Crayotter: Traceable Multi-Agent Workflows for Long-Form Video Editing

Lecheng Yan[1*†], **Yichong Zhang**[2*†], **Xiantao Xu**[3*], **Jianze Lin**[4*], **Ben Pan**[5*], **Xiaoyu Zheng**[6*], **Jiawei Qian**[6*], **Anqi Wu**[6*], **Jiahui Geng**[7], **Ruizhe Li**[8], **Fengyu Cai**[9], **Jingcheng Niu**[9], **Raymond Li**[10], **Wenxi Li**[6], **Chenyang Lyu**[11]

[1]University of Science and Technology of China [2]Jilin University
[3]Dmall Inc. [4]Beijing Normal University [5]Tianjin University
[6]East China Normal University [7]Linköping University
[8]University of Birmingham [9]Technische Universität Darmstadt
[10]University of British Columbia [11]Alibaba Group
lyuchenyang.dcu@gmail.com

Abstract

Long-form video editing over heterogeneous footage requires agents to coordinate source selection, multimodal analysis, timeline construction, narration and subtitle alignment, rendering, and revision while exposing intermediate state for inspection and repair. We present **Crayotter**, an open-source multimodal multi-agent demo system for prompt-driven long-form video editing. Crayotter organizes production around coverage-aware material preparation, artifact-grounded editing research, and tool-grounded timeline execution. Across these stages, retrieval reports, video analyses, editing blueprints, scheduler events, tool calls, intermediate renders, and final exports are treated as first-class artifacts rather than hidden transient state. The workbench supports local assets, agent-assisted retrieval, progress monitoring, artifact preview, failure diagnosis, interrupted-job resumption, and resource-aware asynchronous execution for long-running workflows. In a 23-theme evaluation, Crayotter achieves the highest human overall score (3.40/5) among the compared systems, with its largest margins in theme alignment, narrative coherence, and editing smoothness. These results show that long-horizon video editing agents can be made traceable, inspectable, and practically controllable through observable production artifacts.

1 Introduction

Long-form video editing transforms a high-level brief and heterogeneous footage into a coherent audiovisual narrative. Unlike short-form generation, it requires joint decisions over material selection, clip understanding, narrative structure, timeline operations, and audio–text alignment. LLM agents enable interleaved reasoning and tool use (Yao et al., 2022) and role-structured collaboration (Wu et al., 2024; Hong et al., 2024); visual-production systems extend these capabilities to planning and specialized filmmaking roles (Lin et al., 2023; Xu et al., 2025b). However, they primarily synthesize shots, whereas editing real or retrieved footage requires source-grounded decisions and executable timecodes.

Editing agents increasingly support hierarchical media compilation (Li et al., 2026) and long-horizon, music-synchronized execution (Zhao et al., 2026). Practical systems must also expose why footage was selected, how operations changed the timeline, and where failures occurred. Without such evidence, users cannot inspect intermediate decisions, revise local segments, or resume interrupted runs. We therefore formulate prompt-driven long-form editing as a *traceable* agent workflow whose production state is observable and reusable.

We present **Crayotter**, an open-source multimodal multi-agent demo system. Unlike pipelines that assume fixed inputs or synthetic assets, Crayotter automatically searches multiple video sources when local material is insufficient. It constructs a coverage-verified material pool, derives a time-grounded blueprint, and executes it through registered timeline tools. Retrieval reports, analyses, plans, tool events, and intermediate renders remain first-class artifacts. Building on artifact-centered research and presentation systems (Yang and Weng, 2025; Zheng et al., 2026), the workbench exposes them for inspection, checkpointed resumption, and localized re-execution, as illustrated in Figure 1.

Across 23 editing themes, Crayotter obtains the highest aggregate scores under both human and GPT-5.4 evaluation, including a human score of 3.40/5. Its resource-aware asynchronous runtime further improves end-to-end throughput by approximately $1.6\times$ over the earlier unoptimized scheduling path.

Our contributions are three-fold:

[*]Core Contributor. [†] Project Lead.
[‡]Our code is publicly available at https://github.com/idwts/Crayotter and our live demo at https://idwts.github.io/Crayotter/#demo.

1

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1–8
November 6-10, 2026 ©2026 Association for Computational Linguistics

