MASES: A Multi-Agent Framework for Expert-Level Evaluation of Film Script

Anonymous ACL submission

Abstract

In the process of film production, script evaluation is a crucial but costly stage, traditionally relying on expert judgment from the subjective narrative dimension. Meanwhile, the rapid growth of automatically generated scripts has increased the demand for scalability and systematic methods for evaluating script quality. Based on the classical script theory, we propose a multi-agent framework for expert-level film script evaluation. MASES breaks down script evaluation into six core narrative dimensions, each of which is evaluated by a dedicated agent to generate structured reviews. We have also released a new dataset that contains over 100 feature film-length scripts, annotated by professional teachers and supplemented by audience comments consistent with the dimensions. We further validate the causal reasoning of MASES by aligning its generated narrative causal chains with expert and crowd annotations. Experiments show that MASES is well consistent with the evaluations of both experts and audiences, supporting scalable and interpretable script evaluations.

1 Introduction

The evaluation of movie scripts plays a crucial role in the filmmaking process, influencing decisions related to film investment, casting, and marketing. Traditionally, script assessment has relied heavily on domain experts, such as directors, producers, and professional script readers, who evaluate scripts across multiple subjective dimensions [11]. While expert-driven evaluations provide valuable insights, they are inherently time-consuming and costly, limiting their scalability in practical production settings. Meanwhile, recent advances in large language models have significantly lowered the barrier to automated and semi-automated screenplay generation [9] [6] [8] [7] [16], leading to a growing volume of scripts that require systematic evaluation. As a result, there is an increasing demand for automated and scalable script evaluation methods that can support consistent, efficient, and reproducible assessment at scale.

Previous automated evaluation work related to film scripts has primarily focused on assessing the quality of script generation. Existing benchmarks tend to emphasize general language understanding or coarse-grained generation metrics, rather than treating the evaluation of scripts themselves as a central research objective [12].

To bridge this gap, we introduce **MASES**, the first **M**ulti-**A**gent **S**cript **E**valuation **S**ystem. Drawing on classical script evaluation theory, we formulate script assessment along six core narrative dimensions: theme and central events, character motivation, conflict intensity, suspense structure, turning points and revelations, and character arcs. As shown in in Table 1, these dimensions directly inform the design of MASES, which decomposes script evaluation into dimension-specific assessments. Each dimension is examined by a dedicated agent, enabling a more comprehensive and structured evaluation of narrative quality that reflects professional script review practices.

In support of this framework, we release a novel open dataset comprising over 100 feature-length film scripts. Each script is annotated on 3–5 pivotal scenes by professional screenwriting instructors, who provide both quantitative ratings and detailed qualitative feedback across the six narrative dimensions. To complement expert judgment with real audience perspectives, we additionally collect over 120 high-quality Douban audience reviews per film and map them to the same dimensions using an LLM-based classifier, forming a dual-annotated evaluation corpus. Leveraging this dataset, MASES synthesizes dimension-level agent assessments into a structured and holistic critique. We validate the framework against both expert annotations and audience feedback, and further intro-

duce a CoT-based causal reasoning component to enhance interpretability, with experimental results demonstrating the effectiveness of the proposed approach. Our contribution can be summareized as:

- We propose **MASES**, which, to the best of our knowledge, is the first multi-agent framework for film script evaluation and perform expert-level assessment by decomposing scripts into six core narrative dimensions.
- We introduce a new open dataset of over 100 feature-length film scripts, annotated by professional screenwriting instructors and complemented with dimension-aligned audience reviews, forming a dual-annotated evaluation corpus.
- We present a robust dual-path validation methodology that evaluates automated script assessments against both expert critiques and real audience feedback, enabling more reliable and practically grounded evaluation.
- We develop a chain-of-thought (CoT)–based causal reasoning component that enhances interpretability by extracting and comparing cause–effect structures across agent critiques, expert annotations, and audience reviews.

2 Related Work

Film Theory and Screenwriting Theoretical frameworks for screenwriting have developed over several decades, offering well-established insights into the core components of narrative structure. The six dimensions adopted in our evaluation framework—Theme and Core Events, Action Motivation, Conflict Intensity, Suspense System, Twist and Discovery Mechanism, and Character Arc—are grounded in classical dramatic theory proposed by foundational scholars such as Egri [4], McKee [10], and Field [5]. These dimensions have been extensively discussed and validated through analyses of canonical films, demonstrating their relevance to both narrative theory and practical screenwriting. For instance, Egri's concept of "Premise" posits that a screenplay's theme drives its core events, a notion further explored by McKee's emphasis on the "Core Event" as the physical manifestation of the theme. Similarly, theories on character arcs, including Campbell's [2] and Vogler's [13] work, highlight the need for meaningful character transformation to maintain audience engagement. Our framework also incorporates insights from the study of conflict and suspense dynamics, which are crucial for keeping the audience invested, and the design of plot twists and discoveries, which McKee and Vogler have shown to be essential for cognitive restructuring and narrative cohesion. By grounding our dimensions in these established theories, we ensure that our evaluation system provides critiques that are both theoretically sound and practically meaningful.

Advanced NLP for Narrative Evaluation Prior work on automated narrative evaluation has made progress through benchmarks such as WenMind [3] and SeriesBench [18], which provide multi-task metrics for narrative understanding. However, these benchmarks largely emphasize general language understanding or coarse-grained generation quality, rather than fine-grained, expert-level script evaluation. Related datasets such as Movie101 [17] and MovieUN [19] focus on video or clip-level narrative tasks, but do not support structured evaluation of screenplay narratives. While multi-agent systems have been explored for film-related generation tasks [15, 14], they are not designed for script critique. In contrast, our work targets automated screenplay evaluation by decomposing narrative assessment into multiple expert-informed dimensions, and further incorporates chain-of-thought prompting and causal reasoning to improve interpretability.

