CineAGI: Character-Consistent Movie Creation through LLM-Orchestrated Multi-Modal Generation and Cross-Scene Integration

Tianyidan Xie[1], Mingjie Wang[2], Qiang Tang[3], Feixuan Liu[4], Rui Ma[5], Lanjun Wang[6], Zili Yi[1,*]

[1]Nanjing University, [2]Zhejiang Sci-Tech University, [3]University of British Columbia,
[4]Beijing Shuzhimei Technology Co., Ltd, [5]Jilin University, [6]Tianjin University

***Abstract*—Automated movie creation requires coordinating multiple characters, modalities, and narrative elements across extended sequences—a challenge that existing end-to-end approaches struggle to address effectively. We present CineAGI, a hierarchical movie generation framework that decomposes this complex task through specialized multi-agent orchestration. Our framework employs three key innovations: (1) a multi-agent narrative synthesis module where specialized LLM agents collaboratively generate comprehensive cinematic blueprints with character profiles, scene descriptions, and cross-modal specifications; (2) a decoupled character-centric pipeline that maintains identity consistency through instance-level tracking and integration while enabling flexible multi-character composition; and (3) a hierarchical audio-visual synchronization mechanism ensuring frame-level alignment of dialogue, expressions, and music. Extensive experiments demonstrate that CineAGI achieves 40% improvement in overall consistency, 4.4% gain in subject consistency, 5.4% enhancement in aesthetic quality, and 28.7% higher character consistency compared to baselines. Our work establishes a principled foundation for automated multi-scene video generation that preserves narrative coherence and character authenticity.**

I. Introduction

The automation of movie creation represents a fundamental challenge in generative AI, requiring the coordination of narrative structure, visual synthesis, and audio generation across extended temporal sequences. While recent advances in Large Language Models (LLMs) [1], [2] and diffusion-based video synthesis [3]–[5] have shown impressive capabilities in individual content generation tasks, creating coherent long-form movies poses three critical challenges illustrated in Figure 1.

Character Identity Preservation. Quadratic computational complexity necessitates limited processing windows, causing loss of long-range identity information. Current approaches lack effective mechanisms to consistently preserve multi-scale identity features, from global attributes (facial structure, physique) to fine-grained details (expressions, lighting adaptation), across frames and scenes.

Scene-Level Temporal Coherence. Memory constraints prevent global scene consistency enforcement, and existing architectures lack explicit mechanisms for decomposing scene elements at different temporal frequencies, making it difficult to maintain compositional coherence over extended durations.

Cross-Modal Synchronization. Coordinating lip movements, emotional expressions, and musical timing at frame-level precision requires explicit synchronization mechanisms that existing end-to-end methods lack.

To address these challenges, we present CineAGI, a hierarchical movie creation framework that orchestrates multi-modal generation through principled decomposition and specialized agent coordination. Our framework introduces three key technical innovations:

Multi-Agent Narrative Synthesis. Specialized LLM agents (Character Designer, Script Writer, Storyteller, Composer, Quality Inspector) collaboratively generate cinematic blueprints. Unlike prior approaches that treat narrative elements independently, our coordinated agents maintain cross-modal consistency through structured information flow and validation mechanisms.

Decoupled Character-Centric Pipeline. A three-stage approach of text-guided segmentation (Grounded-SAM2), identity-preserving face integration (SimSwap), and talking face synthesis (Wav2Lip) enables independent multi-character processing while maintaining global consistency across diverse scene contexts.

Hierarchical Audio-Visual Synchronization. Explicit coordination at multiple levels, including frame-level lip alignment, emotion-aware music generation, and integrated scene assembly, addresses cross-modal alignment limitations in existing methods. This structured approach ensures synchronized audiovisual integration throughout the production.

