REVI: RECURSIVE RECAP VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Novels tell rich stories, but their textual form cannot directly present the sights and sounds of the worlds they describe. Film adaptations offer an audiovisual experience but require substantial production effort and viewing time; film recaps are shorter but depend on existing footage. Neither route directly provides a concise audiovisual account of a novel. We therefore study *novel-to-recap* video generation: producing a self-contained, source-grounded audiovisual retelling directly from a novel within a fixed viewing-time budget. This task requires compressing the story while preserving causal connections, allocating sufficient time to retained events, and aligning narration with newly generated visuals. We propose **ReVi**, a dual-loop multi-agent framework guided by *edit first, generate second*. Its Inner Loop plans story content, narration time, and corresponding shots before video generation. Across novels, its Outer Loop performs recursive self-improvement by revising both the editing policy and the meta policy that governs its revision. Experiments on two evaluation sets show that ReVi improves recap quality and duration compliance over representative baselines, with human ratings and multi-round improvement results providing further support. Our code is available at `https://anonymous.4open.science/r/ReVi-84EF`.

1 INTRODUCTION

For centuries, novels have conveyed human experience through rich plots, complex characters, and imagined worlds, but their textual form cannot directly present the sights and sounds of those worlds (Forster, 1927). Advances in audiovisual technology brought literary narratives to the screen through film and television adaptations, giving audiences a direct sensory experience. Producing such adaptations requires substantial resources and coordinated effort (Zuo et al., 2026), while watching them in full demands a considerable time commitment. Movie recap videos offer a shorter way to follow these screen narratives by combining selected clips from an existing film with voice-over narration that summarizes its main storyline (Shaar et al., 2026). Unlike the original film, which tells the story through scenes and dialogue, a recap uses narration to summarize key events and explain what happens in scenes it leaves out. Clips from the film help illustrate the story. These recaps have become popular on video-sharing platforms such as YouTube (Shaar et al., 2026; Ding et al., 2025).

Existing agentic work on narrative video production focuses on two tasks: *novel-to-film* and *film-to-recap*. In *novel-to-film*, multi-agent workflows turn literary narratives into cinematic sequences by developing scripts, storyboards, and shot plans (Hu et al., 2024; Wu et al., 2025; Wang et al., 2026). They coordinate visual and audio generation through cinematic planning, continuity management, and iterative refinement (Huang et al., 2026a;b; Zuo et al., 2026). In *film-to-recap*, existing work covers different parts of condensing a film or television episode into a short video. Narrative video summarization focuses on identifying key story events and selecting the scenes or shots that depict them (Papalampidi et al., 2021; Xie et al., 2023; Singh et al., 2024), while narrated-video editing pipelines write summaries, synthesize voice-overs, and align narration with footage to assemble a recap (Barua et al., 2025; Ding et al., 2025; Zhou et al., 2026).

However, neither route directly meets the need to help audiences grasp a novel's story quickly and intuitively within a fixed viewing-time budget. As shown in Figure 1, *Novel-to-film* turns literary material into a full dramatized adaptation, but producing that adaptation is costly and time-consuming (Huang et al., 2026b; Zuo et al., 2026). *Film-to-recap* produces an edited video, but it is limited to the scenes and shots available in the source video (Ding et al., 2025; Zhou et al., 2026). This motivates *novel-to-recap* video generation: producing a self-contained audiovisual retelling directly

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.

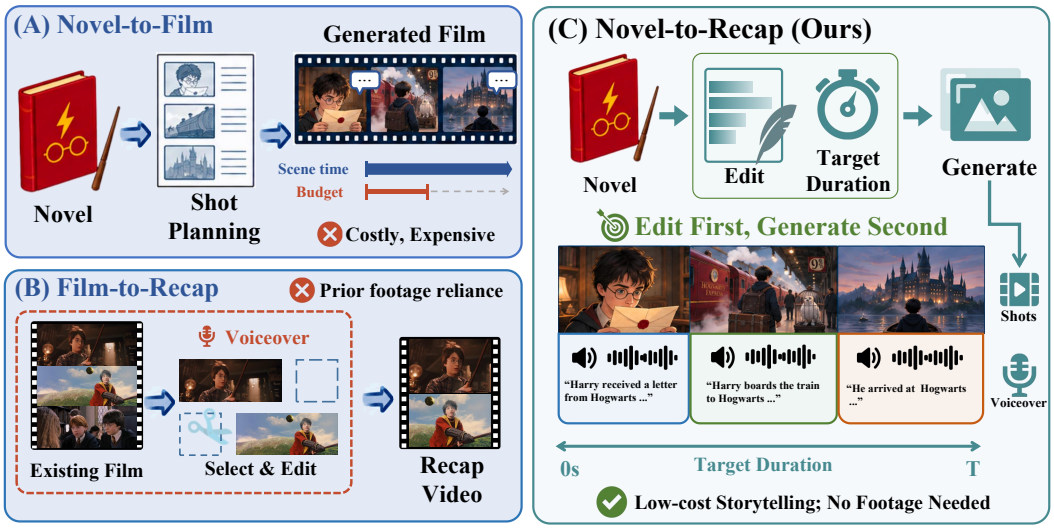

Figure 1: **From novels to recap videos.** (A) Novel-to-film generates a full video adaptation. (B) Film-to-recap selects and edits scenes from a source video, combining them with voice-over narration. (C) Novel-to-recap produces a narrated audiovisual retelling directly from a novel.

from a novel under a fixed viewing-time budget. Compared with *novel-to-film* and *film-to-recap*, *novel-to-recap* compresses the storyline through narration and generates only the new shots needed to visualize it, without requiring either a full dramatized adaptation or a pre-existing source video.

Realizing this goal poses three challenges. Take the early chapters of *Harry Potter and the Philosopher's Stone* as an example. First, **narrative compression** should preserve the causes, motivations, and setups needed to understand retained events even after large parts of the story are omitted. For example, Harry's journey to Hogwarts becomes difficult to follow if the recap omits the arrival of his Hogwarts letter. Second, **temporal allocation** should give retained events enough time for intelligible narration and perceptible visual action within the fixed viewing-time budget. A recap may, for instance, compress Harry's childhood with the Dursleys into a brief introduction while reserving more time for his arrival at Hogwarts. Third, **audiovisual alignment** should coordinate what is narrated with what is shown, preserving source fidelity and temporal correspondence. Consider the Sorting Hat scene: when the narration says that Harry is placed in Gryffindor, the visuals should show the sorting rather than a generic view of Hogwarts.

To address these challenges, we propose **ReVi**, a dual-loop multi-agent framework for **Re**cursive **Re**cap **Vi**deo Generation, guided by the principle of *edit first, generate second*: determine what to retain, how much time to allocate, and what to show before generating video clips. The two loops operate at different scopes. The Inner Loop constructs a recap video for the current novel, whereas the Outer Loop uses experience across novels to improve both the reusable editing policy and the meta policy that governs its revision. Within the Inner Loop, the **Paper Editor** is guided by the Causal Network Model of narrative comprehension (Trabasso & van den Broek, 1985): it selects, compresses, omits, or merges source events while preserving their causal dependencies. The **Radio Editor** converts the edited story into timed narration and allocates the viewing-time budget across retained events, operationalizing Genette's theory of narrative duration (Genette, 1980). The **AV Planner** organizes the timed narration into Dramatic Beats and maps them to corresponding shots, adapting Event Segmentation Theory (Zacks et al., 2007). These editing agents form a pipeline for generating the recap video, progressing from story selection to timed narration and shot planning. Across novels, the Outer Loop performs bounded recursive self-improvement by aggregating production and repair traces to refine both the editing policy and the meta policy governing failure diagnosis and policy revision. Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to formulate and study *novel-to-recap* video generation: the task of producing a fixed-duration, source-grounded audiovisual recap directly from a novel through concise narration and newly generated visuals.

- We propose **ReVi**, a dual-loop multi-agent framework with feedback at two scopes: the Inner Loop plans and repairs each recap, while the Outer Loop uses experience across novels to improve both the editing policy and the meta policy that governs its revision.
- Experiments show that ReVi improves recap quality and duration compliance over agentic video-generation baselines. Component ablations examine its editing stages, while paired repeated searches show separate gains from editing-policy and meta-policy revision.

2 METHOD

Given a source novel x and a recap contract c, ReVi produces a self-contained audiovisual recap y. The contract specifies a target duration B, audience, narrative focus, spoiler policy, output language, and visual style. For example, c may request a 180-second English recap for viewers unfamiliar with the novel, covering the full plot including the ending, and rendered in a Studio Ghibli-inspired anime style. As shown in Figure 2, ReVi follows two coupled loops: the Inner Loop constructs and executes an audiovisual edit for the current novel, while the Outer Loop uses production experience and improvement history to revise the reusable policies governing both editing and policy improvement.

2.1 INNER LOOP: RECAP VIDEO GENERATION

The Inner Loop follows an *edit first, generate second* workflow for each novel. Before generating clips, the Paper Editor selects the events needed for a coherent recap, the Radio Editor writes and times the narration, and the AV Planner specifies the corresponding on-screen actions and shots. These stages must be coordinated under the fixed duration budget: retaining an event may require time to explain its causes, while showing its consequences requires enough screen time for the action to be understood. The resulting story, narration, and shot plans guide clip generation and allow coverage or timing problems to be addressed before footage is produced.

