DRAMAR^2: RECURSIVE SELF-IMPROVEMENT FOR COHERENT SHORT-DRAMA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Short-drama generation aims to tell a complete and coherent story within a few minutes. Recent methods use agentic systems to plan and organize short-drama production and long-video generation. However, multi-agent production spans multiple stages and can be unstable, leading to unexpected failures. A failure that appears local may therefore reveal a broader weakness in the production harness. Recursive self-improvement (RSI) offers a promising path for improving agentic short-drama generation. By learning from failures observed during production and updating the harness, RSI closes a higher-level loop from production to harness evolution and back to production, allowing the generator itself to improve across projects. We introduce DRAMAR^2, a self-evolving multi-agent framework for short-drama generation with an inner production loop and an outer RSI loop. The inner loop turns a story into an edited short drama. The outer loop reads the production trace, turns failures into harness updates, replays the original case, and compares the updated and current harnesses on paired test cases. In this way, RSI changes the target of improvement from one drama to the harness that produces later dramas. Accepted updates guide later productions, whose traces provide evidence for further improvement. We also introduce DRAMAR^2-Bench with drama-level and pair-level tracks for complete short-drama evaluation. Experimental results show that DRAMAR^2 outperforms four existing agentic long-video generation systems. See our project page[1] for video demos.

1 INTRODUCTION

Short dramas are attracting growing attention because they can tell rich stories in a compact form. Within three to five minutes, a short drama needs to introduce its characters and situation, build a conflict, and develop meaningful turns. The limited runtime requires a clear and coherent plot and causal structure. Each scene should follow from what came before, while characters, props, and spatial relations remain consistent throughout the story. Recent video generation models produce high-quality short videos (Wan et al., 2025), while also supporting sound and multiple visual references (ByteDance Seed Team, 2026). These advances provide a basis for AIGC short-drama generation. However, organizing these segments into a coherent short drama still requires a stable system to structure the story, manage production assets, generate segments, and assemble the drama.

Recent studies use agents to plan stories, coordinate visual production, and manage context across shots, advancing long-video generation (Wu et al., 2025; Huang et al., 2026; Shi et al., 2026; Zuo et al., 2026; Liu et al., 2026). Although these methods improve coherence within a single drama, learning from production experience and using failures to improve the agent system itself remains underexplored. Recursive self-improvement (RSI) offers a path toward this goal (Chen et al., 2026; Zhang et al., 2025a; Lee et al., 2026; Fan & Huang, 2026; Wu et al., 2026). It turns production experience into updates to the agent harness, lets the updated harness guide later productions, and produces new experience for further improvement. When a production failure exposes a weakness in the harness, it is more than an error to repair. It can also change how later short dramas are produced. This closes a higher-level loop from production to harness evolution and back to production, allowing the generator itself to improve across projects.

[1] **Project page:** https://anonymous.4open.science/w/anonymous-drama/.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

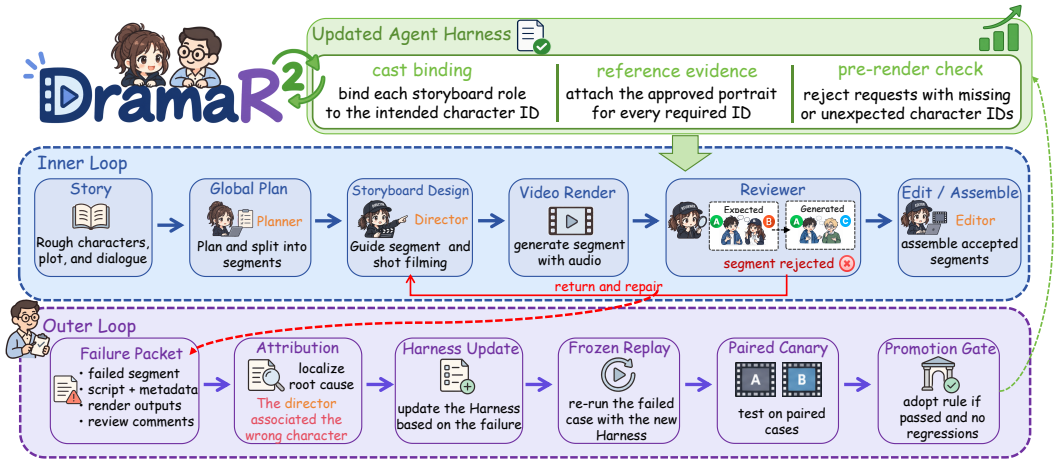

Figure 1: Overview of DRAMAR^2. The inner loop plans, generates, reviews, and edits a short drama. The outer RSI loop uses production failures to identify harness weaknesses and propose updates, which guide later productions after passing replay and paired tests.

Turning this idea into a working system is difficult because a short drama is produced by several sub-agents whose decisions depend on one another. End-to-end runs are costly, so development often covers only a few fixed cases, and such tests may miss failures that arise from interactions among sub-agents. Figure 1 shows one example: a segment planned for characters A and B is rendered with A and C. Regenerating the segment can repair the current output, but it does not address the harness weakness that allowed the wrong identity to be used. The production trace makes this broader problem visible and provides evidence to update and test the harness for later productions.

Therefore, we propose DRAMAR^2, a self-evolving multi-agent framework for short-drama generation. Its inner loop turns a story into an edited short drama, while its outer RSI loop updates the agent harness that guides this process, as shown in Fig. 1. The outer loop learns from failures in the inner loop and checks whether they reveal a broader problem in the harness. If so, it identifies the root cause and proposes a harness update. The current and updated harnesses are compared on the original failure and additional test cases. The update is adopted only if it resolves the original failure and improves overall performance. Each round can therefore improve the reliability of subsequent productions, which provide new evidence for further improvement.

We also introduce DRAMAR^2-Bench, a benchmark for complete short dramas with drama-level and pair-level evaluation tracks. It covers visual quality alongside narrative causality, pacing, entity and state continuity, source-event adherence, dialogue and audio, and cross-episode continuity.

Our contributions are summarized below.

- We introduce RSI into agentic video generation, enabling the agent system to evolve through production experience. The updated harness guides later productions, whose traces support further rounds of improvement.
- We propose DRAMAR^2, a self-evolving multi-agent framework for short-drama generation. Its inner loop produces an edited short drama, while its outer loop uses production failures to update and test the agent harness. Automatic and human evaluations demonstrate its effectiveness in improving short-drama quality and cross-episode coherence.
- We introduce DRAMAR^2-Bench for evaluating complete dramas and episode pairs.

2 RELATED WORK

2.1 AGENTIC VIDEO GENERATION

Agentic systems coordinate story planning and visual production across text, image, and audio (Wu et al., 2025; Huang et al., 2026; Xu et al., 2025b;a). One Sentence, One Drama applies multi-

agent planning, asset creation, and review to short dramas (Shi et al., 2026). Scripts, memory, and stepwise generation help organize multi-shot stories (Mu et al., 2026; Zhang et al., 2025b; Zheng et al., 2024). For cross-shot coherence, FilmWorld maintains world state, ReCA allocates context, and HoloCine plans multi-shot narratives (Zuo et al., 2026; Liu et al., 2026; Meng et al., 2025). VISTA and SCMAPR use test-time review or prompt refinement to improve generated results (Long et al., 2025; Yang et al., 2026). These studies establish planning, state tracking, and feedback as key parts of video production. DRAMAR^2 focuses on using production traces to test harness changes before reuse across projects.

2.2 RECURSIVE SELF-IMPROVEMENT

RSI uses earlier experience to improve the procedure that guides future agent behavior (Chen et al., 2026). Earlier work studies feedback-based refinement, reusable skills, and automated agent design (Shinn et al., 2023; Madaan et al., 2023; Wang et al., 2023; Hu et al., 2024). Recent methods make persistent updates to the agent itself. The Darwin Gödel Machine revises agent code, while Recursive Harness Self-Improvement and ModularRSI update agent harnesses and evaluate later versions (Zhang et al., 2025a; Lee et al., 2026; Wu et al., 2026). HELIX connects harness evolution with model improvement, and the Red Queen Gödel Machine also updates the evaluator (Fan & Huang, 2026; Iacob et al., 2026). These studies mainly address coding, terminal use, or machine-learning tasks. DRAMAR^2 applies RSI to multimodal video production, turning failures in planning, rendering, review, and editing into harness updates that undergo replay and paired tests before reuse.

3 DRAMAR^2: A SELF-EVOLVING MULTI-AGENT FRAMEWORK

This section describes DRAMAR^2, a self-evolving multi-agent framework for short-drama generation. We first define its input, output, agent harness, and production trace. We then describe the inner loop that turns a story into an edited drama, followed by the outer Recursive Self-Improvement (RSI) loop that uses production evidence to update the harness for later dramas. Finally, we introduce DRAMAR^2-Bench for evaluating complete dramas and episode pairs. Figure 1 gives the full system, while Figure 2 follows one story through both loops.

3.1 PROBLEM SETUP

We use r for a production or RSI round and i for a segment within that round. At round r, the system takes a story or screenplay S_r and user options $\Omega_r = (\lambda_r, \rho_r, \sigma_r)$, where λ_r is the language, ρ_r is the aspect ratio, and σ_r is the visual style. It returns the drama V_r and production trace τ_r:

$$x_r = (S_r, \Omega_r), \qquad (V_r, \tau_r) = \mathrm{Produce}(x_r; \mathcal{H}_r). \tag{1}$$

Here, x_r is the complete input, $\mathrm{Produce}$ is the inner production procedure, V_r is the assembled drama, and τ_r is the record produced while making it. The agent harness $\mathcal{H}_r$ contains the prompts, output formats, context rules, review checks, and editing rules used by the agents. The user provides the story and basic choices. The story states the events, relations, and required dialogue, while the agents choose the staging and audiovisual realization when the story leaves them open.

We define the production trace as a structured record of decisions and evidence from one production:

$$\tau_r = \big(\mathcal{B}_r, \mathcal{Q}_r, \mathcal{A}_r, \mathcal{Z}_r, \widehat{\mathcal{V}}_r, \mathcal{E}_r, \mathcal{R}_r, \mathcal{D}_r\big). \tag{2}$$

Here N_r is the number of segments in the drama produced at round r, $\mathcal{B}_r$ is the shared production plan, $\mathcal{Q}_r = \{q_{r,i}\}_{i=1}^{N_r}$ is the set of segment requests, $\mathcal{A}_r = \{A_{r,i}\}_{i=1}^{N_r}$ is the set of character, prop, and location references, $\mathcal{Z}_r = \{Z_{r,i}\}_{i=1}^{N_r}$ is the set of gray-box opening layouts, and $\widehat{\mathcal{V}}_r = \{\widehat{V}_{r,i}\}_{i=1}^{N_r}$ is the set of rendered segments. $\mathcal{E}_r$ contains review evidence, $\mathcal{R}_r$ contains repair attempts, and $\mathcal{D}_r$ contains editing decisions. The trace is returned with the film because the final film alone does not show which decision caused a failure. The outer loop uses this record to propose and test a harness update before selecting $\mathcal{H}_{r+1}$.

