Design and Application of an Intelligent Short Drama Creation System Based on Multi-Agent Collaboration

Wu Nan, Yang Xiaoyu, Liu Jun

2026-08-25

ChinaRxiv

Abstract

With the development of generative artificial intelligence technologies, intelligent creation tools have been widely applied to script generation, visual design, video production, and music generation in short-form film and television dramas. However, existing creation approaches mostly rely on independent models or tools, leading to problems such as poor coordination between generation stages, difficulty in reusing visual assets, and insufficient continuity across segmented videos. To address these issues, this paper designs and implements a collaborative creation system composed of seven agents. Following the division of labor in film and television production, the system includes an art director, a screenwriter, a character designer, a scene designer, a storyboard artist, a video generation specialist, and a music director. It adopts Across-agent and Across-VAM, post-trained for film and television tasks, as the model foundation, respectively undertaking language generation and image, video, and music generation tasks. By sharing scripts and settings, multi-view visual assets, reference materials from historical projects, and video boundary frames among agents, the system can complete the full generation workflow from creative requirement input to final video output, and can also perform visual effects reconstruction, scene extension, and narrative insert-shot generation based on existing materials. To alleviate visual discontinuities caused by the independent generation of multiple video segments, the system uses the terminal boundary frame of the preceding shot as the visual generation condition for the subsequent shot, transmitting character states, scene content, and composition information. The system has been applied to multiple actual broadcast projects. Project records show that, in applicable complex visual effects tasks, the production cost per minute decreased from tens of thousands of yuan in traditional workflows to approximately 1000 yuan, and the production cycle was shortened from 7–10 days to about 1 day. Application results demonstrate that the system can achieve coordinated integration of multimodal generation tasks and, in applicable scenarios, reduce the production cost and shorten the production cycle of short-form film and television dramas.

Full Text

Design and Application of an Intelligent Short Drama Creation System Based on Multi-Agent Collaboration

Wu Nan[1], Yang Xiaoyu[1], Liu Jun[2]

1. Across Physical (Beijing) Technology Co., Ltd., Beijing 100083

2. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876

Note: Yang Xiaoyu and Wu Nan are co-first authors.

Abstract With the development of generative artificial intelligence technology, intelligent creation tools have been widely applied to such stages of short film and television drama production as script generation, visual design, video production, and music generation. However, existing creation methods mostly rely on independent models or tools, leading to problems such as insufficient coordination between generation stages, difficulty in reusing visual assets, and inadequate continuity across segmented videos. In response to these problems, this paper designs and implements a collaborative creation system comprising seven agents. Following the division of labor in film and television production, the system sets up an art director, screenwriter, character designer, scene designer, storyboard artist, video generation specialist, and music director. Using Across-agent and Across-VAM, post-trained for film and television tasks, as the underlying models, these agents respectively undertake language generation and image, video, and music generation tasks. By sharing scripts and settings, multi-view visual assets, reference materials from previous projects, and video boundary frames among agents, the system can complete the entire generation process from creative requirement input to finished-video output, and can also reconstruct visual effects, expand scenes, and generate supplementary shots for existing materials. To alleviate visual jumps caused by independently generated multiple video clips, the system uses the terminal boundary frame of the preceding shot as the visual generation condition for the next shot, transferring information about character states, scene content, and composition. The system has been applied in multiple real broadcast projects. Project records show that, in applicable complex visual-effects tasks, the per-minute production cost was reduced from tens of thousands of yuan under traditional workflows to about 1,000 yuan, and the production cycle was shortened from 7–10 days to about 1 day. The application results show that the system can achieve collaborative coordination of multimodal generation tasks and, in applicable scenarios, reduce the production cost and shorten the production cycle of short film and television dramas.

[**Keywords**] multimodal generation; visual asset reuse; shot continuity

[**CLC Classification Numbers**] J91; TP391

Design and Application of an Intelligent Short Drama Creation System Based on Multi-Agent Collaboration

Nan Wu (Across Physical AI Beijing Co., Ltd.)
YANG Xiaoyu (Across Physical AI Beijing Co., Ltd.)
Jun Liu (School of Artificial Intelligence, Beijing University of Posts and Telecommunications)