Novel Preprocessing. A novel distributes story information across long passages and may reveal events out of chronological order. Therefore, the three editors need a compact, source-verifiable representation that exposes the dependencies among events before editing begins. The Story Mapper preprocesses the source novel x into a source-linked StoryGraph $G = (V, E)$. First, it identifies story events as nodes $e_i \in V$, each recording its participants, motivations, actions, outcomes, and supporting passages. Second, it adds edges in E for causal dependencies, prerequisites, setup–payoff relations, and temporal or character-state transitions. Finally, inspired by Bordwell's three-dimensional model of narrative, which distinguishes the story world, plot structure, and narration (Bordwell, 2008), it records both the chronological order in which events occur and the disclosure order in which the novel presents them. The resulting StoryGraph provides shared evidence for deciding what can be omitted and which explanations must survive compression. Because every downstream artifact retains links to the source events and upstream decisions from which it was derived, an observed failure can be traced through the narration and visual plan to its editorial origin.

Paper Editor. The StoryGraph exposes the novel's complete event structure, but a fixed-duration recap cannot present every event. The Paper Editor must therefore decide what survives compression without removing the causes, motivations, or setups that make the retained storyline intelligible. Given the StoryGraph G and recap contract c, it produces an ordered Paper Edit P whose units p_i retain or combine source events, or supply source-supported bridges and externalized explanations. This process is guided by the Causal Network Model of narrative comprehension, in which an event becomes important through its causal connections and its position in the story's causal chain (Trabasso & van den Broek, 1985). For example, if a recap retains Harry's arrival at Hogwarts, it must also retain the arrival of his Hogwarts letter and Hagrid's revelation that he is a wizard, or replace the omitted material with a brief, source-supported bridge; otherwise, his journey lacks an explanation.

First, the Paper Editor identifies events central to the requested narrative focus and traces their causal prerequisites, motivations, and setup–payoff links in G. Second, it assigns KEEP, COMPRESS, OMIT, or MERGE to each event or event chain. If an omission breaks a necessary connection, BRIDGE inserts a concise, source-supported explanation; if understanding depends on an internal state, EXTERNALIZE assigns that information to narration. Finally, a fixed closure checker verifies source support, prerequisite coverage, setup–payoff relations, and reveal order, returning any broken

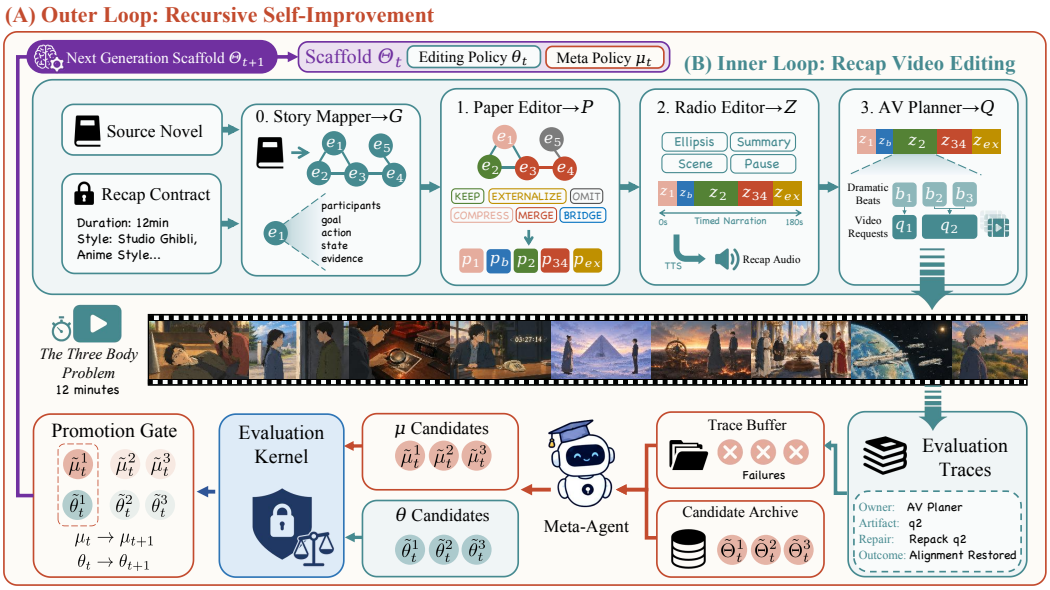

Figure 2: **Overview of the ReVi framework.** (A) The Outer Loop proposes revisions to the editing policy θ_t and meta policy μ_t for recursive self-improvement. (B) The Inner Loop passes each recap through the editing agents in sequence before video generation.

dependency for revision. Once these checks pass, P records the selected event sequence, its source links, and the operation applied to each unit. This gives the Radio Editor a compact, causally coherent account to schedule, preventing missing explanations from surfacing only after timing or video generation. Appendix B defines the operations.

Radio Editor. The Paper Edit determines which events survive compression, but it does not determine how much narration time each event should receive. Under a fixed viewing-time budget, treating all retained events equally would either exceed the budget or compress important developments until they become difficult to follow. Given the Paper Edit P and recap contract c, the Radio Editor produces timed narration Z as source-linked segments z_i anchored to the relevant edit units in P. This process is grounded in Genette's theory of narrative duration, which distinguishes *ellipsis, summary, scene, and pause* according to the relation between story time and narrative time (Genette, 1980). For example, a recap of the first *Harry Potter* novel may use ellipsis to skip an uneventful interval, summarize Harry's childhood with the Dursleys, present the Sorting Ceremony as a scene, and pause briefly to explain the Hogwarts houses.

First, the Radio Editor assigns a narration mode to each unit in P and allocates the available time according to its narrative function and information density. Here, SCENE denotes comparatively detailed treatment rather than literal equality between story time and playback time. Second, it writes the narration, with each segment recording its source events, introduced facts, assigned mode, and text. Finally, text-to-speech converts the draft into measured timestamps and checks whether its total duration satisfies the budget B. If the narration is overlong, the Radio Editor first removes redundant wording; persistent overload is returned to the Paper Editor for structural compression. If time remains, it expands necessary context or gives key events fuller treatment. The resulting Z is a source-linked, budget-compliant narration with measured timestamps, providing the AV Planner with a stable temporal backbone for shot planning.

Audio-Visual Planner. Timed narration specifies what is said and when, but its sentence boundaries do not necessarily correspond to perceptible events or feasible shots. The AV Planner must therefore decide how each narrated development should unfold on screen while keeping the visuals synchronized with the narration. Given the StoryGraph G, timed narration Z, and recap contract c, it produces an audiovisual plan Q. This process is grounded in Event Segmentation Theory, which holds that viewers maintain a representation of the current event and perceive a boundary when pre-

diction error triggers an update to that representation (Zacks et al., 2007). For example, a single sentence may narrate the Sorting Hat considering Slytherin, announcing Gryffindor, and the hall erupting in applause. Each change in action or outcome updates the viewer's understanding of the event, so the planner represents these moments as separate *Dramatic Beats*.

First, the AV Planner segments Z at perceptible changes in goals, actions, outcomes, character focus, time, or location, forming dramatic beats b. Each Beat records its source events, initial and outcome states, narration anchor, screen interval, and continuity group. Second, the planner determines the division of information between modalities: narration carries internal states, background, and elapsed time, while visuals depict selected observable actions and state changes. It then maps each Beat to one or more shot intents. Finally, it groups compatible shots into timestamped generation requests q with visual references, checking compatibility in characters, setting, state transitions, timing, and continuity. If an action cannot be presented intelligibly within its assigned interval, the planner requests revised timing or upstream compression. The resulting Q provides a source-grounded, temporally aligned blueprint that guides a video generation model in producing clips.

Repair and Trace Collection. After clip generation and assembly, ReVi repairs the current recap and records each production attempt in a trace buffer $\mathcal{T}$. First, a fixed temporal critic checks event coverage, event order, narration–visual alignment, and compliance with the target runtime. Second, each detected failure is converted into a repair request containing the supporting evidence and affected artifact. The orchestrator routes this request to the earliest responsible stage: source reconstruction, content selection, temporal allocation, or visual planning. Third, the responsible stage revises its artifact under the fixed editing policy θ_t, after which dependent outputs are recomputed and affected clips are regenerated while unaffected artifacts are reused. This cycle continues until all checks pass or the repair budget is exhausted. Finally, ReVi records the decisions, intermediate artifacts, critic feedback, repairs, and outcomes from every attempt as a production trace. The repairs update only the current recap; the collected traces provide evidence for the Outer Loop.

2.2 Outer Loop: Recursive Self-Improvement

The Inner Loop's production traces feed the Outer Loop. At generation t, ReVi maintains the system state in Equation equation 1:

$$\Theta_t = (\theta_t, \mu_t), \tag{1}$$

where θ_t is the editing policy and μ_t guides its improvement. Both consist of versioned prompt instructions, examples, and configuration rules. First, the Meta-Agent uses the trace buffer $\mathcal{T}$ and archive $\mathcal{A}$ of prior revision attempts to propose candidate states $\widetilde{\Theta}_t^{(i)}$, each changing one policy at a time. Second, the frozen Evaluation Kernel E^* supplies scores for the fixed promotion rules. Finally, ReVi records each attempt and its outcome in $\mathcal{A}$; accepted revisions form the next state Θ_{t+1}.