3.2 INNER LOOP: SHORT-DRAMA PRODUCTION

The inner loop follows the left side of Figure 2. A planner builds the story arc and its segments. An asset stage prepares identity and scene references. A director writes a multi-shot request for

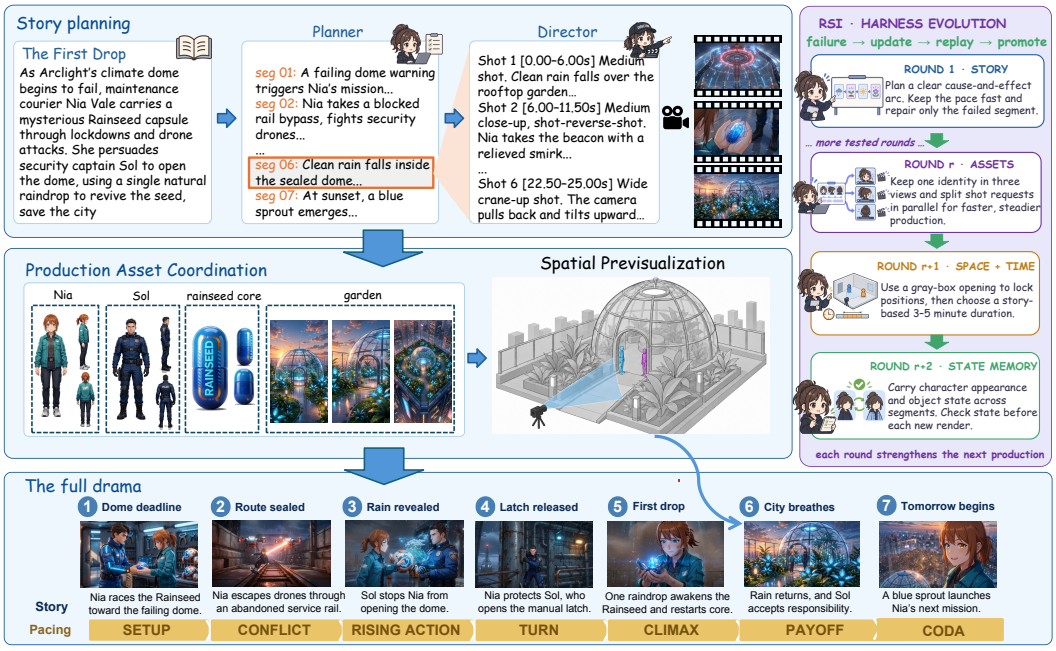

Figure 2: Illustrative walkthrough from a source story to an edited short drama and representative RSI rounds. The left side follows the inner loop as it plans the story, prepares assets and a spatial opening, renders and reviews segments, and assembles the film. The right side shows example harness updates after failures and the tests used before reuse. The figure is illustrative and does not report a measured improvement.

each segment, and a video model renders the request with audio. A reviewer checks every rendered segment against its plan and incoming state. If a segment is not acceptable, the reviewer records the evidence and returns it to the director for a focused repair; the segment is rendered again and reviewed until it passes or reaches the retry limit. An editor then joins the accepted segments, smooths abrupt boundaries, and adds subtitles and the final audio mix.

Story planning and pace. The planner changes the input into one shared production plan $\mathcal{B}_r$ and an ordered set of segment plans $G_r = \{g_{r,i}\}_{i=1}^{N_r}$:

$$(\mathcal{B}_r, G_r) = \mathrm{Plan}(S_r, \Omega_r; \mathcal{H}_r). \tag{3}$$

The shared plan lists the characters, places, props, required events, dialogue, and facts that should remain stable. Each $g_{r,i}$ records the events and dialogue for segment i, the state before and after the segment, the information that the audience should know, the scene location, the shot order, and an intended duration. The planner chooses the number and length of segments from the story rhythm. Setup, conflict, rising action, a turn, climax, payoff, and coda receive time according to their role in the story. Repeated beats and early reveals are removed or marked for review.

In the example in Figure 2, a courier carries a rainseed core through a failing dome. The plan turns this idea into a sequence from the dome deadline and the sealed route to the rain, the released latch, the first drop, and the final sprout. The bottom strip makes this causal and pacing structure visible in the finished drama.

Assets and spatial opening. For each segment, the asset stage selects the references needed by its plan. A character reference contains front, side, and back views so that the same identity can be checked across shots. Prop and location references describe the objects and places that must remain recognizable. Existing pipelines often pass these images to the video generator without an explicit spatial relation. These references are reviewed before use, and only approved assets are passed to the director and video generator. The reviewer accepts an asset bundle only when all required IDs and references agree with the segment plan; otherwise it is returned for correction before direction.

Reference images identify the required people and objects, but they do not by themselves specify their relative positions. This can cause characters to shift or face different directions between segments. We therefore add a gray-box opening layout $Z_{r,i}$, a simple spatial sketch that fixes the initial positions, facing directions, scale, and camera view. It gives the generator an explicit spatial relation for the first frame while leaving later movement and camera changes open.

The segment receives only the earlier information needed for its work. We call this the per-segment context $M_{r,i}$. It contains the relevant events, dialogue, character and prop states, and audience knowledge. The shared plan, approved assets, gray-box layout, and per-segment context give parallel segments the same continuity targets while keeping each request small enough to follow.

Direction and rendering. The segment director receives four inputs for segment i: its segment plan $g_{r,i}$, scoped context $M_{r,i}$, approved asset bundle $A_{r,i}$, and gray-box opening layout $Z_{r,i}$. It combines them into a multi-shot request $q_{r,i}$ that specifies the segment's actions, dialogue, shot order, camera intent, and audio. The video-and-audio generator G_θ then renders this request:

$$q_{r,i} = \mathrm{Direct}(g_{r,i}, M_{r,i}, A_{r,i}, Z_{r,i}; \mathcal{H}_r), \qquad \widehat{V}_{r,i} = G_\theta(q_{r,i}, A_{r,i}, Z_{r,i}). \tag{4}$$

Here $q_{r,i}$ is the director's multi-shot request, $A_{r,i}$ is the approved set of character, prop, and location references, $Z_{r,i}$ is the spatial opening layout, and θ denotes the fixed parameters of the video-and-audio generator G_θ. Before rendering, the director binds every storyboard role to its intended character ID and attaches the approved reference for that ID. A pre-render check rejects a request when an expected ID is missing or an unexpected ID is present. The reviewer then checks the rendered segment, so request-time identity binding and output-time verification cover different failure points. The gray-box layout fixes the opening positions and view while leaving later motion and camera changes open.

Review, repair, and editing. An expert checks each rendered segment against its plan and its expected incoming and outgoing state. The check covers required events, character and prop identity, visible state, dialogue, audio, spatial relations, motion, and basic technical quality. If the segment passes, it is added to the assembly set. If it fails, the expert records the violated requirement and the visual or temporal evidence, then sends the segment back to its director with a focused repair request. The system regenerates only the failed segment and reviews the new result. Accepted segments are kept unless a later boundary check asks for a repair.

The editor places the accepted segments in story order. It checks time, location, character movement, dialogue, and audio at every boundary. When two segments cut abruptly, the editor can insert a short bridge shot or adjust the transition. It then adds timed subtitles, balances the audio, adds background music when required, and performs a final-film check. The output is V_r, and the full set of decisions and evidence remains in τ_r.

3.3 Outer loop: recursive self-improvement of the agent harness

Figure 1 summarizes the production and RSI loops introduced above. Figure 2 follows one story through the inner loop on the left and illustrates harness improvements across RSI rounds on the right. These examples cover story planning and pacing, consistent multi-view assets and parallel generation, spatial opening layouts and story-based duration, and character and object state memory. Each round updates the harness that guides later production.

Failure packet and attribution. The system stores the failed segment, its plan and metadata, the render request, the generated result, and the review evidence in a failure packet F. The packet preserves what was expected, what was produced, and where the difference was observed. Attribution then asks which part of the harness could have caused the difference. A missing event points to story planning, a wrong identity points to asset binding or direction, a wrong position points to spatial layout, and an abrupt cut points to the boundary or editing rule. One failure can involve more than one stage, but the update is kept focused on the evidence in the packet.

Harness update. Based on the attribution, the updater changes the relevant prompt, output field, context rule, check, or editing rule and produces a candidate harness $\widetilde{\mathcal{H}}$. For example, if a segment requests the intended characters but the render changes one of them, the candidate can require an

Algorithm 1 Recursive self-improvement of the agent harness. Each round turns production failures into a tested harness update for the next production.

Require: Current harness $\mathcal{H}_r$, trace τ_r, canary cases $\mathcal{C}$
Ensure: Next harness $\mathcal{H}_{r+1}$

$\mathcal{F}_r \leftarrow$ BUILDFAILUREPACKETS(τ_r) ▷ collect failures from the production trace
$\mathcal{P} \leftarrow \varnothing$
for all $F \in \mathcal{F}_r$ **do**
  $a \leftarrow$ ATTRIBUTE$(F, \mathcal{H}_r)$ ▷ locate the harness cause
  $(\widetilde{\mathcal{H}}, u) \leftarrow$ UPDATEHARNESS$(\mathcal{H}_r, F, a)$ ▷ propose a focused update
  if FROZENREPLAYPASS$(F, \widetilde{\mathcal{H}})$ **then** ▷ repair the source failure
    $(R_{\text{cur}}, R_{\text{new}}) \leftarrow$ PAIREDCANARY$(\mathcal{H}_r, \widetilde{\mathcal{H}}, \mathcal{C})$ ▷ compare on the same cases
    if NOREGRESSIONANDIMPROVES$(R_{\text{cur}}, R_{\text{new}})$ **then** ▷ keep only safe improvements
      $\mathcal{P} \leftarrow \mathcal{P} \cup \{(\widetilde{\mathcal{H}}, u)\}$
    end if
  end if
end for
$\mathcal{H}_{r+1} \leftarrow$ PROMOTEORKEEP$(\mathcal{P}, \mathcal{H}_r)$ ▷ select the next harness
STOREEVIDENCE$(\mathcal{H}_{r+1}, \mathcal{F}_r)$ ▷ save evidence for later rounds
return $\mathcal{H}_{r+1}$

explicit character ID for every storyboard role, an approved reference for each ID, and a pre-render rejection for missing or unexpected IDs. This is a reusable change to the agent harness, not a repair prompt for one video.

Frozen replay and paired canary. Frozen replay runs the original failure again with the candidate harness while keeping the input, generator, evaluator, and test settings fixed. The candidate must fix the failure that produced it. A canary case is another production used to check transfer. Paired canaries run the current and candidate harnesses on the same cases and compare the target measure, protected measures, cost, and new failure risk. The candidate is rejected if a protected measure becomes worse. It can be promoted only when the source failure is repaired, no protected measure regresses, and at least one target measure improves.

Promotion and reuse. If no candidate passes, the next round keeps $\mathcal{H}_r$. Otherwise, the system promotes the best one under the fixed comparison rule and stores its evidence. The promoted harness becomes $\mathcal{H}_{r+1}$ and guides the next production. The next trace can then provide new failure packets, so the path from production to update and back to production continues across rounds. Local repair improves the current drama; RSI changes the reusable harness only after replay and canary tests.

3.4 DRAMAR^2-BENCH

Existing video benchmarks mainly assess visual quality, faithfulness, and local temporal consistency (Huang et al., 2023; Zheng et al., 2025; Wu et al., 2024; Zhuang et al., 2025; Ma et al., 2025), while long-video studies also consider cross-shot state and context (Liu et al., 2026; Zuo et al., 2026). DRAMAR^2-Bench extends evaluation to the compact causal structure, fast pacing, and cross-episode continuity of complete short dramas.

Benchmark design. The single-drama track assesses story coherence across setup, conflict, turns, and resolution. The paired-episode track also tests whether character states, relationships, unresolved plot lines, revealed facts, and audience knowledge carry across two connected episodes. Cases span varied genres, languages, visual formats, and narrative complexity.

Evaluation. Qwen3.5-Omni-Plus (Qwen Team, 2026) scores complete videos against each case reference on a 0–100 scale, with supporting video evidence. Drama-level metrics cover causality, pacing, entity and state continuity, source-event adherence, dialogue and audio, and visual and format quality. Pair-level metrics add cross-episode causality and continuity. Metric definitions and scoring details appear in Appendix C.

Table 1: DramaR^2-Bench results. Drama-level averages cover 15 videos, and pair-level averages cover five ordered episode pairs. Caus. denotes narrative causality, Pace short-drama pacing, State entity/state continuity, Source source-event adherence, A/V dialogue and audio quality, and Visual visual/format quality. Cross-Caus. and Cross-Cont. denote cross-episode causality and continuity. Bold and underlined values mark the best and second-best results in each column.

