs.

4 Experiment

To evaluate the effectiveness of MAViS2, we benchmark it against five representative long video generation baselines and conduct detailed ablation studies. We use the same test set as MAViS (Wang et al., 2025), which consists of 20 user prompts, each requiring the generation of a video of at least one minute with a specific style and content. Consistent with MAViS, we use CLIP, Inception Score, and VBench (Huang et al., 2024) for quantitative evaluation, along with a user study for subjective assessment. Further experimental settings and results are provided in Appendix D, E, and F. Due to page and file size limits, refer to "Software" and "Data" supplementary files for qualitative comparisons (*qualitative_comparisons.html*) and applications (*sequel_generation_application.html*).

4.1 Performance Comparisons

We evaluate each method via automatic metrics and a user study. Since MAViS2 adopts different generation strategies for different shots, we extract the first frame from each shot's video clip and use these frames for comparison in keyframe generation. To ensure fairness, the generated results of MAViS2 have not undergone any manual intervention.

4.1.1 Quantitative Evaluation

Table 2 presents shot-level evaluation results for keyframe and video clip generation. MAViS2 achieves the highest CLIP (*34.45*) and Inception (*12.92*) scores, benefiting from its internal

Method	Narrative ↑	Visual ↑	User Align. ↑	Sub. Cons. ↑	Sub. Natural. ↑	Bg. Cons. ↑	Bg. Real. ↑
Baseline MAViS (Wang et al., 2025)	0.58	6.30	10.58	11.25	9.62	17.25	8.77
+ Scriptwriting Workflow	17.62	6.18	19.12	6.34	7.43	0.62	7.42
+ Adaptive Generation Planning	21.41	26.43	21.44	26.16	23.39	25.35	22.91
+ Post-production	26.52	28.50	19.34	26.20	23.42	25.65	22.56
+ Fine-grained HITL Control (**MAViS2**)	**33.87**	**32.59**	**29.52**	**30.05**	**36.14**	**31.13**	**38.34**

Table 4: **Comparison of incremental variants built upon MAViS, culminating in MAViS2.** (Voting Results). Underlined values indicate a significant improvement compared to the previous variants.

Ablation	Narrative ↑	Sub Cons. ↑	Bg. Cons. ↑
w/o Structure Enhancer	3.24	15.47	12.06
w/o Content Enhancer	24.26	6.37	5.81
w/o Narration Enhancer	8.17	25.42	26.94
w/o Ending Designer	9.85	26.07	26.61
MAViS2 (ours)	**54.48**	**26.67**	**28.58**

Table 5: **Ablation study on the Scriptwriting Workflow** (Voting Results). Gray-shaded cells indicate values that are significantly lower than the others.

Ablation	CLIP ↑	Inception ↑	Natural. ↑	Sub. Cons. ↑	Bg. Cons. ↑
w/o IL	31.16	12.51	62.37	93.25	94.76
w/o Adp.	33.95	12.74	92.43	93.86	93.25
w/o IL+Adp.	31.08	12.42	58.62	93.11	93.19
MAViS2	**34.45**	**12.92**	**94.85**	**95.81**	**96.27**

Table 6: **Ablation study on Adaptive Generation Planning**. "IL" refers to the internal loop of generation–evaluation–refinement. "Adp." refers to adaptive generation strategy adjustment. "Natural." refers to the Naturalness score.

generation–evaluation–refinement loop. Although MAViS2 shows slightly lower temporal flickering than MAViS—likely due to rich lighting effects and dynamic camera motions—it outperforms all baselines in motion smoothness and image quality. MAViS2 performs slightly worse than VGoT in subject consistency, background consistency, and aesthetic quality, but outperforms all baselines in motion smoothness and image quality, as VGoT favors low-motion dynamics and highly saturated styles that enhance consistency and aesthetics at the expense of realism. Overall, MAViS2 achieves the best comprehensive performance.

4.1.2 User Study

Table 3 reports the voting results of 50 evaluators across multiple evaluation dimensions for all baseline methods. As shown, MAViS2 outperforms all baselines on every metric, achieving particularly strong results in narrative expressiveness and background realism, with scores of 61.80% and 60.00%, respectively. These gains can be attributed to the Scriptwriting Workflow, which enriches the narrative structure and enhances expressive storytelling, and to Adaptive Generation Planning, which maintains consistency through diversified generation strategies. Overall, MAViS2 attains an average vote share of 57.92% across all dimensions, demonstrating its superiority and effectiveness in long-sequence video storytelling.

4.2 Ablation Study and Analysis

We conduct ablation studies and analyses to verify the contributions of each component in MAViS2.

4.2.1 Comparison with MAViS

Table 4 compares incremental variants starting from MAViS, where components are added one by one, culminating in the full MAViS2 (MAViS2). Adding the Scriptwriting Workflow significantly improves narrative expressiveness, but degrades subject and background consistency. This is because it enriches the script, while the original MAViS pipeline cannot fully accommodate the newly expanded scripts during generation. Introducing Adaptive Generation Planning leads to substantial improvements in Visual Quality as well as subject and background consistency, as it adaptively selects appropriate generation strategies for each shot. Incorporating Post-production further improves multiple metrics. Finally, the Fine-grained Human-in-the-loop Control further enhances subject naturalness and background realism, since it allows the system to refine generation results based on real-time user feedback.

4.2.2 Scriptwriting Workflow

Table 5 presents the effects of removing each agent from the Scriptwriting Workflow. Removing the Structure Enhancer severely degrades narrative expressiveness, as the script lacks sufficient buildup and narrative progression. Removing the Content Enhancer negatively affects both subject consistency and background consistency, since contradictions and inconsistencies in the script are no longer identified and corrected. Excluding the Narration Enhancer and Ending Designer also harms narrative expressiveness, as the narration becomes abrupt and the story lacks a coherent narrative closure at the end. Optimal script generation is achieved only when all agents collaborate.

4.2.3 Adaptive Generation Planning

We report the ablation results of Adaptive Generation Planning in Table 6, where the Naturalness

Duration (seconds)	Narrative ↑	Sub. Natural ↑	Bg. Real. ↑
0	6.35	2.51	1.97
2	18.62	19.35	20.73
4 (ours)	**44.85**	**46.72**	**45.21**
6	30.18	31.42	32.09

Table 7: **Analysis of Maximum Video Trimming Duration** (Voting Results). **Bold** indicates the best results.

Ablation	TS. Rate ↑	OC. Rate ↑	API calls ↓
w/o manager feedback	83.50	72.30	15.50
w/o manager	71.65	69.55	**8.50**
MAViS2 (ours)	**98.60**	**99.85**	36.00

Table 8: **Ablation study on Fine-grained Human-in-the-loop Control**. "TS. Rate" and "OC. Rate" refer to "Task Success Rate" and "Output Completeness Rate", respectively. ↑ indicates that a higher value is more desirable. ↓ indicates that a lower value is more desirable.

score is defined as the proportion of generated images or videos that appear physically plausible and naturally rendered. As shown, removing the internal loop of generation–evaluation–refinement leads to a noticeable degradation in visual quality, with both CLIP and Inception scores being significantly lower than those of other ablation settings; meanwhile, naturalness also declines. Without adaptive generation strategy adjustment, subject consistency and background consistency are adversely affected. When the entire Adaptive Generation Planning module (IL + Adp.) is removed, all metrics drop substantially. These results demonstrate the essential role of Adaptive Generation Planning throughout the overall pipeline.

4.2.4 Video Editing

We investigate the impact of the maximum trimming duration of the Video Editor on generated video clips in Table 7. As shown, all metrics increase as the trimming duration grows, reaching their peak when the duration is set to 4 seconds. This improvement arises because the Video Editor can remove more redundant content and segments that are misaligned with the script. When the trimming duration exceeds 4 seconds, the metrics begin to decline, likely due to over-trimming that removes informative or semantically important content. Therefore, setting the maximum trimming duration to 4 seconds is the most reasonable choice.

4.2.5 Fine-grained Human-in-the-loop Control

The key distinction between MAViS2's HITL and conventional HITL lies in the introduction of group managers, which enable fine-grained task decomposition and provide feedback on the outputs of specialized agents. In Table 8, we report the Task Success Rate (the proportion of completed tasks over the total number of tasks), Output Completeness Rate (the proportion of outputs that are complete and correctly formatted), and the number of API calls required to complete a task under different ablation settings. As shown, disabling manager feedback leads to a noticeable drop in output completeness, indicating that without group-manager supervision, the output quality of specialized agents degrades. Removing the group managers entirely results in a substantial decrease in the task success rate because the User Proxy alone cannot effectively manage a large number of specialized agents and fine-grained outputs when handling complex tasks. Although introducing group managers increases the number of API calls, this trade-off is worthwhile for improving the overall stability and reliability of the system.