Abstract: This article presents a seven-agent collaborative system for short-drama production to address Abstract: With the rapid development of generative artificial intelligence, AI-assisted creative tools have been widely adopted in short-form drama production, including script generation, visual design, video production, and music generation. However, existing workflows often rely on isolated models or tools, resulting in poor coordination between generation stages, limited reuse of visual assets, and insufficient continuity across separately generated video segments. To address these issues, this study designs and implements a collaborative creation system comprising seven agents. Following conventional film and television production roles, the system includes an art director, screenwriter, character designer, scene designer, storyboard artist, video generation specialist, and music director. Across-agent and Across-VAM, both post-trained for film and television tasks, serve as the foundation models: Across-agent supports language generation, while Across-VAM supports image, video, and music generation. By sharing scripts and settings, multi-view visual assets, reference assets from previous projects, and video boundary frames among the agents, the system supports end-to-end production from creative briefs to final outputs. It can also reconstruct visual effects, extend scenes, and generate supplementary shots for existing footage. To mitigate visual discontinuities caused by generating multiple video segments independently, the terminal boundary frame of each shot is used as a visual condition for generating the next shot, thereby carrying forward character states, scene content, and compositional information. The system has been deployed in several productions that have been broadcast. Project records show that, for applicable complex visual-effects tasks, the production cost per minute was reduced from tens of thousands of yuan under conventional workflows to approximately RMB 1,000, while the production time was shortened from 7–10 days to about one day. The application results demonstrate that the system can effectively coordinate multimodal generation tasks and, in applicable scenarios, reduce both the production cost and production time of short-form dramas.

Keywords: Multimodal generation; Visual asset reuse; Shot continuity

1 Introduction

In recent years, artificial intelligence generated content (AI-generated content, AIGC) has been able to stably generate multimodal content such as text, images, video, and music[1]; the development of video diffusion and multimodal generation technologies has further improved the visual quality and content controllability of generated videos[2-5]. Related technologies are gradually ···

enter such film and television stages as scriptwriting, concept design, storyboard creation, visual effects, and sound production[6-8], driving production processes to reduce costs and shorten cycles[9-10]. At the same time, online audiovisual content and micro-short dramas maintain a high level of user activity. The *China Online Audiovisual Development Research Report (2026)* shows that, as of December 2025, the scale of online audiovisual users in China had reached

1.099 billion, with average daily use of 201 minutes per person, and 129 minutes for micro-short-drama users; in 2025, AI-generated videos and audios exceeded 2 billion in total, an increase of more than 14 times over 2024[11]. Against this background, demand for the application of AIGC in the creation and production of micro-short dramas continues to grow[12].

Existing film and television AIGC production still mostly adopts a workflow in which production personnel separately invoke different models: after a language model generates the script, character and scene descriptions must be transferred into an image model; before video generation, prompts must again be reorganized, and finally the clips are manually spliced. Such processes fragment the creative stages, information is easily lost as it passes across tools, and different models understand elements such as character appearance and scene environment inconsistently. Meanwhile, a video model can usually generate only clips of several seconds to several dozen seconds at a time, and independently generated shots struggle to maintain continuity of characters, scenes, and actions. Production personnel therefore need to repeatedly adjust parameters and screen results, making it difficult to form a stable automated process. Existing studies have likewise pointed out that long-video generation remains constrained by problems such as cross-shot consistency, temporal coherence, and shot transitions[13-14]. Thus, the bottleneck of film and television AIGC lies not only in the performance of a single model, but also in the way multimodal stages are organized.

In response to the above problems, this paper constructs a seven-agent collaborative creation system for film-and-television short-drama production (hereinafter referred to as the "collaborative system"). Role-based multi-agent systems have demonstrated capabilities for division of labor and collaboration in complex tasks such as software engineering[15-17], and have begun to be applied in the film domain[18-20]. The system is based on Across-agent and Across-VAM, which have been trained on film-and-television tasks, and refers to film-and-television positions by setting up an art director, screenwriter, character designer, scene designer, storyboard artist, video generation engineer, and music director. Among them, Across-agent is respectively adapted to language tasks such as art coordination, scriptwriting, and storyboard planning; Across-VAM sets up three branches—image, video, and music—and undertakes the generation of visual assets, shot clips, and background music. The agents share production information through scripts and settings, multi-view reference images, and video boundary frames; in the renovation of existing projects, original shots, keyframes, and modification requirements are also incorporated into the shared context.

The system design mainly includes two aspects. First, it uses role-based agents and structured intermediate assets to connect scripts, characters, scenes, storyboards, video, and music generation, reducing the loss of production information during transitions between stages. Second, when continuously generating short-video clips, it passes the boundary frame of the previous shot to the next shot,

thereby constraining the character state, scene, and composition of adjacent shots. Thus, the system can generate a complete short drama from creative requirements, and can also carry out visual-effect reconstruction, expansion of scenes and plot elements, multi-shot fill-in shooting, and narrative extension on the basis of existing materials. Combining projects such as pure-AI short dramas, visual-effect replacement, scene reconstruction, and reconstruction of live-action content, this paper analyzes the system' s production usability, production cost, and cycle.