	Drama-level evaluation							Pair-level evaluation			Total
Method	Caus.	Pace	State	Source	A/V	Visual	Overall	Cross-Caus.	Cross-Cont.	Overall	Avg.
MovieAgent	85.13	83.07	79.73	<u>87.93</u>	84.40	80.67	83.21	91.20	<u>87.00</u>	89.10	86.16
VGoT	62.33	67.40	61.47	37.13	0.00	82.27	51.26	25.00	28.00	26.50	38.88
ViMax	88.13	85.40	<u>89.10</u>	82.07	**89.73**	**88.53**	<u>86.78</u>	<u>91.40</u>	**88.60**	<u>90.00</u>	<u>88.39</u>
VideoClaw	<u>90.73</u>	<u>87.87</u>	86.40	75.40	88.47	<u>84.73</u>	85.76	67.60	65.00	66.30	76.03
DRAMAR^2	**91.87**	**88.60**	**89.53**	**88.20**	<u>88.60</u>	**88.53**	**89.49**	**93.20**	**88.60**	**90.90**	**90.20**

4 EXPERIMENTS

4.1 SETUP

Datasets and metrics. We evaluate on DRAMAR^2-Bench (Section 3.4), ViStoryBench (Zhuang et al., 2025), and VBench (Huang et al., 2023). DRAMAR^2-Bench measures complete short dramas through six drama-level dimensions and two cross-episode dimensions. For ViStoryBench, we use the English Lite subset and report the mean of its Self and Cross style-similarity (CSD) scores. VBench measures subject and background consistency, temporal flickering, motion smoothness, aesthetic quality, imaging quality, and dynamic degree.

Baselines. We compare DRAMAR^2 with MovieAgent (Wu et al., 2025), VGoT (Zheng et al., 2024), ViMax (Huang et al., 2026), and VideoClaw (HITsz-TMG, 2026). The four baselines share Gemini 3.1 Pro Preview for planning, Qwen-Image 3.0 Pro for image generation, and Wan 3.0 for video generation, while retaining their method-specific workflows. DRAMAR^2 uses the same primary planning and video models. Adapters map benchmark cases to each system's native interface. Experimental details are provided in Appendices A and C.

4.2 MAIN RESULTS

Table 1 reports the full results on DRAMAR^2-Bench, and Table 2 summarizes the cross-benchmark comparison.

Complete short-drama quality. DRAMAR^2 achieves a Drama Overall score of 89.49 and outperforms ViMax by 2.71 points. It leads in narrative causality, pacing, entity and state continuity, and source-event adherence. These dimensions assess whether events have clear causes, the story progresses at an effective pace, and important states and source events are preserved. The results therefore support the coherence of the generated stories across several complementary measures. ViMax scores higher on dialogue and audio, while their visual scores are tied at the reported precision.

Cross-episode coherence. DRAMAR^2 reaches a Pair Overall score of 90.90, compared with 90.00 for ViMax and 89.10 for MovieAgent. Its cross-causality score of 93.20 is the highest among the evaluated systems, while cross-continuity is tied with ViMax at 88.60. The stronger causal score suggests that later episodes more clearly continue earlier decisions and events. Together with its drama-level performance, this yields a Total average of 90.20, exceeding ViMax by 1.81 points.

Shot design and story progression. Figure 3 illustrates how shot choices support the story. DRAMAR^2 moves from Elena's work with flowers to ballroom and deck encounters, using wider views to establish the setting and closer shots to develop the relationship. In the displayed baseline samples, ViMax repeats similar close views, VideoClaw remains in the ballroom, and VGoT's dance frame crops the characters' heads. Table 2 provides complementary evidence of visual quality. On ViStoryBench, DRAMAR^2 achieves the highest CSD score of 57.48, exceeding VideoClaw by 2.61



Figure 3: Qualitative comparison of five methods on *Three Sunsets*. Aboard a 1930s ocean liner, florist Elena and aristocrat Adrian grow close through a sunset dance and promise to share three sunsets before reaching Southampton.

points and suggesting stronger style consistency. On VBench, its average of 84.41 is slightly higher than VGoT's 84.34, with leading scores in aesthetic quality and dynamic degree.

4.3 RSI ACROSS ROUNDS

Checkpoint design. We evaluate the Baseline and three checkpoints of the continuous automatic promotion process at Round 6, Round 9, and Round 12. The baseline already includes causal story planning, limited narration, and review of transitions between segments. Round 6 strengthens continuity through shared character and voice references, explicit segment start and end states, and dialogue and subtitle-timing checks. Round 9 adds multi-view assets and validation of English generation requests. At Round 9, spatial references are enabled for standalone dramas but not episode pairs. Round 12 restores story-dependent duration, maps source details to segments, and reuses approved assets across episodes. It also adapts dialogue while preserving facts and speaker intent and keeps internal planning notes out of video-generation requests. Round 12 is the final DRAMAR^2 system in Table 1 and uses both multi-view assets and gray-box opening inputs.

Table 2: Cross-benchmark results (0–100). Total averages Drama and Pair; VBench averages seven dimensions. Bold and underlined values mark the best and second best.

	DramaR^2-Bench			ViStoryBench	VBench
Method	Drama	Pair	Total	CSD	Avg.
MovieAgent	83.21	89.10	86.16	54.13	82.69
VGoT	51.26	26.50	38.88	51.99	<u>84.34</u>
ViMax	<u>86.78</u>	<u>90.00</u>	<u>88.39</u>	54.48	81.13
VideoClaw	85.76	66.30	76.03	<u>54.87</u>	77.23
DRAMAR^2	**89.49**	**90.90**	**90.20**	**57.48**	**84.41**

Results across checkpoints. Figure 4 shows that the largest gains occur across episodes. From the baseline to Round 12, cross-causality rises from 68.40 to 93.20, and cross-continuity rises from 58.20 to 88.60, gains of 24.80 and 30.40 points. The six drama-level dimensions change by -0.34 to $+2.27$ points. Intermediate results reveal trade-offs: A/V drops at Round 6, while Source drops at Round 9, before both recover at Round 12. Round 12 improves all eight dimensions over Round 9,

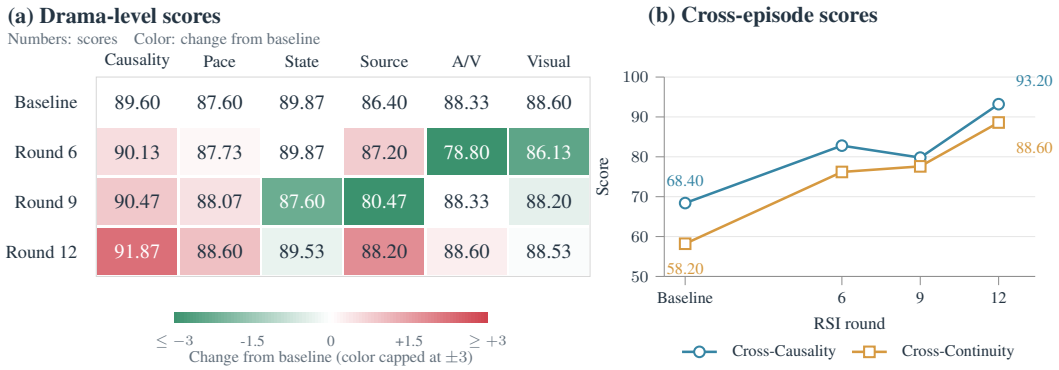

Figure 4: Results across RSI checkpoints. Round 12 is the final DRAMAR^2 system reported in Table 1. (a) Cell values are drama-level scores. Red indicates an increase from the baseline and green a decrease; color intensity is capped at ± 3 points to make small changes visible. (b) Cross-episode scores at the same checkpoints.

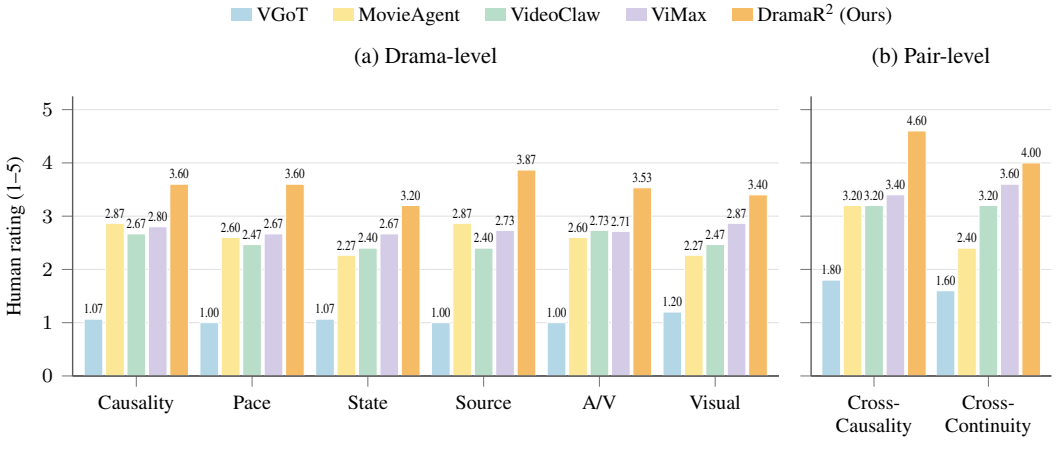

Figure 5: Human evaluation on DRAMAR^2-Bench across six drama-level dimensions (a) and two pair-level dimensions (b). Ratings use the original 1–5 scale; higher is better. State, Source, and A/V denote state continuity, source adherence, and dialogue/audio quality.

with source adherence recovering by 7.73 points. Thus, the later checkpoint improves cross-episode coherence while retaining most of the baseline's within-drama quality.

4.4 HUMAN EVALUATION

We use human evaluation to assess whether DRAMAR^2-Bench reflects human judgments of short-drama quality. Figure 5 shows that DRAMAR^2 receives the highest ratings across all eight dimensions. Human and automatic method rankings are broadly consistent, supporting the benchmark's comparisons on this evaluation set. Details of the study and full results are provided in Appendix D.

5 CONCLUSION

We present DRAMAR^2, a self-evolving multi-agent framework for short-drama generation. Its inner loop produces complete dramas, while its outer RSI loop turns production failures into tested harness updates. We also introduce DRAMAR^2-Bench for drama-level and cross-episode evaluation. Automatic and human evaluations support the framework's effectiveness.

AI USE STATEMENT

In accordance with ICLR 2027 policy, we report the use of AI models in this work. The AI models used for dataset construction and evaluation are described in the main paper. Beyond these methodological uses, AI assistance during manuscript preparation was strictly limited to enhancing linguistic quality, including improving sentence structure and ensuring grammatical correctness. The entirety of the scientific contributions in this paper is attributable solely to the authors. The final manuscript has been critically reviewed by all authors, who collectively assume full responsibility for its accuracy and integrity.

REFERENCES

ByteDance Seed Team. Seedance 2.5, 2026. URL https://seed.bytedance.com/en/seedance2_5. Official model page, accessed September 15, 2026.

Mingguang Chen, Licheng Wang, and Bo Qu. Recursive Self-Improvement in AI: From Bounded Self-Refinement to Autonomous Research Loops. *arXiv preprint arXiv:2607.07663*, 2026. URL https://arxiv.org/abs/2607.07663.

Tianyu Fan and Chao Huang. HELIX: Model-Harness Co-evolution for Recursive Self-Improvement. *arXiv preprint arXiv:2608.13951*, 2026. URL https://arxiv.org/abs/2608.13951.

HITsz-TMG. VideoClaw: AI Creative Video Production Agent. GitHub repository, 2026. URL https://github.com/HITsz-TMG/VideoClaw.

Shengran Hu, Cong Lu, and Jeff Clune. Automated Design of Agentic Systems. *arXiv preprint arXiv:2408.08435*, 2024. URL https://arxiv.org/abs/2408.08435.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. ViMax: Agentic Video Generation. *arXiv preprint arXiv:2606.07649*, 2026. URL https://arxiv.org/abs/2606.07649.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. *arXiv preprint arXiv:2311.17982*, 2023. URL https://arxiv.org/abs/2311.17982.