Editing Policy. The editing policy in Equation equation 2 is

$$\theta_t = (\pi_t^S, \pi_t^P, \pi_t^R, \pi_t^{AV}, \rho_t), \tag{2}$$

where the first component governs source interpretation, the next three govern the editing agents, and ρ_t governs coordination and repair. Editable rules determine which motivations and dependencies to recover, which events to retain or compress, how to assign narration modes and time, how to define and group Beats, and when to request upstream revision. A patch may strengthen the condition for retaining a payoff's setup, change how explanatory passages receive time, or revise when dense Beats are split. Such changes persist across novels. Evolution revises how the editors apply them.

Editing-Policy Evolution. The Meta-Agent aggregates traces across novels and diagnoses recurring failures by inspecting source evidence, intermediate edits, and attempted repairs. It proposes editing-policy candidates $\widetilde{\theta}_t^{(i)}$, each specifying a target rule, triggering condition, revised instruction or configuration, expected effect, and constraints. An alignment failure, for example, can originate in visual realization, insufficient screen time, or excessive narration density. This use of execution feedback follows reflective prompt optimization (Agrawal et al., 2026), with revisions restricted to ReVi's editable policies. Candidates undergo structural and narration checks before expensive rendering; changes affecting visual coverage or timing are evaluated through generated outputs. The

fixed kernel assesses source support, causal closure, comprehension, duration, audiovisual alignment, and cost. Only candidates satisfying its acceptance constraints are eligible for promotion under a fixed quality–cost rule and margin. The selected editor becomes θ_{t+1}; if none passes, then $\theta_{t+1} = \theta_t$. Every attempt is archived with its diagnosis, patch, outcome, regressions, and cost, so later searches can examine whether an earlier intervention supported its diagnosis.

Meta Policy. The meta policy μ_t determines how to diagnose failures, propose patches, and allocate a fixed search budget. Its updates address recurring problems in improvement itself: incorrect attribution, ineffective revisions, or repeated spending on unproductive candidates. A diagnosis rule may require checking missing causal information before blaming rendering; a proposal rule may favor conditional instructions over generic prompt expansion; an allocation rule may favor testing alternative Paper Edits over repeated video generation. A new diagnosis for one failure is an output of μ_t; changing the reusable diagnostic procedure updates μ_t itself. The current meta policy uses production traces and archived improvement attempts to propose these revisions, following self-referential improvement approaches (Fernando et al., 2024; Zelikman et al., 2024).

Meta-Policy Evaluation. To evaluate a candidate meta policy $\widetilde{\mu}_t^{(j)}$, ReVi compares it with μ_t under matched inputs and search budget; its index j is independent of the editing-policy candidate index i. First, each comparison episode $h \in \mathcal{E}_{\text{meta}}$ fixes a parent editor θ_h, production traces $\mathcal{T}_h$, improvement history $\mathcal{A}_h$, and search budget K. Holding the tested meta policy μ fixed within the episode, the search selects an editor using development and promotion data under the fixed acceptance procedure:

$$\widehat{\theta}_h(\mu) = \text{Search}_\mu(\theta_h, \mathcal{T}_h, \mathcal{A}_h; K). \tag{3}$$

Second, the editor $\widehat{\theta}_h(\mu)$ selected in Equation equation 3 is frozen and evaluated on the episode's held-out novels $\mathcal{D}_h^{\text{held}}$. Writing J_{E^*} for the frozen kernel's utility relative to the parent editor θ_h, ReVi averages these scores across episodes:

$$V(\mu) = \frac{1}{|\mathcal{E}_{\text{meta}}|} \sum_{h \in \mathcal{E}_{\text{meta}}} J_{E^*}\left(\widehat{\theta}_h(\mu); \theta_h, \mathcal{D}_h^{\text{held}}\right). \tag{4}$$

Thus Equation equation 4 evaluates a search procedure through the downstream improvement of the editors it produces, as in Hyperagents (Zhang et al., 2026b). Finally, the promotion gate compares the candidate with μ_t using this score. The candidate becomes μ_{t+1} only if it passes the fixed promotion criterion without increasing permitted constraint violations; otherwise $\mu_{t+1} = \mu_t$. Together with the accepted editor θ_{t+1}, it forms $\Theta_{t+1} = (\theta_{t+1}, \mu_{t+1})$, which becomes the current scaffold in the next generation. Its accepted meta policy governs subsequent editor searches and further revisions to the improvement procedure. Task contracts, evaluation definitions, and final acceptance rules remain fixed, with a sealed test partition reserved for assessing the resulting system. If a proposed editor revision fails acceptance, $\theta_{t+1} = \theta_t$.

3 Experiments

3.1 Experimental Settings

Datasets and duration contracts. We evaluate recap generation across difficulty-graded narratives and longer novels using FilmEval and Recap32. **FilmEval** (Zuo et al., 2026) provides source narratives grouped into Easy, Medium, and Hard tiers by narrative complexity. We use these source texts as recap inputs and assign a duration target to each tier. **Recap32** draws on longer English Gutenberg novels from NarrativeQA (Kočiský et al., 2018). The protocol assigns separate novels to policy evolution, editor promotion, meta-policy evaluation, and testing. We analyze the 12 designated test novels under a 180-second duration target.

Metrics. Three Gemini 3.7 Flash judgments score each completed video on three quality dimensions. **Recap Fidelity (RF)** combines weighted coverage of essential source events, the logical consistency of their relationships, and resistance to unsupported content, using a shared source reference. **Narrative Pacing (NP)** scores how clearly the finished recap develops its story, including plot progression, transitions, and the distribution of information over playback time. **Audio-Visual**

Table 1: **Recap quality and duration compliance.** RF/NP/AVG are quality scores; DE/DP are percentages (DP pass counts in parentheses). Best and second-best are bold and underlined.

Method	RF ↑	NP ↑	AVG ↑	DE ↓	DP ↑
FilmEval: 15 sources; 90/120/150 s					
Direct LLM	74.46	66.51	61.62	226.70	0.0% (0/15)
ViMax	82.09	78.44	<u>78.66</u>	<u>108.71</u>	0.0% (0/15)
VideoClaw	81.11	<u>79.20</u>	77.30	145.24	0.0% (0/15)
MovieAgent	<u>83.72</u>	78.61	74.73	160.34	0.0% (0/15)
MM-StoryAgent	75.57	73.77	69.90	219.40	0.0% (0/15)
VGoT	77.48	76.20	72.65	163.53	0.0% (0/15)
ReVi	**91.58**	**90.20**	**88.67**	**2.37**	**100.0% (15/15)**
Recap32: 12 test novels; 180 s					
Direct LLM	68.95	66.40	63.25	215.72	0.0% (0/12)
ViMax	78.30	76.35	<u>74.07</u>	<u>125.20</u>	0.0% (0/12)
MovieAgent	<u>80.99</u>	<u>77.50</u>	72.30	191.31	0.0% (0/12)
ReVi	**90.68**	**88.25**	**82.88**	**2.07**	**100.0% (12/12)**

Table 2: **Human evaluation of recap quality.** Mean ± SD of individual FilmEval ratings (0–5). Best and second-best means are bold and underlined.

Method	RF ↑	NP ↑	AVG ↑
Direct LLM	3.44 ± 0.51	3.00 ± 0.34	3.06 ± 0.42
ViMax	<u>4.17</u> ± 0.51	<u>3.94</u> ± 0.64	<u>4.00</u> ± 0.49
VideoClaw	3.78 ± 0.43	3.83 ± 0.62	3.78 ± 0.43
MovieAgent	4.06 ± 0.42	3.50 ± 0.62	3.39 ± 0.50
MM-StoryAgent	3.33 ± 0.49	3.44 ± 0.51	2.94 ± 0.42
VGoT	3.67 ± 0.49	3.67 ± 0.49	3.56 ± 0.51
ReVi	**4.44** ± 0.51	**4.56** ± 0.51	**4.50** ± 0.62

Grounding (AVG) measures the share of expected narration segments whose visuals support or complement the corresponding spoken content. We assess temporal control separately: **Duration Error (DE)** is the absolute difference between final and target playback times, expressed as a percentage of the target, and **Duration Pass Rate (DP)** is the share of evaluated outputs that both complete and satisfy the prescribed duration tolerance. Keeping timing separate from quality makes coverage gains interpretable under the same viewing-time contract.

Baselines and scoring. ViMax (Huang et al., 2026b) is the primary system-level baseline. The FilmEval comparison also includes VideoClaw (HITsz-TMG, 2026), MovieAgent (Wu et al., 2025), MM-StoryAgent (Xu et al., 2025), and VGoT (Zheng et al., 2025). Direct LLM creates narration and a storyboard in a single pass before common rendering. Recap32 compares Direct LLM, ViMax, MovieAgent, and ReVi. All methods receive the same output request; baselines receive the target duration only in the prompt. The appendix gives more details.

3.2 Main Comparison

Recap quality and duration compliance. Table 1 compares ReVi with Direct LLM and agentic video-generation baselines on FilmEval and Recap32. Each method follows the same recap task and output specification, and we assess completed videos for Recap Fidelity, Narrative Pacing, Audio-Visual Grounding, and Duration Compliance. ReVi leads on the quality dimensions across both collections. The strongest baseline varies by dimension, indicating that plot accuracy, intelligibility, and visual correspondence place distinct demands on recap generation. ReVi's joint gains are consistent with its connected sequence of causal selection, timed narration, and shot planning. The appendix reports a shared-narration control (Table 7) and component interventions.