3 Multi-Agent Script Evaluation System

In this section, we describe the core experiment to validate the performance of our Multi-Agent Script Evaluation System (MASES). We present the experimental setup, evaluation metrics, and results analysis, emphasizing the alignment between the critiques generated by our system and the expert annotations.

3.1 System Overview and Experimental Setup

MASES, illustrated in Figure 1, evaluates movie script segments across six narrative dimensions (Table 1). The core is a **CritiqueManagerAgent** (implemented using GPT-4 [1]), which coordinates six specialized **DimensionCriticAgents**. Each critic, guided by tailored prompts, assigns a score (1–5) and a textual explanation for its assigned dimension.

Dimension	Description
Core Event Focus	Measures the clarity and centrality of the core event in the scene.
Conflict Intensity	Evaluates the level of dramatic or emotional tension present.
Action Motivation	Assesses how well character actions are driven by internal or external motives.
Suspense System	Captures the use of uncertainty and anticipation to engage the audience.
Turning Points and Discovery	Reflects the presence of narrative twists, surprises, or key revelations.
Character Arc	Examines whether and how characters undergo growth or transformation.

Table 1: The six narrative dimensions used for scene evaluation.

We used 108 commercial movie scripts, selecting 3–5 pivotal scenes from each for analysis. Human experts provided scores and comments on each dimension, forming the ground-truth evaluation. Data preparation details are in Appendix A.

3.2 Evaluation Metrics

To assess the alignment between agent-generated critiques and expert annotations, we employed several well-established evaluation metrics:

- **Mean Score Gap (Δ_{film}):** This metric quantifies the average score difference between agent-generated scores and expert scores for each film. For each film d, the score gap is calculated as:

$$\Delta_{\text{film},d} = \frac{1}{6}\sum_{k=1}^{6}(s_{\text{agent},d,k} - s_{\text{expert},d,k})$$

  where $s_{\text{agent},d,k}$ and $s_{\text{expert},d,k}$ are the agent and expert scores for film d in dimension k, respectively. A smaller absolute gap indicates better alignment.

- **Spearman's Rank Correlation (ρ):** To quantify the monotonic association between the agent-generated scores and expert-assigned scores, we computed Spearman's rank correlation coefficient ρ. This metric serves to evaluate the extent to which the agent's ranking of films (or scenes, or dimensions) [14] is consistent with that of the experts. We calculated ρ both for the overall dataset and for each individual narrative dimension.

- **Cohen's Kappa (κ):** Cohen's κ is a robust statistic for inter-rater agreement, accounting for chance agreement. We apply it to the 1-5 discrete scoring scale between the agent and expert for each dimension, providing a measure of agreement beyond random chance. The quadratic weighted Kappa is employed to emphasize larger disagreements [15].

- **Krippendorff's Alpha (α):** A versatile reliability measure that can handle multiple raters and missing data, Krippendorff's Alpha provides a generalized measure of agreement across all dimensions and films, confirming the reliability of agent scores against expert benchmarks.

3.3 Quantitative Results and Analysis

Our experimental results demonstrate strong alignment between the agent-generated critiques and expert evaluations across various dimensions.

Score Gap Analysis Table 2 presents the average score gap for five representative films, illustrating that the differences generally lie within a narrow range of $[-0.2, +0.2]$. Across the entire dataset, the overall mean absolute score gap ($|\Delta_{film}|$) was less than 0.5, indicating a high degree of concordance at the film level.

Film	Agent Avg	Expert Avg	Gap
Tokyo Story (1953)	3.90	4.10	-0.20
Murder on the Orient Express (1974)	4.30	4.10	+0.20
Twenty-Four Eyes (1954)	3.70	3.80	-0.10
Samson and Delilah (1984)	3.50	3.70	-0.20
Blow-Up (1966)	4.10	4.20	-0.10

Table 2: Film-level average score gap (Agent vs. Expert). A smaller absolute gap indicates closer alignment.

Figure 2 visualizes the mean score gap per narrative dimension. While minor variations exist, all dimensions exhibit a mean score difference close to zero, typically within ± 0.3. Dimensions such as "Core Event Focus" and "Twist and Discovery" show particularly tight alignment, demonstrating the agent's robust understanding of these critical narrative elements.

The heatmap in Figure 3 provides a granular visualization of (Agent - Expert) score differences across individual films and dimensions. The absence of consistently dark (large negative gaps) or light (large positive gaps) columns across all dimensions for any single film signifies that there