2 Multi-Agent Film-and-Television Short-Drama Creation System Design

The collaborative system is oriented toward minute-level film-and-television short-drama production and accommodates both creation from scratch and reconstruction using existing materials. For creation from scratch, after the user inputs requirements such as theme, characters, visual style, emotional tone, and target duration, the system sequentially completes scriptwriting, visual asset generation, storyboard planning, video generation, music creation, and audiovisual synthesis. For existing projects, the user simultaneously provides original shots, keyframes, and modification requirements, and the seven types of agents complete visual-effect reconstruction, expansion of scene and plot elements, shot supplementation, or reconstruction of live-action content under the constraints of the original plot, character relationships, and visual style. The two types of tasks share the collaboration framework and creative-asset management mechanism, reducing manual transfer and repeated operations. The system organizes model capabilities through a unified interface and reserves extension interfaces for adding professional agents.

2.1 Overall System Architecture

The overall system architecture is shown in Figure 1. The user interaction layer transforms natural-language requirements into structured creative tasks; the original shots and keyframes of existing projects can be input together as reference assets. The agent collaboration layer maintains a shared creative context, schedules the seven types of agents according to task progress, and invokes the language, image, video, and music generation capabilities of the model capability layer. The scripts, character and scene settings, visual reference images, storyboard instructions, original materials, and video clips formed at each stage uniformly enter the creative asset layer for use in subsequent stages. The output synthesis layer is responsible for shot splicing, music alignment, and finished-film packaging.

Figure 1 Overall architecture of the multi-agent film-and-television short-drama creation system

Figure 1. Overall architecture of the multi-agent film-and-television short-drama creation system

Figure 1: Figure 1. Overall architecture of the multi-agent film-and-television short-drama creation system

2.2 Model Capabilities and Mapping to Professional Agents

In the collaborative system, Across-agent and Across-VAM provide model capabilities. Across-agent is responsible for language understanding, content generation, and task planning; Across-VAM sets up three branches: image, video, and music. Both types of models are post-trained on film-and-television creation tasks and are connected to the agent collaboration layer through a unified interface. The correspondence between agents and model capabilities is shown in Table 1.

Language-based agents do not adopt the same universal output; instead, task objectives are set according to positions: the artistic director is responsible for subject analysis, visual style, and overall constraints, and extracts the aesthetic and image features of the original footage from existing projects; the screenwriter is responsible for minute-level narratives, character relationships and roles, and scene descriptions, while also determining the narrative functions of added elements and supplementary shots; the shot designer is responsible for shot breakdown, scene and camera-position planning, and the organization of video prompts. The Across-VAM image branch generates multi-view assets for characters and scenes, and can also generate extended assets coordinated with the original footage by combining the original images; the video branch receives conditions such as reference images, original shots, key frames, and boundary frames, and generates added, replacement, or supplementary shots; the music branch generates background music according to the script emotions and target duration.

This paper discusses the organization and application of model capabilities in the multi-agent film-and-television workflow, and does not elaborate on training data, parameters, or low-level implementation.

Table 1 Mapping between agents, model capabilities, and task adaptation

Agent	Model capability	Task adaptation focus	Main output
Artistic director	Across-agent	Subject analysis, visual style, extraction of constraints from existing materials	Overall creative plan, visual and modification constraints

Agent	Model capability	Task adaptation focus	Main output
Screenwriter	Across-agent	Minute-level narrative, plot context, character and scene descriptions	Script, character and scene settings, supplementary-shot requirements
Shot designer	Across-agent	Shot breakdown, multiple scenes and camera-position planning, prompt organization	Shot list, shot prompts
Character designer	Across-VAM image branch	Character appearance, multi-view perspectives, and age-stage extension	Character multi-view images, stage-based character references
Scene designer	Across-VAM image branch	Scene space, multi-view perspectives, and extension of original-footage style	Scene multi-view images, extended scene references
Video generator	Across-VAM video branch	Generation conditioned on reference images, original materials, and boundary-frame conditions	Added, replacement, and supplementary shot clips
Music director	Across-VAM music branch	Adaptation to script emotion and duration	Background music

2.3 Agent Responsibilities and Collaborative Creation Mechanism

The system connects the agents through position-task definitions and structured outputs, so that the results of the previous stage can directly enter the next stage. For creation from scratch, the art director and screenwriter share the creative context: the former refines the theme, visual style, emotional tone, and overall constraints, while the latter uses these to form the initial script draft, character relationships, and scene requirements. After the initial script

draft is completed, the art director proposes adjustment suggestions around target duration, character behavior, scene settings, and overall style; after the screenwriter revises it, the final script and descriptions of characters and scenes are formed. For existing projects, the original shots, key frames, and revision requirements enter the same context. The art director extracts character states, spatial relationships, lighting colors, and artistic style; the screenwriter combines the plot to clarify the narrative function of newly added elements, supplementary shots, or flashback segments, jointly forming task descriptions executable by downstream agents. Both types of tasks write the final script, character and scene descriptions, and visual constraints into the shared context, for direct invocation by the character-design, scene-design, and shot-division agents.