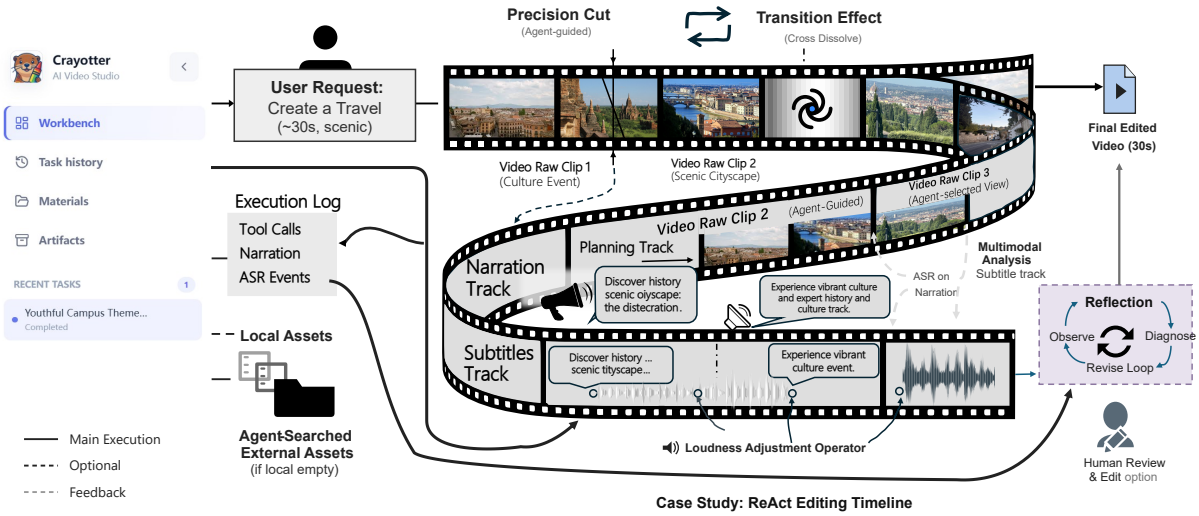


Figure 1: Crayotter workbench and an editing trajectory. From left to right: task and asset specification; tool-grounded timeline construction for a travel-video request; and artifact-based review and revision before export.

1. We formulate prompt-driven long-form video editing as a traceable workflow whose source evidence, plans, timeline states, renders, and logs support inspection and localized repair.
2. We develop a three-phase multi-agent system that integrates coverage-aware retrieval, time-grounded planning, and tool-grounded execution within an observable and resumable workbench.
3. We release the code, live demo, and execution traces; evaluate the system across 23 themes with human and AI judges; and introduce a resource-aware runtime for long-running production.

2 Related Work

Agentic visual production. VideoDirectorGPT decomposes multi-scene generation through LLM-based planning (Lin et al., 2023). DreamFactory and MovieAgent extend role-based coordination to long-form narrative production (Xie et al., 2024; Wu et al., 2025a). MM-StoryAgent integrates text, image, and audio for narrated storybooks (Xu et al., 2025a), whereas FilmAgent simulates specialized filmmaking roles in virtual 3D environments (Xu et al., 2025b). AniMaker further introduces MCTS-based clip exploration for animated storytelling (Shi et al., 2025). These systems establish effective agent specialization for content synthesis; Crayotter instead coordinates post-production decisions over heterogeneous source footage and preserves the corresponding executable tool trajectory.

Long-form video generation and editing. FIFO-Diffusion extends video duration through an autoregressive diffusion process (Kim et al., 2024), while Long Context Tuning adapts video models to longer temporal contexts (Guo et al., 2025). MinT targets explicit multi-event temporal control (Wu et al., 2025b), and test-time training supports minute-scale generation (Dalal et al., 2025). Story-visualization research studies multi-subject consistency and iterative refinement (He et al., 2025; Mao et al., 2026); complementary benchmarks evaluate story-level consistency (Zhuang et al., 2026) and generated-video quality (Liu et al., 2024). Closer to our setting, DIRECT performs hierarchical, intent-guided mashup creation (Li et al., 2026), CineAgents studies instruction-driven cinematic compilation (Zhang et al., 2026), and CutClaw targets music-synchronized hours-long editing (Zhao et al., 2026). FireRed-OpenStoryline combines conversational media search, planning, tool orchestration, and human control,[1] while NarratoAI automates script, narration, and video assembly for film commentary.[2] Crayotter complements these systems with coverage evidence, time-grounded blueprints, and localized failure traces throughout execution.

Intervenable agent workflows. ResearStudio externalizes plans, file changes, and tool activity to support real-time user intervention (Yang and Weng, 2025). DeepPresenter grounds iterative re-

[1] https://github.com/FireRedTeam/FireRed-OpenStoryline
[2] https://github.com/linyqh/NarratoAI

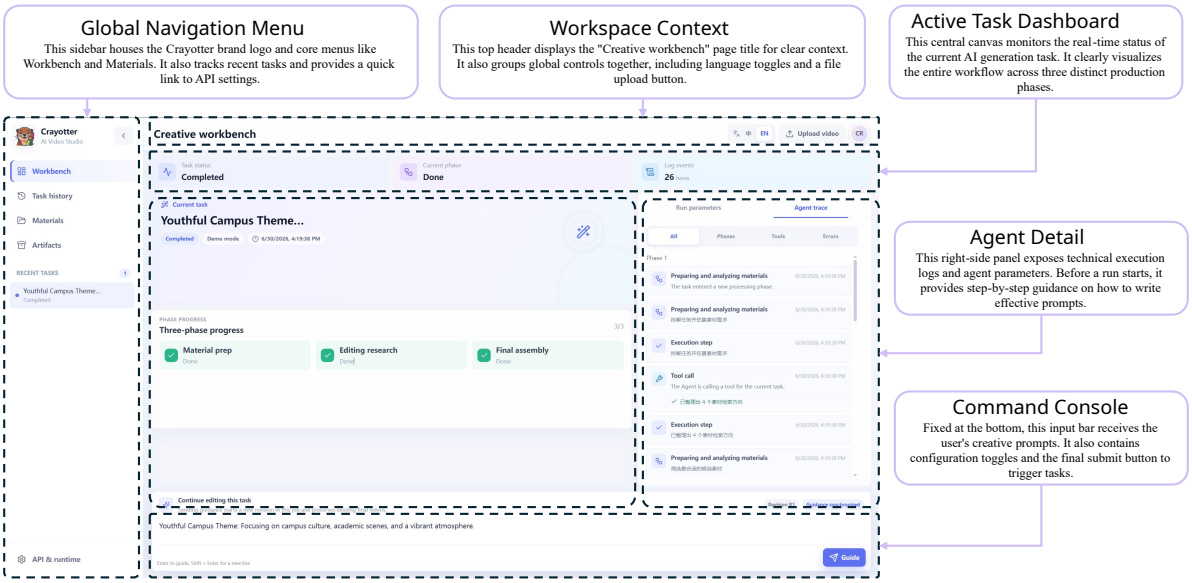


Figure 2: Crayotter workbench interface. The workspace exposes global navigation, task context and status, agent details, command logs, and intermediate artifacts for monitoring and inspecting long-running editing workflows.

finement in rendered presentation artifacts (Zheng et al., 2026). Crayotter transfers this artifact-centered principle to temporal video editing, where users must inspect clip evidence, time-grounded plans, timeline operations, and render diagnostics. Its contribution therefore lies not only in the final video, but also in an executable production trajectory that can be audited, resumed, and selectively repaired.