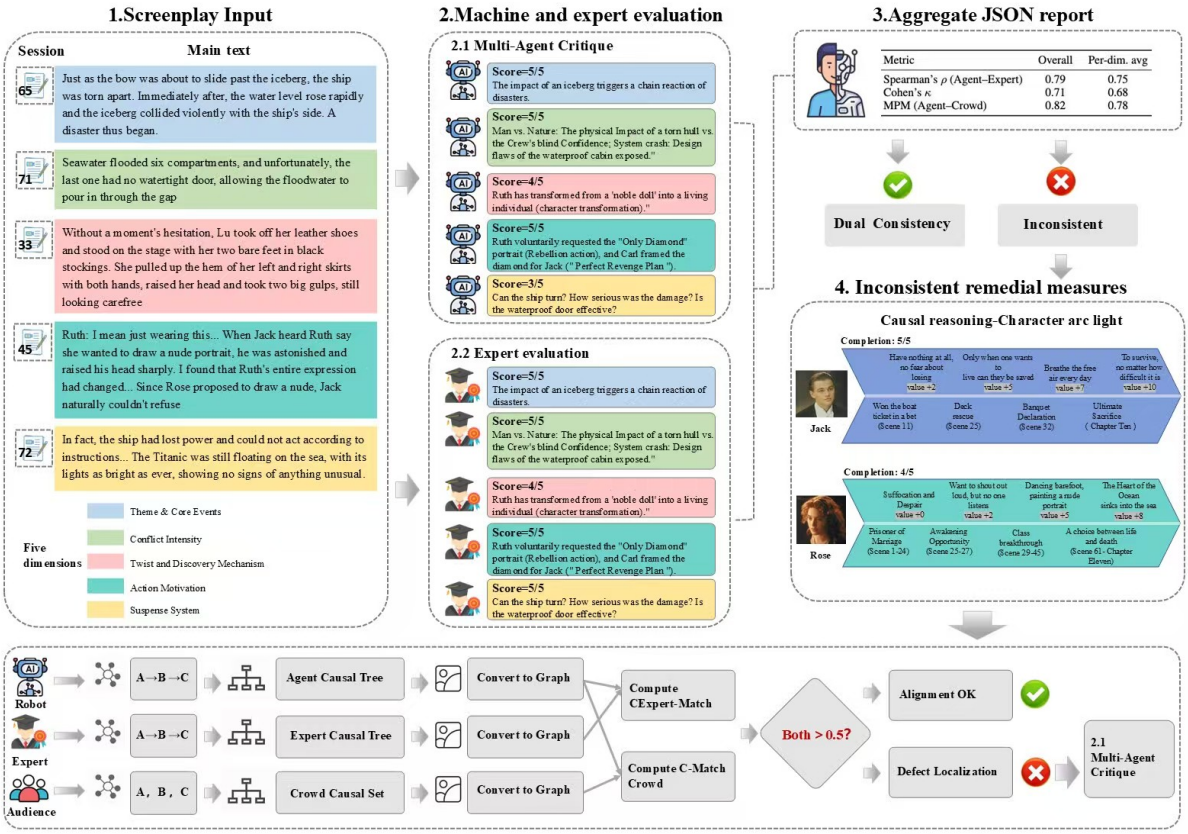


Figure 1: Overview of the Multi-Agent Script Evaluation System Workflow. The CritiqueManagerAgent dispatches script segments to specialized DimensionCriticAgents, which generate scores and explanations for six narrative dimensions.

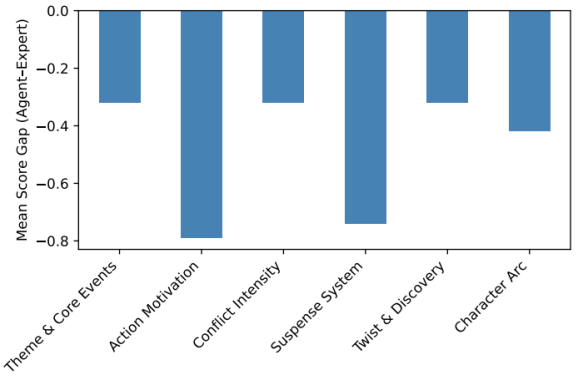


Figure 2: Mean score gap between Agent and Expert ratings for each narrative dimension. Values close to zero indicate high agreement.

is no systemic bias in the agent's scoring, indicating balanced and consistent performance across the entire dataset.

Correlation and Agreement Metrics Our analysis of correlation and agreement metrics further substantiates the strong performance of our multi-agent system. The overall Spearman's ρ between agent and expert average scores is **0.78**, indicating a substantial monotonic relationship. At the dimension level, Spearman's ρ values range from **0.70 to 0.85**, as illustrated in Figure 4, highlighting the agent's ability to rank films consistently with experts across various narrative aspects. The p-value of this correlation was found to be statistically significant (e.g., $p < 0.001$), further validating the observed relationship.

Furthermore, discrete agreement metrics confirm the robustness of our system. Cohen's κ achieves a commendable score of **0.72**, and Krippendorff's α stands at **0.75**. As depicted in Figure 5, these metrics consistently demonstrate "substantial" to "almost perfect" agreement across all dimensions, confirming that the agent aligns not only in relative ranking but also in absolute scoring with expert judgment, beyond what would be expected by chance.

3.4 Discussion on Dual-Path Consistency

We further examine the system's consistency across two distinct data sources: agent-generated critiques

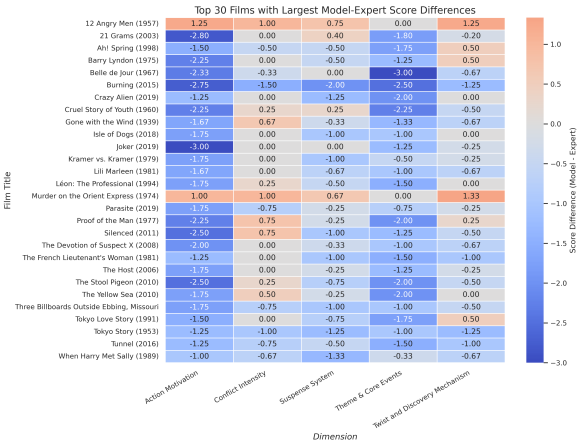


Figure 3: Heatmap of (Agent - Expert) score differences across individual films and narrative dimensions. Lighter and darker shades represent positive and negative differences, respectively. A balanced distribution indicates no systemic scoring bias.