5 Applications

To further extend its video storytelling capability, MAViS2 leverages video captioning and RAG to support a broader range of long-sequence video storytelling tasks. When users provide an input video, the Captioner performs video understanding and collaborates with the scriptwriting group to extract a script, enabling video remaking. When users request a prequel or sequel, MAViS2 retrieves relevant external knowledge from Wikipedia to generate the corresponding script and videos. Example applications are provided in the Appendix G.

6 Conclusion

In this work, we present MAViS2, a multi-agent framework for interactive and adaptive long-sequence video storytelling. By orchestrating the Scriptwriting Workflow, Adaptive Generation Planning, Post-production, and Fine-grained Human-in-the-loop Control, MAViS2 effectively addresses key limitations of prior works, including underdeveloped scriptwriting, reliance on a fixed T2I+I2V paradigm, limited post-production support, and insufficient user interaction. Leveraging RAG and video captioning, MAViS2 supports a wide range of storytelling tasks, including end-to-end generation, video understanding, external knowledge retrieval from Wikipedia, prequel and sequel creation, and video remaking. Experimental results demonstrate that MAViS2 achieves state-of-the-art expressiveness in long-sequence video storytelling.

Limitations

One limitation of this work stems from the limited reliability of the Generation Evaluator. Since current multimodal large language models still exhibit limited accuracy in evaluating visual content, the Generation Evaluator does not achieve high accuracy when scoring generated candidates. In practical usage, this issue can be mitigated by incorporating additional evaluation operators—such as the prompt consistency and image quality metrics from VBench–to assist the Evaluators.

Moreover, because MAViS2 supports Fine-grained Human-in-the-loop Control, users can freely select alternative candidates if they are dissatisfied with the one chosen by the agent, or modify prompts and regenerate images and videos. This interactive mechanism partially compensates for the limited reliability of automated Evaluators.

Similarly, due to the limited accuracy of current multimodal large models in evaluating visual content, another limitation of this work lies in the effectiveness of the Video Editor during video trimming. To mitigate this issue, a maximum trimming duration is imposed to prevent excessive or inappropriate edits by the video editor. Moreover, due to the model-agnostic nature of MAViS2, additional editing models can be incorporated into the Video Editor's model pool in the future, enabling more diverse and advanced video editing functionalities.

Ethical Considerations

MAViS2 is a research framework for controllable long-sequence video storytelling based on multi-agent coordination. As it integrates multiple pre-trained multimodal models and external tools, the generated outputs may inherit biases, inaccuracies, or safety limitations from these components. We encourage transparent disclosure of model usage and responsible practices when applying MAViS2 in research or creative contexts. In addition, MAViS2 supports fine-grained human-in-the-loop interaction, placing responsibility on users to guide, review, and refine generated content in accordance with ethical norms.

Potential Risks

Due to its ability to generate coherent and realistic long-form videos with synchronized audio and post-production refinement, MAViS2 may be misused to create misleading or deceptive synthetic media, such as disinformation or fabricated narratives. Furthermore, reliance on large pretrained models and external knowledge sources may unintentionally propagate societal biases or unverified information. MAViS2 is intended for academic research and creative exploration, not for deployment in deceptive, high-stakes, or adversarial scenarios, and responsible human oversight is essential to mitigate potential risks.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. *Preprint*, arXiv:2211.10435.

GenmoTeam. 2024. Mochi 1. `https://github.com/genmoai/models`.

DeepMind Google. 2025. Veo 3.1.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. 2025. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*.

Tiankai Hang, Shuyang Gu, Dong Chen, Xin Geng, and Baining Guo. 2024. Cca: Collaborative competitive agents for image editing. *arXiv preprint arXiv:2401.13011*.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. 2024. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *arXiv preprint arXiv:2407.12899*.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*.

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. *arXiv preprint arXiv:2403.14773*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

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

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

KlingAI. 2024. Kling-v1.5.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Zhengyang Liang, Daoan Zhang, Huichi Zhou, Rui Huang, Bobo Li, Yuechen Zhang, Shengqiong Wu, Xiaohan Wang, Jiebo Luo, Lizi Liao, and 1 others. 2025. Univa: Universal video agent towards open-source next-generation video generalist. *arXiv preprint arXiv:2511.08521*.

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

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2025. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision*, pages 126–142. Springer.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

MiniMax. 2024. Hailuoai.

The ModelScopeTeam. 2023. Modelscope: bring the notion of model-as-a-service to life.

Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. 2024. Human-in-the-loop or ai-in-the-loop? automate or collaborate? *Preprint*, arXiv:2412.14232.

OpenAI. 2024. Creating video from text.

OpenAI. 2025. Sora 2. https://openai.com/index/sora-2.

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. arxiv. *arXiv preprint ArXiv:2304.03442*.

Runway Research. 2025. Gen 4.5. https://runwayml.com/research/introducing-runway-gen-4.5.

Runway. 2024. Gen-3.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jiuniu Wang, Zehua Du, Yuyuan Zhao, Bo Yuan, Kexiang Wang, Jian Liang, Yaxi Zhao, Yihen Lu, Gengliang Li, Junlong Gao, and 1 others. 2024a. Aesopagent: Agent-driven evolutionary system on story-to-video production. *arXiv preprint arXiv:2403.07952*.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. 2025. Mavis: A multi-agent framework for long-sequence video storytelling. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, and 1 others. 2024b. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*.

Yaohui Wang, Xin Ma, Xinyuan Chen, Cunjian Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. 2024c. Leo: Generative latent image animator for human video synthesis. *International Journal of Computer Vision*, pages 1–13.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. 2024d. Genartist: Multimodal llm as an agent for unified image generation and editing. *arXiv preprint arXiv:2407.05600*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, and 1 others. 2024. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7395–7405.

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

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

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 preprint arXiv:2408.11788*.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2025. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer.

Deshun Yang, Luhui Hu, Yu Tian, Zihao Li, Chris Kelly, Bang Yang, Cindy Yang, and Yuexian Zou. 2024a. Worldgpt: a sora-inspired video ai agent as rich world models from text and image inputs. *arXiv preprint arXiv:2403.07944*.

Hui Yang, Sifu Yue, and Yunzhong He. 2023a. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024b. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. 2025. Longlive: Real-time interactive long video generation. *Preprint*, arXiv:2509.22622.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *Preprint*, arXiv:2303.11381.

Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, and Bill Yuchen Lin. 2023. Lumos: Learning agents with unified data, modular design, and open-source llms. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

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

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

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. 2024. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and 1 others. 2025. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2503.15138*.

Cailin Zhuang, Ailin Huang, Wei Cheng, Jingwei Wu, Yaoqi Hu, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, and 1 others. 2025. Vistorybench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*.

A Positioning and Design Rationale

In this section, we restate the motivation of MAViS2, explain the relationship between MAViS2 and the generative models, and elaborate on the improvement over MAViS (Wang et al., 2025).

A.1 Motivation

MAViS2 is motivated by a set of long-standing yet insufficiently addressed challenges in practical long-sequence video generation with generative models. These challenges include:

- The lack of rich and filmable scripts.
- Cross-shot consistency of subjects and backgrounds.
- Generation stability.
- Fully automated, end-to-end execution from scriptwriting to narratively complete videos.
- Fine-grained human-in-the-loop control.

Despite their critical importance in real-world video storytelling, these issues have not been systematically resolved by existing approaches.

A.2 Relationship to Generative Models

The aforementioned challenges cannot be effectively addressed by introducing a new generative model alone; instead, they require solutions at the system level. MAViS2 is therefore designed as a more meta-level framework that is agnostic to any specific generative model and can seamlessly integrate with existing or future generation models. This model-agnostic design enables MAViS2 to naturally evolve alongside advances in next-generation GenAI models, ensuring long-term scalability and sustained performance improvements. Moreover, MAViS2 amplifies the capabilities of current generative models, significantly enhancing their performance in long-sequence video storytelling. In contrast, incorporating a specially tailored new model into the framework would reduce MAViS2's adaptability and undermine the generality and long-term value of this work.