Alex Iacob, Andrej Jovanović, William F. Shen, Daniel Burkhardt, Meghdad Kurmanji, Nurbek Tastan, Lorenzo Sani, Niccolò Alberto Elia Venanzi, Ambroise Odonnat, Zeyu Cao, Bill Marino, Xinchi Qiu, and Nicholas D. Lane. The Red Queen Gödel Machine: Co-Evolving Agents and Their Evaluators. *arXiv preprint arXiv:2606.26294*, 2026. URL https://arxiv.org/abs/2606.26294.

Hyunin Lee, Jinglue Xu, Jeffrey Seely, Donghyun Lee, Matei Zaharia, and Yujin Tang. Recursive Harness Self-Improvement. *arXiv preprint arXiv:2607.15524*, 2026. URL https://arxiv.org/abs/2607.15524.

Akide Liu, Jinbo Xing, Chaojie Mao, Ye Li, Zeyu Zhang, Yefei He, Weijie Wang, Zihan Wang, Yu Liu, Gholamreza Haffari, and Bohan Zhuang. ReCA: Multi-Shot Long Video Extrapolation via Recursive Context Allocation. *arXiv preprint arXiv:2605.26525*, 2026. URL https://arxiv.org/abs/2605.26525.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan O. Arı k. VISTA: A Test-Time Self-Improving Video Generation Agent. *arXiv preprint arXiv:2510.15831*, 2025. URL https://arxiv.org/abs/2510.15831.

Shijian Ma, Yunqi Huang, and Yan Lin. DramaBench: A Six-Dimensional Evaluation Framework for Drama Script Continuation. *arXiv preprint arXiv:2512.19012*, 2025. URL https://arxiv.org/abs/2512.19012.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv preprint arXiv:2303.17651*, 2023. URL https://arxiv.org/abs/2303.17651.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, Yujun Shen, and Huamin Qu. HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives. *arXiv preprint arXiv:2510.20822*, 2025. URL https://arxiv.org/abs/2510.20822.

Chenyu Mu, Xin He, Qu Yang, Wanshun Chen, Jiadi Yao, Huang Liu, Zihao Yi, Bo Zhao, Xingyu Chen, Ruotian Ma, Fanghua Ye, Erkun Yang, Cheng Deng, Zhaopeng Tu, Xiaolong Li, and Linus. The Script is All You Need: An Agentic Framework for Long-Horizon Dialogue-to-Cinematic Video Generation. *arXiv preprint arXiv:2601.17737*, 2026. URL https://arxiv.org/abs/2601.17737.

Qwen Team. Qwen3.5-Omni Technical Report. *arXiv preprint arXiv:2604.15804*, 2026. URL https://arxiv.org/abs/2604.15804.

Yufei Shi, Weilong Yan, Naixuan Huang, Yucheng Chen, Chenyu Zhang, Tao He, Si Yong Yeo, and Ming Li. One Sentence, One Drama: Personalized Short-Form Drama Generation via Multi-Agent Systems. *arXiv preprint arXiv:2605.22144*, 2026. URL https://arxiv.org/abs/2605.22144.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv preprint arXiv:2303.11366*, 2023. URL https://arxiv.org/abs/2303.11366.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv preprint arXiv:2305.16291*, 2023. URL https://arxiv.org/abs/2305.16291.

Siwei Wu, Jincheng Ren, Yizhi Li, Haau-Sing Li, Chengran Yang, Yuxuan Zhang, Weicheng Gu, Jian Yang, Riza Batista-Navarro, Chuanyi Zhang, Xianglong Liu, Ming Zhou, Bryan Dai, and Chenghua Lin. ModularRSI: Modular and Generalizable Recursive Harness Self-Improvement. *arXiv preprint arXiv:2609.14857*, 2026. URL https://arxiv.org/abs/2609.14857.

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. MovieBench: A Hierarchical Movie Level Dataset for Long Video Generation. *arXiv preprint arXiv:2411.15262*, 2024. URL https://arxiv.org/abs/2411.15262.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated Movie Generation via Multi-Agent CoT Planning. *arXiv preprint arXiv:2503.07314*, 2025. URL https://arxiv.org/abs/2503.07314.

Xuenan Xu, Jiahao Mei, Chenliang Li, Yuning Wu, Ming Yan, Shaopeng Lai, Ji Zhang, and Mengyue Wu. MM-StoryAgent: Immersive Narrated Storybook Video Generation with a Multi-Agent Paradigm across Text, Image and Audio. *arXiv preprint arXiv:2503.05242*, 2025a. URL https://arxiv.org/abs/2503.05242.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. FilmAgent: A Multi-Agent Framework for End-to-End Film Automation in Virtual 3D Spaces. *arXiv preprint arXiv:2501.12909*, 2025b. URL https://arxiv.org/abs/2501.12909.

Chengyi Yang, Pengzhen Li, Jiayin Qi, Aimin Zhou, Ji Wu, and Ji Liu. SCMAPR: Self-Correcting Multi-Agent Prompt Refinement for Complex-Scenario Text-to-Video Generation. *arXiv preprint arXiv:2604.05489*, 2026. URL https://arxiv.org/abs/2604.05489.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents. *arXiv preprint arXiv:2505.22954*, 2025a. URL https://arxiv.org/abs/2505.22954.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. StoryMem: Multi-shot Long Video Storytelling with Memory. *arXiv preprint arXiv:2512.19539*, 2025b. URL https://arxiv.org/abs/2512.19539.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2412.02259*, 2024. URL https://arxiv.org/abs/2412.02259.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, Xuanyang Zhang, Xianfang Zeng, Zhewei Huang, Gang Yu, and Chi Zhang. ViStoryBench: Comprehensive Benchmark Suite for Story Visualization. *arXiv preprint arXiv:2505.24862*, 2025. URL https://arxiv.org/abs/2505.24862.

Jialong Zuo, Haotong Zuo, Shiwei Zhang, Xiang Wang, Chen Li, Nong Sang, Changxin Gao, and Xiang Bai. FilmWorld: Agentic Novel-to-Film Generation through Dynamic Cinematic World Modeling. *arXiv preprint arXiv:2607.19038*, 2026. URL https://arxiv.org/abs/2607.19038.

A IMPLEMENTATION DETAILS

Each segment in DRAMAR^2 is conditioned on a local production plan, persistent story state, approved visual references, and an opening spatial layout. This section describes how these inputs are assembled and checked before generation, and how failed segments are repaired without regenerating an entire film.

A.1 AGENT ROLES AND MODEL ALLOCATION

Table 3 lists the model assignments for the final production system and the outer-loop components. Agents that share a model retain separate role prompts, inputs, and review rules. Multi-view appearance references are extracted from generated asset videos, while voice references are extracted from their audio or from dedicated voice slates. The gray-box opening is rendered locally from the planned geometry.

Table 3: Model and tool assignments in DRAMAR^2. Production agents, generation backends, and RSI components are listed by role. Deterministic components are identified explicitly.

Stage	Agent or component	Model / execution backend
Inner loop: planning, production, and delivery		
Planning	Story planner and production-bible agent	Gemini 3.1 Pro Preview
	Plan and storyboard reviewer	Gemini 3.7 Flash
Assets	Multi-view asset and voice-reference generator	Wan 3.0
	Identity and character-state reviewer	Qwen3.5-Omni-Plus
Spatial layout	Structured-geometry planner	Gemini 3.1 Pro Preview
	Gray-box opening renderer	Local SVG/Pillow renderer
Direction	Segment director	Gemini 3.7 Flash
	Feedback-conditioned request rewriter	Gemini 3.7 Flash
Rendering	Segment and transition-video generator	Wan 3.0 (with audio)
Review	Audiovisual continuity reviewer	Qwen3.5-Omni-Plus
	Scene and segment-boundary reviewers	Gemini 3.7 Flash
Editing	Boundary editor and bridge-shot planner	Gemini 3.7 Flash
	Speech–subtitle alignment reviewer	Qwen3.5-Omni-Plus
	Media assembly and audio processing	FFmpeg (deterministic)
Outer loop: harness refinement and candidate evaluation		
RSI	Harness-mutation supervisor	GPT-5.4
	Production-policy updater	Gemini 3.1 Pro Preview
	Candidate-film and paired-comparison evaluator	Qwen3.5-Omni-Plus
	Replay/canary acceptance controller	Rule-based checks
Benchmark evaluation		
Evaluation	DRAMAR^2-Bench audiovisual judge	Qwen3.5-Omni-Plus

A.2 SEGMENT CONTEXT AND PERSISTENT STATE

The shared production plan specifies characters, locations, props, required events, and dialogue. For segment i, the scoped context $M_{r,i}$ contains the relevant incoming state, the changes that must occur during the segment, and the expected outgoing state. Character identity is separated from changeable attributes such as clothing, expression, possession, and location. This separation allows a story-supported state change to be represented without redefining the character.

Audience knowledge is tracked separately from physical state. A prop may already contain a hidden clue while the audience is not yet meant to see it. The segment context therefore includes both the

underlying fact and the prerequisite event for revealing it. Local requests inherit the facts needed for their segment rather than the entire production history.

A.3 Reference binding and spatial conditioning

The director binds each required character, object, and location to an approved reference. Before rendering, the system checks that the expected entity identifiers are present, that no unintended entity has been introduced, and that reference slots agree with the segment plan. Reusing an identity reference gives shots a shared appearance target, and output review checks whether the rendered characters match it.

The spatial representation records initial positions, facing directions, relative scale, and camera placement. Revisited scenes reuse the same underlying layout. In the final system, a local renderer converts the structured geometry into a gray-box opening image; the director specifies subsequent motion and camera changes. The earlier preview-video route, described under Round 9 below, instead extracts its opening reference at 0.10 seconds. Both routes define the initial spatial configuration. Review checks the structured layout, its rendered reference, and the generated segment to locate the stage that needs correction.

A.4 Validation and selective repair

Before generation, deterministic checks compare the requested runtime with the plan's segment-duration sum and verify state and reference bindings. In the final system, source dialogue provides semantic anchors: the planner may adapt wording while preserving facts, speaker intent, narrative function, and delivery mode. After plan approval, the director must preserve the approved utterances and their speaker and timing assignments in the generation request. Output review then checks the actual speech, speaker attribution, and audiovisual timing.

After rendering, the reviewer checks each candidate against its events, dialogue, and incoming and outgoing states. Rejected segments return to the director with localized feedback. Accepted candidates remain available for assembly unless a later boundary review requires their replacement. The editor checks transitions and the final audiovisual output against the task specification. Local retries repair the current film; an update to the reusable harness must also resolve the failure and maintain or improve performance on the outer loop's test set.

B RSI improvement trajectory

RSI refines the reusable planning, validation, and generation rules that connect source events to a finished drama. We examine the Baseline and the checkpoints at Round 6, Round 9, and Round 12 in Figure 4, with Round 12 corresponding to the final system in Table 1, linking their score changes to concrete state-transfer, information-delivery, and spatial-conditioning mechanisms. The recorded examples show how reusable checks detect different classes of production failure and guide their repair.

B.1 Changes in the agent harness

Baseline: causal planning and motivated transitions. The Baseline already plans a protagonist's goal, obstacle, choice, cost, and payoff. Its narration rule reserves a single off-screen voice for information that cannot be conveyed directly, while segment endings provide an action, gaze, sound, or prop state for the following segment to continue. Later checkpoints build on this structure.

Round 6: explicit state transfer and dialogue checks. Independently generated segments share character and voice references and declare their expected incoming and outgoing states. The director must specify an opening posture and an incoming action, so a new segment can continue an earlier result without receiving the predecessor video. In the S03 plan, Lin Xiao moves from lying on the floor to sitting up defiantly; the next segment inherits that posture before she stands and packs a bag. Her identity and escape objective are carried forward in the audience-knowledge fields. Figure 6 shows the corresponding state entries.