The character-design agent invokes the Across-VAM image branch to convert character descriptions in the script into visual assets usable for subsequent generation, outputting multi-view character images such as front, left-side, and right-side views. These images are used both for character concept design and as identity references for video generation. For existing live-action characters, original shots or actor key frames may be introduced; when age or state changes are involved, reference assets for corresponding stages are generated while retaining the main identity features. Image-generation technology has already been used for film and television character concept design[^21].

The scene-design agent likewise invokes the Across-VAM image branch and, based on descriptions such as spatial type, time, weather, lighting, color, architectural furnishings, cinematographic style, and atmosphere, generates multi-view scene images such as front views, side views, wide-angle panoramas, and local compositions. For existing projects, the agent can also extract spatial layout, lighting colors, period features, and architectural style from original shots or key frames, and on this basis generate extended scenes coordinated with the original footage. These assets are used to constrain the spatial relationships of continuous shots, and also provide visual references for scenes not yet filmed in the original project or for which materials are insufficient.

The shot-division agent converts the literary script or revision requirements for an existing project into shot tasks executable by the Across-VAM video branch. A complete script of about 1 minute is usually divided into approximately 12 shots, each about 5 seconds, with the specific number adjusted according to the plot; existing projects also require analysis of the original shot structure and localization of content to be replaced, supplemented, or expanded. Each task records shot duration, frame content, character actions, scene position, scene scale, camera movement, reference assets, video prompts, and the linkage relationship with preceding and following shots, and binds the corresponding character images, scene images, and necessary original shots or key frames.

The video-generation agent invokes the Across-VAM video branch by shot. When creating from scratch, the input includes shot prompts of about 5 seconds, character images, scene images, boundary frames from the previous shot, and parameters such as duration and aspect ratio; when adapting an

existing project, original shots or key frames and specific revision requirements are also added to constrain the characters, space, and composition of the newly generated content. The output may be a complete new shot, or may be used to replace, supplement, or expand an original segment. For the same shot task, multiple candidate clips may be retained for production staff to compare character completeness, action performance, and scene matching. Because generation results are stochastic, for problems such as character deformation, abnormal actions, or scene drift, the system regenerates or screens other candidate results, and enters the synthesis stage only after meeting usage requirements.

The music-director agent invokes the Across-VAM music branch to generate background music close to the final film duration according to the complete script, story theme, emotional changes, visual style, and target duration. After generation is completed, the system makes basic adjustments according to the actual final film duration and completes audio-video synthesis. At the current stage, this agent mainly handles background music and does not generate complete dialogue or sound effects.

Four types of production data are transmitted among agents: first, the script and character and scene settings, used to determine subsequent creative content; second, original shots, key frames, and revision requirements from existing projects, used to provide character, space, style, and plot context; third, generated assets such as multi-view images of characters and scenes, serving as visual conditions for image and video generation; fourth, boundary frames of adjacent video clips, used to transmit the visual state between shots. The above data are stored uniformly and flow with the tasks, avoiding the need to repeatedly organize inputs in subsequent stages.

2.4 Multi-Agent Collaborative Creation Process

The collaborative process is shown in Figure 2. When creating from scratch, the seven types of agents sequentially complete script and setting, visual assets, shot division, video, and soundtrack generation, and then form the final film through unified asset management and output synthesis. When adapting an existing project, the original shots, key frames, and revision requirements are incorporated into the shared context as additional inputs, and are used to constrain characters, scenes, supplementary shots, and replacement content. The two types of tasks adopt the same task scheduling, asset transmission, and synthesis mechanisms.

Figure content: User creative requirements → Art Director + Screenwriter (collaborative creation and consultation; script | character setting | scene setting) → Character Designer (character multi-view images) and Scene Designer (scene multi-view images) → Shot-Division Specialist (shot-unit splitting / shot-prompt generation) → Video Generator (shot-clip generation / boundary-frame extraction) → Video splicing / audio-video synthesis → Final film output. A

"boundary-frame transfer" link connects the video-generation stage back toward the shot-division/video-generation workflow. The Music Director (background-music generation) feeds into video splicing / audio-video synthesis.