A.3 Improvements over MAViS

Compared to MAViS, MAViS2 introduces a Scriptwriting Workflow that produces richer and more filmable scripts. Through Adaptive Generation Planning, MAViS2 preserves scriptwriting flexibility, improves visual diversity, and removes the reliance on training task-specific LoRA models. More importantly, MAViS2 introduces the Fine-grained Human-in-the-loop Control that allows users to intervene in and refine the generated content in real time, thereby facilitating flexible exploration of diverse visual narratives and creative directions. In addition, MAViS2 extends the framework with video editing, video captioning, and RAG, enabling broader applications and delivering a more coherent and engaging viewing experience for long-sequence video storytelling. Despite these extensions, MAViS2 retains the 3E principle (Explore–Examine–Enhance) inherited from MAViS, which continues to play a central role in improving the quality and stability of generation results across all stages.

B Details of the Scriptwriting Workflow

Scripts from real-world productions are typically not designed with the capability boundaries of GenAI tools or script–tool compatibility in mind, making them difficult to directly use for long-sequence video storytelling. Meanwhile, scripts for AI short films are scarce and hard to obtain, which makes training a dedicated scriptwriting model impractical in the short term. Moreover, scripts written directly by LLMs often lack sufficient buildup and filmability. Therefore, a dedicated mechanism is required to generate rich and directly usable scripts for GenAI tools, and the Scriptwriting Workflow is proposed to address this need.

Owing to space constraints in the main text, the Scriptwriting Workflow is described only at a high level. In this section, we elaborate on the script structure, the supplementary shots added by the Structure Enhancer, and the details of content augmentation performed by the Content Enhancer.

B.1 Script Structure

A generated script must include the following items:

- Title.
- Summary: A brief overview of the script, summarizing the background, development, and conclusion of the event.
- Style: The style of the story, such as "action," "sci-fi," "historical," etc.
- Narrator: The character who tells the entire story.
- Subject Appearance.
- Background Description.
- Shot-by-Shot Script: Each shot must include the following fields:

Agent/Model	Role
HITL	
User Proxy	Receive user instructions, decompose them into group-level tasks and dispatch them to the corresponding Group Managers.
Captioner	Video understanding and captioning.
Script Group	
Script Manager	Decompose script task, dispatch task to a suitable specialized agent, and provide feedback on the agents' outputs.
Scriptwriter	Generate a raw script.
Structure Enhancer	Structure enhancement for the script.
Content Enhancer	Content augmentation for the script.
Narration Enhancer	Narration optimization for the script.
Ending Designer	Design an ending shot for narrative closure.
Generation Group	
Generation Manager	Decompose generation tasks, dispatch tasks to a suitable specialized agent, and provide feedback on the agents' outputs.
Prompt Generator	Generate prompts for candidate generation.
Generation Model	Generate image and video candidates.
Generation Evaluator	Evaluate the generated candidates.
Post-production Group	
Post-production Manager	Decompose post-production tasks, dispatch task to a suitable specialized agent, and provide feedback on the agents' outputs.
Video Editor	Edit the video clips, removing redundant content and segments that are misaligned with the script.
Voice Designer	Configure voice ID and emotion for voice-overs.
Voice Model	Text-to-speech generation.
BGM Retriever	Select a background music for the final video storytelling.

Table 9: Summary of All Agents in MAViS2

• Shot Number: Starting from Shot 1 and incrementing sequentially (Shot 2, Shot 3, etc.).

• Subject: The subject appearing in the shot (e.g., "Jack").

• Background: The location, scene, or environment where the shot is filmed.

• Shot Content: Describes what happens in this shot, including the time and location, as well as details about the shot scale, perspective, camera movement, and lighting.

• Subtitle: The narration corresponding to the shot content.

B.2 Supplementary Shots

The Structure Enhancer is designed to add the following shots to enhance the atmosphere, drive the story forward, improve the continuity of the plot, and enrich the development of events:

• Multiple Angles: Use the same subject or background as an existing shot but vary the shot scale, camera angle, position, or movement. These shots can set the atmosphere and make the video more dynamic.

• Reverse Shots: Insert shots between two existing shots. The subject and background in the inserted shots can match one of the existing shots. These shots show both sides of a conversation or interaction, allowing the audience to see the reactions from each perspective.

• Long Takes: Add a shot immediately following another, using the same subject, background, shot scale, camera angle, and position, but with different content, creating a seamless long take. These shots maintain continuity in time and space, immersing the audience in a complete background.

• Bridging Shots: Add a shot between two backgrounds with large plot jumps to explain the development of events between them. These shots increase plot continuity, helping the audience understand the connection between the two backgrounds.

B.3 Script Content Augmentation

The Content Enhancer examines the script from the following aspects:

• Whether the descriptions of each subject and background are sufficiently detailed and remain consistent throughout the entire script.

• Whether the descriptions of subjects and backgrounds conflict with the content of specific shots.

• Whether sub-backgrounds are consistent with the descriptions of their corresponding parent backgrounds, including both semantic coherence and the alignment of details.

• For multiple shots occurring within the same background, whether any inconsistencies in environmental descriptions across shots are reasonable and well justified.

After identifying conflicts and inconsistencies, the Content Enhancer adds or removes necessary details to resolve these issues and refine the script.

C Details of Adaptive Generation Planning

In this section, we elaborate on generation planning, the prompt structure, and candidate evaluation.

C.1 Generation Planning

The Generation Manager selects an initial generation strategy for each shot based on the following criteria:

• T2V: Used when no reference images are available, or when the subject and background of the current shot appear for the first time.

• Image Retrieval + I2V: Applied when the current shot is an extension of a previously generated shot.

• T2I + I2V: Adopted when T2V produces unsatisfactory results, or when the shot involves specific textual elements that require explicit visual grounding.

• Video Extension: Used when the current shot can be merged with the preceding shot to form a continuous long take.

• Image Retrieval + I2I + I2V: Employed when suitable reference images can be retrieved from previously generated shots to ensure consistency of subjects and backgrounds.

C.2 Prompt Structure

Each image generation prompt must specify six elements: shot scale, subject, background, initial state of the shot, lighting, initial camera position, and the usage of reference images. Example:

Create a cinematic medium shot from a side follow perspective. The image shows Ella, a female adventurer around 30 years old, white, with an athletic build and long brown hair tied in a high ponytail. Her face is marked with dirt and sweat. She is dressed in a khaki tactical vest, dark green cargo pants, dark brown, worn leather boots, and carries a worn canvas backpack. She is captured in the action of swinging a machete to forcefully cut through thick vines blocking her path, walking firmly. The setting is a primeval dense forest on an island, characterized by extremely dense vegetation where giant ferns and twisting vines obscure the sky. The atmosphere is humid and hot, with dappled sunlight streaming through the leaf gaps and illuminating her sweaty face. My first image shows Ella.

Each prompt for video generation must specify six elements: shot scale, subject, background, event, lighting, and camera movement. Example:

A photorealistic close-up shot with an over-the-shoulder angle, set inside a dimly lit army command tent. A general, age 45, white-skinned, with silver-white tied-back hair and a brown woolen coat, is studying a large military map spread across a wooden table. His hand rests firmly on the map as he analyzes the current battlefield situation with intense focus. Around the map are a teacup, a military compass, and a measuring ruler. The camera slowly tilts upward from the map to reveal the general's serious, contemplative expression. A single lantern with a lit candle serves as the primary light source, casting warm, soft illumination and delicate shadows across the scene. The visual style is historical realism — evoking nostalgia, quiet intensity, and the thoughtful mood of wartime decision-making. Outside the tent, the distant whinnies of warhorses and the rhythmic drills of Continental soldiers can be heard.

Methods	FID↓	Inception↑	CLIP↑
DreamFactory	$7.03^{\dagger}$	$169.71^{\dagger}$	$30.92^{\dagger}$
MAViS2	-	12.92	34.45

Table 10: Compare with DreamFactory (Xie et al., 2024). † denotes the reported score in the paper of the baseline method.

Methods	CLIP↑	Inception↑	AS↑	CLIP-sim↑
MovieDreamer	$19.52^{\dagger}$	$8.64^{\dagger}$	$6.05^{\dagger}$	$0.70^{\dagger}$
MAViS2	34.45	12.92	7.36	0.73

Table 11: Compare with MovieDreamer (Zhao et al., 2024). † denotes the reported score in the paper of the baseline method. "AS" denotes Aesthetic Score used in (Schuhmann et al., 2022).