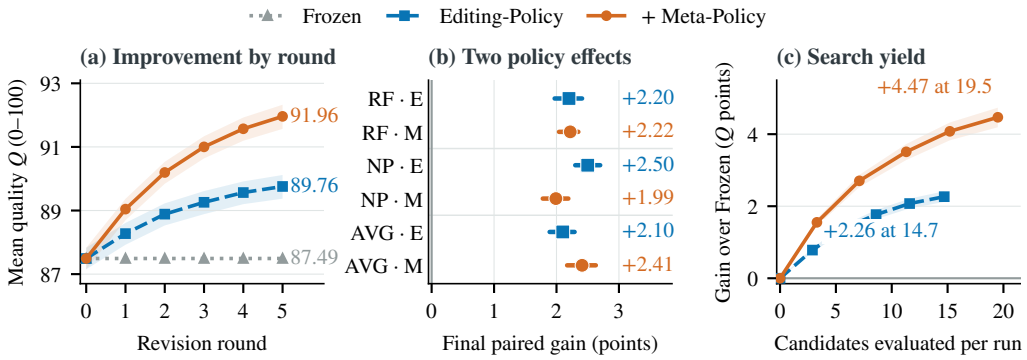

Figure 3: **Outer Loop evaluation.** Ten paired searches are compared over five revision rounds. (a) Mean quality $Q = (\mathrm{RF} + \mathrm{NP} + \mathrm{AVG})/3$ for Frozen, Editing-Policy, and Editing-Policy + Meta-Policy. (b) Final run-paired differences for editing-policy updates (E: Editing-Policy minus Frozen) and added meta-policy updates (M: Editing-Policy + Meta-Policy minus Editing-Policy). (c) Mean Q gain over Frozen versus cumulative candidates evaluated per run.

Duration under a fixed viewing time. To assess duration control, we compare each completed video's playback length with its assigned target and report DE and DP alongside the quality scores in Table 1. ReVi meets the requested duration in the supplied comparisons and achieves higher quality with substantially shorter playback than the baselines. Longer outputs offer additional room to mention events and extend scenes; ReVi's joint quality and duration results point to efficient allocation of the available viewing time. This pattern is consistent with the Radio Editor's allocation of narration time before rendering and its use of measured speech timing to refine the plan. Under a fixed budget, causal links require deliberate selection and enough narration time to remain intelligible.

3.3 Human Evaluation

To test whether the automatic quality gains carry over to human viewing, we ask raters to evaluate source-matched FilmEval recaps from ReVi and the evaluated baselines. Every method is judged on the same source set. Raters have access to the source text and shared event reference, while method identities are hidden and presentation order is randomized. They score recap fidelity, narrative pacing, and audio-visual grounding on a common scale; Table 2 summarizes their ratings across methods. ReVi receives the highest ratings in each dimension, with the most pronounced separation in narrative pacing. This pattern is consistent with the editing pipeline: causal event selection preserves the context needed to understand later developments, timed narration regulates how that context unfolds, and beat-aligned shots reinforce the spoken account. The joint preference across dimensions supports the interpretation that the automatic gains reflect qualities perceptible in the finished recaps. The appendix details the rating aggregation, source-paired comparisons, and alignment between automatic scores and human judgments.

3.4 Recursive Policy Improvement

Figure 3(a) shows that recap quality rises as the editing policy is revised, while the Frozen condition remains unchanged. Allowing meta-policy revision produces a steeper trajectory and a higher final quality level. The gains on six held-out Recap32 novels are consistent with policy changes that carry beyond the recaps used during search. The run-paired comparisons in Figure 3(b) show positive effects from editing-policy revision across RF, NP, and AVG. Meta-policy revision adds gains in each dimension. Figure 3(c) further shows greater quality improvement at comparable candidate counts when the meta policy can evolve, indicating more productive editor search under the measured protocol. Duration error also declines across the conditions while evaluated outputs continue to satisfy the duration contract. Together, these results support ReVi's bounded recursive self-improvement claim at both the editing-policy and meta-policy levels.



Figure 4: **Story progression in *The Necklace*.** ReVi completes the story within the duration contract; ViMax exceeds it and repeats shots marked by red outlines.

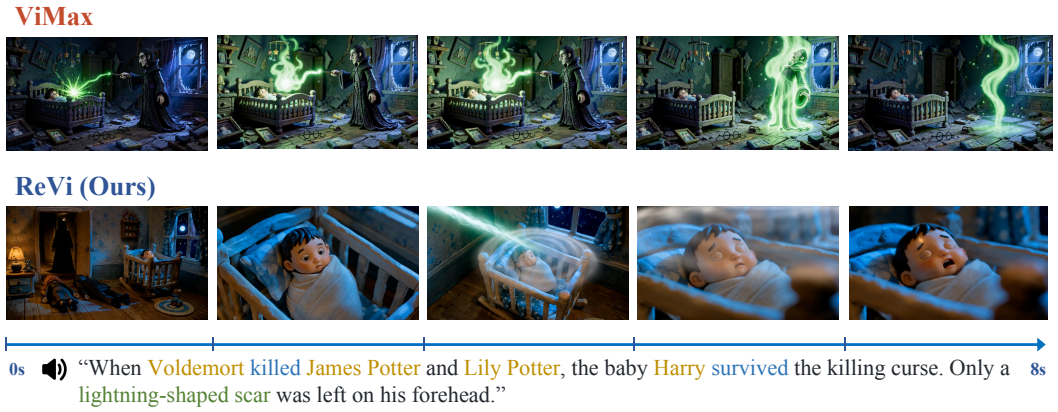

Figure 5: **Comparison of shots in generated recaps.** Frames partly selected from the generated recap video of *Harry Potter and the Philosopher's Stone*.

3.5 Qualitative Case Study

Figure 4 compares complete recaps of *The Necklace* against a 150-second target. ReVi reaches the concluding encounter within the budget; ViMax repeats highlighted shots and runs over time. This illustrates how event selection and timing affect story progression.

Figure 5 shows an eight-second passage from *Harry Potter and the Philosopher's Stone*. ViMax holds a wide view of Voldemort and the crib, while ReVi cuts from the fallen adults and infant Harry to the curse and Harry's reaction, matching distinct narrated Dramatic Beats.

4 Conclusion

We introduced ReVi for fixed-duration audiovisual recaps of novels. Its inner loop plans causal coverage, timed narration, and matching shots; its outer loop revises reusable editing and meta policies. Across two evaluation sets, ReVi improves measured quality and duration compliance. Human ratings and repeated policy-search results provide further evidence. Future work should test more novels and independent generations, and strengthen human assessment of recap fidelity.

AI USE STATEMENT

Generative AI models are components of the video-production pipelines and automated evaluation described in Section 3.1. We use AI to polish paper writing and assist with finding related work. The authors take responsibility for the methodology, reported results, figures, and final text.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.19457.

Aadit Barua, Karim Benharrak, Meng Chen, Mina Huh, and Amy Pavel. Lotus: Creating short videos from long videos with abstractive and extractive summarization. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712090. URL https://arxiv.org/abs/2502.07096.

David Bordwell. Three dimensions of film narrative. In *Poetics of Cinema*. Routledge, 2008. URL https://www.davidbordwell.net/books/poetics_03narrative.pdf.

Zihan Ding, Xinyi Wang, Junlong Chen, Per Ola Kristensson, and Junxiao Shen. Prompt-driven agentic video editing system: Autonomous comprehension of long-form, story-driven media. *arXiv preprint arXiv:2509.16811*, 2025. URL https://arxiv.org/abs/2509.16811.

Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13481–13544. PMLR, 2024. URL https://proceedings.mlr.press/v235/fernando24a.html.

E. M. Forster. *Aspects of the Novel*. Harcourt, Brace and Company, New York, 1927. URL https://www.gutenberg.org/ebooks/70492.

Gérard Genette. *Narrative Discourse: An Essay in Method*. Cornell University Press, Ithaca, NY, 1980.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Interactive story visualization with multiple characters. In *SIGGRAPH Asia 2023 Conference Papers*. Association for Computing Machinery, 2023. doi: 10.1145/3610548.3618184. URL https://doi.org/10.1145/3610548.3618184.

HITsz-TMG. VideoClaw: AI creative video production agent. GitHub repository, 2026. URL https://github.com/HITsz-TMG/VideoClaw.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. StoryAgent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*, 2024. URL https://arxiv.org/abs/2411.04925.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/36b7acf6f6010652b3f2a433774a66fe-Abstract-Conference.html.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. FilMaster: Bridging cinematic principles and generative AI for automated film generation. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fb8e52adcd9b59bad73f109c53afc43a-Abstract-Conference.html.

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

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6536–6558. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.488. URL https://aclanthology.org/2022.findings-emnlp.488/.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. StoryGAN: A sequential conditional GAN for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6329–6338, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Li_StoryGAN_A_Sequential_Conditional_GAN_for_Story_Visualization_CVPR_2019_paper.html.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via LLM-guided planning. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2309.15091.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arık. VISTA: A test-time self-improving video generation agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6021–6032, 2026. URL https://arxiv.org/abs/2510.15831.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, Yanhong Zeng, Xing Zhu, Yujun Shen, and Huamin Qu. HoloCine: Holistic generation of cinematic multi-shot long video narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 461–471, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Meng_HoloCine_Holistic_Generation_of_Cinematic_Multi-Shot_Long_Video_Narratives_CVPR_2026_paper.html.

Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1920–1933. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.174. URL https://aclanthology.org/2020.acl-main.174/.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie summarization via sparse graph construction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13631–13639, 2021. doi: 10.1609/aaai.v35i15.17607. URL https://ojs.aaai.org/index.php/AAAI/article/view/17607.