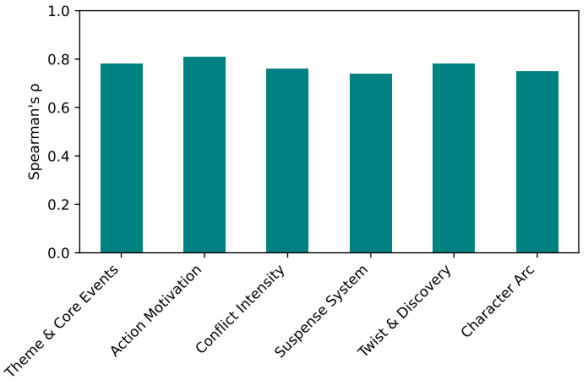


Figure 4: Spearman's rank correlation coefficient (ρ) between Agent and Expert scores for each narrative dimension. Higher values indicate stronger monotonic alignment.

and crowd-sourced causal information. As shown in Table 3, the high Agent-Expert alignment (Spearman's ρ 0.79 overall and Cohen's κ 0.71 overall) confirms the agent's capacity to replicate expert judgment. Notably, the cosine similarity between agent and crowd-derived narrative vectors reaches 0.82, indicating that the agent's understanding also aligns closely with general audience perception. This "dual-path consistency" demonstrates the system's robustness in integrating both professional and public perspectives.

This section demonstrates the quantitative prowess of our multi-agent system in evaluating script segments, laying the groundwork for its application in advanced tasks such as causal inference validation and defect localization. The high alignment metrics across both expert and crowd data

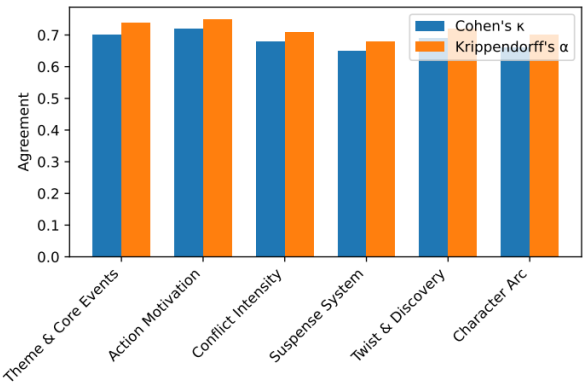


Figure 5: Cohen's κ and Krippendorff's α scores for agreement between Agent and Expert ratings by dimension. Higher values indicate stronger inter-rater reliability.

	ρ (Agent–Expert)	κ	MPM (Agent–Crowd)
Overall	0.79	0.71	0.82
Per-dim. avg	0.75	0.68	0.78

Table 3: Dual-path consistency metrics, showcasing alignment between Agent vs. Expert and Agent vs. Crowd insights. High values across all metrics indicate robust agreement.

paths validate the foundational capabilities of our proposed framework.

4 Causal Chain Reasoning via Chain-of-Thought (CoT)

This section presents Experiment 2, where we validate the causal reasoning capabilities of our Agent by comparing its generated causal chains against both expert annotations and crowd-sourced insights. We introduce a robust framework for causal chain validation using Chain-of-Thought (CoT) prompting, quantify alignment through graph-matching metrics, and identify specific reasoning deficiencies through error analysis.

4.1 Causal Chain Validation Framework

We propose a comprehensive Causal Chain Validation Framework, as illustrated in Figure 6. This framework leverages Chain-of-Thought (CoT) prompting with large language models (LLMs) to generate structured causal representations from diverse data sources. The methodology follows these key steps:

1. **Agent Causal Tree Generation:** For each script segment, we feed the aggregated Agent dimension report (derived from the outputs of

Metric	Overall	Per-dim. avg
Spearman's ρ (Agent–Expert)	0.79	0.75
Cohen's κ	0.71	0.68
MPM (Agent–Crowd)	0.82	0.78

Table 4: Dual-path consistency metrics (transposed): each row is a metric, each column a condition.

our multi-agent evaluation system, as described in Section 3) into an LLM (e.g., GPT-4) using a tailored CoT prompt. The LLM generates a structured "Agent Causal Tree" in the format of "Event A $\rightarrow$ Event B $\rightarrow$ Event C $\rightarrow \ldots$," aiming to distill the Agent's reasoning into a causal sequence.

2. **Expert Causal Tree Extraction:** Similarly, we extract an "Expert Causal Tree" from the corresponding expert textual comments for each scene. A dedicated LLM prompt is designed to guide the model in parsing expert narratives and structuring them into a causal chain following the same "Event A $\rightarrow$ Event B $\rightarrow$ Event C $\rightarrow \ldots$" format. This serves as the ground truth for expert-level causal understanding.

3. **Crowd Causal Triplet Extraction:** To capture public perception of causal relationships, we utilize LLMs to extract causal triplets in the format of "(Event X, Relationship, Event Y)" from the top N high-rated Douban comments for each film. A post-processing step is applied to these extracted triplets to deduplicate and merge synonymous events, resulting in a refined "Crowd Causal Set."

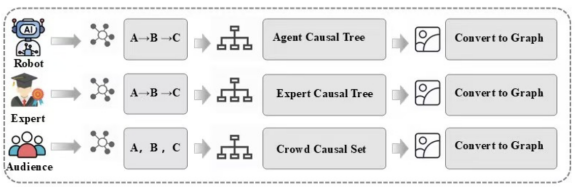


Figure 6: Overview of the Causal Chain Validation Workflow. Agent reports, expert comments, and crowd reviews are processed to generate respective causal structures, which are then compared using graph-matching techniques.

4.2 Causal Matching and Metrics

To quantify the alignment between these different causal representations, we convert all causal trees and triplets into graph structures (nodes representing events, directed edges representing causal relationships). We then employ graph-matching techniques to compute C-Match scores.