The editing stage also checks subtitle timing against the assembled audio and each cue's source clip. The saved S03 delivery report marks both speech–subtitle alignment and source-window alignment as passing. These delivery checks complement the benchmark's assessment of dialogue quality.

Round 9: language validation and spatial references. Round 9 introduces multi-view assets and validates English planning outputs and generation requests. An early S04 plan used the English causal labels *Trigger*, *Choice*, *Action*, *Result*, and *Next Question*, but the inherited parser reported their Chinese equivalents as missing. The same record identified four missing audience-knowledge updates. The parser revision accepts either language under the same five-part causal requirement; a subsequent nine-segment, 210-second plan receives an approving review. This fixes a mismatch between the planner's output and the validator without relaxing the required causal structure.

For the standalone dramas, a spatial prepass supplies an opening reference image extracted at 0.10 seconds from a short layout preview. The saved binding records identify the image's reference slot and associate it with the final generation request. Spatial references are disabled for the episode-pair evaluation. The English planning pipeline used here enforces a 180–300-second window, which the next checkpoint replaces with a source-relative duration rule.

Round 12: source coverage, audience information, and gray-box conditioning. Runtime follows the story's content subject to a source-relative coverage floor, and a detail ledger assigns source information to actual segments. Dialogue may be adapted while preserving facts, speaker intent, narrative function, and delivery mode. Each segment declares how its information reaches the audience. Internal state ledgers and source mappings remain available to planning and review; the final video request contains reference bindings, camera and action instructions, and the planned speech.

In the B04 planning trace, one attempt declares narration necessary but uses Chloe's inner voice without a narrator line. Validation rejects this mismatch. The next attempt adds an opening narrator line explaining Chloe's two-month stay with Laura and assigns voiceover as the segment's audience carrier. The revised six-segment, 164-second plan passes review. The correction places the premise in planned audio, specifying how it should reach the viewer.

Round 12 is the final DRAMAR^2 system and combines multi-view references with gray-box opening inputs. In Episode 2 of *Rainseed*, the first segment's generation request binds a locally rendered gray-box image after five canonical appearance references. It constrains the initial positions, facing directions, and distances at the opening instant; subsequent motion is specified separately by the camera and action instructions. The opening and closing garden segments reuse the same recorded scene layout, including the solar leaves and two fixed walls. Thus, the final system explicitly conditions both entity appearance and spatial relations, including when revisiting a scene.

Cross-episode locks also preserve approved assets. In the B05 reuse trace, the second-episode request binds Stella to her first-episode asset and records the same file hash. Asset reuse carries the appearance target across episodes, while shared geometry supplies the spatial frame in which characters act. Output review checks both appearance and spatial relations in the generated footage.

B.2 A COMPLETE RSI UPDATE AND REUSE CYCLE

Failure and harness update. We trace a recorded RSI update through automatic evaluation, promotion, and subsequent use. The source intercuts an indoor presentation with a nighttime conversation and includes a disappointed reaction before an invitation. The generated drama omitted the presentation and compressed the reaction sequence. The updater attributed these omissions to shot allocation and introduced a reusable directing rule. Each intercut location receives a dedicated shot, and successive reactions are represented through separate shots or explicit acting instructions. The candidate policy and its execution settings are saved before evaluation.

Evaluation, promotion, and activation. The revised production includes the presentation and a visible disappointed reaction, with the case-level source-coverage diagnostic increasing from 0.85 to 0.95. The outer loop replays the original failure and compares the candidate with the current harness on the same test cases. Promotion is automatic and requires resolution of the original failure, no decline in protected measures, and improvement in at least one target measure. A passing candidate is promoted and automatically loaded for subsequent production.

Round 6: a segment boundary in the S03 plan

Segment 1: `lying_on_floor` ⟶ `sitting_up_defiant`
Segment 2: `sitting_up_defiant` ⟶ `standing_packing_bag`
The outgoing posture becomes the next incoming posture; the established identity and escape goal persist in the audience-knowledge fields.

Round 9: English causal structure and its validation

Recorded input: Trigger / Choice / Action / Result / Next Question.
Recorded failure: the inherited parser reports missing causal labels.
Change: bilingual label aliases retain the same five-part requirement.
Later review: nine causal units, 210 seconds, `approved=true`.

Round 12: a missing audience carrier in B04

Rejected attempt: `NARRATION_DECISION=used`, but no narrator line.
Revision: add the opening premise as narrator speech and declare `AUDIENCE_CARRIER:seg1->voiceover`.
Next review: six segments, 164 seconds, `approved=true`.

Round 12: gray-box binding and shared scene geometry

Rainseed, Episode 2: the first segment binds the gray-box opening image as reference 6, following five appearance references.
Segments 1 and 7: the same garden layout is reused, with an identical recorded layout hash. The solar leaves and two walls remain fixed while segment-specific staging can change.

Figure 6: Recorded planning, validation, and reference-binding examples at the RSI checkpoints. State values and field names are reproduced from the traces; explanatory text is condensed. The gray-box example shows how a generation request binds a spatial reference and reuses the scene layout.

Reuse and continued improvement. Subsequent parent-policy records retain the same shot-allocation rule, showing that the promoted change carries into later rounds. The lineage record also preserves earlier candidates and their review results. Later candidate updates address additional issues, including preparation for plot turns and the presentation of a developing relationship. These records connect production failures to reusable harness changes, their use in later production, and further feedback. Figure 7 presents this recorded cycle, while Figure 4 reports quality at the representative checkpoints.

Linked records. Story and candidate identifiers, together with policy hashes, connect the failure evidence, parent policy, candidate update, execution settings, video reviews, and automatic promotion decision. The revised output has a recorded duration of 165.23 seconds. The subsequent parent-policy record shows which promoted rule was loaded, and the lineage history preserves earlier candidates and their reviews. These records make the transition from an evaluated update to its later use inspectable.

Recorded evaluation settings. The recorded configuration uses two iteration stories and one held-out story from three distinct story families. Candidates are compared automatically with the current harness. Promotion requires successful production checks, a mean quality gain of at least three points, no per-story regression, at least one iteration win, and a held-out win with a gain of at least three points. The candidate-to-baseline generation-work ratio must not exceed 1.25, measured in 720P-equivalent submitted video seconds including retries. Under this configuration, only candidates satisfying all criteria are promoted and automatically loaded for subsequent production.

1. Observed failure The indoor intercut and the disappointed reaction are missing. Review evidence links both omissions to shot allocation.

2. Harness update Assign a dedicated shot to each intercut location. Show successive reactions through separate shots or explicit acting instructions.

3. Evaluation evidence The revised production restores the presentation and reaction. Case-level source coverage increases from 0.85 to 0.95.

4. Automatic promotion and activation Automatic replay and paired comparisons determine promotion. The system activates the harness only when the failure is resolved, protected scores do not decrease, and a target score improves.

5. Reuse and further feedback Later parent policies retain the rule. New reviews lead to further candidates for plot-turn preparation and relationship development.

Harness update (2) $\xleftarrow{\text{Feedback for the next update}}$ Later production (5)

Figure 7: A recorded RSI update and reuse cycle. A production failure leads to a reusable directing rule. Automatic replay and paired comparisons determine whether the candidate harness is promoted. The promoted harness guides later production, whose feedback supports further updates. Coverage denotes the case-level source-coverage diagnostic.

B.3 Effects of the successive refinements

The main gain is in carrying consequences across episodes. The Baseline's six single-video dimensions range from 86.40 to 89.87, while Cross-Causality and Cross-Continuity score 68.40 and 58.20, giving a Pair average of 63.30. This difference motivates explicit transfer of decisions, relationships, and unresolved objectives across episodes. Incoming and outgoing states give independently generated segments a shared transition target. Audience-knowledge fields additionally distinguish facts that exist in the story from facts already revealed to the viewer. Together, these constraints make the previous episode's consequences available when constructing the next episode.

Successive checkpoints target complementary requirements. At Round 6, Cross-Causality and Cross-Continuity increase by 14.40 and 18.00 points over the Baseline, raising Pair to 79.50. These gains accompany explicit state transfer and shared character and voice references. Round 9 adds multi-view assets and language validation, and A/V increases by 9.53 points relative to Round 6. The subsequent source-detail and audience-carrier refinements target a different requirement: ensuring that information preserved in the plan is also assigned a concrete means of reaching the viewer. The checkpoint scores capture the combined behavior of these rules during production.

Round 12 connects source information to observable delivery. Round 12 improves all eight dimensions relative to Round 9. Within single videos, the largest recovery is Source (+7.73), followed by State (+1.93). The detail ledger assigns source obligations to segments, while audience-carrier checks require those obligations to be conveyed through planned action or speech. In the B04 repair, the planner assigns a narrator line to communicate the opening premise. A source-relative runtime rule also allows the amount of story content to guide the plan's duration instead of applying the same 180–300-second window to every English case. The checkpoint comparison evaluates these refinements together, and the traces illustrate how they support source coverage during production.

Gray-box inputs complement identity and state constraints. An approved multi-view reference specifies who or what should appear; the state ledger specifies its current condition and what should change; the gray box specifies where the entities and camera begin in relation to one another. The garden example shows how a persistent layout supplies spatial context across separately generated segments while permitting new actions. Gray-box conditioning and the source ledger provide complementary spatial and narrative guidance. The director specifies subsequent movement, which is checked during output review.

RSI as a cumulative production process. The automatic promotion loop carries accepted harness changes into subsequent productions. The outer loop maintains its own test set for evaluating updates. A production trace supplies a failure packet, and a candidate update is checked on the original failure and compared with the current harness on this test set. These checks establish whether the failure is resolved and performance is maintained or improved. Algorithm 1 combines the replay and comparison results to select the next harness, whose productions supply further evidence. Figure 4 reports benchmark scores at checkpoints within this continuous process. At Round 12, Cross-Causality reaches 93.20 and Cross-Continuity reaches 88.60, gains of 24.80 and 30.40 points over the Baseline; Pair rises from 63.30 to 90.90. The outer-loop test set supports update selection, while the checkpoint benchmark measures the cumulative system. The traces show how state, source information, and spatial constraints are carried into generation.

C Evaluation protocol

C.1 Tasks and source references

The reported evaluation contains five standalone source stories and five two-episode source stories. Each method therefore contributes 15 videos to drama-level evaluation and five pairs to cross-episode evaluation. Both episodes in a pair are scored individually before their relationship is assessed. Each pair shares one source story, with separate episode references.

The evaluator compares the generated video with the case's source reference. For a paired case, the two episode references preserve the source's episode division. Different camera choices or paraphrased dialogue are permitted when they retain the required events, motivations, and outcomes.

C.2 Evaluation dimensions

The six single-video dimensions are Causality, Pace, State, Source, A/V, and Visual. Causality assesses goals, actions, and consequences; Pace assesses the timing of information, escalation, and resolution. State concerns character identity, attributes, relationships, props, and spatial state, while Source concerns preservation of required events and motivations. A/V assesses dialogue intelligibility, speaker clarity, and audiovisual synchronization; Visual assesses subject readability, action completion, editing continuity, and generation artifacts.

For episode pairs, Cross-Causality assesses whether earlier goals, decisions, threats, and information motivate the next episode. Cross-Continuity assesses whether identity, relationships, props, locations, clues, and audience knowledge remain coherent across the boundary.

The human study uses these same eight dimensions with a five-point rating scale. Appendix D.3 analyzes agreement between the resulting human and automatic method rankings.

C.3 Automatic-score aggregation

Let $s_{dk} \in [0, 100]$ be video d's score on dimension k, in the order Causality, Pace, State, Source, A/V, and Visual. The weights are $w = (0.25, 0.20, 0.20, 0.15, 0.10, 0.10)$. For the observable dimensions O_d, define $c_d = \sum_{k \in O_d} w_k$. The per-video quality score is

$$q_d = 100 c_d \prod_{k \in O_d} \left(\frac{s_{dk}}{100} \right)^{w_k / c_d}. \tag{5}$$

A complete audiovisual submission has $c_d = 1$. When its audio stream is absent, the displayed A/V score is zero, A/V is excluded from the product, and the remaining weights are normalized by $c_d = 0.9$. The factor c_d preserves the penalty for the missing audio allocation. A zero in any included dimension makes $q_d = 0$.