3 System Design: Crayotter Demo Architecture

The complete system architecture and artifact flow are illustrated in Appendix A (Figure 5). Crayotter maps a user request q and optional local assets $\mathcal{A}_0$ to an exported video y through three artifact-producing phases. Phase 1 constructs a coverage-verified material pool $\mathcal{P} = (\mathcal{S}, \Gamma, \mathcal{M})$, where $\mathcal{S}$ contains selected videos, Γ records coverage evidence, and $\mathcal{M}$ stores multimodal analyses. Phase 2 converts $\mathcal{P}$ into a time-grounded editing blueprint $\mathcal{B}$, and Phase 3 realizes that blueprint as an observable execution history $\mathcal{H}$ and final render.

$$(q, \mathcal{A}_0) \xrightarrow{\text{Phase 1}} \mathcal{P} \xrightarrow{\text{Phase 2}} \mathcal{B} \xrightarrow{\text{Phase 3}} (\mathcal{H}, y).$$

The decomposition exposes the information required for inspection and repair rather than retaining it in a latent agent state. Specifically, $\mathcal{B} = \{b_m\}_{m=1}^{M_z}$ contains shot-level decisions, while $\mathcal{H} = \{(s_k, a_k, s_{k+1}, \zeta_k)\}_{k=1}^{K}$ records each tool action and its artifact-level diagnostic ζ_k.

Figure 2 shows the corresponding workbench. Users may upload source clips or invoke agent-assisted retrieval, submit an editing brief, monitor phase-level progress, inspect intermediate artifacts, and access the final export within a single workspace. Interrupted jobs retain validated artifacts and scheduler checkpoints, enabling execution to resume without reconstructing the complete trajectory.

3.1 Phase 1: Coverage-Aware Multimodal Footage Retrieval

Effective timeline planning requires source material that covers the requested scenes, actions, narrative functions, shot types, and styles. Phase 1 therefore expands q into weighted production tags $T = \{(t_i, w_i)\}$ and retrieves a high-recall candidate set. For each candidate v, the multimodal analyzer estimates tag support $g(v, t_i)$ from sampled frames and temporal windows. Let $G_{\mathcal{S}}(t_i) = \max_{v \in \mathcal{S}} g(v, t_i)$ denote the evidence already supplied by the selected pool. Candidate selection combines request relevance with the marginal evidence it contributes:

$$\begin{aligned} D_i(v, \mathcal{S}) &= \max(0, g(v, t_i) - G_{\mathcal{S}}(t_i)), \\ \text{score}(v \mid \mathcal{S}, T) &= \lambda s_{\text{rel}}(q, v) \\ &\quad + (1 - \lambda) \sum_i w_i D_i(v, \mathcal{S}), \\ \text{Cov}(\mathcal{S}, T) &= \frac{\sum_i w_i \mathbb{I}[G_{\mathcal{S}}(t_i) \geq \eta_i]}{\sum_i w_i}. \end{aligned}$$

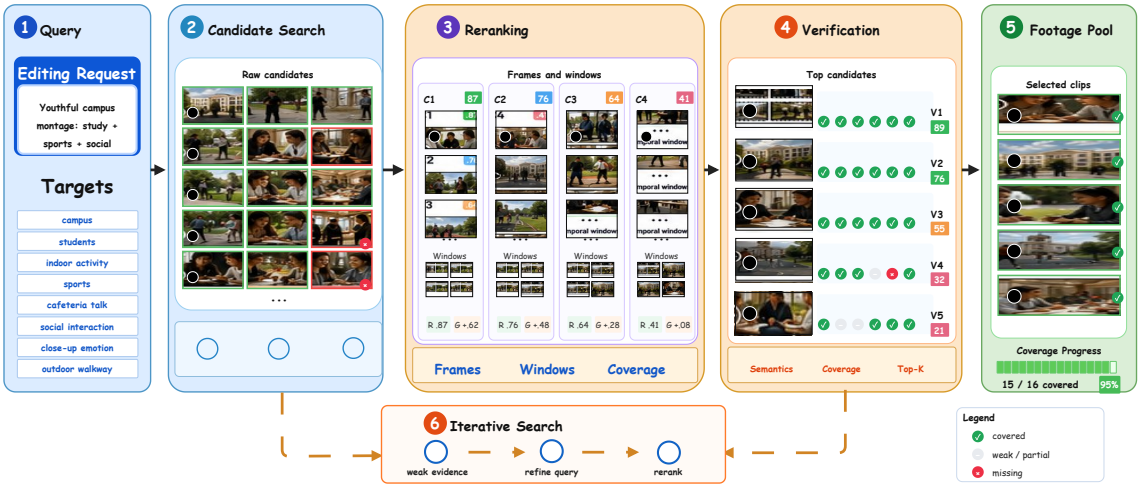


Figure 3: Coverage-aware multimodal footage retrieval. Crayotter expands an editing request into production tags, retrieves a high-recall candidate pool, reranks candidates using frame- and temporal-window evidence, verifies top-ranked videos with full-video analysis, and issues follow-up queries for uncovered tags until coverage is sufficient or the retrieval budget is exhausted.