Figure 2 Collaborative workflow of the seven types of agents

3 Multimodal Collaboration and Continuous-Shot Generation Method

3.1 Conversion from Script to Shot Prompts

A literary script is usually narrative-oriented, for example: "The male protagonist enters the restaurant, sees the female protagonist sitting by the window, and looks surprised." A video-generation model, however, requires explicit conditions such as character appearance, scene layout, action sequence, expression changes, shot scale, camera movement, lighting, color, style, and duration. The shot-division agent therefore

The system combines information such as shot images, characters entering/exiting the scene, actions, and expressions with the character and scene settings, generates structured prompts according to the input specifications of the Across-VAM video branch, and binds corresponding reference images to each shot. This conversion does not change the narrative events in the script; rather, it decomposes them into visual conditions executable by individual shots, and records the start and end states of shots as well as the preceding and following connection relationships. Existing projects also need to add the scene, composition, character status, scene features, and modification requirements of the original shots, so that the generated shots can connect with the original materials.

3.2 Generation of the First Shot

The first generated shot has no preceding segment for reference, and is mainly based on the shot prompt and multi-view images of the characters and scene. When reconstructing or adding shots, the original shot or key frames may also be added as visual conditions. After the shot is generated, the system saves the video segment and extracts the first and last frames. The first frame records the character position, scene state, and composition at the beginning of the shot; the last frame records the relationship among the characters, environment, and camera position at the end of the shot. Together, the two constitute the continuity conditions for subsequent shot generation.

3.3 Sequential Shot Generation Process Based on Boundary-Frame Propagation

The duration of a single video generation is limited, and independently generating multiple shots is prone to visual jumps. Therefore, the system introduces

Figure 3 Sequential shot generation process based on boundary-frame propagation

Figure 2: Figure 3 Sequential shot generation process based on boundary-frame propagation

boundary-frame propagation into shot scheduling for the Across-VAM video branch, with the process shown in Figure 3. The complete script is first divided into several shot units; the first shot is generated from text prompts and character and scene assets; the system then extracts the first and last frames of this shot, and inputs them together with the prompt of the next shot, the character images, and the scene images into the video branch. This process is repeated shot by shot until all shots are completed.

Boundary frames provide the visual states of character posture, scene lighting, and picture composition at the end of the previous segment to the next shot. Compared with using only text prompts, this condition enables the subsequent shot to reference the already formed relationships among characters, scenes, and camera positions, and supplements local visual information that is difficult to describe completely in text, helping to reduce obvious jumps between adjacent shots. This method does not change the basic generation process of the video model; it only adds visual conditions between shots. Continuous generation and consistency maintenance of long videos remain current research focuses[^22-23].

Figure 3 Sequential shot generation process based on boundary-frame propagation

3.4 Multi-level Consistency Control

The system controls consistency at three levels: settings, visual assets, and adjacent shots. Characters and scenes take unified settings and multi-view images as the main constraints, and the art director writes the visual style into sub-shot and video prompts. When reconstructing existing content, the original shot or key frames are further used to constrain character identity, scene space, lighting color, and image style. When characters undergo age or status changes, actor references and character settings are also needed to maintain the main identity features. Adjacent shots transmit action, space, and composition states through boundary frames. The above conditions cannot completely eliminate generative randomness; shots involving rapid movement, multi-person interaction, and contact with complex objects may still need to be regenerated or manually selected.

3.5 Video Stitching and Audio-Video Synthesis

After all video segments are generated, the system stitches them in shot-number order, unifies the resolution, frame rate, and frame width of each segment, and adjusts duration deviations. Stitching is mainly handled automatically, while production staff review shot connections. If obvious jumps exist between adja-

Figure 4: Visual-effects reconstruction results of the collaborative system in *Lost in the Stars*

Figure 3: Figure 4: Visual-effects reconstruction results of the collaborative system in *Lost in the Stars*

cent segments, the relevant shots are regenerated or manually adjusted. After the background music generated by the music director is basically adjusted according to the duration of the finished film, it is synchronously packaged and output with the video track.

4 Project Application and Effect Analysis

To investigate the application effects of the collaboration system in actual film and television production, this paper selects three completed projects: *The Mystery of Three Threads*, *Marriage Branch*, and *Rebirth 2000: Starting Again from Chasing the Campus Belle at the Same Desk*. All three adopt the seven-agent collaboration process, but the input materials and production goals differ, respectively involving visual-effects reconstruction, expansion of scene and plot elements, and reconstruction of live-action content. The applicability of the system in different production tasks can be observed from aspects such as visual content adjustment, shot planning, scene expansion, and narrative supplementary shots.