C.3 Candidate Evaluation

The candidate evaluation guide assesses candidates based on three axes:

• Visual Quality: clarity, detail resolution, and aesthetic harmony;

• Naturalness: identifying distortions, anatomical inaccuracies, or proportion inconsistencies;

• Prompt Consistency: whether the generated content aligns with the input prompt.

• Subject Consistency: temporal stability in facial features, attire, hairstyle, and skin tone across frames;

• Reference Image Consistency: whether the generated content aligns with the reference images.

The candidate evaluation guides work in concert to ensure the completeness of the final generation.

C.4 Video Editor

The preceding sections primarily describe how the Video Editor performs intelligent video trimming. Beyond video trimming, the Video Editor also supports invoking multiple video editing models to perform video editing according to user requirements. These models include object swapping / editing, video enhancement, repainting, background swapping / editing, and style transfer.

Methods	Aesthetic↑	Image Quality↑	Consistency (avg.)↑	P. Consistency↑
LCT	60.79†	67.44†	95.65†	30.14†
MAViS2	64.25	73.08	96.04	37.31

Table 12: Compare with LCT (Guo et al., 2025). † denotes the reported score in the paper of the baseline method. "Consistency (avg.)" represents the average score of subject and background consistency. "P. Consistency" denotes Prompt Consistency.

D Implementation Details

In this section, we provide the implementation details of MAViS2 and baseline methods.

D.1 MAViS2

All agents in our framework are powered by Gemini-3-pro (Anil et al., 2023). We use Gemini-2.5-flash-image for image generation, Veo3 (Google, 2025) for video generation, and Minimax Hailuo (MiniMax, 2024) for text-to-speech generation. Each generated script undergoes two iterations of structure enhancement and one iteration of content augmentation. The number of internal generation–evaluation–refinement loops is set to two, with each loop generating five candidates. The maximum number of 3E iterations is set to 3 for the Script Manager, 2 for the Generation Manager, and 2 for the Post-production Manager. Each script undergoes 2 rounds of structure enhancement and 1 round of content augmentation. The maximum video trimming duration is set to 4. All generated videos have a resolution of 1280×720.

D.2 Baselines

We compare MAViS2 with five long video generation baselines: VGoT, Mora, MovieAgent, Univa, and MAViS. Among them, Mora, MovieAgent, Univa, and MAViS adopt modular designs and are equipped with the same video generation models as MAViS2, and are also powered by Gemini-3-pro. Since Mora and MovieAgent do not support scriptwriting or storyboard generation, we provide them with scripts generated by ChatGPT. For MovieAgent, which requires manual training of LoRA models, we generate character images based on the provided scripts and train LoRA models for each script.

D.3 User Study

In the user study, evaluators were asked to vote for the best video in each group based on seven evaluation criteria listed in Table 13. The criteria are consistent with those used in MAViS (Wang et al., 2025), including narrative expressiveness, Visual Quality, User Prompt Alignment, Character Consistency, Character Naturalness, Background Consistency, and Background Realism. Before voting, we provided detailed guidelines for each metric to ensure consistency and reliability in the evaluation. Screenshots of the user study interface are shown in Figures 8, 9, and 10.

E Dataset

We use the same test set as MAViS (Wang et al., 2025), which consists of 20 user prompts, each requiring the generation of a video of at least one minute with a specific style and content. In this section, we explain the reasons for using such a dataset and report the experimental results on other public datasets.

E.1 Dataset for Testing

Each long-sequence video generated by MAViS2 contains an average of 18 shots. Each shot undergoes 3 cycles of the Group Manager's 3E iteration, each involving 2 rounds of generation–evaluation–refinement, and we generate 5 candidates per round. Our test data consists of 20 ChatGPT-generated random user inputs, covering a wide range of video styles. As a result, the total number of generated videos is $20 \times 18 \times 2 \times 5 \times 3 = 10800$. This represents a substantial computational workload. When combined with image generation, TTS generation, and the videos generated by baseline methods, the required time and computational resources are considerable.

For long-sequence videos, evaluation focuses primarily on viewer experience: if even a single shot fails, the entire video's coherence is broken. Thus, 20 test samples—equivalent to about 50 minutes of video—are entirely sufficient for users to provide an objective assessment of MAViS2.

E.2 Evaluation on the Existing Dataset

Existing public benchmarks for long-sequence video generation (e.g., MovieBench (Wu et al., 2025a), ViStoryBench (Zhuang et al., 2025)) evaluate per-shot similarity between generated videos and ground-truth clips using metrics such as FID and CLIP. They do not assess the overall viewing experience of a long-sequence video and ignore the script generation and scripts' compatibility with existing T2I and I2V models' capabilities. The scripts used in these benchmarks are adapted from

Metric	Question
Narrative Expressiveness	Which video performed best overall in terms of narrative engagement, story coherence, and viewing experience?
Visual Quality	Which video was the most visually expressive in terms of emotion, atmosphere, and cinematic feel?
User Prompt Alignment	Which video best aligns with the original user prompt (e.g., character setup, plot elements, or keywords)?
Character Consistency	Which video had the most consistent character appearances across scenes? (clothing, hairstyle, facial features)
Character Naturalness	Which video had the most natural and correct character generation? Consider anatomical realism, physical motion plausibility, and the absence of duplicate or mistakenly generated characters.
Background Consistency	Which video had the most consistent and logically coherent background style across shots?
Background Realism	Which video had the most natural background elements and physically plausible movements?

Table 13: Metrics and Questions for User Study.

Method	Keyframe Generation		Video Clip Generation					
	CLIP ↑	Inception ↑	T. Flick. ↑	M. Smooth. ↑	Sub. Cons. ↑	Bg. Cons. ↑	Aesthetic ↑	I. Quality ↑
VGoT (Zheng et al., 2025)	21.96	8.43	99.01	99.26	**98.95**	**98.74**	**79.96**	64.33
Mora (Yuan et al., 2024)	34.05	12.69	97.26	99.15	94.62	95.37	64.02	71.44
MovieAgent (Wu et al., 2025b)	31.70	10.83	97.02	99.18	93.43	94.68	63.40	91.31
Univa (Liang et al., 2025)	34.22	12.84	99.06	99.31	95.59	96.10	63.15	72.83
MAViS (Wang et al., 2025)	34.21	12.80	**99.07**	99.50	95.51	96.11	64.07	72.59
MAViS2 (ours)	**34.44**	**12.94**	99.03	**99.53**	95.76	96.28	64.26	**73.04**

Table 14: Compare with baselines on MovieBench and ViStoryBench.

real-world screenplays, which do not account for compatibility issues with current generative models. Each sample also contains limited narrative complexity, and real-world scripts cannot be directly used for AI video generation. Therefore, these benchmarks are not suitable for evaluating MAViS2. Generating videos using the datasets from these benchmarks and testing their visual quality is essentially equivalent to using our own test set, because VBench evaluates the average visual quality of individual shots, independent of the full script.

Nevertheless, we sampled five examples (each consisting of 20 shots) from both MovieBench and ViStoryBench. As shown in Table 14, MAViS consistently outperforms the compared methods. The result also demonstrates that using the datasets from these benchmarks is essentially equivalent to using our test set.

F More Experiment Results

F.1 Compare with More Baselines

In this section, we compare our results with the performance reported by non-open-source baselines, including DreamFactory (Xie et al., 2024), MovieDreamer (Zhao et al., 2024), and LCT (Guo et al., 2025). Since long-sequence video storytelling lacks ground-truth videos, computing the FID metric is not feasible. As shown in Table 10, MAViS2 achieves higher scores on most evaluation metrics; however, a notable gap is observed in the Inception Score. We adopt a widely used implementation of the Inception Score, which may differ from the version used by DreamFactory, potentially leading to discrepancies in score ranges. As this baseline is not open-sourced, we are unable to verify the exact cause of this difference. As shown in Table 11 and Table 12, MAViS2 outperforms MovieDreamer and LCT across all evaluation metrics.