As illustrated in Figure 3, top-ranked videos undergo full-video verification before entering $\mathcal{S}$. Tags below their support thresholds form a coverage gap Δ_r that conditions the next retrieval query. The loop terminates when coverage is sufficient or the retrieval budget is exhausted, yielding selected assets $\mathcal{S}$, tag-level evidence Γ, and per-video analyses $\mathcal{M}$.

3.2 Phase 2: Deep Editing Research

Phase 2 performs planning without invoking editing tools. Its principal task is temporal localization: each narrative beat must be associated with both a source video and an executable in/out interval. Crayotter overlays human-readable temporal coordinates on sampled evidence, producing $\tilde{v}_j = R_{\text{time}}(v_j)$. This interface binds semantic observations to absolute source timestamps without modifying the underlying multimodal model.

For narrative beats $\mathcal{Z} = \{z_m\}_{m=1}^{M_z}$ derived from q and Γ, the research agent constructs a structured blueprint:

$$b_m = (v_{\hat{j}_m}, \hat{\tau}_m, c_m, \delta_m, n_m),$$
$$\mathcal{B} = \{b_m\}_{m=1}^{M_z}.$$

Here $\hat{\tau}_m$ is the selected interval, c_m specifies its cinematic role, δ_m encodes transition and pacing intent, and n_m specifies narration or subtitle intent. The resulting plan is temporally addressable rather than free-form: users and downstream tools can inspect, revise, or execute each beat independently.

3.3 Phase 3: Tool-Grounded Timeline Execution

Phase 3 translates $\mathcal{B}$ into concrete timeline operations. State s_k comprises the current timeline, media assets, narration and subtitle tracks, rendered previews, and tool logs; an action a_k applies a registered editing operator such as trimming, transition insertion, narration alignment, loudness normalization, or export. Execution and diagnosis are coupled as follows:

$$s_{k+1} = E(s_k, a_k; \mathcal{B}),$$
$$\zeta_k = D_{\text{tool}}(s_k, a_k, s_{k+1}, \mathcal{B}, q).$$

The diagnostic ζ_k summarizes observable properties including tool validity, timestamp accuracy, coverage preservation, narration alignment, transition continuity, and render quality. When a check fails, the agent grounds reflection in the affected artifact and re-executes the corresponding segment, rather than restarting the complete edit.

3.4 Agent Roles and Artifact Contracts

The implementation separates planner, researcher, and executor responsibilities. The planner derives phase goals and coverage requirements; the researcher synthesizes multimodal evidence into $\mathcal{B}$; and the executor converts blueprint entries into validated tool calls. Their interface is a stable artifact contract spanning input evidence, planning outputs, and execution products such as timelines, intermediate renders, and logs. Before narration, the merged video undergoes composite re-analysis so

Table 1: Tool taxonomy with functional grouping.

Category	Tools
Search	`search_video`
Ranking	`rank_video_candidates`
Download	`download_video`
Analysis	`analyze_video`, `recall_semantic_segments`
Cutting	`cut_video`, `batch_cut_video`
Merging	`merge_videos`
Inspection	`inspect_video_duration`
Transitions	`add_transition`, `plan_transition_timeline`, `list_transition_presets`
Continuity	`score_cut_continuity`, `recommend_transition_for_cut`
Timeline	`build_edit_timeline_from_segments`, `align_narration_to_timeline`, `validate_timeline_constraints`
Narration	`add_narration_segments`, `validate_narration_timeline`
Audio Post	`duck_background_audio`, `normalize_loudness`
Subtitles	`add_subtitles`
Export	`export_video`

that scripts are conditioned on rendered content rather than source analyses alone. This contract supports replay, phase-level diagnosis, and selective re-execution.

3.5 Modular Tool Taxonomy

Crayotter exposes 23 registered tools for material preparation and timeline execution. Table 1 presents the complete functional taxonomy.

The transition subsystem provides configurable basic, motion, and cinematic effects, while continuity tools estimate visual discrepancies around cut points and recommend an appropriate transition. Phase routing supports full, local-first, and direct-execution modes. These components share the same artifact contract, allowing their outputs and validation signals to participate in the reflection loop without altering the three-phase abstraction.

3.6 Asynchronous Scheduling and Demo Runtime

Long-horizon editing contains substantial parallelism but also shared-write constraints. Crayotter represents each phase as a resource-aware task DAG whose nodes declare dependencies, required resources, output artifacts, conflicts, and retry policies. The scheduler validates the graph, executes independent search, download, analysis, research, cutting, and narration tasks with bounded concurrency, and serializes operations that modify shared timelines or final outputs.

Resource limits are configurable from the workbench to accommodate different hardware profiles. Relative to the earlier unoptimized scheduling path, asynchronous execution improves throughput by approximately 1.6× by overlapping independent model, network, and media-processing work.

The scheduler checkpoints complete per-phase task state, dependency fingerprints, artifacts, and validity, enabling localized resume. It validates produced artifacts before reuse, while the workbench visualizes phase progress, events, and artifacts for diagnosis.