4.1 Application of the Collaboration System in Visual-Effects Reconstruction —*The Mystery of Three Threads*

The production tasks for *Lost in the Stars* include adjusting the original visual-effects plan, supplementing shots, and expanding action sequences. While preserving the original plot, actors' performances, and overall artistic style, the system uniformly analyzes the shots to be revised and decomposes the requirements into the stages of visual design, shot planning, and shot generation, so that the reconstruction of visual effects and subsequent shot supplementation adopt the same creative constraints. This organizational approach allows all stages to share information about the characters, scenes, and art design of the original footage, reducing repeated communication when shots are handled separately. Figure 4 shows before-and-after comparisons for two representative shots —the unknown life form and the laser gun. The left side shows the original effect, and the right side shows the reconstruction result produced by the collaborative system. The two shots demonstrate that the system can complete localized visual-effects replacement on the basis of existing materials, and that the reconstructed results are largely consistent with the main characters, composition, and artistic style of the original footage.

Figure 4 Visual-effects reconstruction results of the collaborative system in *Lost in the Stars*

Figure 5: Shot-division effects for action scenes generated by the collaborative system in *Lost in the Stars*

Figure 4: Figure 5: Shot-division effects for action scenes generated by the collaborative system in *Lost in the Stars*

Figure 6

Figure 5: Figure 6

In addition to visual-effects adjustment for individual shots, the system also decomposes and supplements shots according to the rhythm of the plot and the relationships among actions. In the action sequences of *Lost in the Stars*, the storyboard- and video-generation agents generate shots with different scenic units and camera positions around the characters' positions, movement directions, and scene space (Figure 5). These materials share the constraints of the plot, characters, and scenes, enabling the same action sequence to obtain multiple viewing angles and providing more sufficient choices for subsequent editing.

Figure 5 Shot-division effects for action scenes generated by the collaborative system in *Lost in the Stars*

4.2 Application of the Collaborative System in Scene and Content Expansion—*A Gilded Bough*

The production task set for *A Gilded Bough* focuses on supplementing plot elements in existing shots and expanding scene space. The system plans new content according to the original shot content, plot requirements, and overall visual style, and completes localized generation while maintaining the relationships among characters, scenes, and composition. Compared with processing each shot independently, the shared visual constraints allow newly added content in different shots to continue using the same project settings, reducing obvious stylistic discrepancies. Figure 6 shows two typical processes: the left side is an expansion of the water area, in which the system preserves the characters, buildings, and original composition while supplementing the water surface; the right side is supplementation of a white-snake plot element, adding a white-snake figure related to the plot into the original character shot. The two cases correspond respectively to changes in scene space and plot content, and the generated portions continue to follow the main colors, lighting, and compositional characteristics of the original project.

Figure 6. Effects of expanding scene and plot elements by the co-creation system in *The Glory of Tang Dynasty*

The co-creation system can also, according to the scene requirements in the script, supplement environmental shots that were not filmed in the original project or for which the available material is insufficient. For this type of task,

Figure 7

Figure 6: Figure 7

Figure 8

Figure 7: Figure 8

the scene-design and shot-splitting agents plan by combining the story location, historical background, art style, and shot purpose, and the video-generation agent then generates empty-shot material. As shown in Figure 7, four shots cover different spaces and scene types and can be directly used in subsequent editing. This case shows that the system can transform textual scene requirements into environmental shots whose visual style is close to that of the overall project.

Figure 7. Scene shots generated by the co-creation system in *The Glory of Tang Dynasty*

4.3 Application of the co-creation system in reconstructing live-action content—*Reborn 2000*

Reborn 2000: Starting from Pursuing the Campus Belle at the Same Desk contains a relatively large amount of live-action content. Compared with generation from scratch, adding new shots still requires continuity with the original footage in terms of characters, period features, spatial environment, and image style; therefore, greater emphasis is placed on examining the system' s ability to understand and expand existing materials. Using the characters and scene information in the finished footage as shared creative constraints, the system completes scene reconstruction and narrative supplementary shots. During scene reconstruction, the art-director and scene-design agents analyze the spatial form, lighting, color, and period characteristics already present in the original footage, and then regenerate the indoor spaces required for the subsequent plot, such as the kitchen, bedroom, and living room, as shown in Figure 8. The added scenes retain the main visual tone of the original footage and expand the scene assets available for subsequent shot generation.

Figure 8 The scene-reconstruction effect of the collaborative system in *Reborn 2000: Starting from Chasing the Green Campus Belle at the Same Desk*

In addition to scene expansion, the flashback segments of this project also need to depict the character' s childhood. Using the appearance features and character settings of the adult actor as references, the system generates the character' s childhood image and corresponding life scenes while retaining the main identity features, and on this basis supplements multiple dramatic shots (Figure 9). This case shows that when the character's age and the scene change, the character reference assets can still provide identity constraints for non-live-action segments, enabling newly added shots to unfold around the same character setting.