Shaden Shaar, Bradon Thymes, Sirawut Chaixanien, Claire Cardie, and Bharath Hariharan. MovieRecapsQA: A multimodal open-ended video question-answering benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4537–4546, 2026. URL https://arxiv.org/abs/2601.02536.

Aditya Kumar Singh, Dhruv Srivastava, and Makarand Tapaswi. "Previously on . . . " From Recaps to Story Summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13635–13646, 2024. URL https://arxiv.org/abs/2405.11487.

Yale Song, Yiwen Song, Nick Losier, Nathan Hodson, Ye Jin, Rhyard Zhu, Yan Xu, Daniel Vlasic, Carina Claassen, Jasmine Leon, Khanh G. LeViet, Zack Chomyn, Joe Timmons, Brett Slatkin, Scott Penberthy, and Tomas Pfister. Co-Director: Agentic generative video storytelling. *arXiv preprint arXiv:2604.24842*, 2026. URL https://arxiv.org/abs/2604.24842.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Tom Trabasso and Paul van den Broek. Causal thinking and the representation of narrative events. *Journal of Memory and Language*, 24(5):612–630, 1985. doi: 10.1016/0749-596X(85)90049-X.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vOEXS39nOF.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. MAViS: A multi-agent framework for long-sequence video storytelling. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2273–2295. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.101. URL https://aclanthology.org/2026.eacl-long.101/.

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.

Jiehang Xie, Xuanbai Chen, Tianyi Zhang, Yixuan Zhang, Shao-Ping Lu, Pablo Cesar, and Yulu Yang. Multimodal-based and aesthetic-guided narrative video summarization. *IEEE Transactions on Multimedia*, 25:4894–4908, 2023. doi: 10.1109/TMM.2022.3183394. URL https://ir.cwi.nl/pub/31736.

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*, 2025. URL https://arxiv.org/abs/2503.05242.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.06072.

Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event perception: A mind/brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007. doi: 10.1037/0033-2909.133.2.273.

Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. Self-taught optimizer (STOP): Recursively self-improving code generation. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.02304.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin gödel machine: Open-ended evolution of self-improving agents. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2505.22954.

Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob Foerster, Jeff Clune, Minqi Jiang, Sam Devlin, and Tatiana Shavrina. Hyperagents. *arXiv preprint arXiv:2603.19461*, 2026b. URL https://arxiv.org/abs/2603.19461.

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*, 2025. URL https://arxiv.org/abs/2512.19539.

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. In *NeurIPS Workshop on What Makes a Good Video: Next Practices in Video Generation and Evaluation*, 2025. URL https://what-makes-good-video.github.io/assets/10_VideoGen_of_Thought_Step_by.pdf.

Hengji Zhou, Lingxuan Huang, Jian Wang, Bing Zhou, Si Wu, Lianghao Xia, and Chao Huang. VideoAgent: All-in-one framework for video understanding and editing. *arXiv preprint arXiv:2606.23327*, 2026. URL https://arxiv.org/abs/2606.23327.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3501. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c7138635035501eb71b0adf6ddc319d6-Abstract-Conference.html.

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 Related Work

Narrative Video Generation Narrative video generation extends prompt-conditioned synthesis (Yang et al., 2025; Team Wan et al., 2025) toward variable-length sequences (Villegas et al., 2023), holistic multi-shot generation (Meng et al., 2026), and memory-conditioned shot synthesis (Zhang et al., 2025). Story visualization addresses visual consistency across textual narratives (Li et al., 2019; Gong et al., 2023; Zhou et al., 2024). Language-guided planning (Lin et al., 2024) and multi-agent production (Hu et al., 2024; Wu et al., 2025; Wang et al., 2026; Huang et al., 2026b) connect these capabilities to coordinated storytelling workflows, with specialized approaches to cinematic control (Huang et al., 2026a), world-state modeling (Zuo et al., 2026), and narrated storybooks (Xu et al., 2025). Feedback-driven systems further refine video prompts (Long et al., 2026) or combine global creative-direction search with local multimodal refinement (Song et al., 2026). ReVi shares these planning and feedback mechanisms, but focuses on compressing an existing novel into a self-contained recap, jointly constraining causal content, narration duration, and visual coverage.

Narrative Summarization Narrative summarization selects essential content while accounting for long-range causal and temporal dependencies. BOOKSUM studies summaries at paragraph, chapter, and book levels (Kryscinski et al., 2022), while screenplay summarization uses narrative structure to identify salient content (Papalampidi et al., 2020). In video, PlotSnap and TaleSumm use television recaps to study summarization of existing episodes (Singh et al., 2024), and MovieRecapsQA evaluates open-ended question answering from recap visuals and aligned movie dialogue (Shaar et al., 2026). ReVi extends the content-selection problem to generating new narration and footage directly from a novel, requiring explicit temporal allocation and audiovisual realization within a fixed runtime.

Recursive Self-Improvement. Self-improvement ranges from revising individual outputs (Madaan et al., 2023; Long et al., 2026) to optimizing reusable prompts (Agrawal et al., 2026) and agent designs (Hu et al., 2025). STOP and the Darwin Gödel Machine explore recursive program improvement and coding agent evolution (Zelikman et al., 2024; Zhang et al., 2026a). Promptbreeder improves both task prompts and their mutation prompts (Fernando et al., 2024), while Hyperagents makes both the task agent and its improvement mechanism editable (Zhang et al., 2026b). ReVi instantiates this distinction for recap editing: the Inner Loop repairs the current video, and the Outer Loop revises editing and meta policies using experience across novels. Given reported limitations of intrinsic self-correction on reasoning tasks (Huang et al., 2024), a frozen external Evaluation Kernel governs successor acceptance, and meta policies are compared through the editors they produce under matched search budgets.

B Additional Method Details

B.1 Complete Algorithm

Algorithm 1 ReVi: editing and recursive policy improvement

Require: Development/promotion data $\mathcal{D}_{\text{dev}}, \mathcal{D}_{\text{prom}}$; initial policies Θ_0;
meta-policy episodes $\mathcal{E}_{\text{meta}}$; frozen kernel E^*;
production/evolution budgets and search budget K
Ensure: Accepted policy state $\Theta_t = (\theta_t, \mu_t)$
$t \leftarrow 0$; $\Theta_t = (\theta_t, \mu_t) \leftarrow \Theta_0$; trace buffer $\mathcal{T} \leftarrow \emptyset$; candidate archive $\mathcal{A} \leftarrow \emptyset$
while evolution budget remains **do**
  for all $(x, c) \in \mathcal{D}_{\text{dev}}$ **do** ▷ Inner Loop: fixed θ_t
    $G \leftarrow \textsc{StoryMapper}(x, c; \theta_t)$ ▷ source events e_i
    $P \leftarrow \textsc{PaperEditor}(G, c; \theta_t)$ ▷ edit units p_i
    $Z \leftarrow \textsc{RadioEditor}(P, c; \theta_t)$ ▷ timed segments z_i
    $Q \leftarrow \textsc{AVPlanner}(G, Z, c; \theta_t)$ ▷ beats b_i, requests q_j
    $(y, F) \leftarrow \textsc{RenderCheck}(G, P, Z, Q, c)$
    while $F \neq \emptyset$ and repair budget remains **do**
      Revise the earliest responsible stage under ρ_t; recompute dependent artifacts
      $(y, F) \leftarrow \textsc{RenderCheck}(G, P, Z, Q, c)$
    end while
    Record the production trace in $\mathcal{T}$, including unsuccessful attempts
  end for
  $\theta_{t+1} \leftarrow \text{Search}_{\mu_t}(\theta_t, \mathcal{T}, \mathcal{A}; K)$
    ▷ test $\widetilde{\theta}_t^{(i)}$ with E^* on $\mathcal{D}_{\text{prom}}$; retain θ_t if none passes
  $\mu'_t \leftarrow \textsc{ProposeMeta}(\mu_t, \mathcal{T}, \mathcal{A})$ ▷ a $\widetilde{\mu}_t^{(j)}$ candidate
  Evaluate $V(\mu'_t)$ and $V(\mu_t)$ on $\mathcal{E}_{\text{meta}}$ with matched inputs and budget K
  $\mu_{t+1} \leftarrow \text{Promote}_{E^*}(\mu_t, \mu'_t; V)$
  Archive tried candidate states $\widetilde{\Theta}_t^{(i)}$ and their outcomes in $\mathcal{A}$
  $\Theta_{t+1} \leftarrow (\theta_{t+1}, \mu_{t+1})$; $t \leftarrow t + 1$
end while
return Θ_t

B.2 Editorial Operations and Narration Modes

Table 3 defines the transformations recorded in the Paper Edit. Removing presentation detail is permitted when the information needed to understand retained developments survives elsewhere in the account. In particular, BRIDGE supplies a source-supported connection across omitted material, while EXTERNALIZE assigns an internal state or implicit motivation to narration without inventing an observable action. The closure checker examines prerequisites, setup–payoff relations, source support, and disclosure order after revision.

Table 4 gives the operational adaptations of Genette's duration categories. These modes determine the granularity of narration under the recap budget. The SCENE adaptation permits comparatively detailed treatment without requiring exact equality between story time and playback time. The Radio Editor may elaborate retained material when time remains, but cannot independently restore events excluded by the Paper Edit.

Table 3: Paper Edit operations and the information they preserve.