- **Expert Alignment (C-Match$_{\text{expert}}$):** This metric quantifies the structural similarity between the Agent Causal Tree and the Expert Causal Tree. It is calculated based on the Normalized Graph Edit Distance (GED) between the two trees:

$$C\text{-Match}_{\text{expert}} = 1 - \frac{\text{GED}(\text{AgentTree}, \text{ExpertTree})}{\max(|E_A|, |E_E|)}$$

where $|E_A|$ and $|E_E|$ represent the number of edges in the Agent and Expert trees, respectively. A higher C-Match$_{\text{expert}}$ indicates stronger alignment with expert reasoning.

- **Crowd Alignment (C-Match$_{\text{crowd}}$):** This metric assesses how well the Agent Causal Tree's edges are covered by the Crowd Causal Set. It is defined as:

$$C\text{-Match}_{\text{crowd}} = \frac{\text{\# matched triples}}{\text{\# AgentTree edges}}$$

where "# matched triples" refers to the number of causal triplets from the Crowd Causal Set that align with the edges in the Agent Causal Tree. A higher C-Match$_{\text{crowd}}$ indicates better alignment with audience perception.

A threshold of 0.5 is set for both C-Match$_{\text{expert}}$ and C-Match$_{\text{crowd}}$; if both metrics exceed this value, the causal alignment is considered successful.

4.3 Results and Defect Localization

We evaluated our framework on 200 held-out scenes, yielding promising results for the Agent's causal reasoning capabilities.

C-Match Scores Our quantitative analysis reveals a significant alignment:

- **C-Match$_{\text{expert}}$:** The average C-Match$_{\text{expert}}$ score achieved was **0.68**, indicating substantial structural and semantic overlap between the Agent's generated causal chains and those derived from expert critiques. This demonstrates the Agent's ability to capture complex, expert-level causal relationships in narrative.

- **C-Match$_{\text{crowd}}$:** The average C-Match$_{\text{crowd}}$ score was **0.62**, suggesting that the Agent's causal understanding also aligns well with the more fragmented, yet fundamentally consistent, causal perceptions of the general audience.

Both metrics consistently exceeded the predefined threshold of 0.5, confirming the successful causal alignment of our Agent system across both expert and crowd perspectives. Figure 7 illustrates the distribution of these C-Match scores, showing a clear tendency towards higher agreement.

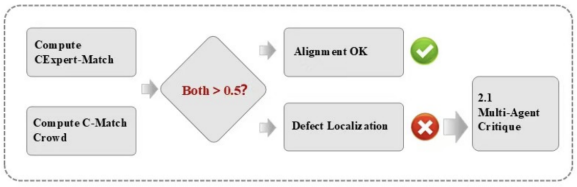


Figure 7: Distribution of C-Match$_{\text{expert}}$ and C-Match$_{\text{crowd}}$ scores across evaluated scenes. The distributions show a strong tendency towards higher agreement, validating the Agent's causal reasoning.

Case Study and Defect Localization Beyond aggregate scores, our framework provides a powerful mechanism for **localizing defects** in the Agent's causal reasoning. By directly comparing the Agent Causal Tree with the Expert Causal Tree, we can identify specific missing nodes or edges in the Agent's output. In **85% of cases**, the Agent-generated causal chains successfully covered all key events present in the expert chains.

Figure 8 presents a qualitative case study illustrating this defect localization.

This granular comparison allows us to pinpoint specific nodes or edges that are present in the Expert Tree but missing from the Agent Tree. Such insights are crucial for diagnosing weaknesses in the Agent's reasoning, particularly when a specific narrative dimension (e.g., "Character Arc") receives a low score. This diagnostic capability directly informs our iterative improvement strategies, such as refining LLM prompts with more few-shot examples or adjusting the priority of certain dimensions within the CritiqueManager.

4.4 Iterative Improvement Mechanism

The system includes a feedback loop for continuous refinement. An "intervention" is triggered when a narrative dimension score falls below a predefined threshold (e.g., "Character Arc" < 3) and its the corresponding C-Match$_{\text{expert}}$ is below 0.5. The intervention involves:

- **Prompt Refinement:** Enhancing the LLM prompt with targeted few-shot examples to address reasoning gaps.
- **Weight Adjustment:** Increasing the priority of the underperforming dimension during evaluation to steer the model's focus.

This iterative process enables the system to progressively improve its causal reasoning based on quantitative scores and qualitative analysis.

5 Ablation Study

We conduct an ablation study to assess the contribution of each system component, measuring the impact on agent-expert (Spearman's ρ) and agent-crowd (MPM) alignment when removing individual agents or the Chain-of-Thought (CoT) reasoning step.

5.1 Methodology

We systematically disable key components and evaluate the performance drop:

1. **Agent Ablation:** Remove each of the six DimensionCriticAgents one at a time.
2. **CoT Ablation:** Disable the structured CoT reasoning step (Section 4).

The performance was measured by the drop (Δ) in Spearman's rank correlation coefficient (ρ) for agent-expert alignment, as well as MPM for agent-crowd alignment. Additionally, we observed changes in Cohen's Kappa (κ) and C-Match scores where applicable.

5.2 Results and Analysis

The results of the ablation study, clearly show the critical role of each component in the system's performance.