Figure 9 Flashback shots of the childhood period generated by the collaborative system in *Reborn 2000: Starting from Chasing the Green Campus Belle at the Same Desk*

4.4 Production Cost and Cycle Analysis

According to the company' s production records for completed projects, in complex VFX shots suitable for generative methods, the traditional computer graphics (CG) workflow usually includes concept design, modeling, texture mapping, rigging, animation, lighting, rendering, and compositing; the collaborative system mainly includes plan and prompt generation, image-asset generation, model invocation, result screening, local adjustment, and video compositing. The per-minute cost is reduced from tens of thousands of yuan to about 1,000 yuan, and the cycle is shortened from 7–10 days to about 1 day (Table 2). These results come from records of specific projects and do not represent all production tasks; work such as high-precision live-action compositing still requires cooperation with traditional workflows.

Table 2 Comparison of the Costs and Cycles of the Traditional Computer Graphics (CG) VFX Workflow and the Collaborative System

Comparison Item	Traditional CG Workflow	Collaborative System
Per-minute production cost	Tens of thousands of yuan	About 1,000 yuan
Production cycle	About 7–10 days	About 1 day
Main production stages	Concept design, modeling, texture mapping, rigging, animation, lighting, rendering, compositing	Plan and prompt generation, image-asset generation, model invocation, result screening, local adjustment, video compositing
Labor input	Multiple professional production personnel	A small number of production personnel

4.5 Content Quality and Production Usability

The project team invited 5 actual production personnel to conduct internal 5-point evaluations of 4 completed films including the above project. The indicators included character consistency, scene consistency, shot continuity, narrative completeness, and production usability. The mean scores of all items were 3.8–4.3 (Table 3), among which narrative completeness was the highest and production usability was relatively low. Statistics on the generation records of 48 shots showed that 5 shots required 3 or more generations, while the remaining

shots were usable after 1–2 generations; all projects passed the review of the corresponding platforms and were broadcast. Because the number of evaluators and samples was limited, these results only reflect the production performance of the current projects.

Table 3 Content Quality Evaluation Results

Evaluation Indicator	Evaluation Content	Average Score
Character consistency	Whether characters in different shots maintain the same identity features	3.9
Scene consistency	Whether the scene layout, lighting, and color are stable	4.2
Shot continuity	Whether the actions and space in adjacent segments connect naturally	4.1
Narrative completeness	Whether the short drama can express the story clearly	4.3
Production usability	Whether the generated results can enter the actual broadcast workflow	3.8

5 Conclusion and Outlook

This paper constructs a seven-agent collaborative creation system for film-and-television short-drama production. Based on Across-agent and Across-VAM as models, the system connects artistic coordination, screenplay, characters, scenes, shot breakdown, video, and music through structured creative assets, existing project materials, and boundary frames, and can complete the full-process creation from requirements to finished film. It can also be used for VFX reconstruction, scene expansion, and narrative supplementary shots. Practical projects show that, in applicable tasks, the system can reduce production costs and cycles, but its effects are still affected by the stability of generative models and manual screening.

Shots involving rapid movement, multi-person interaction, and contact with objects are still prone to limb deformation, abnormal actions, and detail drift; when the time span is long or the scene changes greatly, consistency between characters and the environment will also decline. Current music scoring is mainly

generated according to the complete script and overall emotion, and still cannot sufficiently match shot-level action points and rhythmic cues.

ensemble. Subsequent work will improve lens quality evaluation and rhythm analysis, strengthen the reuse of character, scene, and lens assets, and supplement the dialogue, sound-effect, and mixing modules. Before release, the finished film still requires manual review of quality, copyright, and platform compliance.

References:

[1] Huang Dongjin, Liu Jinhua, Wang Qian, et al. Innovations of AI-generated content in film creation: ECCV 2024 AIGC technical review[J]. *Advanced Motion Picture Technology*, 2024(11).

[2] OpenAI. Video generation models as world simulators[EB/OL]. (2024-02-15)[2026-08-11]. https://openai.com/index/video-generation-models-as-world-simulators/.

[3] XING Z, FENG Q, CHEN H, et al. A survey on video diffusion models[EB/OL]. (2023-10-16)[2026-08-11]. https://arxiv.org/abs/2310.10647.

[4] CAO H, TAN C, GAO Z, et al. A survey on generative diffusion models[EB/OL]. (2022-09-06)[2026-08-11]. https://arxiv.org/abs/2209.02646.