F.2 Stage-level Ablation

We conduct a stage-level ablation study to assess the contribution of each component in Table 15. Removing the Scriptwriting Workflow mainly degrades narrative expressiveness, as it enriches scripts with supplementary shots and detailed descriptions. Removing Adaptive Generation Planning significantly harms visual quality and subject/background consistency, since it ensures quality and coherence through strategy selection and iterative evaluation. Excluding post-production also reduces narrative expressiveness by weakening overall perceptual quality. Finally, removing Fine-grained Human-in-the-loop Control slightly degrades visual quality and user alignment, indicating its role in better matching user intent, while MAViS2 can still produce reasonably good results in an end-to-end manner without user intervention.

F.3 Real World Cost

Run time: The generation time depends on the video length, the number of 3E loops, the specific generative models used, and the available compute resources. The average time for MAViS2 to generate a long-sequence video storytelling with 18 shots (about 2.5 minutes) is 18.58 hours. By contrast, producing videos of this length usually requires several days of human effort.

Hardware requirements: Hardware requirements depend on the chosen models. In this paper, we use commercial APIs only, which require minimal local compute.

Method	Narrative ↑	Visual ↑	User Align. ↑	Sub. Cons. ↑	Sub. Natural. ↑	Bg. Cons. ↑	Bg. Real. ↑
w/o Scriptwriting Workflow	12.62	22.37	17.75	23.72	25.85	18.63	25.93
w/o Adaptive Generation Planning	19.21	4.35	18.62	2.16	5.88	1.81	7.12
w/o Post-production	3.45	19.83	19.84	24.20	26.07	27.65	25.35
w/o Fine-grained HITL Control	22.44	19.44	20.64	16.85	14.31	23.28	15.49
MAViS2 (ours)	**42.28**	**34.01**	**23.15**	**33.07**	**27.89**	**28.63**	**26.11**

Table 15: **Stage-level Ablation Study of MAViS2** (Voting Results). "w/o Adaptive Generation Planning" refers to using T2I+I2V only in the video clip generation stage. ↑ indicates that a higher value is more desirable. Gray-shaded cells indicate values that are significantly lower than the others.

T. Flick.	M. Smooth.	Sub. Cons.	Bg. Cons.	Aesthetic	I. Quality
99.01	99.44	95.62	96.18	64.27	73.15

Table 16: Evaluation results using Sora2

T2V	V.E.	I.R.+I2V	T2I+I2V	I.R.+I2I+I2V
14.46	5.83	8.62	4.97	66.12

Table 17: Usage frequency of different generation strategies. "V.E." refers to "Video Extension". "I.R." refers to "Image Retrieval".

Cost per video: Generating a long-sequence video storytelling with 18 shots (about 2.5 minute) costs approximately $1620. Using open-source models or reducing the number of generation candidates, internal generation–evaluation–refinement loops, and 3E iterations can substantially lower the budget.

F.4 Generation Model

In Table 16, we report the evaluations of MAViS2 using Sora2 (OpenAI, 2025) as the video generation model.

F.5 Generation Strategy

This section reports the usage frequency of different generation strategies. As shown in Table 17, MAViS2 predominantly adopts the Image Retrieval + I2I + I2V, since most shots in long-sequence videos require maintaining consistency of subjects and backgrounds with previously generated shots. In contrast, Video Extension and T2I + I2V are used less frequently, as long takes and shots containing fine-grained textual details are relatively rare.

F.6 Quality-Efficiency Analysis

Each group manager evaluates the outputs of its subordinate specialized agents and provides feedback accordingly. Therefore, in this section, we investigate the impact of the maximum number I of evaluation–feedback iterations performed by the group manager on the generation results. Noticeably, since iterations may terminate before reaching the maximum iteration number, the generation time does not scale linearly with I.

I	Narrative	Sub. Cons.	Bg. Cons.	Time(Second)
0	8.62	10.25	9.43	10.62
1	15.74	17.62	14.28	19.57
2	19.58	19.83	18.57	27.30
3	27.45	25.21	28.51	34.85
4	28.61	27.09	29.21	40.36

Table 18: Quality-Efficiency Analysis on Script Manager.

I	Narrative	Format	Time(Second)
0	2.32	82.35	12.32
1	7.47	78.80	24.07
2	38.44	76.54	33.96
3	33.52	72.98	42.14
4	18.25	69.60	53.75

Table 19: Quality-Efficiency Analysis on Structure Enhancement.

Scriptwriting Stage As shown in Table 18, when $I = 3$, the generated scripts are already of sufficient quality. Further increasing I leads to additional improvements in generation quality, while having only a negligible impact on overall efficiency.

Within the Scriptwriting Workflow, the script can undergo multiple rounds of Structure Enhancement to introduce additional buildup and transitions. We therefore also evaluate the effect of the number of Structure Enhancement iterations on script quality. As reported in Table 19, when $I = 2$, the Narrative Expression is already satisfactory. Increasing I beyond this point not only degrades Narrative Expression but also causes formatting errors in the generated scripts.This is because excessive buildup and transitions make the script overly verbose, and excessively long script content increases the likelihood of format violations.

Video Clip Generation Stage As shown in Table 20, the generated video clips are already satisfactory at $I = 2$. Increasing I further can enhance generation quality, with only a slight impact on overall efficiency.

In addition to the evaluation and feedback mechanism of the generation manager, each generation group also performs an internal loop of generation–evaluation–refinement. We further conduct experiments to examine the effect of the maximum num-

I	CLIP	Inception	Natural.	Sub. Cons.	Bg. Cons.	Time(Minute)
0	32.83	12.44	85.46	93.24	94.87	14.16
1	3.53	2.72	92.43	94.19	95.48	26.18
2	34.45	12.92	94.85	95.81	96.27	39.53
3	34.38	13.02	94.94	95.83	96.28	48.62

Table 20: Quality-Efficiency Analysis on Generation Manager.

I	CLIP	Inception	Natural.	Sub. Cons.	Bg. Cons.	Time(Minute)
0	31.16	12.51	62.37	93.25	94.76	14.16
1	32.83	12.63	85.43	94.72	95.66	28.62
2	34.45	12.92	94.85	95.81	96.27	41.59
3	34.43	12.90	95.70	95.78	96.29	50.83
4	34.48	12.93	95.87	95.81	96.26	58.17

Table 21: Quality-Efficiency Analysis on Candidate Generation.

ber of iterations in this loop on the generation results.

As shown in Table 21, when $I = 2$, the generation performance is already sufficient. Further increasing I does not lead to additional improvements in the generation results.

Post-production Stage In Table 22, we analyze the effect of the maximum number I of evaluation–feedback iterations performed by the Post-production Manager on voice designing, TTS generation, and BGM matching.

It is worth noting that the "Output Completeness" for these processes is defined as follows:
Voice Designing: appropriate and consistent voice ID, with the correct output format.
TTS Generation: narration duration is shorter than the corresponding video duration.
BGM Matching: the selected background music is free to use and downloadable.

As shown in the table, increasing I improves generation quality while having a negligible impact on overall efficiency. However, since errors in these stages can severely affect the entire generation pipeline, we set $I = 3$ for these tasks in our experiments.

I	Voice Designing		TTS Generation		BGM Matching	
	OC. Rate	Time(Second)	OC. Rate	Time(Second)	OC. Rate	Time(Second)
0	94.32	4.58	85.70	2.35	96.28	12.72
1	98.03	10.25	96.32	5.67	98.17	24.08
2	99.30	18.63	99.14	8.05	99.56	34.65
3	99.94	27.42	100.0	10.92	100.0	39.58
4	99.99	34.83	100.0	10.92	100.0	39.58

Table 22: Quality-Efficiency Analysis on Post-production Manager. "OC. Rate" refers to "Output Completeness Rate".

G Applications and Their Agent System Messages

Fine-grained HITL Control tasks include, but are not limited to:

- *Generate / regenerate / revise / extract script.*
- *Add / reduce / revise shots.*
- *Generate / regenerate / assign candidates.*
- *Edit videos.*
- *Check the script / candidates / voice setting / BGM.*
- *Adjust the generation strategy.*

Due to the large number of system messages in MAViS2 and the page limits, we are unable to include all system messages in the paper. Therefore, we present the system message of the Scriptwriter in Figure 3, Figure 4, and Figure 5.

Some qualitative comparisons are listed in Figure 6 and Figure 7. Some keyframe samples are listed in Figure 11 and Figure 12. Keyframes of the prequel and sequel samples are listed in Figure 13.