Operation	Editorial function	Information to preserve
KEEP	Retain an event as a distinct unit.	Its necessary action, motivation, or outcome.
COMPRESS	Shorten an event or a contiguous event chain.	Entry conditions, consequential changes, and outcomes.
OMIT	Remove material from the recap.	The prerequisites of subsequent retained events.
MERGE	Combine overlapping or repetitive material.	Distinct causal contributions of the combined events.
BRIDGE	Insert a concise, source-supported connection across omitted material.	The relation needed to understand the transition.
EXTERNALIZE	Assign an internal state or implicit motivation to narration.	Its narrative role without inventing an observable action.

Table 4: Narration modes adapted from Genette's account of narrative duration.

Mode	Role in recap narration
ELLIPSIS	Pass over an interval already excluded by the Paper Edit, using a brief temporal connection when needed.
SUMMARY	Condense a retained event group into economical narration.
SCENE	Give a key event a more detailed, comparatively direct presentation.
PAUSE	Explain a necessary motivation, rule, or background fact without advancing the narrated action.

B.3 POLICY REVISIONS AND SEARCH ACCOUNTING

Table 5 summarizes the persistent targets of editing-policy revision. A candidate patch records its target component, triggering condition, changed instruction or configuration, expected effect, and constraints. Single-component changes are preferred when they isolate the suspected cause; coordinated changes to adjacent stages are considered when evidence implicates their interaction.

Structural and narration checks can reject candidates with broken dependencies or infeasible speech duration before rendering. These checks allocate evaluation resources; they do not replace assessment of the final video. The frozen Evaluation Kernel applies fixed acceptance constraints, a quality–cost selection rule, and a promotion margin. A search that finds no qualifying candidate retains its parent policy.

Production traces retain inputs, decisions, artifacts, critic feedback, repairs, and outcomes. The improvement archive separately records the parent policy, diagnosed cause, patch, expected benefit, evaluation outcome, regressions, and cost of each attempt. The Meta-Agent can inspect these records but cannot rewrite them. This separation allows meta-policy revision to use evidence about both production failures and unsuccessful attempts to correct them.

For meta-policy comparisons, each candidate receives the same parent editor, accessible traces, improvement history, and total search budget. Search accounting includes diagnosis, proposal generation, candidate evaluation, and rendering. Meta-policy development and comparison costs are recorded separately. Editor selection uses the episode's development and promotion data; the selected editor is frozen before evaluation on its held-out novels. The protocol reserves a sealed test partition outside policy search for assessing the resulting system. Foundation models, source novels, task contracts, evaluation definitions, and final acceptance rules remain outside the mutable policies throughout evolution.

C ADDITIONAL EXPERIMENTAL DETAILS

Datasets. We use FilmEval's 15 narrative texts and its Easy, Medium, and Hard difficulty tiers (Zuo et al., 2026), with five sources per tier. For the recap-generation protocol, we assign

Table 5: Persistent modification targets of the editing policy.

Policy component	Persistent modification target
Source interpretation and causal compression	Rules for identifying necessary motivations and dependencies, prioritizing retained events, and choosing KEEP, COMPRESS, OMIT, MERGE, BRIDGE, or EXTERNALIZE.
Temporal allocation	Rules for assigning duration modes, balancing information density, distributing telling time, and requesting structural compression.
Audiovisual coordination	Rules for dividing information between narration and visuals, defining Beat boundaries, selecting visual coverage, and grouping Beats into feasible requests.
Repair	Conditions for local correction, upstream revision, regeneration, and reuse of unaffected artifacts.

Table 6: Recap32 novel allocation. Test IDs N01–N12 follow the listed title order; main-comparison results cover these 12 novels.

Partition	Titles
Evolution (12)	*Percival Keene*; *The Rise of Silas Lapham*; *Lorna Doone*; *Five Children and It*; *Amelia*; *Jennie Gerhardt*; *Anne of Green Gables*; *The Prairie*; *The Odd Women*; *Moonfleet*; *The Riddle of the Sands*; *The Fifth Queen*.
Editor promotion (4)	*Brewster's Millions*; *Beyond the Rocks*; *The Mysterious Affair at Styles*; *Rebecca of Sunnybrook Farm*.
Meta evaluation (4)	*The Golden Bowl*; *The Iron Heel*; *The Mysterious Island*; *The Mill on the Floss*.
Test (12)	`N01` *Marianela*; `N02` *The Hour of the Dragon*; `N03` *Lothair*; `N04` *Indiscretions of Archie*; `N05` *Seventeen*; `N06` *McTeague*; `N07` *Comrades: A Story of Social Adventure in California*; `N08` *Moll Flanders*; `N09` *The Fortune of the Rougons*; `N10` *The House of the Seven Gables*; `N11` *Tarzan of the Apes*; `N12` *Mary Barton*.

duration targets of 90, 120, and 150 seconds to the respective tiers. For Recap32, we selected 32 book-length English Gutenberg entries from NarrativeQA (Kočiský et al., 2018). The 12/4/4/12 allocation specifies evolution, editor promotion, meta-policy evaluation, and testing, with a 180-second target. Table 6 lists the selected titles and maps test IDs N01–N12 to novels in the listed order. The reported main comparison covers these 12 test IDs; the other 20 titles are assigned to policy-development partitions.

Metrics. All three quality scores lie in $[0, 100]$. **Recap Fidelity (RF)** is defined as the mean of weighted essential-event coverage (WEC), logic reliability (LR), and hallucination resistance (NHR), using a shared duration-specific source reference. WEC gives full, half, or zero credit to covered, partially covered, or missing events, with event weights of 1, 2, or 3. LR and NHR assess logical errors and unsupported content, respectively. **Narrative Pacing (NP)** evaluates narrative clarity, progression, transitions, and pacing. It accepts fine-grained scores from 0 to 100; the five rubric anchors are not a five-value output scale. **Audio-Visual Grounding (AVG)** is the percentage of narration segments labeled *Support* or *Complement*, relative to all expected segments. It is not the average of the three quality metrics. For final duration L_i and target T_i, **Duration Error (DE)** is $100|L_i - T_i|/T_i$. In the reported data, **Duration Pass Rate (DP)** is the percentage of all records with a complete output and DE at most 5%. All reported outputs completed.

Baselines and implementation. ViMax (Huang et al., 2026b) is the primary system-level baseline. The FilmEval comparison additionally includes VideoClaw (HITsz-TMG, 2026), MovieAgent (Wu et al., 2025), MM-StoryAgent (Xu et al., 2025), and VGoT (Zheng et al., 2025). *Direct LLM* generates narration and a storyboard in one pass before using the common rendering pipeline. Recap32 compares Direct LLM, ViMax, MovieAgent, and ReVi. Baselines receive the requested duration in their prompts, but their original planning and rendering pipelines do not enforce a duration constraint. The experimental configuration uses `gpt-5.6-terra` for planning, `gpt-image-2` for images, Seedance 2.0 (`doubao-seedance-2-0-260128`) for video, and Mureka TTS (`mureka/tts`, Emma, default rate) for narration. The shared output specification is 720p, 16:9, and a Ghibli-style prompt, with video requests of 1–15 seconds. ReVi permits two regen-

erations per asset, first frame, or video segment, and three cross-stage repairs per source. Adapters supply plot briefs and narration where required by a baseline.

Scoring and uncertainty. We use three Gemini 3.7 Flash evaluations of each final output. These calls are repeated judgments, not independent generations. Individual RF component judgments are unavailable for recomputation. Means weight sources equally and are rounded only for display. The source-paired analyses below quantify variation across source texts, not repeated-generation uncertainty.

Inner-loop interventions. We ablate five mechanisms on C1, C3, C6, C8, C11, and C13, two sources per difficulty tier. Each variant changes one mechanism at the same target duration. Each variant is compared with the six Full ReVi records from the same ablation batch. *Causal Editing* replaces causal-chain selection and dependency repair with ordinary summary editing; *TTS Feedback* replaces measured speech duration with word-count estimates; *Narrative Allocation* replaces function-aware modes and time allocation with a single length budget while retaining TTS feedback; *Beat Planning* groups visuals by sentence boundaries instead of event changes; and *Repair Routing* confines revisions to the stage reporting the error.

Record linkage and source metadata. The main comparison comprises 105 FilmEval and 48 Recap32 video-level observations. The inner-loop analysis comprises 36 observations, and the repeated-search analysis comprises 1,080 observations across conditions, rounds, runs, and sources. These are distinct experimental batches: the six Full ReVi scores in the ablation batch differ from those for the same FilmEval IDs in the main comparison. We therefore make ablation contrasts within that batch and analyze repeated-search initial scores separately from the main comparison.

The title mapping links N01–N12 to the test novels in Table 6.

RF aggregation and segment scoring. Each Gemini evaluation scores RF from weighted essential-event coverage, logic reliability, and hallucination resistance. The reported video-level RF is the median of the three complete RF judgments. NP uses a fine-grained 0–100 scale, with 0/25/50/75/100 as qualitative anchors. AVG counts *Support* and *Complement* segments as positive; its denominator is all expected narration segments. Segment labels and per-call judgments are unavailable for independent rescoring.