Our main contributions are:

- A novel multi-agent movie generation framework with specialized LLM agents for coordinated narrative planning, enabling fine-grained control over character consistency and temporal coherence that surpasses existing approaches.
- A systematic character-centric pipeline that maintains multi-level identity preservation through decoupled processing, achieving robust character consistency across diverse contexts while preserving creative flexibility.
- Comprehensive evaluation demonstrating 40% improvement in overall consistency, 4.4% gain in subject consistency,

*Corresponding author: yi@nju.edu.cn

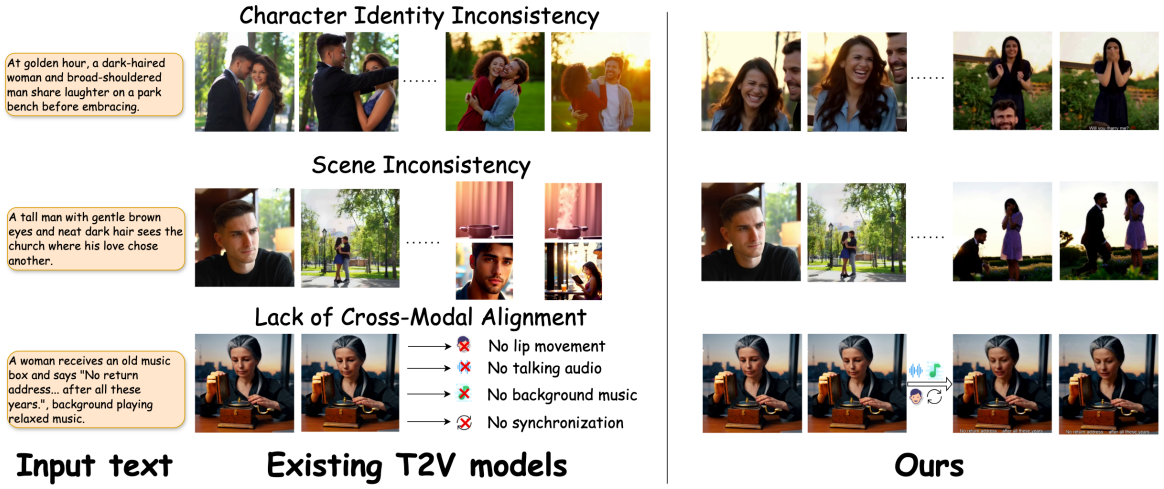

Fig. 1. **Key challenges in automated movie generation.** Existing text-to-video models face three fundamental limitations: (Top) *Character identity inconsistency* across scenes and contexts; (Middle) *Scene inconsistency* in visual styles and transitions; (Bottom) *Cross-modal misalignment* between visual elements, dialogue, and audio. Our hierarchical framework addresses these challenges through specialized agent coordination and decoupled processing pipelines.

5.4% enhancement in aesthetic quality, and 28.7% higher character consistency compared to state-of-the-art baselines.

II. Related Work

A. Text-to-Video Generation

Text-to-video synthesis has evolved rapidly through diffusion-based architectures. Recent systems such as Sora [6], Lumiere [7], and Stable Video Diffusion [4] have significantly advanced generation duration and visual fidelity. CogVideoX [8] introduced expert transformers with 3D VAE for spatial-temporal compression, while VideoCrafter [9] focused on temporal consistency through dedicated frame-level coherence modules. However, these methods face inherent limitations due to localized processing windows and lack of explicit narrative structure modeling.

Recent advances have addressed specific challenges in video generation. For character consistency, ID-Animator [10] and ConsisID [11] introduced identity preservation techniques, while StoryDiffusion [12] and Animate Anyone [13] advanced long-range generation and character animation. For extended sequences, StreamingT2V [14] and FIFO-Diffusion [15] enabled multi-minute to theoretically infinite video generation. LLM integration has emerged as a powerful approach, with VideoDirectorGPT [16] pioneering LLM-guided planning, VideoPoet [17] demonstrating unified multimodal generation, and MovieGen [18] advancing joint video-audio synthesis. Despite these advances, maintaining multi-character consistency, cross-scene narrative coherence, and coordinated multi-modal production in long-form content remains challenging.

Our work differs fundamentally by introducing hierarchical decomposition through multi-agent orchestration that decouples character generation from scene synthesis, enabling robust character consistency while preserving narrative flexibility across extended sequences.