4 Evaluation

4.1 Evaluation Setup

We compare Crayotter with three baselines: CapCut-Mate, CutClaw, and ChatCut, a proprietary prompt-based browser editor.[3] ChatCut was run with proprietary GPT-5.5, while the open-source Crayotter used Qwen3.5-122B-A10B (Qwen Team, 2026). The benchmark includes prompt metadata and final videos for all methods and, for Crayotter, configuration files and tool traces for process-level inspection. The themes span five scenario families: pet (3), campus (5), travel (5), scenery (5), and food (5). Each output is rated on a five-point scale for theme alignment, content richness, narrative coherence, editing smoothness, and visual quality; the overall score is the mean across dimensions. Each run records phase durations, state transitions, tool calls, checkpoint/artifact status, and export duration; released traces target 60 s and export 50.2–59.0 s.

4.2 Scoring Protocol

We report both human judgments and GPT-5.4 judge scores. For GPT-5.4 scoring, we use the same multidimensional rubric across all 23 themes and all methods. For human scoring, three annotators independently score the same outputs under blinded conditions; we first average their scores for each theme–method pair, then average across the 23 themes. This case-level aggregation avoids giving extra weight to any annotator. H2 gave CutClaw 1.0 on all 23 cases, reflecting that annotator's judgment that its editing quality was consistently poor; H1 and H3 provided varying scores. Inter-annotator reliability is measured using a two-way random-effects, absolute-agreement intraclass correlation coefficient for the mean of three annotators, ICC(A,3) (Koo and Li, 2016). Across the 92 case–method outputs, the overall-score ICC is 0.86 (95% cluster-bootstrap CI: 0.78–0.91), while dimension-level ICCs range from 0.77 to 0.83. Table 2 reports

[3] https://chatcut.io

Table 2: Aggregate evaluation on 23 themes. Dimension columns report mean scores on a 1–5 scale; Overall reports mean ± standard deviation across themes. GPT-5.4 uses the same rubric as human evaluation.

Judge	Method	Theme	Richness	Narrative	Smoothness	Visual	Overall
Human	**Crayotter**	**3.59**	**3.36**	**3.22**	**3.29**	**3.54**	**3.40** ± 0.59
Human	ChatCut	3.42	3.22	3.03	3.14	3.35	3.23 ± 0.75
Human	CapCut-Mate	2.59	2.71	2.01	2.13	2.74	2.44 ± 0.72
Human	CutClaw	1.59	1.64	1.72	1.86	1.67	1.70 ± 0.62
GPT-5.4	**Crayotter**	**2.78**	2.61	**2.13**	**2.17**	2.26	**2.39** ± 0.62
GPT-5.4	ChatCut	2.39	**2.70**	1.91	1.87	**2.52**	2.28 ± 0.80
GPT-5.4	CapCut-Mate	2.09	2.57	1.57	1.74	**2.52**	2.10 ± 0.64
GPT-5.4	CutClaw	1.57	1.48	1.52	1.65	1.65	1.57 ± 0.40

aggregate results; Appendix B provides case-level human and GPT-5.4 scores for Crayotter, CapCut-Mate, CutClaw, and ChatCut.

4.3 Results

Table 2 places Crayotter first overall under both human and GPT-5.4 evaluation. Human raters rank it highest across all dimensions, with ChatCut closest overall. GPT-5.4 ranks ChatCut first in content richness and tied in visual quality, while Crayotter leads the other dimensions and overall. These results compare end-to-end systems rather than controlled backbone substitutions. Figure 4 illustrates a campus case: Crayotter follows the requested progression and exposes its assets–inspect–select–edit–verify trace, whereas the displayed baselines contain partial or drifted segments.

5 Discussion

The evaluation traces complement final-video scores by showing where long-horizon execution succeeds or fails. All 23 Crayotter runs completed under the evaluation protocol, with an average of 221 observable events spanning source analysis, semantic recall, cutting, timeline construction, narration and subtitle alignment, audio processing, and export. These artifacts localize failures that aggregate scores cannot explain, including damaged source intervals, missing requested events, vertical-material leakage, and weak continuity.

Figure 1 connects these diagnostics to the user-facing workbench. Assets, analysis outputs, timeline operations, intermediate renders, and tool events remain inspectable throughout execution, allowing users to trace a failure to the responsible stage and rerun only the affected work. This observe–diagnose–revise loop makes Crayotter a controllable production workspace rather than a prompt-only generator. Released traces target 60 s; multi-minute scale and storage/preview overhead remain unmeasured.

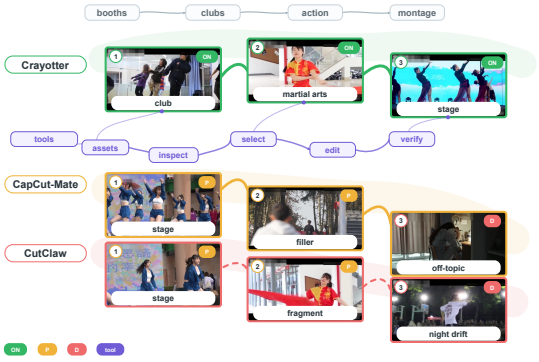


Figure 4: Case-level output and tool-trajectory comparison for the campus club fair scenario. The top path gives the target requirements, each row shows the method's output trajectory, and the purple stream marks Crayotter's replayable tool artifacts. ON/P/D denote aligned, partial, and drifted frames.

6 Conclusion

We presented Crayotter, an open-source multimodal multi-agent demo system that makes long-form video editing traceable through coverage-aware material preparation, artifact-grounded research, tool-grounded execution, and reflection. A 23-theme comparison with CapCut-Mate, CutClaw, and ChatCut under human and AI evaluation demonstrates the practical value of this workflow, while resource-aware asynchronous scheduling improves demo usability through concurrent execution, checkpoints, and artifact validation. Crayotter thus turns opaque one-shot generators into inspectable, controllable workspaces.