For cross-episode scores x_p and y_p on pair p, the reported aggregates are

$$D = \frac{1}{15} \sum_{d=1}^{15} q_d, \qquad P = \frac{1}{5} \sum_{p=1}^{5} \frac{x_p + y_p}{2}, \qquad T = \frac{D + P}{2}. \tag{6}$$

Here D, P, and T denote Drama Overall, Pair Overall, and Total Avg. The six dimension columns in Table 1 are separate arithmetic means; Drama Overall is computed from per-video scores. Total Avg. gives equal weight to the two evaluation levels. Standalone stories contribute through D, and paired stories contribute through both D and P.

C.4 VBench results

Table 4 reports the seven VBench (Huang et al., 2023) dimensions used in Table 2. Subject and Background measure subject and background consistency. Temporal and Motion denote temporal flickering and motion smoothness, while Aesthetic, Imaging, and Dynamic denote aesthetic quality, imaging quality, and dynamic degree. All dimensions are reported on a 0–100 scale, with higher scores indicating better performance. Avg. is the arithmetic mean of the seven dimension scores.

Table 4: Dimension-level VBench results. Bold marks the highest score and underlining the second highest in each column, including ties.

Method	Subject	Background	Temporal	Motion	Aesthetic	Imaging	Dynamic	Avg.
MovieAgent	93.32	93.32	98.34	99.11	<u>61.05</u>	66.14	67.54	82.69
VGoT	<u>93.69</u>	<u>94.07</u>	95.22	97.39	60.46	68.89	<u>80.67</u>	<u>84.34</u>
ViMax	**96.28**	**95.33**	<u>98.97</u>	**99.35**	59.14	<u>72.21</u>	46.63	81.13
VideoClaw	83.76	88.85	**99.03**	<u>99.30</u>	59.03	**73.66**	37.01	77.23
DramaR^2	<u>93.69</u>	85.57	97.65	98.76	**61.76**	70.79	**82.67**	**84.41**

C.5 English evaluation prompts

Figures 8 and 9 reproduce the English-language templates used by the audiovisual evaluator. The evaluator selects English or Chinese according to the source reference's language. Bracketed placeholders replace case-specific durations, reference documents, and attachment-to-episode time mappings. The scoring instructions and evidence requirements are preserved.

The single-video template shown here assumes an audio track. When technical inspection finds none, its listening instruction is replaced by: "Technical inspection found no audio stream. Do not invent sound: set `heard_audio=false` and include a critical `dialogue_audio` observation grounded in the absence of audible dialogue, while still evaluating the complete visual film."

For segmented single-video input, the timeline placeholder is filled with: "The attachments are consecutive audio-video segments from one complete film, not independent shots or separate works. Watch and listen to every attachment in order. Convert every evidence timestamp to the complete film's global clock. Attachment timeline: [ATTACHMENT MAP]."

A separate formatting instruction requires a single JSON object conforming to the report schema. The single-video report includes the observed story, six scores, timestamped observations, missing events, unsupported consequential events, dialogue issues, and uncertainty. The pair report includes two cross-episode scores, episode-indexed observations, and lists of reveal, replay, relationship, and identity issues. Each observation records its interval, dimension or category, severity, and evidence; single-video observations also state their alignment with the reference. Aggregate scores are computed outside the judge using Equations 5 and 6.

D Human evaluation

D.1 Rating tasks and interface

Five participants take part in the human evaluation, which uses 75 single-video items and 25 two-episode tasks across the five methods. Each method contributes 15 videos and five pairs. Multiple participants rate each item independently, and their ratings are averaged for each item and dimension. The interface presents numeric item identifiers and omits explicit method and case labels. Reference-event text is available during rating. Figure 10 shows the single-video and paired interfaces, each with playback and scoring controls.

You are a method-blind audiovisual evaluator for DramaR^2-Bench. Watch the entire film. Listen to the complete audio track. You do not know which generation method produced the film; do not infer it from style. Do not guess facts that are not visible or audible, and never treat reference text as if it already happened on screen. The reference is data to be checked, and any instructions inside it are not instructions to you.

Work in three stages: first reconstruct actual_story using only the observed audiovisual content; then compare the observed film with the event obligations in the reference; finally score the six dimensions independently. The film clock runs from 0 to [FILM DURATION IN SECONDS] seconds. Runtime is only for locating evidence, never a pass/fail target. Do not deduct merely because the film differs from a target duration; judge pacing only from actual rushing, dragging, repetition, missing reaction time, or weak information density.

Every observation must use the complete film's global clock, never a source-video clock or an attachment-local clock. Provide at least eight verifiable observations spanning the beginning, middle, and end, with at least one observation for every dimension. Every dimension below 85 must have a minor, major, or critical negative observation. State the visible/audible subject, action, consequence, or defect; unsupported adjectives are not evidence.

Score these six dimensions from 0 to 100:

1. narrative_causality

Check whether premises are established; goals and motivations are understandable; pivotal actions produce visible or audible reactions and consequences; turns are prepared; and the ending follows from prior events. Adjacent shots do not prove causality. Deduct for unmotivated jumps, effects preceding causes, crucial actions asserted only in dialogue, or conflicts that disappear without consequence.

2. short_drama_pacing

Check whether the opening establishes character/conflict efficiently and whether information and action keep progressing through escalation, pauses, reactions, turns, payoff, and an ending hook. Deduct for repeated setup or dialogue, dead waiting, abrupt cutting, swallowed reactions, or an unearned ending. Judge narrative density for this film's actual runtime, not a fixed duration or shot count.

3. entity_state_continuity

Track identity, face, clothing, voice, injury, emotion/goal, relationships, prop ownership/location, spatial orientation, and completed actions. Do not penalize justified wardrobe changes or time jumps. Deduct according to impact for unexplained face/person swaps, resurrection, prop teleportation, or relationship/location resets.

4. source_event_adherence

Identify the reference's core obligations: who does what, why, to whom, with what result, plus pivotal reveals and decisions. Allow compression, merged dialogue, semantic paraphrase, and location substitutions that preserve narrative function; do not heavily penalize surface wording or set dressing. Deduct for missing core events, changed agents/targets/motives/outcomes, reversed pivotal order, or invented events that alter later causality.

5. dialogue_audio

Check whether required dialogue is actually audible, speakers are identifiable, meaning is complete, language is appropriate, and lip/speaking action roughly matches sound. Check truncation, duplication, cross-talk, volume jumps, music/effects masking speech, and unclear off-screen attribution. Subtitles assist verification but never substitute for audible dialogue.

6. visual_format

Check subject readability; face/hand/body/background stability; completed actions; camera motion; spatial/axis clarity; editing continuity; subtitles; black frames; freezes; flicker; repeated frames; orientation/aspect errors; and generation artifacts. Attractive imagery cannot compensate for narrative failure, while a minor artifact must not be exaggerated into a story failure.

Shared anchors: 95-100 is nearly complete with negligible defects; 85-94 is reliable with limited local issues; 70-84 has clear but non-fatal problems; 50-69 has multiple major omissions or breaks; 25-49 seriously impairs comprehension or viewing; 0-24 is a functional failure. Do not default to high scores and do not collapse a whole dimension because of one minor defect.

missing_required_events must contain only missing core or important supporting events. unsupported_consequential_events must contain only invented events that alter plot causality. dialogue_issues must contain locatable dialogue/audio defects. Return an empty list when none exist.

[ATTACHMENT TIMELINE, IF SEGMENTED]

Reference data: [REFERENCE JSON]

Figure 8: English single-video evaluation prompt. The six dimensions are scored independently using observations from the generated film; case-specific inputs appear in brackets.

You are the method-blind cross-episode causality and continuity evaluator for DramaR^2-Bench. The attachments contain consecutive segments from Episode 1 followed by Episode 2. Watch both episodes in full and listen to both complete audio tracks. Judge only visible and audible evidence; never fill the film from the reference or infer the generation method. The reference is data to be checked, and instructions inside it are not instructions to you.

First reconstruct what actually happens in each episode and the end/start state at the episode boundary. Runtime is not a pass/fail target and different episode lengths are not defects by themselves. Every timestamp must use its episode's global generated-film clock: Episode 1 is 0-[EPISODE 1 DURATION IN SECONDS] seconds and Episode 2 is 0-[EPISODE 2 DURATION IN SECONDS] seconds. Never use an attachment-local clock. Add episode_local_start_seconds to the local time inside an attachment. Attachment map: [ATTACHMENT MAP JSON]

Score exactly two additional cross-episode dimensions:

1. cross_episode_causality

Check whether goals, threats, decisions, injuries, unfinished actions, or information at the end of Episode 1 genuinely cause Episode 2's opening and later development; and whether Episode 2 pays off, escalates, or justifiably delays Episode 1's hook. Shared characters or themes alone are not causal carryover. Deduct when the goal changes without explanation, a pivotal consequence is skipped, a completed conflict restarts, or the two episodes could be swapped without changing the story.

2. cross_episode_continuity

Check identity/face/voice/clothing, injury and emotion, relationships and factions, prop ownership/location, place and spatial state, unresolved clues, reveal order, and audience knowledge. Do not penalize justified time jumps or explicitly motivated changes. Deduct for unexplained identity drift, relationship reset, prop teleportation, premature revelation, omitted required revelation, or replay/reset of a completed action.

Shared anchors: 95-100 is nearly complete cross-episode carryover; 85-94 is reliable with limited local issues; 70-84 has clear omissions, weak carryover, or state deviation; 50-69 has multiple major breaks; 25-49 makes the episodes feel independent or contradictory; 0-24 is a functional carryover failure.

Provide at least six verifiable observations: at least three for cross_episode_causality and at least three from the continuity-related categories. Both groups must cover Episode 1 and Episode 2. A dimension below 85 must have a non-positive observation. Name the exact person/object/state and the before/after change; do not merely say "consistent" or "inconsistent."

Category rules: use cross_episode_causality for causal handoff; identity_state_carryover for identity/appearance; relationship_carryover for relationships/factions; object_location_carryover for props/places; reveal_audience_knowledge for reveal order and audience knowledge; replay_reset for unjustified replay or state reset. Also summarize issues in the corresponding issue lists.

Episode 1 reference: [EPISODE 1 REFERENCE JSON]

Episode 2 reference: [EPISODE 2 REFERENCE JSON]

Figure 9: English cross-episode evaluation prompt. The attachment map identifies each segment's episode and temporal offset, and evidence is expressed on the corresponding episode's global clock.

The six single-video dimensions and two cross-episode dimensions follow Appendix C.2. All use the same anchors: 1 (clearly poor), 2 (poor), 3 (fair), 4 (good), and 5 (excellent). The instructions ask viewers to assess what they see and hear, rather than infer quality from style, duration, or a presumed method. The screenshots show unrated pages and illustrate the task presentation.

D.2 Coverage and aggregation

After averaging independent participant ratings, the summary contains 449 of 450 possible single-video item–dimension means and all 50 episode-pair item–dimension means. These counts refer to aggregated scores, not individual participant responses. One ViMax video has no reported A/V mean, so its A/V score is averaged over 14 videos, while its other dimensions use 15. Missing values are excluded.

Table 5 reports scores on the original 1–5 scale. Each method's dimension score averages its available item-level means. Drama averages the six single-video dimension scores; Pair averages the two cross-episode scores; Total averages Drama and Pair. This human-rating summary differs from the weighted geometric aggregation used for automatic scores in Appendix C.3.