B. Multi-Agent Systems

Recent LLM advances have enabled sophisticated multi-agent systems for complex task automation. MetaGPT [19] and AgentVerse [20] demonstrated foundational principles for agent-based collaboration through specialized roles and structured coordination protocols.

In movie production, agent-based approaches have explored automated content creation. AutoDirector [1] introduced multi-sensory composition through scheduled agents, while Anim-Director [2] focused on animation generation. AesopAgent [21] pioneered agent-driven evolutionary workflows combining RAG-based optimization with utility-layer execution, and StoryAgent [22] introduced specialized agents mirroring professional production roles. However, these systems struggle with temporal consistency and multi-modal synchronization over extended sequences due to insufficient mechanisms for managing complex inter-dependencies.

Our hierarchical architecture advances beyond existing approaches through explicit coordination protocols inspired by professional pipelines, with specialized agents (Character Designer, Script Writer, Storyteller, Composer, Quality Inspector) ensuring both creative consistency and temporal coherence through structured information flow and validation mechanisms.

III. Method

A. Overview

Our CineAGI framework introduces hierarchical multi-agent orchestration for automated movie creation, as illustrated in Figure 2. Unlike end-to-end approaches that process entire sequences holistically, our hierarchical decomposition enables targeted optimization of individual components while maintaining global coherence through structured agent coordination.

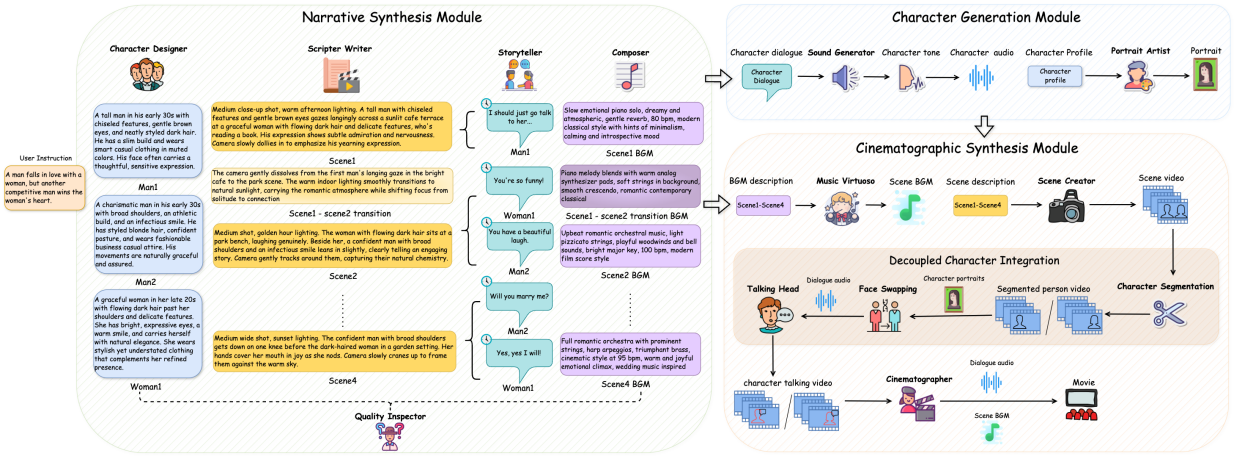

Fig. 2. **CineAGI framework overview.** Given a story concept, our framework proceeds through three modules: (1) *Narrative Synthesis*: LLM agents collaboratively develop character profiles and production scripts ensuring narrative coherence; (2) *Character Generation*: creates visual and audio assets with consistent character appearances and voice profiles; (3) *Cinematographic Synthesis*: orchestrates video generation, character integration, and audio synchronization for complete movie production.

The framework leverages complementary strengths of different generative models: LLMs for high-level creative planning and narrative structure, specialized models for consistent character generation, and targeted synthesis models for precise audiovisual integration. The *Narrative Synthesis Module* serves as the creative foundation where LLM agents collaboratively develop rich, coherent story specifications. The *Character Generation Module* bridges abstract descriptions and concrete assets through specialized portrait and voice synthesis. The *Cinematographic Synthesis Module* coordinates these elements through decoupled processing streams, maintaining both local coherence and global consistency while precisely synchronizing visual, audio, and emotional elements.