Ethics Statement

Media shown in the paper and used for evaluation were manually curated from public-domain, openly licensed, or otherwise reuse-authorized sources; private or restricted content was excluded. Public accessibility was not treated as authorization, and users remain responsible for the licensing, attribution, privacy, and permitted use of uploaded or retrieved media.

Three human annotators evaluated generated videos with the same five-dimensional rubric; no demographic or sensitive attributes were collected, and ratings are reported in aggregate or under anonymous identifiers. GPT-5.4 was applied uniformly as an auxiliary judge, with its results reported separately from human judgments rather than treated as ground truth. Because automated retrieval and editing may raise copyright, privacy, or misuse risks, users should verify permissions and review all retrieved assets and generated edits before publication.

References

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. 2025. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17702–17711.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. 2025. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17281–17291.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. 2025. DreamStory: Open-domain story visualization by LLM-guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. 2024. FIFO-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37:89834–89868.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Ke Li, Maoliang Li, Jialiang Chen, Jiayu Chen, Zihao Zheng, Shaoqi Wang, and Xiang Chen. 2026. DIRECT: Video mashup creation via hierarchical multi-agent planning and intent-guided editing. *arXiv preprint arXiv:2604.04875*.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. 2023. VideoDirectorGPT: Consistent multi-scene video generation via LLM-guided planning. *arXiv preprint arXiv:2309.15091*.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22139–22149.

Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, Zeyu Zheng, Zirui Wang, Cihang Xie, and Yuyin Zhou. 2026. Story-Iter: A training-free iterative paradigm for long story visualization. In *The Fourteenth International Conference on Learning Representations*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. 2025. AniMaker: Multi-agent animated storytelling with MCTS-driven clip generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *First Conference on Language Modeling*.

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

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. 2025b. Mind the time: Temporally-controlled multi-event video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23989–24000.

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*.

Xuenan Xu, Jiahao Mei, Chenliang Li, Yuning Wu, Ming Yan, Shaopeng Lai, Ji Zhang, and Mengyue Wu. 2025a. MM-StoryAgent: Immersive narrated storybook video generation with a multi-agent paradigm across text, image and audio. *arXiv preprint arXiv:2503.05242*.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. 2025b. FilmAgent: A multi-agent framework for end-to-end film automation in virtual 3D spaces. *arXiv preprint arXiv:2501.12909*.

Linyi Yang and Yixuan Weng. 2025. ResearStudio: A human-intervenable framework for building controllable deep research agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 896–905.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Peixuan Zhang, Chang Zhou, Ziyuan Zhang, Hualuo Liu, Chunjie Zhang, Jingqi Liu, Xiaohui Zhou, Xi Chen, Shuchen Weng, Si Li, and Boxin Shi. 2026. A benchmark and multi-agent system for instruction-driven cinematic video compilation. *arXiv preprint arXiv:2604.10456*.

Shifang Zhao, Yihan Hu, Ying Shan, Yunchao Wei, and Xiaodong Cun. 2026. CutClaw: Agentic hours-long video editing via music synchronization. *Preprint*, arXiv:2603.29664.

Hao Zheng, Guozhao Mo, Xinru Yan, Qianhao Yuan, Wenkai Zhang, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2026. DeepPresenter: Environment-grounded reflection for agentic presentation generation. *arXiv preprint arXiv:2602.22839*.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, and 1 others. 2026. Vistorybench: Comprehensive benchmark suite for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9455–9467.

A Extended System Architecture

Figure 5 details the full system architecture and artifact flow across the three editing phases and observable runtime.

B Case-Level Evaluation Scores

Tables 3 and 4 report case-level scores for all 23 themes and all four systems. Human overall scores average the five dimensions for each annotator–case–method tuple; GPT-5.4 scores report the five dimensions and their mean.

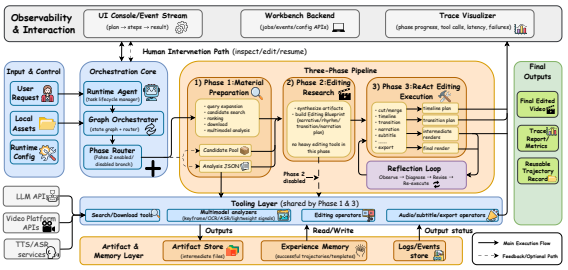


Figure 5: Full Crayotter architecture, including the three-phase pipeline, observable artifacts, runtime services, and reusable trajectory logs.

Table 3: Case-level human overall scores. H1–H3 denote the annotators; Mean averages their scores for each case–method pair.