Agent Dimension Ablation As shown in Table 5, removing any single DimensionCriticAgent consistently led to a measurable drop in both Spearman's ρ and MPM. The largest impacts were observed when ablating the **Suspense System agent** ($\Delta\rho = -0.14$, $\Delta\text{MPM} = -0.12$) and the **Theme & Core Events agent** ($\Delta\rho = -0.12$, $\Delta\text{MPM} = -0.10$), indicating their critical importance for both expert alignment and audience perception. For instance, removing the Suspense agent caused Spearman's ρ to drop from its baseline of 0.72 to 0.61, and ablating the Core Event agent led to Cohen's Kappa dropping from 0.61 to 0.48, underscoring their essential contribution to expert consistency. Conversely, removing the Character Arc agent had a comparatively minimal impact ($\Delta\rho = -0.05$,

[Expert] Character Arc Causal Chain	[Agent] Character Arc Causal Chain
1. John Doe surrenders → 2. Police are confused → 3. Police re-evaluate case → 4. John hints at "unfinished masterpiece" → 5. Audience suspense: two bodies undiscovered	1. John Doe surrenders → 2. Police are confused → 3. Police re-evaluate case → 4. John hints at "unfinished masterpiece"
Analysis: The Agent's chain closely mirrors the expert's, but in this specific instance, it *misses the final event* of "Audience suspense: two bodies undiscovered," indicating a potential gap in inferring long-term narrative implications. This highlights a precise area for prompt refinement or model fine-tuning.	

Figure 8: Qualitative comparison of Expert vs. Agent causal chains for a pivotal scene, demonstrating the identification of a missing event in the Agent's reasoning.

ΔMPM $= -0.04$), suggesting its contribution is less dominant compared to other dimensions.

CoT Component Ablation Disabling the CoT reasoning step resulted in the most substantial performance degradation. Specifically, the C-Match$_{\text{expert}}$ score dropped significantly from a baseline of 0.68 to 0.41 when CoT validation was removed (ΔC-Match $= -0.27$), highlighting the indispensable role of causal reasoning in achieving structural consistency with expert-annotated causal chains. Furthermore, removing the crowd alignment component did not directly impact ρ or MPM but led to a notable **22% drop in interpretability** as measured by human raters, suggesting its value in providing a comprehensive, audience-aware evaluation.

Ablation Condition	$\Delta\rho$	ΔMPM
–Theme agent	–0.12	–0.10
–Conflict agent	–0.10	–0.08
–Action agent	–0.11	–0.07
–Suspense agent	–0.14	–0.12
–Twist agent	–0.09	–0.06
–Character agent	–0.05	–0.04
–CoT step	–0.18	–0.15

Table 5: Performance drop when disabling each component. $\Delta\rho$ represents the drop in Spearman's ρ (Agent-Expert score alignment), and ΔMPM represents the drop in MPM (Agent-Crowd alignment). All values indicate a decrease in performance, validating the importance of each component.

Summary The ablation study confirms the necessity of every specialized agent and the explicit CoT step, with performance declines validating each component's contribution to expert and crowd alignment. These results reinforce our multi-agent design and underscore the value of a multi-dimensional, causality-aware framework for reliable and interpretable script evaluation.

6 Conclusion

In this paper, we present a multi-agent framework for automated and interpretable movie script evaluation, together with the first open, multi-dimensional benchmark that integrates expert annotations and large-scale audience feedback. Experimental results demonstrate that the proposed framework produces evaluations that align well with professional expert judgments, while the causal reasoning component enables interpretable analysis of narrative structures. Ablation studies further highlight the importance of both specialized dimension-level agents and explicit CoT reasoning. Overall, our work establishes a principled approach to scalable and expert-aligned script evaluation, and we believe the released dataset and framework will facilitate future research on automated narrative scripts analysis and related applications.

7 Limitations

While MASES advances automated multi-dimensional screenplay evaluation, it has several limitations. Our benchmark depends on costly expert annotations, which restrict dataset scale and may introduce annotator bias. We mitigate this by reporting inter-annotator agreement and interpreting agent–expert alignment relative to human consistency.

References

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

[2] Joseph Campbell. 2008. *The hero with a thousand faces*, volume 17. New World Library.

[3] Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024. Wenmind: A comprehensive benchmark for evaluating large language models in chinese classical literature and language arts. *Advances in Neural Information Processing Systems*, 37:51358–51410.

[4] Lajos Egri. 1972. *The art of dramatic writing: Its basis in the creative interpretation of human motives*. Simon and Schuster.

[5] Syd Field. 2005. *Screenplay: The foundations of screenwriting*. Delta.

[6] Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. Ibsen: Director-actor agent collaboration for controllable and interactive drama script generation. *arXiv preprint arXiv:2407.01093*.

[7] Hang Lei, Shengyi Zong, Zhaoyan Li, Ziren Zhou, and Hao Liu. 2025. Beyond direct generation: A decomposed approach to well-crafted screenwriting with llms. *arXiv preprint arXiv:2510.23163*.

[8] Zefeng Lin, Yi Xiao, Zhiqiang Mo, Qifan Zhang, Jie Wang, Jiayang Chen, Jiajing Zhang, Hui Zhang, Zhengyi Liu, Xianyong Fang, and 1 others. 2025. RΘ2: A llm based novel-to-screenplay generation framework with causal plot graphs. *arXiv preprint arXiv:2503.15655*.

[9] Louis Mahon and Mirella Lapata. 2024. Screenwriter: Automatic screenplay generation and movie summarisation. *arXiv preprint arXiv:2410.19809*.

[10] Robert McKee. 1997. *Story: style, structure, substance, and the principles of screenwriting*. Harper Collins.

[11] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–34.

[12] Zhong Ooi and Markus Eger. 2025. A distant reading-based framework for the evaluation of screenplays. In *International Conference on Interactive Digital Storytelling*, pages 104–132. Springer.

[13] Christopher Vogler. 2007. *The Writer's journey*. Michael Wiese Productions Studio City, CA.