B. Narrative Synthesis Module

Our narrative synthesis module introduces a hierarchical LLM agent framework that systematically decomposes movie creation into interconnected creative tasks. Unlike previous approaches [1], [2], [23] that treat narrative elements independently, our coordinated agents maintain cross-modal consistency through structured information flow and validation mechanisms.

Character Designer analyzes the input story to establish detailed character profiles, decomposing identities into hierarchical attributes covering appearance, personality, and behavioral patterns. These specifications serve as foundational references across all generation aspects, ensuring consistent character portrayal throughout the pipeline.

Script Writer develops shooting scripts centered around character profiles, maintaining strict alignment with Character Designer outputs. For each scene, it produces technical descriptions including visual composition, character positioning, camera movements, and detection keywords that guide downstream video synthesis.

Storyteller crafts narrative flows by analyzing character-scene relationships. It decomposes stories into coherent scenes, ensuring character actions align with established personalities while preserving dramatic arcs. It develops dialogue content with precise frame-level timing specifications to maintain consistent pacing and emotional progression.

Composer synthesizes character profiles, scene descriptions, and dialogue content to generate background music specifications. Rather than isolated scene-by-scene generation, it creates cohesive musical direction that enhances audiovisual synchronization and emotional resonance.

Quality Inspector ensures end-to-end consistency by validating interconnections between all agent outputs, preventing cascading errors and verifying relevance to the original input. It outputs structured JSON results that standardize inputs for subsequent processing stages. *Detailed multi-agent coordination algorithms, protocol specifications, and convergence analysis are provided in Appendix A.*

C. Character Generation Module

The character generation module transforms abstract character profiles into concrete audiovisual assets. Operating on character profiles and dialogue scripts from the narrative stage, this module employs two specialized components:

Portrait Artist utilizes RealVisXL 3.0 to generate high-fidelity visual representations of each character. Taking detailed character profiles as input, it produces reference portraits capturing the character's appearance from multiple physical features. These portraits serve as primary references for maintaining visual consistency during face swapping in the cinematographic synthesis stage.

Sound Generator implements multi-identity voice synthesis via ChatTTS, combining character-specific voice profiles with dynamic emotional modulation. The framework processes comprehensive character specifications to generate voice assets that maintain speaker identity while supporting nuanced emotional expression. *Technical specifications of the identity*

preservation mechanism and embedding consistency analysis are provided in Appendix B.

D. Cinematographic Synthesis Module

The cinematographic synthesis module introduces a decoupled integration pipeline that systematically addresses character consistency and multi-modal synchronization:

Scene Creator employs HunyuanVideo-13B for text-driven scene generation. Unlike previous methods relying on character reference images, our framework encodes character specifications through rich textual descriptions, enabling flexible multi-character scene composition while providing a foundation for subsequent character-specific processing.

Decoupled Character Integration introduces a key technical innovation that transcends limitations of end-to-end single-character reference approaches. Our three-stage pipeline systematically decomposes multi-character scenes into individual segments while maintaining precise spatial-temporal relationships:

Character Segmentation leverages Grounded-SAM2 for text-guided character isolation, processing scene-specific detection keywords to identify and track individual characters. This precise segmentation enables independent character processing while preserving contextual scene information.

Face Swapping utilizes SimSwap [24] for identity-preserving face integration. By applying character-specific portrait references to segmented regions, this stage ensures consistent visual identity across diverse scene contexts.*Comparative analysis of face swapping methods and identity preservation metrics are provided in Appendix C.*

Talking Face synchronizes visual-audio modalities through Wav2Lip [25], using frame-level timing markers to coordinate character-specific lip movements with dialogue, ensuring natural facial dynamics while preserving both visual identity and audio synchronization.

Music Virtuoso leverages MusicGen to generate scene-specific background music based on musical directions from the Composer, synthesizing background music that aligns with each scene's emotional context while maintaining thematic coherence. The generated music adapts to scene durations and dramatic progression, complementing narrative structure without interfering with dialogue clarity.