Resampling and agreement. The source ID is the resampling unit, and each replicate retains all paired methods or rounds. FilmEval resamples five sources within each difficulty tier; inner-loop and human comparisons resample two per tier. Recap32 main comparisons resample its 12 source IDs without strata. Outer-loop intervals resample independent search runs while preserving paired condition outcomes. Each analysis uses 10,000 replicates, NumPy's default generator with seed 20260921, and linear-interpolated 2.5th/97.5th percentiles. Main-comparison intervals are pointwise and quantify source-sampling uncertainty. Outer-loop intervals quantify variation across search runs on the six-source subset. The human analysis keeps all seven methods together within each of six source clusters. Spearman correlations recompute average ranks within each replicate. Regression bands refit human score on automatic score in each replicate and describe the fitted mean, not individual predictions. No independent-video p-values are reported.

Ordinal Krippendorff α uses the ordered domain $\{0, 1, 2, 3, 4, 5\}$ and all three ratings on each of the 42 videos. For pooled category counts n_c, the squared ordinal distance between categories $a < b$ is $[\sum_{c=a}^{b} n_c - (n_a + n_b)/2]^2$. Observed within-video pairs are weighted by $1/(R_i - 1)$, and expected pairs use pooled category counts. We compute $\alpha = 1 - D_o/D_e$. No ratings are missing. Rating SD uses the sample denominator 17 for each method's 18 ratings; it is distinct from the SD of the six video means.

Cost accounting and rounding. Main-comparison expenditure uses per-video generation costs. Repeated-search results report evaluated candidate entries as the measure of search volume in Figure 3(c). Quality means and paired differences are calculated from per-video values before display rounding. Dollar values use decimal half-up rounding, and pass rates retain integer counts. DE uses the reported higher-precision field; runtime rounding to 0.01 seconds accounts for its small recomputation discrepancy.

D Additional Results

D.1 Shared-Narration Duration Control

For each source, a human editor prepared one narration script at the requested duration. All methods used the same audio file for that source to generate their recap videos. This control fixes spoken content and audio timing while retaining each method's video selection, alignment, and rendering. Table 7 reports the completed videos on the same FilmEval and Recap32 sources as the main comparison. Final playback length can still vary across methods, so we report both its mean and the duration-pass count. ReVi retains higher RF, NP, and AVG scores than ViMax under shared narration: the source-paired gains are 6.47/6.12/6.18 points on FilmEval and 6.02/6.72/6.72 points on Recap32, respectively. This control isolates downstream recap-video realization under fixed narration; the main comparison evaluates the complete editing pipeline. As a sensitivity check, restricting to pairs where both methods pass the 5% duration tolerance leaves 13 FilmEval and 11 Recap32 sources; ReVi's RF/NP/AVG gains over ViMax remain 6.63/6.00/6.46 and 6.19/6.68/6.86 points, respectively.

Table 7: Quality under a shared narration track.

Method	RF ↑	NP ↑	AVG ↑	L (s)	DE ↓	DP ↑
FilmEval: 15 sources; 90/120/150 s						
Direct LLM	72.97	69.61	67.80	117.8	4.16	80.0% (12/15)
ViMax	<u>84.33</u>	<u>82.34</u>	<u>81.11</u>	121.3	<u>3.12</u>	<u>86.7% (13/15)</u>
VideoClaw	80.98	79.02	78.29	120.5	4.15	73.3% (11/15)
MovieAgent	82.56	79.66	76.49	118.8	4.23	66.7% (10/15)
MM-StoryAgent	76.46	74.43	72.04	118.1	4.06	66.7% (10/15)
VGoT	78.25	76.58	74.64	120.5	4.69	46.7% (7/15)
ReVi	**90.81**	**88.47**	**87.30**	119.9	**1.95**	**100.0% (15/15)**
Recap32: 12 test novels; 180 s						
Direct LLM	72.38	69.96	67.51	178.9	4.54	41.7% (5/12)
ViMax	<u>83.41</u>	<u>81.74</u>	<u>80.10</u>	178.6	<u>3.15</u>	<u>91.7% (11/12)</u>
MovieAgent	82.18	79.34	76.44	180.7	3.43	83.3% (10/12)
ReVi	**89.43**	**88.46**	**86.82**	180.6	**1.60**	**100.0% (12/12)**

For each source, every method uses the same human-edited narration script and identical audio file. RF/NP/AVG are quality scores; L is mean final playback length; DE is mean absolute duration error in percent; DP is the share within 5% of the target (pass counts in parentheses). Rankings are within each dataset. Bold and underline mark the best and the second-best results, respectively; ties share a rank.

D.2 Inner Loop Ablations

We ablate five mechanisms on six FilmEval texts, two per tier, at the main-comparison target durations. The replacements use ordinary summaries for Causal Editing, word-count timing for TTS Feedback, one narration budget for Narrative Allocation, sentence boundaries for Beat Planning, and stage-local revisions for Repair Routing. Table 8 compares each variant with the Full ReVi records from the same ablation batch. These scores are kept separate from the main-comparison ReVi scores.

Causal selection and narration allocation. Removing Causal Editing produces the largest RF loss, 4.76 points, while NP falls by 1.60 points. This concentration is consistent with the Paper Editor's role of retaining causally necessary events rather than merely shortening the plot. Removing Narrative Allocation most affects NP (-5.35 points), consistent with the need to reserve narration time according to each retained event's function. These interventions affect different aspects of the story plan: deciding what information survives compression and deciding how long the retelling spends on it.

Table 8: Inner-loop component ablations.

Ablation	RF ↑	NP ↑	AVG ↑	DE ↓	DP ↑
Paper Editor					
w/o Causal Editing	87.10	<u>88.92</u>	<u>88.39</u>	<u>2.18</u>	**100.0% (6/6)**
Radio Editor					
w/o TTS Feedback	<u>91.68</u>	87.71	86.82	8.65	0.0% (0/6)
w/o Narrative Allocation	89.84	85.17	86.27	3.69	<u>83.3% (5/6)</u>
AV Planner					
w/o Beat Planning	90.79	88.23	79.99	2.21	**100.0% (6/6)**
Cross-agent orchestration					
w/o Repair Routing	87.23	87.42	85.75	3.44	<u>83.3% (5/6)</u>
Full system					
Full ReVi	**91.87**	**90.52**	**88.92**	**1.86**	**100.0% (6/6)**

Inner-loop ablations on six sources, grouped by the editing agent that owns each local mechanism. Repair Routing evaluates cross-agent orchestration. Bold and underline mark the best and the second-best results, respectively; ties share a rank.

Visual beats and measured speech timing. Replacing event beats with sentence boundaries yields the largest AVG loss (-8.93 points). The spoken sentences remain available, but the visual plan no longer follows event changes as closely. Measured speech timing addresses a different failure mode: without TTS Feedback, RF changes by only -0.18 points, but DE rises by 6.79 percentage points and duration compliance falls from 6/6 to 0/6. Plausible content selection alone therefore does not ensure that the retelling fits the viewing-time contract.

Cross-stage repair. Removing Repair Routing lowers RF and AVG by 4.63 and 3.17 points and leaves one output outside the duration tolerance. This pattern is consistent with errors requiring revisions across stages rather than only at the stage that reports them. The interventions replace one mechanism at a time on a limited source subset and do not identify how pairs of mechanisms interact. Source-paired bounds are reported in Table 17.

D.3 Repeated Outer-Loop Search

Table 9 reports condition–round means for six Recap32 test novels (N02, N04, N06, N08, N10, and N12). The analysis contains ten runs per condition and six sources per run; every reported video completed and passed the duration threshold. Each run reuses its Frozen output for a source across rounds. The search log and policy timeline link each accepted meta-policy version to one candidate that passed checks and was promoted. Editing-Policy raises all three quality dimensions, and Editing-Policy+Meta-Policy adds a further gain in the recorded searches. Table 10 reports final run-paired contrasts. Editing-Policy minus Frozen is 2.20 [1.97, 2.41] RF points, 2.50 [2.30, 2.70] NP points, and 2.10 [1.89, 2.29] AVG points. Editing-Policy+Meta-Policy minus Editing-Policy is 2.22 [2.04, 2.35], 1.99 [1.78, 2.19], and 2.41 [2.17, 2.63] points, respectively. The intervals resample search runs, keeping the paired sources and conditions together; they do not quantify uncertainty over a broader population of novels. Duration pass rate is 100% in all three conditions, so the quality gains do not come from relaxing the stated duration threshold. For the search-yield view in Figure 3(c), we count every candidate entry in the search records, including editor and meta-policy candidates. Editing-Policy evaluates 147 candidates across ten searches, compared with 195 for Editing-Policy+Meta-Policy. At the final checkpoint, their mean gains in $Q = (\mathrm{RF}+\mathrm{NP}+\mathrm{AVG})/3$ over the paired Frozen outputs are 2.26 and 4.47 points, respectively. Figure 3(c) relates quality gains at each checkpoint to the mean cumulative number of candidates evaluated per run.

Table 9: Quality across recursive policy revisions.

Round	RF ↑	NP ↑	AVG ↑	DE ↓
		Frozen		
0	91.05	88.56	82.87	2.35
1	91.05	88.56	82.87	2.35
2	91.05	88.56	82.87	2.35
3	91.05	88.56	82.87	2.35
4	91.05	88.56	82.87	2.35
5	91.05	88.56	82.87	2.35
		Editing policy		
0	91.05	88.56	82.87	2.35
1	91.82	89.42	83.58	2.23
2	92.39	90.12	84.16	2.11
3	92.76	90.54	84.48	2.06
4	93.05	90.86	84.78	2.00
5	93.24	91.05	84.97	1.95
		Editing + meta policy		
0	91.05	88.56	82.87	2.35
1	92.60	90.10	84.43	2.13
2	93.75	91.29	85.55	2.01
3	94.52	92.08	86.41	1.92
4	95.10	92.67	86.95	1.82
5	95.46	93.05	87.38	1.76

Repeated-search evaluation on six Recap32 test novels. Values are pooled means across runs and sources; DE is duration error in percent.