[5] YIN S, FU C, ZHAO S, et al. A survey on multimodal large language models[EB/OL]. (2023-06-23)[2026-08-11]. https://arxiv.org/abs/2306.13549.

[6] Huang Xiaoyu, Chang Le. Application of artificial intelligence technology in film-production development[J]. *Advanced Motion Picture Technology*, 2024(8):14-20.

[7] Zou Kaijie, Sun Lue. Research on the application of artificial intelligence in the field of film production—taking the split-shot script generation tool as an example[J]. *Advanced Motion Picture Technology*, 2024(2):35-42.

[8] Zhu Shiling, Wang Zhenzi. Research on the application of AIGC technology in independent film creation[J]. *Advanced Motion Picture Technology*, 2024(4):59-64.

[9] Shen Huaqing, Wang Wenqiang. Paradigm change and technological prospects of audiovisual production driven by generative artificial intelligence[J]. *Advanced Motion Picture Technology*, 2025(9):12-18.

[10] Huo Xiaoyan, Sun Lijun, Zheng Zilong. Research on full-process AI animated-film creation—taking the short film *The Foolish Old Man Moves Mountains* as an example[J]. *Advanced Motion Picture Technology*, 2025(4):33-40.

[11] China Netcasting Services Association. Research report on the development of China' s online audiovisual industry (2026)[R]. Chengdu, 2026.

[12] Zheng Xiaojun, Liu Ruojin, Li Yaoqi. Subversion and reshaping: the transformation of AIGC for the micro-short-drama industry[J]. *Journal of North University of China (Social Sciences Edition)*, 2025(4):57-66.

[13] ELMOGHANY M, ROSSI R, YOON S, et al. A survey on long-video storytelling generation: architectures, consistency, and cinematic quality[EB/OL]. (2025-07-09)[2026-08-11]. https://arxiv.org/abs/2507.07202.

[14] ZHANG S, LIM S N. Towards chunk-wise generation for long videos[EB/OL]. (2024-11-27)[2026-08-11]. https://arxiv.org/abs/2411.18668.

[15] WANG L, MA C, FENG X, et al. A survey on large language model based autonomous agents[EB/OL]. (2023-08-22)[2026-08-11]. https://arxiv.org/abs/2308.11432.

[16] XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: a survey[EB/OL]. (2023-09-14)[2026-08-11]. https://arxiv.org/abs/2309.07864.

[17] GUO T, CHEN X, WANG Y, et al. Large language model based multi-agents: a survey of progress and challenges[EB/OL]. (2024-01-21)[2026-08-11]. https://arxiv.org/abs/2402.01680.

[18] Liu Da, Wang Cui, Chang Huiqin, et al. From the Global AI Summit and ICLR 2025: the intelligent evolution and upgrading of China' s film industry[J]. *Advanced Motion Picture Technology*, 2025(5):4-12.

[19] Zeng Zhen. Exploration of the application of RAG-based large-model agent technology in the film industry[J]. *Advanced Motion Picture Technology*, 2025(3).

[20] Liu Hanwen, Huang Lingyao. Integrated innovation and development paths of domestic AI large-model technology in the film industry[J]. *Advanced Motion Picture Technology*, 2025(8).

[21] Liu Jingyi. Reconstruction and transformation of AI painting technology in the creation of concept art for film and television characters[J]. *Advanced Motion Picture Technology*, 2025(9):73-80.

[22] WASEEM F, SHAHZAD M. Video is worth a thousand images: exploring the latest trends in long video generation[EB/OL]. (2024-12-24)[2026-08-11]. https://arxiv.org/abs/2412.18688.

[23] KUMAR N, BHANDARI P, MARAGATHAM G. Bridging text and video generation: a survey[EB/OL]. (2025-10-06)[2026-08-11]. https://arxiv.org/abs/2510.04999.

Author information: Author 1 Wu Nan (1997), male, bachelor' s degree, founder and CEO of Zongheng Wuli (Beijing) Technology Co., Ltd.; main research direction: world models. Yang Leiyu (2002), male, master' s degree, algorithm engineer at Zongheng Wuli (Beijing) Technology Co., Ltd.; main research direction: world models. Liu Jun (1976), male, Ph.D., associate professor at Beijing University of Posts and Telecommunications; main research direction: interconnection of artificial-intelligence agents.

Author contribution statement:

Wu Nan was responsible for writing the implementation section and contributed 40% of the manuscript text;

Yang Xiaoyu was responsible for writing the main technical sections, contributing 50% of the manuscript text;

Liu Jun was responsible for writing the agent interconnection section, contributing 10% of the manuscript text.

Note: Figure translations are in progress. See original paper for figures.