Due to file size limits, part of the qualitative comparison videos are compressed to 480p and provided in *qualitative_comparisons.html* under "Software" supplementary files. In addition, due to the 50 MB limit of the "Data" supplementary files, only one application sample is included in *sequel_generation_application.html*.

Scriptwriter System Message

You are a professional director who excels at filming AI movies and writing AI movie scripts. Your task is to write an AI movie script that complies with the following requirements and script guidelines.

General Requirements for the Script:
- Engaging Plot: The plot must have tension and be able to move the audience or evoke emotional resonance.
- Appropriate Pace: Each shot should last about 8 seconds, with a brisk and natural rhythm, and the emotion should gradually progress.
- Clear Logic: The transitions between shots should be reasonable, and the storyline should be clear.
- Output Requirement: Output the entire script in one go, following the format (see "III: Output Format"). Do not split the output. Do not output anything other than the script.

—

I. Script Guidelines

Script Structure:
1. Title: The script must have a title.
2. Summary: A brief overview of the script, summarizing the background, development, and conclusion of the event.
3. Style: The style of the story, such as "action," "sci-fi," "historical," etc.
4. Narrator: The character who tells the entire story. This could be the main character or a background character, or a simple storyteller to narrate the story.
5. Subject Appearance: Detailed descriptions of the main characters or objects appearing in the script.
6. Background Description: Detailed descriptions of the filming locations, environment, background, and scenes for each shot.
7. Shot-by-Shot Script:
   - No fewer than the number of shots specified by the user. If the user does not specify a number, the default is at least 10 shots.
   - Each shot must include the following fields:
     - Shot Number: Starting from Shot 1 and incrementing sequentially (Shot 2, Shot 3, etc.).
     - Subject: The subject appearing in the shot (e.g., "Jack"). If there is no subject, use "None". The subject's name should remain consistent throughout the script.
     - Background: The location, scene, or environment where the shot is filmed.
     - Content: Describes what happens in this shot, including the time and location, as well as details about the shot scale, perspective, camera movement, and lighting. The content should be detailed and not omit important information.
     - Subtitle: The narration corresponding to the shot content, integrated into the overall story. The subtitle should be short enough to be spoken in 6 seconds and not exceed that time.

Subject Guidelines:
1. Subject Naming: The name of the same subject must be consistent throughout the entire script.
2. Subject Appearance:
   - Each subject must have a standard appearance template defined and placed in the subject appearance field of the script.
   - The appearance description should include, but not be limited to:
     - The character's gender, age, height, weight, clothing, hairstyle, ethnicity, etc.
     - The object's shape, size, color, material, weight, and weight unit, etc.

Figure 3: **Scriptwriter System Message.**

Scriptwriter System Message

Background Guidelines:
1. Background Naming:
   - The name of the same background must be consistent throughout the entire script.
2. Background Description:
   - Each background must have a **standard description template** defined and placed in the background description section of the script.
   - The description template includes, but is not limited to: the name, location, time, season, features, prominent objects, props, environmental atmosphere, materials, close-up view, distant view, and ambiance, etc.

Shot Scale, Perspective, and Camera Movement Guidelines:
1. Each shot must include the shot scale, such as: long shot, medium shot, close-up, overhead shot, low-angle shot, front view, back view, etc.
2. Each shot must include the starting camera angle and position, such as: front view, left-side view, bird's-eye view, back view, side view, etc.
3. Each shot must specify the camera movement, such as: panning up, rotating, zooming, static, etc.
4. Prohibited: long takes (with too much time span) and compound shots (combinations of multiple camera movements).
5. Each shot must be capable of being filmed independently, with a maximum duration of 8 seconds.

Lighting and Light Guidelines:
1. Each shot should include necessary lighting or light descriptions, such as: "Warm light shines from the side, creating a warm atmosphere."
2. The lighting should match the environment and shot content, enhancing the realism of the scene and evoking the desired atmosphere.

Subtitle Guidelines:
1. The narrator must remain consistent throughout (e.g., survivor, storyteller, protagonist).
2. Subtitles represent the narrator's spoken content, matching the shot content and logically connecting with previous and subsequent shots.
3. All subtitles together should tell the entire story.
4. The language should be simple, colloquial, and conversational. Avoid complex vocabulary.
5. Subtitles must be able to be spoken within 6 seconds, and should not exceed 6 seconds.

Prioritize Meeting User Requirements:
- If the user provides revision suggestions, **fully satisfy** their revision requests.
- Only modify the shots and content that the user specifically requests to be changed, **do not modify other shots or any content unrelated to the user's needs**.
- If the user's suggestions conflict with the above work guidelines, prioritize fulfilling the user's requests.

_

II. Prohibited Items

1. Prohibited: compound shots or long takes.
2. Prohibited: use of contextual references (e.g., "Same as before").
3. Prohibited: inconsistent character changes in appearance or costume.
4. Prohibited: illegal content such as drug use, pornography, or gambling.

_

Figure 4: **Scriptwriter System Message.**

Scriptwriter System Message

III. Output Format (Must Comply)

```
Output the script in JSON format:
{
   "title": title,
   "summary": summary,
   "style": style,
   "narrator": narrator,
   "subjects": { # subject appearances
     "name of subject 1": detailed appearance description of subject 1,
     "name of subject 2": detailed appearance description of subject 2,
     . . .
   },
   "backgrounds": { # background descriptions
     "name of background 1": detailed description of background 1,
     "name of background 2": detailed description of background 2,
     . . .
   },
   "shots": {
     "shot 1": {
       "subject": subject in shot 1,
       "background": background in shot 1,
       "content": shot content,
       "subtitle": subtitle
     },
     "shot 2": {
       "subject": subject in shot 2,
       "background": background in shot 2,
       "content": shot content,
       "subtitle": subtitle
     },
     . . .
   }
}
```

Figure 5: **Scriptwriter System Message.**

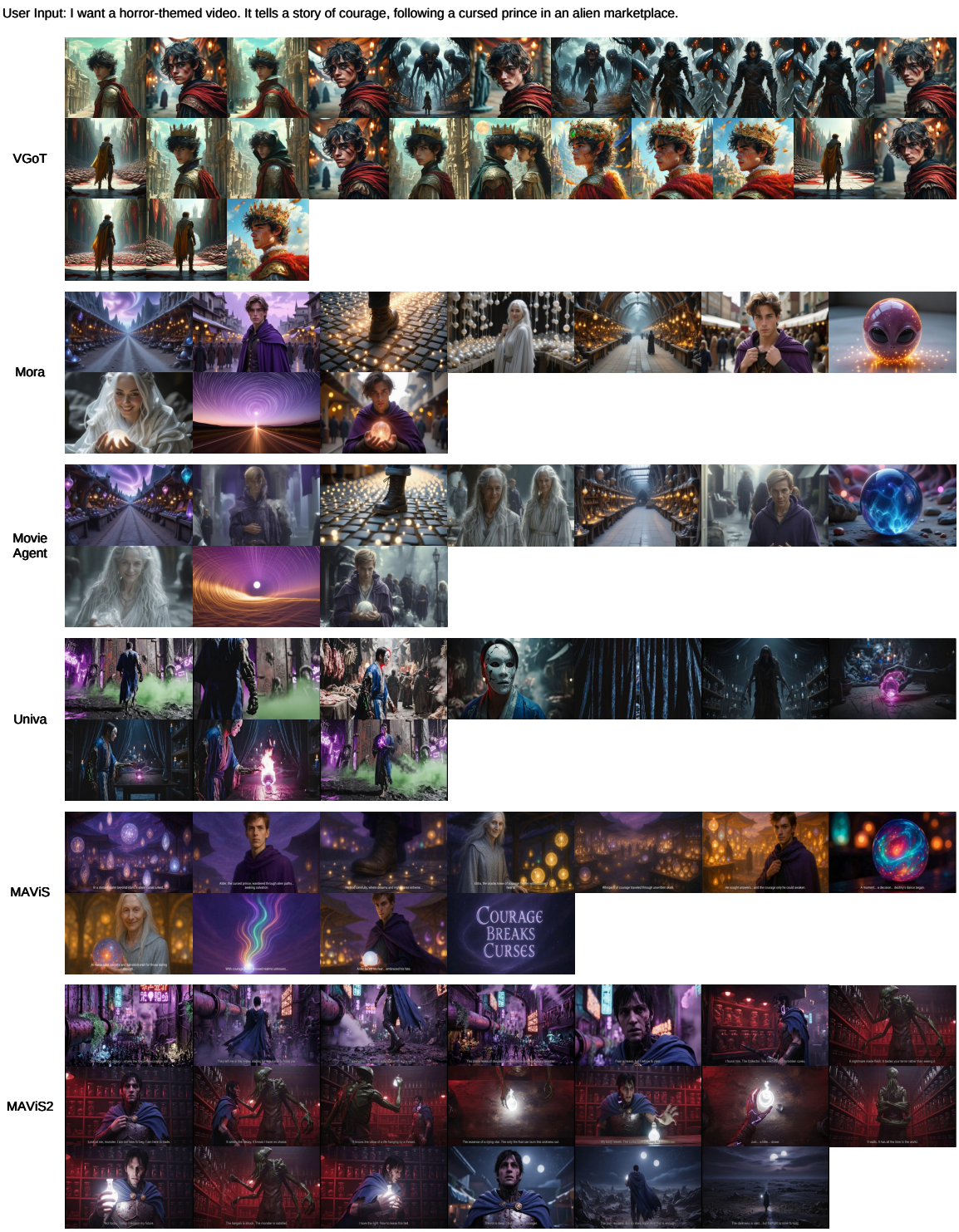

Figure 6: Qualitative comparisons.