Cinematographer executes final assembly through a multi-stage pipeline: integrating segmented characters back into original scene videos, overlaying character dialogue audio within specified frame ranges, adding corresponding subtitles, incorporating background music, and concatenating processed scenes according to narrative sequence. This systematic approach ensures synchronized audiovisual integration in the final output.

IV. Experiments

A. Experimental Setup

Evaluation Benchmark. We construct a comprehensive benchmark comprising 100 diverse story prompts spanning five genres: romantic comedies, action sequences, dramatic narratives, family dramas, and suspense thrillers. Each prompt contains detailed character descriptions, plot elements, and emotional arcs to test different aspects of our system.

TABLE I
Quantitative comparison. Our method achieves superior performance across all metrics. OC: Overall Consistency, SC: Subject Consistency, AQ: Aesthetic Quality, MS: Motion Smoothness.

Method	OC↑	SC↑	AQ↑	MS↑
CogVideoX	0.096	0.823	0.379	0.960
VideoCrafter2	0.076	0.885	0.364	0.920
Hunyuan	0.185	0.909	0.569	0.976
CineAGI	**0.259****	**0.949****	**0.600***	**0.987***
Improvement	**+40.0%**	**+4.4%**	**+5.4%**	**+1.1%**

$^{*}p < 0.05$, $^{**}p < 0.01$ (paired t-test over 100 prompts vs. best baseline).

Generation Settings. For fair comparison, we sample videos at 24 FPS with 5.375-second duration (129 frames per scene) at 512×512 resolution. Baseline models are evaluated using two complementary strategies: (1) generating multiple scenes with different random seeds until matching our total video length; (2) using scene descriptions from our Script Writer to generate equivalent-length videos per scene. Quantitative metrics are computed as average scores across both strategies, providing comprehensive evaluation accounting for both generation approaches.

Evaluation Metrics. We employ the standardized VBench framework to assess multiple aspects: Overall Consistency (OC) through ViCLIP evaluates text-to-video alignment and story flow continuity; Subject Consistency (SC) assesses maintenance of character identities across scenes; Aesthetic Quality (AQ) evaluates visual fidelity and artistic composition; Motion Smoothness (MS) measures fluidity of character movements and scene transitions.

Baselines. We compare against CogVideoX [8], VideoCrafter2 [26], and Hunyuan [27].

Computational Cost. Our full pipeline requires approximately 11.3 minutes per scene (5.375s) on a single NVIDIA A100 GPU; a detailed per-component breakdown is provided in Appendix E.

B. Quantitative Evaluation

Table I presents quantitative results. CineAGI achieves the highest Overall Consistency (0.259, 40% improvement over Hunyuan), Subject Consistency (0.949, 4.4% improvement), Aesthetic Quality (0.600, 5.4% improvement), and Motion Smoothness (0.987, 1.1% improvement), demonstrating superior narrative coherence and character consistency. *Per-genre performance breakdown, failure case analysis, and computational efficiency comparison are provided in Appendix E.*

C. Human Evaluation

We conducted a user study with 20 participants (10 with professional multimedia experience) rating videos on a 5-point Likert scale across five dimensions.