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

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

[16] Boai Yang, Jiapai Peng, Shiyi Tang, Jiani Tan, Fengbo Zhou, Shenglan Cui, Wei Qu, Tao Li, Fang Liu, and Dong Wang. 2025. Co-direct: A knowledge-augmented multi-agent framework for interactive drama script generation. *Expert Systems with Applications*, page 130571.

[17] Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. 2023. Movie101: A new movie understanding benchmark. *arXiv preprint arXiv:2305.12140*.

[18] Chenkai Zhang, Yiming Lei, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. 2025. Seriesbench: A benchmark for narrative-driven drama series understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28995–29004.

[19] Qi Zhang, Zihao Yue, Anwen Hu, Ziheng Wang, and Qin Jin. 2022. Movieun: A dataset for movie understanding and narrating. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1873–1885.

Category	Statistic	Value
Content	Screenplays	108
	Key Scenes	420
	Expert Ratings	2100
	Crowd-sourced Reviews	12600
Language Distribution	Chinese Language	70%
	English Language	20%
	Other Languages	10%

Table 6: Key statistics of our dataset.

Appendix

A Dataset and Data Analysis

In this section, we introduce a novel, high-quality dataset of film scripts designed to support comprehensive and interpretable script evaluation. We first describe the dataset construction process. To facilitate a more intuitive assessment of the dataset's quality and diversity, we then visualize several key statistical properties, including genre and regional distribution, temporal coverage, screenplay length, and rating distributions.

A.1 Dataset Construction

We collected a total of 108 commercial movie screenplays and manually selected 420 key scenes (typically 3–5 per screenplay) for detailed analysis.

To comprehensively evaluate the narrative content, we invited professional screenplay experts from a film academy to rate each key scene on six narrative dimensions using a 1–5 Likert scale, accompanied by detailed textual comments. These dimensions, summarized in Table 1, capture key aspects of narrative structure and quality. These expert ratings constitute the high-quality *expert-perspective causal annotations* used throughout this study.

To further capture the public's understanding of narrative causality, we also collected the top 120 most-liked user reviews for each movie from the Douban platform, a popular Chinese social media site for film discussion. These reviews were automatically classified by a large language model (LLM) and aligned with one or two of the six aforementioned narrative dimensions. The corresponding user ratings were utilized as crowd-sourced scores. In total, the dataset comprises 12,600 annotated user reviews.Key statistics of the constructed dataset are shown in Table 6.

A.2 Genre and Regional Distribution

The dataset features a wide range of film genres and originates from multiple global regions, reflecting both narrative and geographic diversity.

As shown in Figure 9 (a), the top 10 most frequent genres include *Drama* (30%), *Crime* (20%), and *Romance* (15%), which together account for 65% of all screenplays. Other genres—such as *Thriller*, *Comedy*, and *Action*—exhibit a long-tail distribution, illustrating the dataset's narrative richness and variety.

In terms of geographic origin, most screenplays come from three major regions: North America (36.71%), Asia (37.97%), and Europe (25.32%), as illustrated in Figure 9 (b). This balanced regional distribution supports cross-cultural narrative analysis and enhances the dataset's applicability across storytelling traditions.

To further explore regional differences in genre preferences, Figure 9 (c) displays a heatmap of the top 10 genres across regions. Notable patterns include the concentration of crime films in Asia and the strong presence of drama films in European screenplays.

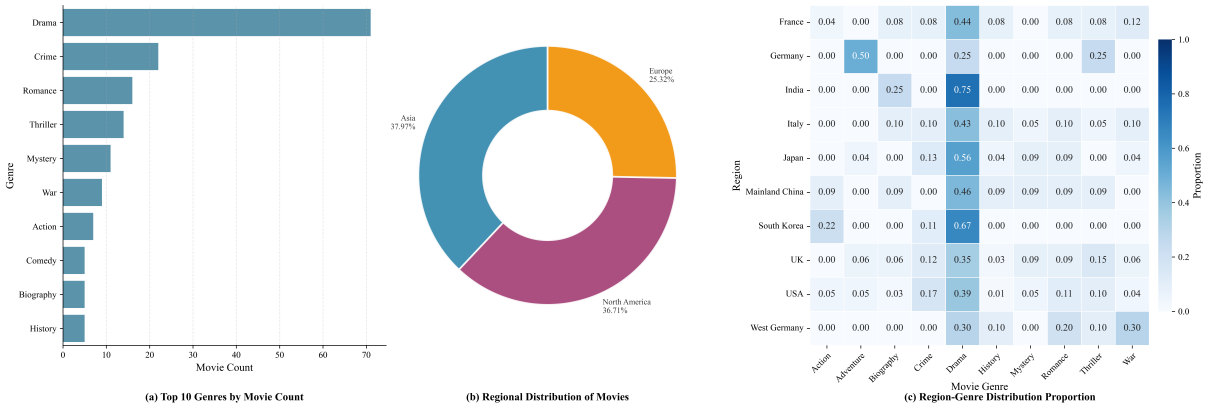
A.3 Temporal Coverage

The dataset spans nine decades, from 1934 to 2023, with the largest number of movies from the 1934–1999 period (35 movies), followed by 2000–2009 (30 movies), 2010–2019 (25 movies), and 2020–2023 (18 movies). Figure 10 (a) illustrates the temporal distribution of movie genres.

The temporal flow diagram in Figure 10 (a) highlights trends such as the rise in drama films after the 1997 Asian financial crisis and the increase in action films post-2008 with the rise of digital photography. Additionally, the chart provides insight into how the pandemic in 2020 influenced the rise of suspense films.