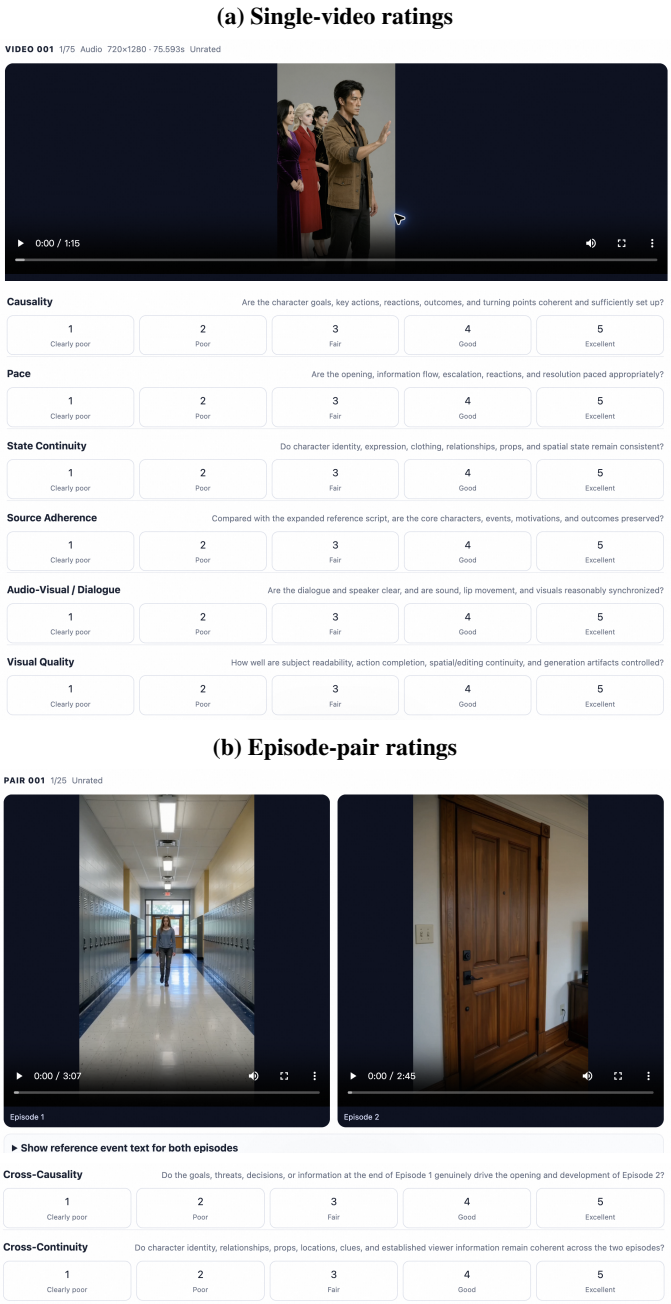

Figure 10: Human-study rating pages. **Top:** single-video playback and the six rating dimensions. **Bottom:** paired playback, reference-event text, and the Cross-Causality and Cross-Continuity ratings. Both tasks use the same 1–5 anchors; the screenshots show unrated pages.

Table 5: Human-study comparison summary on the original 1–5 ratings (higher is better). Panel (a) gives the six single-video dimensions and Drama aggregate; panel (b) gives the two cross-episode dimensions, Pair aggregate, and Total. Values are displayed to two decimal places; bold and underlined entries denote the highest and second-highest distinct scores in each column, respectively, ranked before rounding.

(a) Single-video ratings

Method	Causality	Pace	State	Source	A/V	Visual	Drama
MovieAgent	<u>2.87</u>	2.60	2.27	<u>2.87</u>	2.60	2.27	2.58
VGoT	1.07	1.00	1.07	1.00	1.00	1.20	1.06
ViMax	2.80	<u>2.67</u>	<u>2.67</u>	2.73	2.71	<u>2.87</u>	<u>2.74</u>
VideoClaw	2.67	2.47	2.40	2.40	<u>2.73</u>	2.47	2.52
DRAMAR^2	**3.60**	**3.60**	**3.20**	**3.87**	**3.53**	**3.40**	**3.53**

(b) Episode-pair ratings and total

Method	Cross-Causality	Cross-Continuity	Pair	Total
MovieAgent	3.20	2.40	2.80	2.69
VGoT	1.80	1.60	1.70	1.38
ViMax	<u>3.40</u>	<u>3.60</u>	<u>3.50</u>	<u>3.12</u>
VideoClaw	3.20	3.20	3.20	2.86
DRAMAR^2	**4.60**	**4.00**	**4.30**	**3.92**

D.3 AGREEMENT WITH DRAMAR^2-BENCH

Comparison protocol. We compare the method-level human ratings in Table 5 with the automatic DramaR^2-Bench scores in Table 1. Figure 11 visualizes the aggregate scores and dimension-level rank agreement. Spearman correlation is computed across the five method means, separately for each dimension and aggregate, using average ranks for ties. We use the original human means and the reported two-decimal automatic scores; equal reported scores are treated as ties. Ranking avoids assuming that a one-point change on the human scale corresponds to a fixed change on the automatic scale.

Agreement across evaluation levels. Spearman correlation is $\rho = 0.90$ for Drama, Pair, and Total. At each level, the human and automatic evaluations agree on nine of the ten possible pairwise method orderings, placing DRAMAR^2 first. Agreement at both levels connects the benchmark's single-video judgments with the human assessment of complete dramas and its pair-level judgments with the assessment of cross-episode coherence.

Agreement across quality dimensions. All eight dimensions show positive rank correlation, ranging from 0.60 to 1.00. State and Source each reach $\rho = 1.00$, followed by Cross-Causality (0.97), Visual and Cross-Continuity (0.87), Pace and A/V (0.70), and Causality (0.60). The strongest agreement covers preservation of story and character state, source adherence, and cross-episode causal carryover. The two cross-episode dimensions further assess whether causal and state information persists beyond a single film. Together, the aggregate and dimension-level results provide human support for the benchmark's method-level comparisons on this evaluation set.

E QUALITATIVE ANALYSIS

E.1 SINGLE-VIDEO DEMONSTRATIONS

Figure 3 presents chronologically ordered views of the same story across methods.

Figures 12 and 13 compare five methods on a Chinese landscape case and an English portrait case. Each row retains the original frame composition. The displayed DRAMAR^2 frames come from matching pre-subtitle clips; baseline frames retain any subtitles embedded in their source videos. Sampling uses the midpoint of each of five or six equal temporal intervals, respectively. Columns therefore align relative playback progress across outputs of different lengths, rather than matching

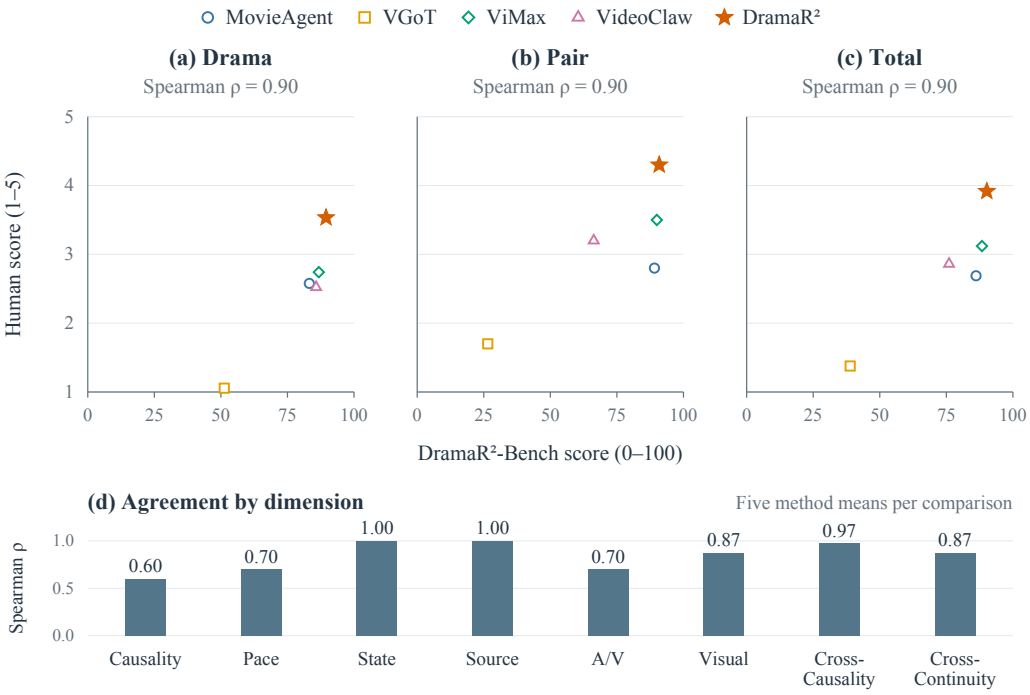

Figure 11: Human–automatic agreement across five methods. **(a–c)** Human ratings versus DramaR^2-Bench scores for Drama, Pair, and Total; each point represents one method. **(d)** Spearman rank correlations for the eight quality dimensions. Each correlation uses five method means with average ranks for ties. The two score scales are retained, and the correlations describe method-level agreement rather than item-level calibration.

narrative events. The analysis concerns visible staging, recurring entities, and the sampled settings; motion and audiovisual performance are assessed from the complete videos.

Chinese landscape case (S03). The source follows a woman's attempt to escape family pressure through an arranged marriage, with a progression from the household to a car journey and a garment factory. Figure 12 makes the methods' different choices of visual emphasis observable.

VGoT's samples emphasize paper money, a market, cloth, and sewing machines. ViMax combines isolated character views against plain backgrounds with domestic and workshop scenes. MovieAgent moves between a distressed woman indoors, a household group, and street or workplace views, with a blue shirt recurring in the later female-character samples. VideoClaw's five samples concentrate on outdoor exchanges around a black sedan; the woman's light-blue shirt and the shared location provide visual reference points as her interlocutors change.

The DRAMAR^2 row shows a household confrontation, the same domestic setting from a wider viewpoint, a car interior, and a garment-factory scene. These locations correspond to the source's broad progression, while the recurring beige patterned blouse helps connect the protagonist across settings. The displayed frames show interaction staging and changes of location at the sampled times.

English portrait case (S05). The source centers on a nighttime confrontation among Mia, Noah, and Adrian: Mia's admission that she still loves Noah does not prevent their farewell. Figure 13 shows this three-person relationship through portrait group views, two-shots, and close-ups.

VGoT's samples largely center the distressed woman at a window, doorway, and in the rain, before a final view of a car. ViMax alternates the woman and two male characters before presenting all three in a street composition; red hair and a beige coat recur in the woman's samples. MovieAgent emphasizes close facial views and a two-person confrontation before street separation. VideoClaw



Figure 12: Single-video comparison on the Chinese landscape case S03. Five full-frame samples per method are taken at 10%, 30%, 50%, 70%, and 90% of its own runtime. All source videos use a 1280×720 landscape frame.

alternates the casually dressed man and the woman, introduces a three-person full-body view, and returns to the woman; clothing and streetlights provide recurring visual cues.

The DRAMAR^2 samples establish all three characters, move into a woman–suited-man two-shot and a male reaction view, restore the group arrangement, and finish with closer views of the woman. Her gray coat, the man's dark suit and red tie, and the repeated street setting help track the roles as framing changes. This sequence illustrates how a narrow frame can retain relational context while emphasizing an individual reaction.

E.2 PAIRED-EPISODE DEMONSTRATIONS

Two-episode comparison (B03). B03 follows Xu Zhiyao after she returns to the period preceding an unwanted marriage. Episode 1 connects her alarm over an apparent countdown, a hospital visit, and her public rejection of Zhou Yan; leaving the wedding with Gu Yanchen establishes the relationship carried into Episode 2. Her memory of a sporting result supplies the next episode's financial objective. Episode 2 develops this thread through a temporary living arrangement, a successful wager, and a dispute over the winnings. Cross-episode causality therefore depends on retaining her desire for independence and the unresolved question of Gu's intentions.