Case	Theme	Crayotter				CapCut-Mate				CutClaw				ChatCut			
		H1	H2	H3	Mean	H1	H2	H3	Mean	H1	H2	H3	Mean	H1	H2	H3	Mean
001	Funny pet fails	2.4	3.2	2.8	2.80	1.6	2.0	1.8	1.80	1.4	1.0	1.6	1.33	1.8	2.6	1.8	2.07
002	Lazy cat afternoon	3.8	3.8	4.0	3.87	2.0	1.6	1.8	1.80	3.0	1.0	3.4	2.47	2.2	3.8	2.6	2.87
003	Human-pet bonding	3.0	3.4	3.6	3.33	1.2	1.0	1.2	1.13	2.0	1.0	2.2	1.73	1.0	1.0	1.4	1.13
004	Library white noise	4.0	3.8	4.2	4.00	1.8	1.0	2.0	1.60	2.6	1.0	2.6	2.07	3.2	4.6	3.4	3.73
005	Campus sports	3.4	3.0	3.4	3.27	3.2	1.6	3.2	2.67	2.4	1.0	2.8	2.07	3.2	2.4	3.6	3.07
006	Graduation farewell	3.6	4.2	3.6	3.80	3.8	2.4	4.0	3.40	2.0	1.0	2.2	1.73	3.2	3.0	3.4	3.20
007	Club fair	4.8	3.6	4.8	4.40	2.6	1.0	2.8	2.13	3.0	1.0	3.2	2.40	2.4	2.8	2.4	2.53
008	Campus canteen	3.8	4.0	3.6	3.80	2.2	1.0	2.6	1.93	1.8	1.0	1.6	1.47	2.6	2.6	2.4	2.53
009	Road trip	3.2	2.8	3.6	3.20	2.8	2.2	2.8	2.60	1.2	1.0	1.2	1.13	3.2	2.6	3.4	3.07
010	City walk	4.6	4.0	4.0	4.20	4.0	1.0	4.0	3.00	1.2	1.0	1.2	1.13	3.8	3.6	3.4	3.60
011	Island holiday	4.2	2.8	4.2	3.73	3.6	2.4	3.4	3.13	1.8	1.0	1.8	1.53	3.4	3.4	3.6	3.47
012	Mountain trek	2.6	3.0	3.0	2.87	4.4	1.0	4.6	3.33	3.8	1.0	4.0	2.93	3.8	3.2	4.0	3.67
013	Ancient-town visit	3.0	3.6	3.6	3.40	3.6	2.4	3.6	3.20	1.4	1.0	1.4	1.27	5.0	4.0	4.8	4.60
014	Yili grassland	2.2	3.6	2.6	2.80	2.6	2.0	2.8	2.47	3.0	1.0	3.0	2.33	4.4	3.8	4.2	4.13
015	Jiuzhaigou waters	2.2	1.0	3.0	2.07	2.8	3.0	2.8	2.87	3.2	1.0	3.4	2.53	3.6	4.0	3.4	3.67
016	Huangshan clouds	3.8	4.4	3.8	4.00	4.0	3.2	3.8	3.67	1.6	1.0	1.4	1.33	4.4	4.4	4.0	4.27
017	Guilin landscape	4.4	4.2	4.4	4.33	3.2	3.2	3.0	3.13	3.8	1.0	4.0	2.93	3.6	4.2	3.6	3.80
018	Meili sunrise	3.2	3.4	3.8	3.47	3.2	2.6	3.0	2.93	1.0	1.0	1.2	1.07	3.6	3.0	3.2	3.27
019	Sichuan hot pot	3.2	3.0	3.2	3.13	2.4	1.4	2.6	2.13	1.0	1.0	1.0	1.00	4.0	4.0	3.8	3.93
020	Cantonese morning tea	3.0	3.2	3.8	3.33	1.8	2.8	1.6	2.07	1.0	1.0	1.2	1.07	3.2	3.0	3.4	3.20
021	Beijing roast duck	2.2	3.0	2.6	2.60	2.0	3.0	2.2	2.40	1.0	1.0	1.0	1.00	3.0	2.6	3.0	2.87
022	Xinjiang barbecue	2.6	2.8	3.2	2.87	1.0	2.0	1.0	1.33	1.6	1.0	1.6	1.40	3.0	2.2	3.4	2.87
023	Jiangnan pastries	2.6	3.4	2.8	2.93	1.0	1.8	1.2	1.33	1.0	1.0	1.2	1.07	2.6	3.2	2.6	2.80
Mean	All cases	3.30	3.36	3.55	3.40	2.64	1.98	2.69	2.44	1.99	1.00	2.10	1.70	3.23	3.22	3.25	3.23

Table 4: Case-level GPT-5.4 scores. TA, CR, NC, ES, VQ, and O denote theme alignment, content richness, narrative coherence, editing smoothness, visual quality, and their mean, respectively.

