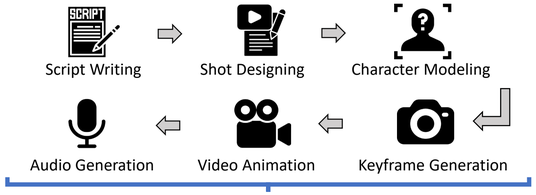
HiGen 2025 workshop #9

MAViS: A Multi-Agent Framework for Long-Sequence Video Storytelling

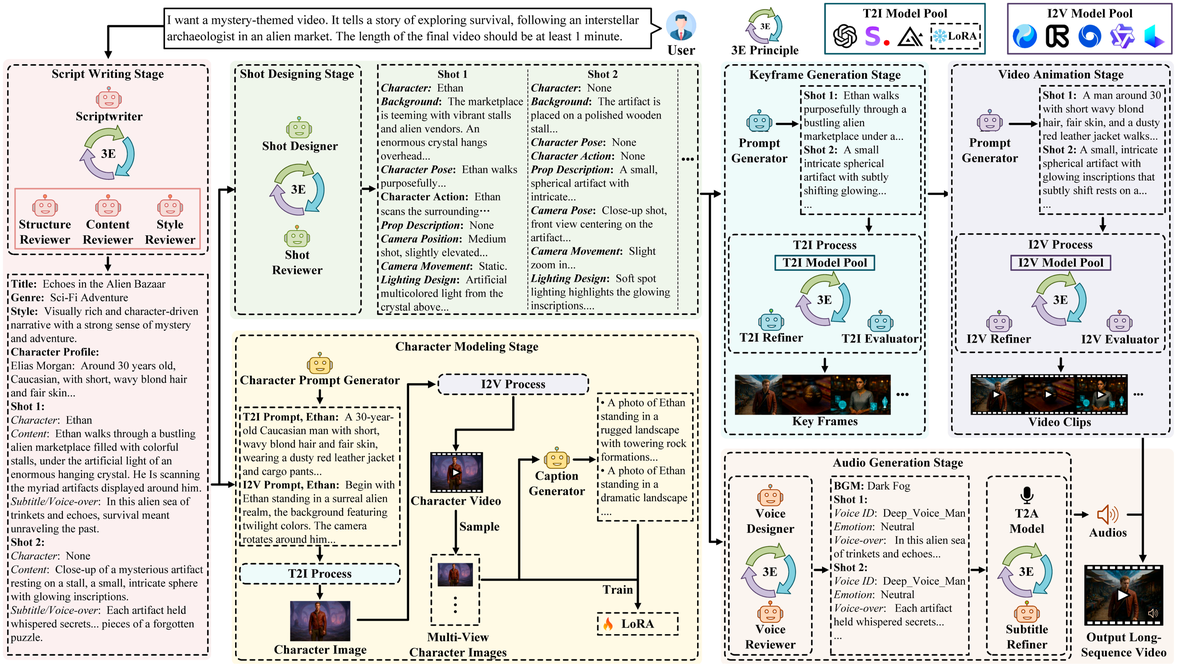
Qian Wang[1] Ziqi Huang[2] Ruoxi Jia[1] Paul Debevec[3] Ning Yu[3]✉
[1]Virginia Tech [2]Nanyang Technological University [3]Netflix Eyeline Studios ✉Corresponding Author and Project Lead

MAViS is an assistive framework that empowers users to efficiently explore diverse visual storytelling directions for sequential video generation. Given only a brief idea description, MAViS can generate high-quality, expressive long-sequence videos.



By orchestrating multiple stages within a unified pipeline, **MAViS** effectively addresses the key limitations of prior approaches—poor assistive capability, limited visual expressiveness, and incomplete outputs.

We propose the **Scriptwriting Guideline** to improve compatibility between narrative structure and generative tools. The **3E Principle (Explore – Examine – Enhance)**, which enables iterative refinement across all stages, helps **MAViS** maximize output quality. To the best of our knowledge, MAViS is the only framework that provides multimodal design output – videos with narratives and background music.



Methods	Agentic	Script Writing	Shot Designing	Dynamic	ID Consistency	Audio	Assistive
Mora (Yuan et al., 2024)	✓	×	×	✓	×	×	✓
AesopAgent (Wang et al., 2024a)	✓	✓	×	×	✓	×	✓
MovieDreamer (Zhao et al., 2024)	×	×	×	✓	✓	×	×
DreamFactory (Xie et al., 2024)	✓	×	✓	✓	✓	×	×
VGoT (Zheng et al., 2025)	×	✓	×	✓	✓	×	✓
LCT (Guo et al., 2025)	×	×	×	✓	✓	×	×
MovieAgent (Wu et al., 2025b)	✓	×	✓	✓	✓	✓	×
MAViS (ours)	✓	✓	✓	✓	✓	✓	✓

Tabel 1. Properties of MAViS vs Other Methods

We summarize seven metrics for long-sequence video storytelling works.
"**Agentic**": whether the system employs a multi-agent architecture;
"**Script Writing**": ability generate scripts from user prompt;
"**Shot Designing**.": support for structured shot designing and controllable narrative flow;
"**Dynamic**": whether the generated video exhibits dynamic features;
"**ID Consistency**": consistency in character identity across multiple shots; "Audio": whether the output includes synchronized audio;
"**Assistive**": whether the pipeline is designed to support users in script writing, script planning, or model fine-tuning by streamlining processes.

Method	Keyframe Generation		Video Animation					
	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
MAViS (ours)	**34.22**	**12.81**	**99.09**	**99.53**	95.72	96.12	63.17	**72.91**

Table 2. Evaluation of Automatic Metrics for Keyframe Generation and Video Animation.
"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.

Method	Narrative ↑	Visual ↑	User Align. ↑	Sub. Cons. ↑	Sub. Natural. ↑	Bg. Cons. ↑	Bg. Real. ↑
VGoT (Zheng et al., 2025)	3.39	3.39	6.79	5.56	3.07	4.14	3.24
Mora (Yuan et al., 2024)	15.18	16.96	13.57	15.77	13.00	13.69	15.29
MovieAgent (Wu et al., 2025b)	9.46	11.79	11.96	12.37	12.27	12.07	10.97
MAViS (ours)	**71.96**	**67.86**	**67.68**	**66.31**	**71.66**	**70.09**	**70.50**

Tabel 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.

Visualizations from MAViS