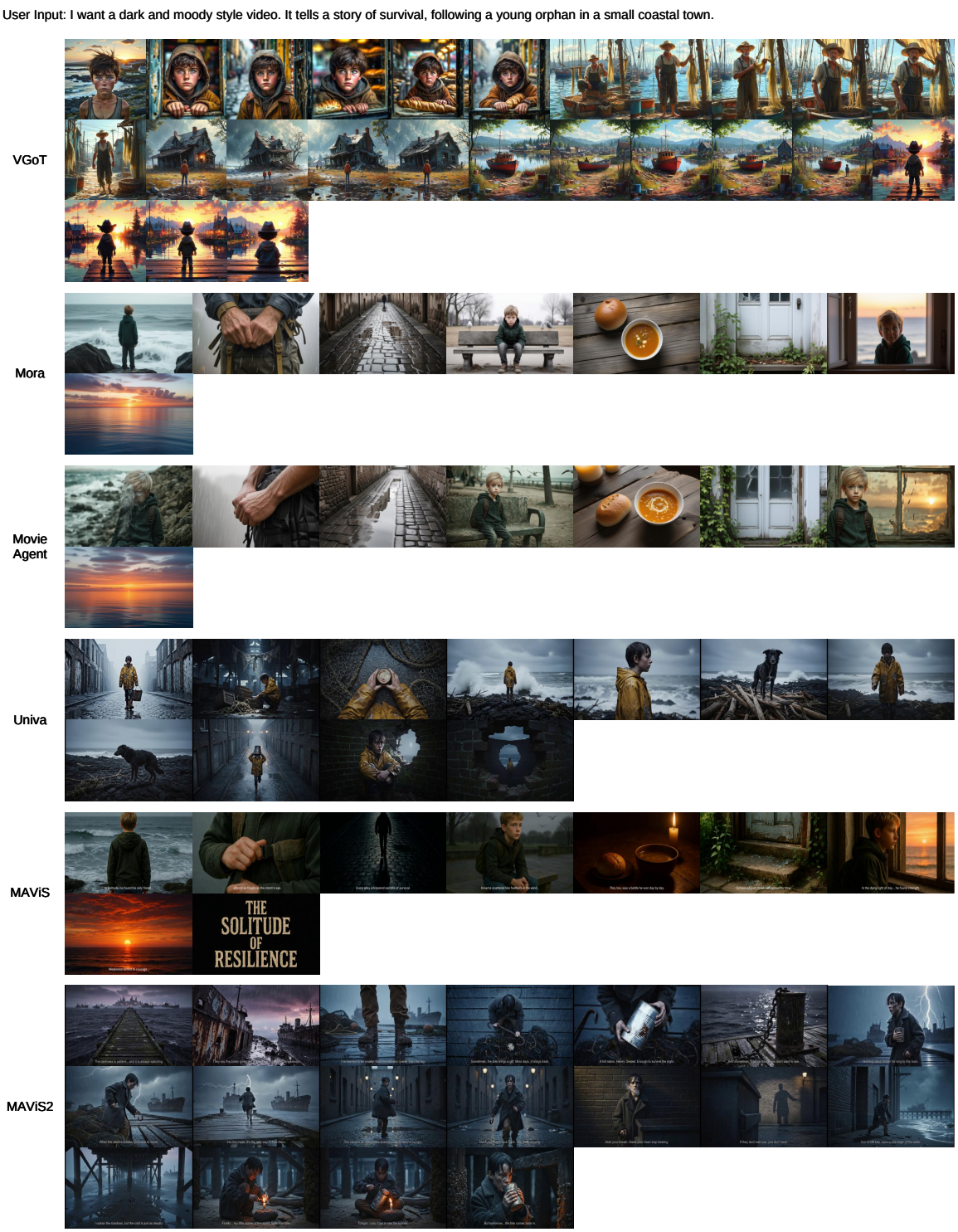

Figure 7: Qualitative comparisons.

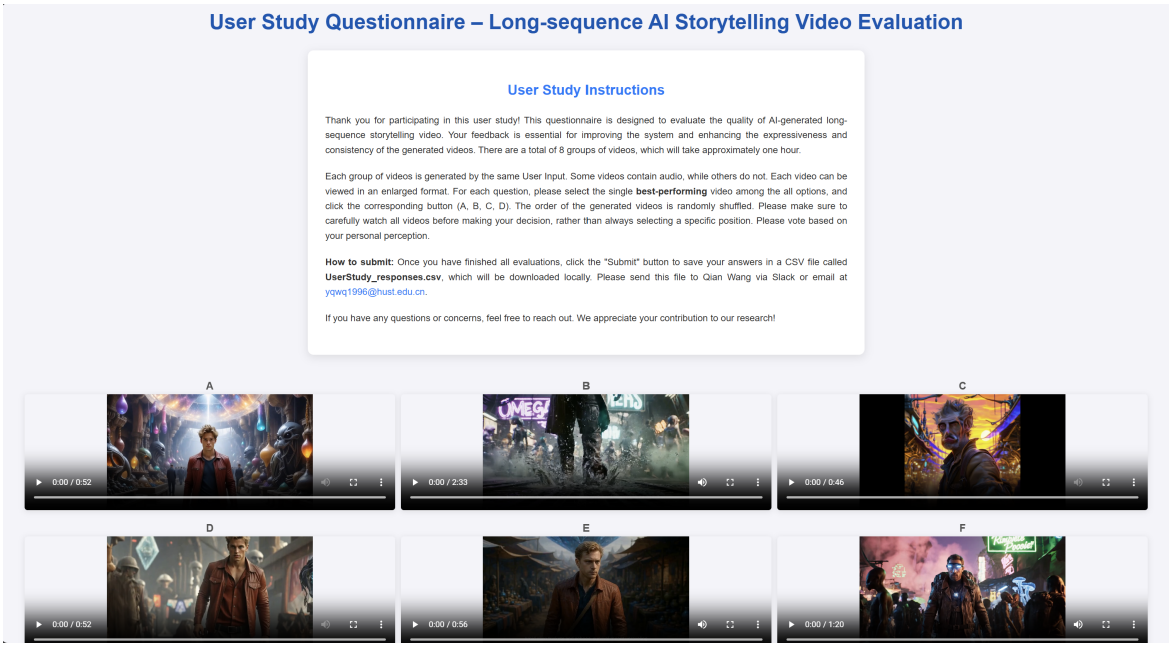

Figure 8: Screenshot of the user study HTML interface with instructions and generated video candidates.