Case	Theme	Crayotter						CapCut-Mate						CutClaw						ChatCut					
		TA	CR	NC	ES	VQ	O	TA	CR	NC	ES	VQ	O	TA	CR	NC	ES	VQ	O	TA	CR	NC	ES	VQ	O
001	Funny pet fails	2.0	3.0	2.0	2.0	1.0	**2.0**	1.0	2.0	1.0	2.0	2.0	**1.6**	2.0	1.0	1.0	2.0	1.0	**1.4**	2.0	4.0	2.0	2.0	2.0	**2.4**
002	Lazy cat afternoon	3.0	3.0	2.0	2.0	3.0	**2.6**	2.0	3.0	1.0	1.0	2.0	**1.8**	1.0	2.0	1.0	1.0	1.0	**1.2**	2.0	2.0	2.0	2.0	3.0	**2.2**
003	Human-pet bonding	2.0	2.0	1.0	2.0	2.0	**1.8**	1.0	1.0	1.0	1.0	2.0	**1.2**	1.0	1.0	1.0	2.0	2.0	**1.4**	1.0	1.0	1.0	1.0	2.0	**1.2**
004	Library white noise	3.0	2.0	2.0	2.0	2.0	**2.2**	1.0	2.0	1.0	1.0	2.0	**1.4**	2.0	1.0	2.0	2.0	2.0	**1.8**	2.0	3.0	2.0	2.0	2.0	**2.2**
005	Campus sports	2.0	2.0	2.0	2.0	1.0	**1.8**	2.0	2.0	1.0	1.0	2.0	**1.6**	2.0	1.0	2.0	2.0	1.0	**1.6**	2.0	2.0	2.0	2.0	2.0	**2.0**
006	Graduation farewell	3.0	2.0	2.0	2.0	2.0	**2.2**	3.0	3.0	2.0	2.0	2.0	**2.4**	2.0	1.0	2.0	2.0	1.0	**1.6**	2.0	3.0	2.0	2.0	2.0	**2.2**
007	Club fair	4.0	5.0	3.0	3.0	4.0	**3.8**	2.0	2.0	1.0	1.0	2.0	**1.6**	2.0	3.0	2.0	2.0	3.0	**2.4**	1.0	1.0	1.0	1.0	2.0	**1.2**
008	Campus canteen	2.0	2.0	2.0	1.0	1.0	**1.8**	2.0	2.0	1.0	1.0	1.0	**1.6**	2.0	2.0	2.0	2.0	1.0	**1.6**	2.0	2.0	1.0	1.0	2.0	**1.6**
009	Road trip	3.0	3.0	2.0	2.0	3.0	**2.6**	2.0	3.0	2.0	2.0	3.0	**2.4**	2.0	2.0	1.0	2.0	2.0	**1.8**	2.0	2.0	1.0	1.0	2.0	**1.6**
010	City walk	4.0	3.0	2.0	2.0	2.0	**2.6**	3.0	4.0	2.0	3.0	4.0	**3.2**	1.0	1.0	1.0	1.0	2.0	**1.2**	3.0	4.0	2.0	2.0	4.0	**3.0**
011	Island holiday	3.0	2.0	2.0	3.0	3.0	**2.6**	2.0	3.0	2.0	2.0	3.0	**2.4**	2.0	1.0	2.0	2.0	2.0	**1.8**	2.0	3.0	2.0	1.0	3.0	**2.2**
012	Mountain trek	2.0	2.0	2.0	2.0	2.0	**2.0**	1.0	2.0	2.0	2.0	2.0	**1.8**	1.0	1.0	1.0	1.0	2.0	**1.2**	2.0	2.0	2.0	3.0	3.0	**2.4**
013	Ancient-town visit	3.0	3.0	2.0	2.0	2.0	**2.4**	4.0	4.0	3.0	3.0	3.0	**3.4**	2.0	2.0	2.0	2.0	2.0	**2.0**	3.0	3.0	2.0	2.0	2.0	**2.4**
014	Yili grassland	4.0	3.0	3.0	4.0	4.0	**3.6**	3.0	3.0	2.0	2.0	3.0	**2.6**	2.0	3.0	2.0	2.0	2.0	**2.2**	4.0	4.0	3.0	3.0	4.0	**3.6**
015	Jiuzhaigou waters	2.0	2.0	2.0	3.0	2.0	**2.2**	2.0	2.0	1.0	2.0	3.0	**2.0**	2.0	2.0	1.0	1.0	1.0	**1.4**	4.0	4.0	3.0	3.0	4.0	**3.6**
016	Huangshan clouds	4.0	4.0	3.0	2.0	3.0	**3.2**	4.0	4.0	3.0	3.0	3.0	**3.4**	1.0	1.0	1.0	1.0	1.0	**1.0**	4.0	4.0	3.0	3.0	4.0	**3.6**
017	Guilin landscape	4.0	3.0	3.0	3.0	4.0	**3.4**	3.0	3.0	2.0	2.0	3.0	**2.6**	2.0	3.0	2.0	2.0	3.0	**2.4**	4.0	3.0	3.0	3.0	3.0	**3.2**
018	Meili sunrise	2.0	2.0	2.0	2.0	2.0	**2.0**	3.0	2.0	2.0	2.0	3.0	**2.4**	1.0	1.0	1.0	2.0	2.0	**1.4**	3.0	3.0	2.0	2.0	3.0	**2.6**
019	Sichuan hot pot	3.0	3.0	2.0	2.0	3.0	**2.6**	2.0	3.0	1.0	2.0	3.0	**2.2**	1.0	1.0	1.0	2.0	2.0	**1.2**	4.0	4.0	3.0	2.0	3.0	**3.2**
020	Cantonese morning tea	2.0	2.0	2.0	1.0	1.0	**1.6**	1.0	2.0	1.0	1.0	2.0	**1.4**	1.0	1.0	2.0	1.0	1.0	**1.2**	1.0	2.0	1.0	1.0	1.0	**1.2**
021	Beijing roast duck	2.0	2.0	1.0	1.0	1.0	**1.4**	1.0	2.0	1.0	1.0	2.0	**1.4**	1.0	1.0	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0	1.0	1.0	**1.0**
022	Xinjiang barbecue	2.0	2.0	2.0	2.0	2.0	**2.0**	2.0	3.0	2.0	2.0	3.0	**2.4**	2.0	2.0	2.0	2.0	1.0	**1.8**	2.0	2.0	1.0	1.0	2.0	**1.6**
023	Jiangnan pastries	3.0	3.0	3.0	2.0	2.0	**2.6**	1.0	2.0	1.0	1.0	2.0	**1.4**	1.0	1.0	2.0	2.0	2.0	**1.6**	2.0	3.0	2.0	2.0	2.0	**2.2**
Mean	All cases	2.78	2.61	2.13	2.17	2.26	**2.39**	2.09	2.57	1.57	1.74	2.52	**2.10**	1.57	1.48	1.52	1.65	1.65	**1.57**	2.39	2.70	1.91	1.87	2.52	**2.28**