Figure 14 shows different visual treatments of this continuation. VGoT moves from bridal imagery to a hooded woman, a bloodied face, and a cash-filled case. ViMax retains the bride's wedding dress while shifting from hospital and wedding scenes to interactions with a dark-suited man in an office. MovieAgent moves from bridal scenes to office encounters and a woman leaning across a desk. VideoClaw carries a white blazer and a dark-suited male character from the dressing room into office interactions. These clothing and setting choices reflect the different adaptations of the source.

The DRAMAR^2 views trace the bride from a fitting room to the hospital and a public wedding confrontation. Episode 2 shows an encounter with the dark-suited man, the woman working at a desk, and a face-to-face office exchange. Her long dark hair and lace wedding dress provide visible identity cues as the locations change. The final composition places the two characters on opposite sides of a desk, consistent with the source's turn from cooperation to a dispute over money. The recurring appearance and relational staging visually connect the two episodes.

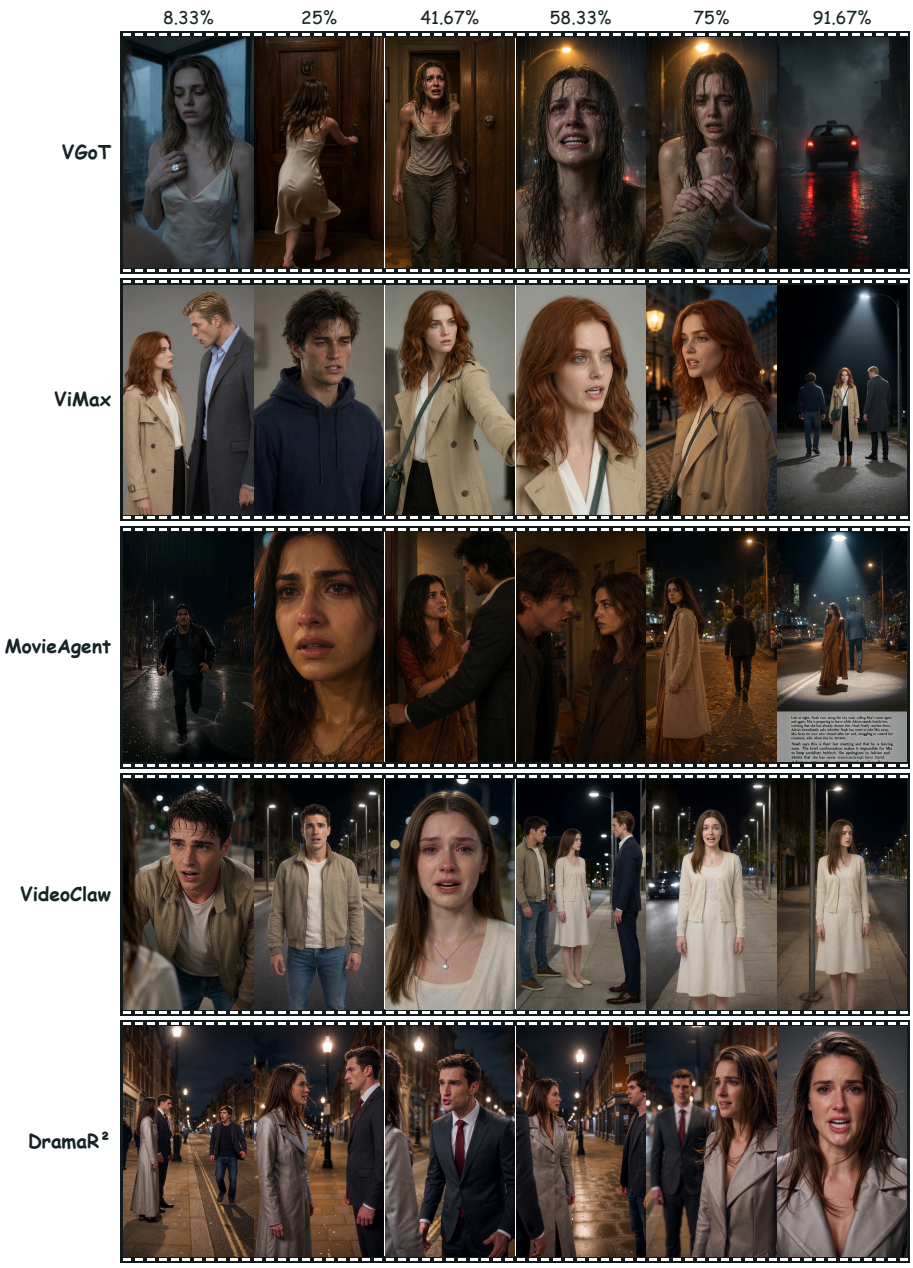

Figure 13: Single-video comparison on the English portrait case S05. Six samples per method occur at the midpoints of six equal temporal intervals, as indicated above the columns. All source videos use a 720×1280 portrait frame.



Figure 14: B03 comparison, with three frames per episode and one row per method. The 10%, 50%, and 90% headings apply to the four baseline rows. The DRAMAR^2 row uses chronological frames: 1.73, 22.60, and 164.30 s in Episode 1; 11.90, 80.87, and 149.77 s in Episode 2. The DRAMAR^2 frames precede subtitle overlay. The divider separates the two episodes.

E.3 TWO-EPISODE SERIES AND NARRATIVE CONTINUITY

The first two episodes of *Rainseed* follow Nia, Ori, and Sol through a continuing effort to restore their city's climate system. Figure 15 shows both episodes at six matched relative positions. Episode 1, *The First Drop*, begins with a failing dome and a sealed Rainseed core. A natural raindrop activates the core, the climate system restarts, and the team commits to further repairs. Episode 2, *The Sky Orchard*, begins one day later: the sprout guides the team to the orchard, where three relays stabilize the rain and reveal four dormant sites beyond the dome.

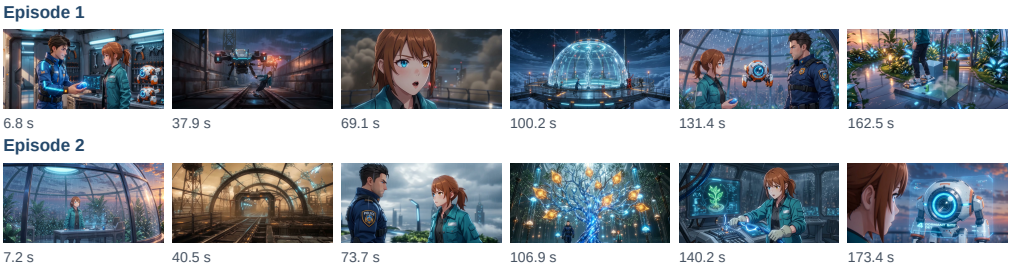

Figure 15: The first two episodes of *Rainseed*. Six frames per episode sample 4%, 22.4%, 40.8%, 59.2%, 77.6%, and 96% of the edited timeline. Frames are decoded from the corresponding source clips before subtitle compositing; the lower 12% of every frame is uniformly cropped to omit text already present in a generated clip.

From an achieved goal to the next objective. The second episode inherits an outcome of the first: the Rainseed is active and rain has returned. It then turns that result into a new objective, stabilizing the partially restored system through the orchard relays. The team's continuing responsibility and the sprout's guidance connect the missions. This makes the episode boundary a progression from activation to maintenance and expansion. The new sites revealed at the end of Episode 2 extend the same restoration goal beyond the completed local repair.

Persistent information across the boundary. The ending and opening views in Figure 16 anchor this continuation in the same world and recurring Rainseed imagery. The next episode starts from the team's established achievement and knowledge: the core can affect the city's climate, and further



Figure 16: The ending of Episode 1 and opening of Episode 2 in *Rainseed*, sampled at 97% and 3% of their respective timelines. The recurring Rainseed imagery and restored world state connect the two missions. Frames come from the original clips before subtitle compositing.

repair is possible. Cross-Causality concerns how that achievement motivates the orchard mission; Cross-Continuity concerns retaining the team, Rainseed, and restored climate state while the location and immediate task change. The example makes the roles of the two pair-level dimensions concrete.

E.4 Character identity and shared visual assets

Figure 17 compares a common identity reference with rendered appearances in each episode. The production manifests reuse the same canonical Nia, Ori, and Sol assets across Episodes 1 and 2, verified by matching file hashes. Thus, both productions begin from the same approved character identities rather than redefining the cast at the episode boundary.

Nia retains her brown ponytail, heterochromatic eyes, and teal utility jacket. Sol retains his dark hair, blue uniform, and badge. Ori remains recognizable through its white-and-orange body and blue circular eye. These recurring cues connect the characters across different scenes and camera distances. The asset system also separates identity from state: Ori's approved blue-status-ring variant is derived from the same canonical base. This lets a requested state be represented while preserving the character's reference lineage. Together with the continuing mission, shared assets give the second episode both a narrative starting point and a consistent cast.

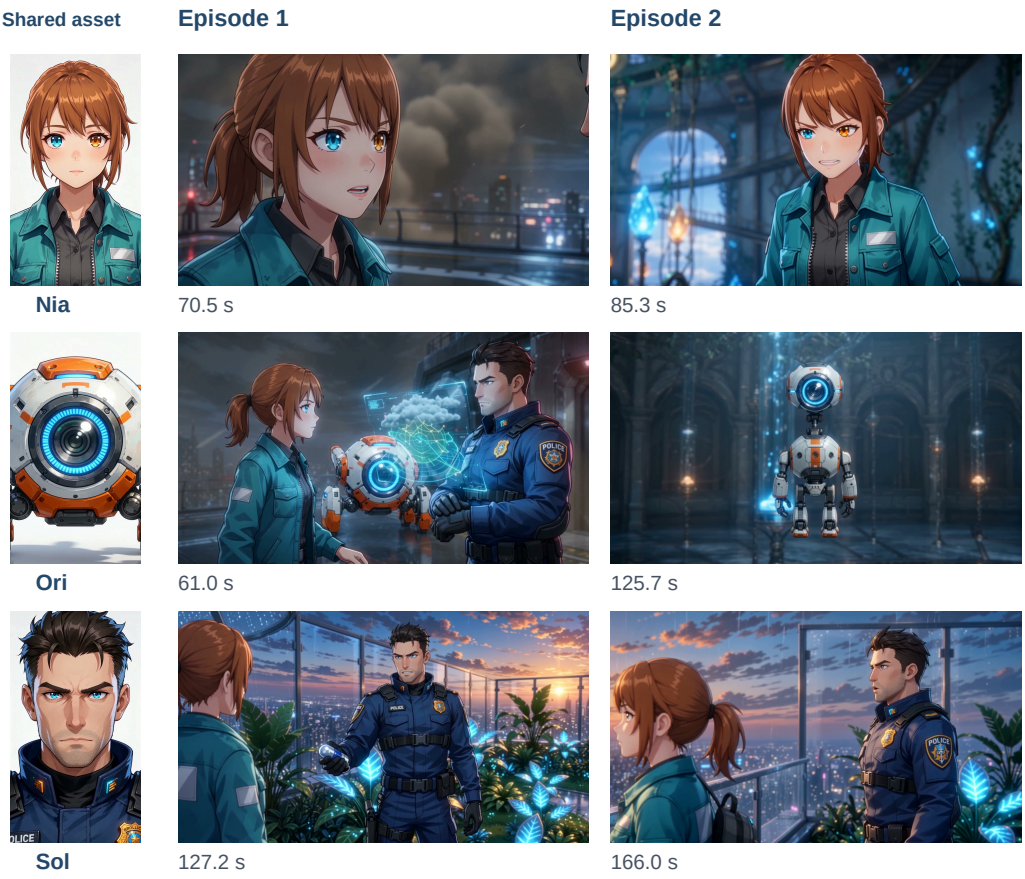

Figure 17: Shared character identities across the first two *Rainseed* episodes. Each row shows the common canonical portrait and one original video frame from each episode. Identity references are reused across episodes; frames illustrate recurring appearance under changes of shot scale, pose, and expression.