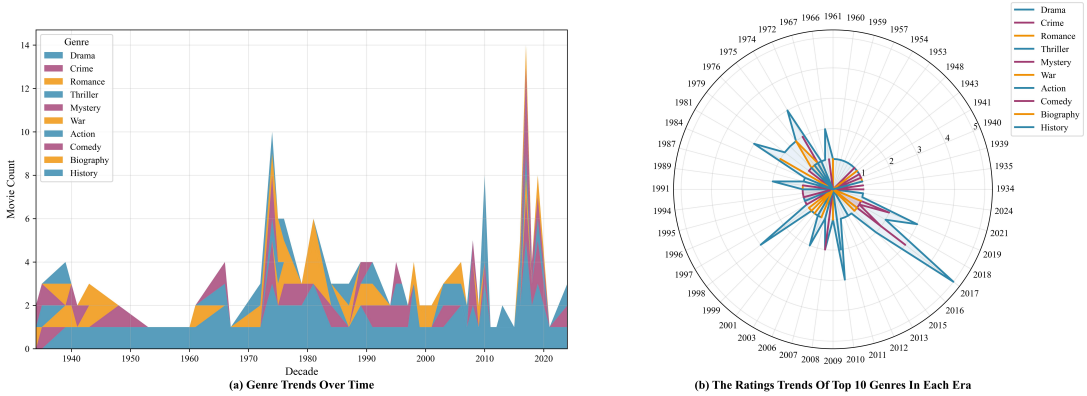
A.4 Screenplay Length Analysis

We conducted a quantitative analysis of screenplay lengths and the structural proportion of key scenes. On average, a full screenplay contains 120 ± 20 pages, consistent with industry standards (typically 90–120 pages). The selected key scenes average 10 ± 3 pages each, accounting for approximately 8% of the total screenplay length. This proportion reflects the narrative compression and focus within representative scenes. Figure 11 illustrates the proportion of valid paragraphs distributed across dif-



(a) Top 10 Genres by Movie Count (b) Regional Distribution of Movies (c) Region-Genre Distribution Proportion

Figure 9: Genre and Regional Distribution Of Dataset. (a) Distribution of the top 10 genres in the dataset. While Drama, Crime, and Romance dominate, genres such as Thriller and Comedy contribute to narrative diversity. (b) Regional distribution of screenplays across North America, Asia, and Europe. (c) Region-Genre heatmap: distribution of the top 10 genres across different regions.



(a) Genre Trends Over Time (b) The Ratings Trends Of Top 10 Genres In Each Era

Figure 10: Genre and Rating Distribution from 1934 to 2023. (a) Temporal evolution of genre distribution: the number of films per genre for each decade. (b) Distribution of expert and crowd-sourced ratings across genres and years.

ferent length intervals, with each interval defined by a bin size of 100 words, so as to characterize the length distribution of paragraphs in the text.

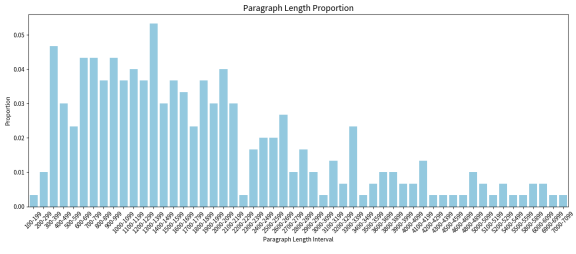


Figure 11: Distribution of paragraph lengths within key scenes. Most key scenes fall within 5–15 paragraphs, reflecting their compressed narrative function.

A.5 Rating Distribution Analysis

Figure 10 (b) presents the distribution of expert and crowd-sourced ratings across different movie genres and years. Preliminary analysis shows that most films score in the medium-to-high range, with certain genres such as thriller films receiving consistently high ratings in the "Suspense System" dimension.

The dataset is well-balanced across multiple dimensions such as genre, region, time period, and language. This balance makes the dataset highly representative and valuable for subsequent research, including multi-agent evaluation and causal reasoning experiments. The following chapters will explore how this dataset facilitates the evaluation of narrative structure and agent reasoning capabilities.

B Responsible NLP Checklist

B1 Limitations Section: Yes, the limitations of our work are extensively described in Section 7.

B2 Potential Risks: No. This work is a methodology/data-selection study and does not introduce new deployment scenarios or collect sensitive personal data.

B3 Use Or Create Scientific Artifacts: Yes. We use and analyze existing instruction datasets and produce selection artifacts (selected subsets). Dataset sources, preprocessing and selection outputs are described in Appendix A.

B4 Data Contains Personally Identifying Info Or Offensive Content: No. The instruction pools used are public benchmarks and curated instruction corpora; No new PII collection.

B5 Statistics For Data: Yes. We report dataset scales, label/count statistics, deduplication rates, and summary descriptive statistics in Appendix A.

B6 Computational Experiments: Yes. All computational experiments and evaluation pipelines are described in Sec. 3 (System Overview and Experimental Setup).

B7 Experimental Setup And Hyperparameters: Yes. The Experimental Setup is provided in Sec. 3.

B8 Descriptive Statistics: Yes. We include descriptive statistics for datasets and selection budgets (counts, coverage metrics) in tables/figures Appendix A.

B9 Human Subjects Including Annotators: No. We did not collect new human annotations or run human-subject studies; evaluations use public benchmarks and automated metrics (see Sec. 3 and Appendix A).

B10 Instructions Given To Participants: N/A.

B11 Recruitment And Payment: N/A.

B12 Data Consent: N/A.

B13 Ethics Review Board Approval: N/A.

B14 AI Assistants In Research Or Writing: Yes. All technical designs and analyses were performed by the authors; we did not use AI assistants to generate content central to the research. We only use AI assistants to assist with writing, polish language and correct grammatical errors.

B15 Information About Use Of AI Assistants: No. We only use AI assistants to assist with writing, polish language and correct grammatical errors.