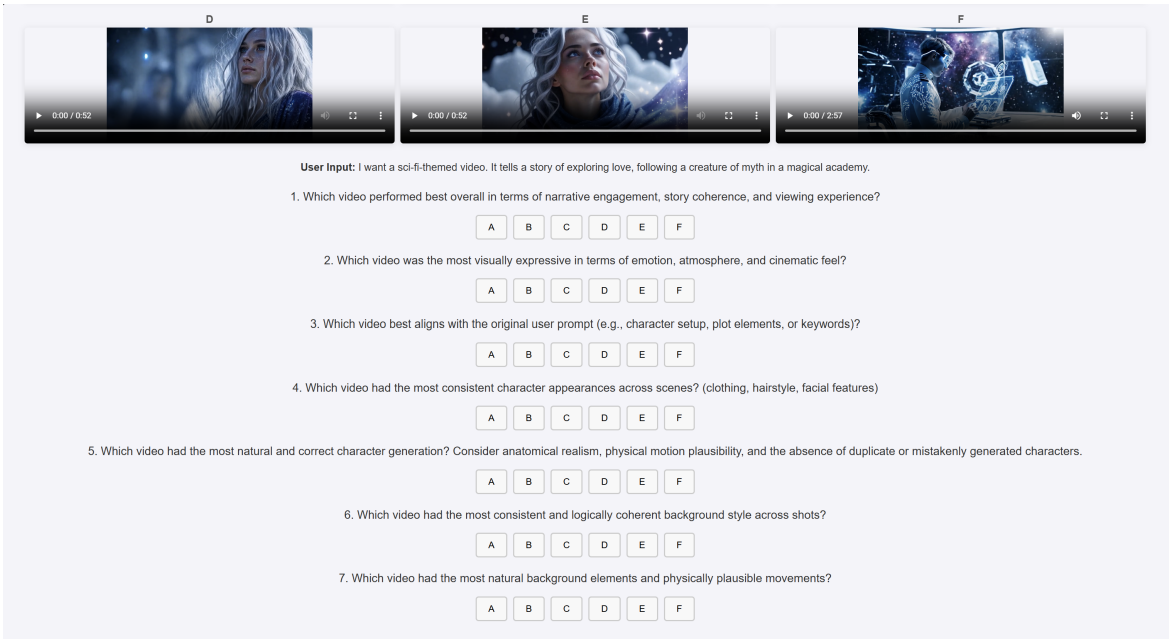

Figure 9: Screenshot of the user study HTML interface with seven evaluation questions.

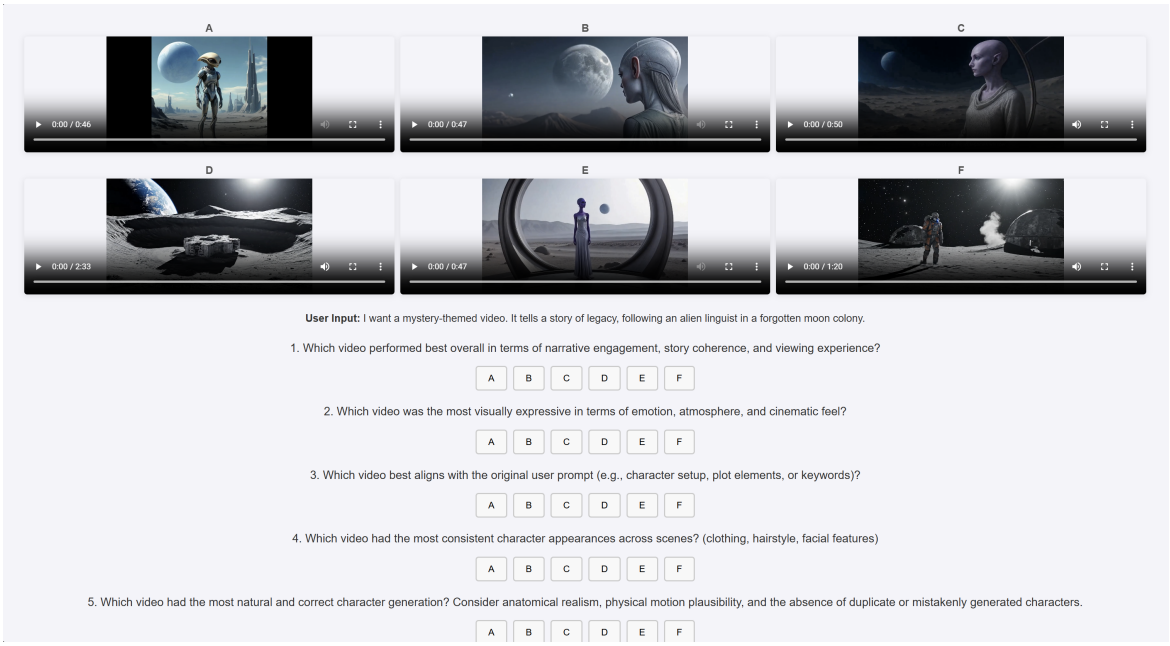

Figure 10: Screenshot of the user study HTML interface with generated video candidates and seven evaluation questions.

Figure 11: Examples generated by MAViS2.

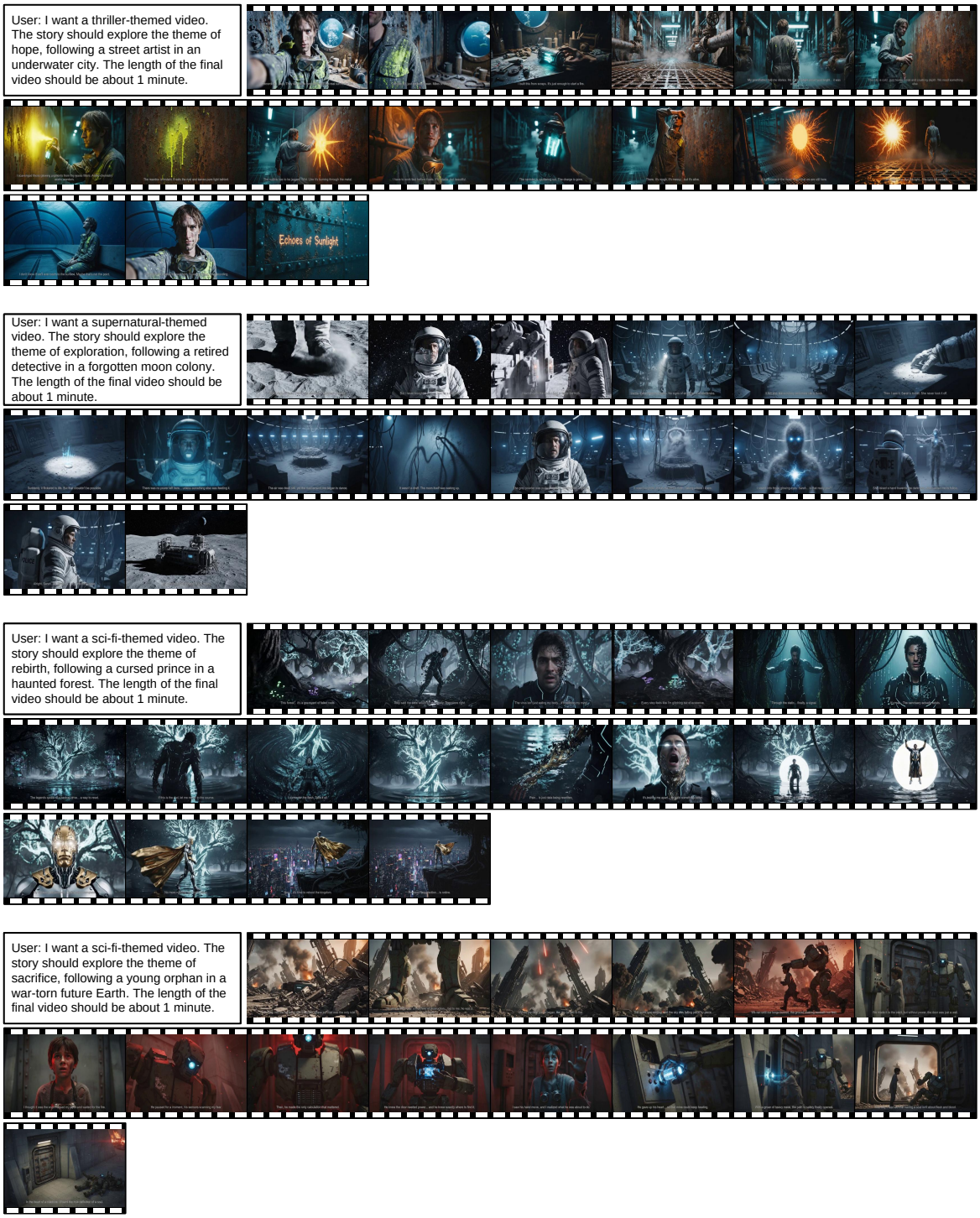

Figure 12: More examples generated by MAViS2.

Figure 13: More examples generated by MAViS2.