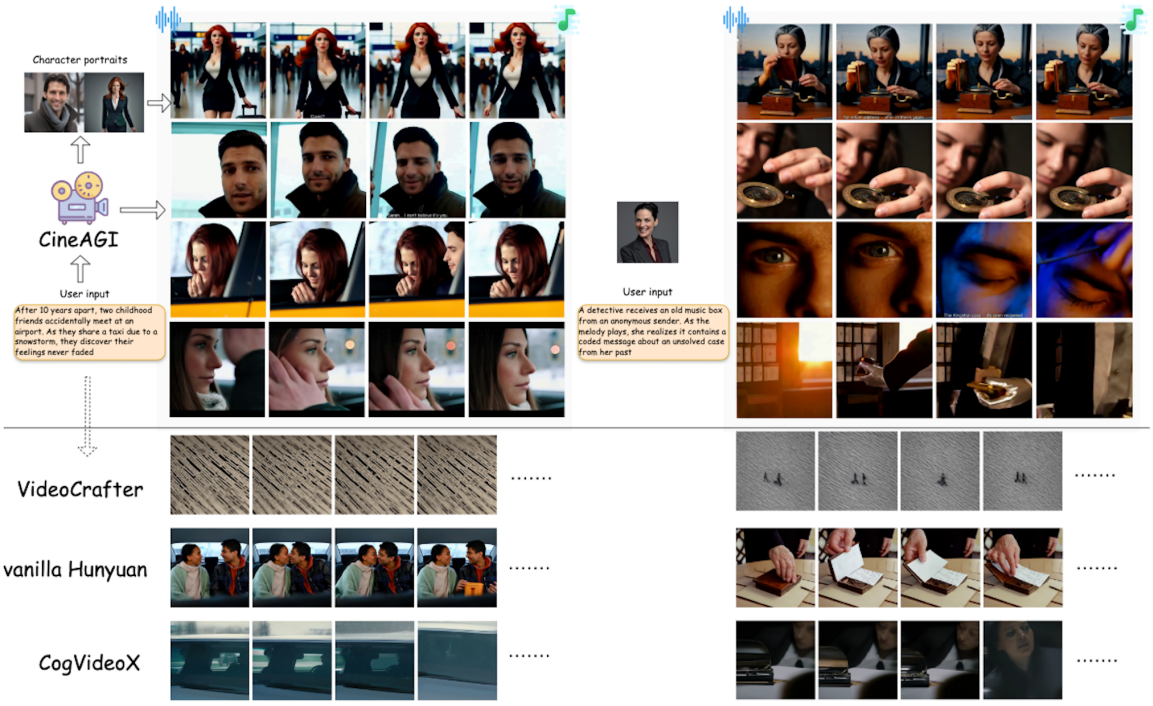

Fig. 3. **Qualitative comparisons with state-of-the-art methods.** Our approach generates more coherent narratives with consistent character portrayals across varied scenes while maintaining superior visual quality and natural character movements.

TABLE II
HUMAN EVALUATION RESULTS. SCORES RATED FROM 1-5 (HIGHER IS BETTER). VQ: VISUAL QUALITY, NC: NARRATIVE COHERENCE, CC: CHARACTER CONSISTENCY, AC: AUDIO COHERENCE, OQ: OVERALL QUALITY. (-) INDICATES UNSUPPORTED FEATURE.

Method	VQ↑	NC↑	CC↑	AC↑	OQ↑
CogVideoX	3.16	2.52	2.21	-	2.63
VideoCrafter2	2.75	2.26	1.98	-	2.45
Hunyuan	3.52	2.91	2.44	-	2.88
CineAGI	**3.83**	**3.57**	**3.14**	**3.26**	**3.37**
Improvement	**+8.8%**	**+22.7%**	**+28.7%**	-	**+17.0%**

Table II shows superior performance across all dimensions. CineAGI achieves highest scores in Visual Quality (3.83), Narrative Coherence (3.57), and Character Consistency (3.14, 28.7% improvement). Audio Coherence (3.26) demonstrates effective multi-modal integration unique to our approach.

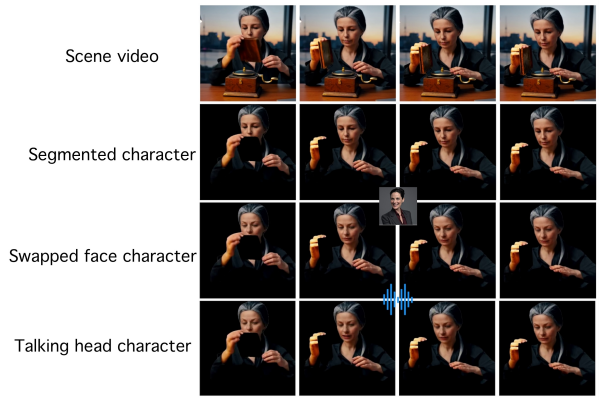

Fig. 4. **Decoupled Character Integration (DCI) pipeline visualization.** From top to bottom: original scene video, segmented character regions, face-swapped results with reference portrait, and final talking head output with audio synchronization. This systematic decomposition enables precise character control while maintaining scene context.

D. Ablation Study

We conduct comprehensive ablation studies to validate our design choices (Table III, Figure 4). We examine three key components:

Table III validates our design choices. Without Narrative Synthesis Module, Overall Consistency drops to 0.232 and Subject Consistency to 0.924. Without Quality Inspector, performance moderately decreases (OC: 0.245, SC: 0.938). Without Decoupled Character Integration, Aesthetic Quality decreases to 0.583 and Motion Smoothness to 0.971, confirming each component's essential role. *Individual agent contribution analysis, alternative architecture comparisons, and sensitivity analysis for key hyperparameters are provided in Appendix F.*

E. Qualitative Results

Figure 3 demonstrates CineAGI's effectiveness in producing coherent movies with consistent character portrayals and appropriate cinematography.

TABLE III
ABLATION STUDY RESULTS. NSM: NARRATIVE SYNTHESIS MODULE, QI: QUALITY INSPECTOR, DCI: DECOUPLED CHARACTER INTEGRATION.

Variant	OC↑	SC↑	AQ↑	MS↑
w/o NSM	0.232	0.924	0.575	0.974
w/o QI	0.245	0.938	0.570	0.982
w/o DCI	0.206	0.911	0.583	0.971
Full CineAGI	**0.259**	**0.949**	**0.600**	**0.987**

V. CONCLUSION

We present CineAGI, a hierarchical framework for automated movie creation that coordinates multi-agent narrative synthesis, decoupled character-centric processing, and audio-visual synchronization. Extensive experiments validate significant gains across both automated metrics and human evaluation, establishing a modular foundation for AI-driven filmmaking.

REFERENCES

[1] Minheng Ni, Chenfei Wu, Huaying Yuan, Zhengyuan Yang, Ming Gong, Lijuan Wang, Zicheng Liu, Wangmeng Zuo, and Nan Duan, "Autodirector: Online auto-scheduling agents for multi-sensory composition," *arXiv preprint arXiv:2408.11564*, 2024.

[2] Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang, "Anim-director: A large multimodal model powered agent for controllable animation video generation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[5] William Peebles and Saining Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh, "Video generation models as world simulators," https://openai.com/research/video-generation-models-as-world-simulators, 2024, OpenAI Technical Report.

[7] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri, "Lumiere: A space-time diffusion model for video generation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[8] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al., "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al., "Videocrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023.

[10] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang, "Id-animator: Zero-shot identity-preserving human video generation," *arXiv preprint arXiv:2404.15275*, 2024.

[11] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan, "Identity-preserving text-to-video generation by frequency decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[12] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou, "Storydiffusion: Consistent self-attention for long-range image and video generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[13] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 8153–8163.

[14] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi, "Streamingt2v: Consistent, dynamic, and extendable long video generation from text," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[15] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han, "Fifo-diffusion: Generating infinite videos from text without training," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[16] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal, "Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning," *arXiv preprint arXiv:2309.15091*, 2023.

[17] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, et al., "Videopoet: A large language model for zero-shot video generation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[18] Adam Polyak, Amit Zohar, et al., "Movie gen: A cast of media foundation models," *arXiv preprint arXiv:2410.13720*, 2024.

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

[20] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al., "Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents," *arXiv preprint arXiv:2308.10848*, vol. 2, no. 4, pp. 6, 2023.

[21] Jiuniu Wang, Zehua Du, Yuyuan Zhao, Bo Yuan, Kexiang Wang, Jian Liang, Yaxi Zhao, Yihen Lu, Gengliang Li, Junlong Gao, Xin Tu, and Zhenyu Guo, "Aesopagent: Agent-driven evolutionary system on story-to-video production," *arXiv preprint arXiv:2403.07952*, 2024.

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

[23] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang, "Vlogger: Make your dream a vlog," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8806–8817.

[24] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2003–2011.

[25] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 484–492.

[26] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7310–7320.

[27] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al., "Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding," *arXiv preprint arXiv:2405.08748*, 2024.