Table 10: Final paired quality differences.

Contrast	RF	NP	AVG
Editor – Frozen	+2.20 [+1.97, +2.41]	+2.50 [+2.30, +2.70]	+2.10 [+1.89, +2.29]
Editor + Meta – Editor	+2.22 [+2.04, +2.35]	+1.99 [+1.78, +2.19]	+2.41 [+2.17, +2.63]

Final-round run-paired quality differences with run-bootstrap intervals. Each run contains the same six Recap32 source IDs under the compared conditions.

D.4 Generation Cost

Table 11 reports expenditure and video volume for the main-comparison outputs. Final denotes completed runtime; Generated includes discarded or regenerated clips. The reported $\overline{G/L}$ averages per-output ratios. Figure 6 shows generation expenditure against RF. The status of outer-loop search accounting is described in Appendix D.3.

Table 11: Generation expenditure and video volume.

Method	Final (s)	Generated (s)	$\overline{G/L}$	Cost ($) ↓	DP ↑
		FilmEval			
Direct LLM	397.7	547.7	1.361	9.04	0.0% (0/15)
ViMax	247.7	339.2	1.362	<u>5.92</u>	0.0% (0/15)
VideoClaw	291.0	392.2	1.342	6.69	0.0% (0/15)
MovieAgent	319.5	436.8	1.348	7.36	0.0% (0/15)
MM-StoryAgent	393.0	516.8	1.316	8.46	0.0% (0/15)
VGoT	312.1	426.3	1.364	7.16	0.0% (0/15)
ReVi	121.8	147.5	1.214	**2.85**	**100.0% (15/15)**
		Recap32			
Direct LLM	537.6	734.8	1.360	11.68	0.0% (0/12)
ViMax	396.2	540.0	1.358	<u>8.76</u>	0.0% (0/12)
MovieAgent	524.4	733.7	1.398	11.36	0.0% (0/12)
ReVi	179.3	210.3	1.172	**3.87**	**100.0% (12/12)**

Main-comparison generation expenditure and video volume. Final and Generated are mean seconds; $\overline{G/L}$ averages per-output ratios. Costs exclude offline search. Rankings apply to cost and DP within each dataset. Bold and underline mark the best and the second-best results, respectively; ties share a rank.

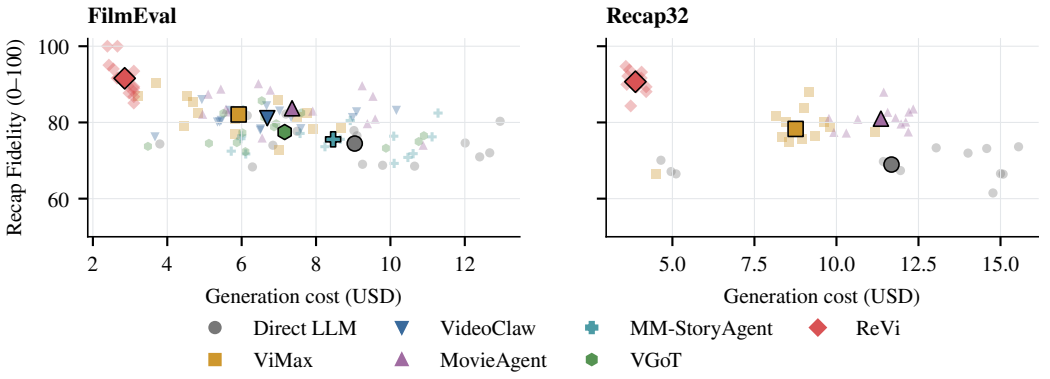

Figure 6: **ReVi has higher mean Recap Fidelity (RF) at lower generation cost than ViMax.** The comparison covers FilmEval and Recap32. Faint markers show individual outputs and outlined markers show method means (15 FilmEval and 12 Recap32 sources per method). Costs exclude offline RSI search.

D.5 Human Evaluation and Metric Alignment

Study set and presentation. The human study uses FilmEval sources C1, C3, C6, C8, C11, and C13: two from each difficulty tier. For each source, raters evaluate one completed recap from each of the seven methods in the main comparison. Method identities are hidden, presentation order is randomized, and raters have access to the source text and common event reference for each rating. Three raters score every source–method pair, giving 42 videos, 126 rating records, and 378 dimension scores.

Rating dimensions and aggregation. The rating records contain three separate integer scores on a 0–5 scale, with higher values indicating better quality. Recap Fidelity (RF) concerns whether the recap preserves important source events and their causal relations without introducing unsupported developments. Narrative Pacing (NP) concerns the clarity and progression of the condensed story, including transitions and narration pace. Audio-Visual Grounding (AVG) concerns whether the displayed shots support or complement the spoken account. A video may therefore score differently across the three dimensions. We first average the three raters' scores for each video, then give the

Table 12: Source-paired human-rating differences.

Baseline	RF	NP	AVG
Direct LLM	+1.00 [+0.67, +1.33]	+1.56 [+1.33, +1.78]	+1.44 [+1.11, +1.78]
ViMax	+0.28 [+0.06, +0.50]	+0.61 [+0.33, +0.89]	+0.50 [+0.11, +0.89]
VideoClaw	+0.67 [+0.44, +0.89]	+0.72 [+0.33, +1.11]	+0.72 [+0.50, +0.94]
MovieAgent	+0.39 [+0.06, +0.72]	+1.06 [+0.94, +1.17]	+1.11 [+0.89, +1.33]
MM-StoryAgent	+1.11 [+0.78, +1.44]	+1.11 [+0.72, +1.50]	+1.56 [+1.22, +1.89]
VGoT	+0.78 [+0.50, +1.06]	+0.89 [+0.56, +1.22]	+0.94 [+0.50, +1.39]

ReVi-minus-baseline human-rating differences and paired intervals on the original 0–5 scale, using six source clusters. Only methods with observed ratings are included.

six sources equal weight when computing each method's mean. Table 2 reports these means and the sample SD of the 18 raw ratings per method.

Figure 7 pairs each automatic score with the same video's mean human rating, scaled by 20. The correlations pool within-method variation and between-method differences. Table 13 reports source-clustered intervals and inter-rater agreement. Pooled Pearson correlations for RF/NP/AVG are 0.831/0.881/0.899, while within-method centered correlations are 0.614/0.688/0.694. Thus, the RF association is not solely due to differences between method means. Ordinal inter-rater α is 0.462/0.535/0.624. For RF, the three raters show partial agreement, and ReVi's source-paired human advantage over ViMax is 0.28 [0.06, 0.50] points (Table 12). Together these results offer limited human corroboration of the automatic RF trend; they do not make automatic and individual human judgments interchangeable. The six-source sample and pooled correlation design limit the strength of the claim.

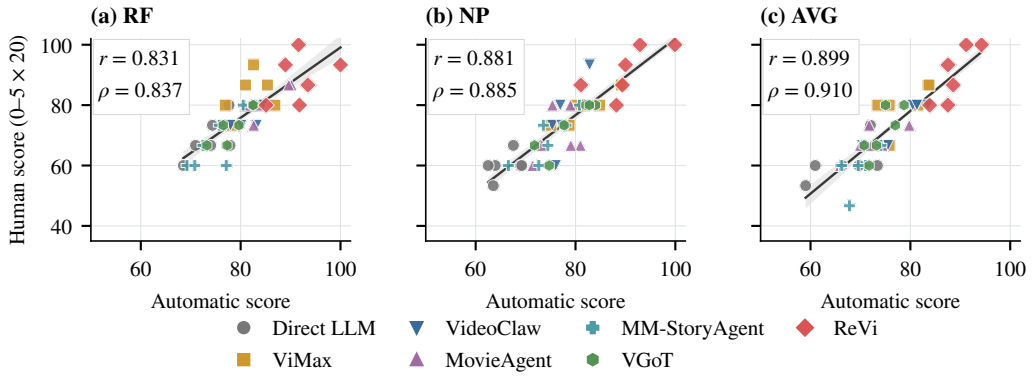

Figure 7: **Automatic scores correlate positively with human judgments.** Results cover 42 videos from seven methods. Panels show Recap Fidelity (RF), Narrative Pacing (NP), and Audio-Visual Grounding (AVG); human means are scaled from 0–5 by 20. Lines and gray bands show linear fits and pointwise 95% source-cluster bootstrap intervals. Insets report Pearson r and Spearman ρ.

Table 12 reports ReVi-minus-baseline rating differences with paired intervals on the original 0–5 scale.

Table 13: Automatic–human metric agreement.

Metric	Pooled r [95% CI]	Within-method r [95% CI]	Pooled ρ [95% CI]	Ordinal α
RF	0.831 [0.819, 0.854]	0.614 [0.401, 0.793]	0.837 [0.782, 0.870]	0.462
NP	0.881 [0.825, 0.919]	0.688 [0.548, 0.770]	0.885 [0.794, 0.932]	0.535
AVG	0.899 [0.875, 0.921]	0.694 [0.615, 0.777]	0.910 [0.872, 0.929]	0.624

Automatic–human association on the 42 rated videos. Within-method r subtracts each method's mean before pooling. Correlation intervals resample the 6 source clusters within difficulty, keeping all 7 methods together and re-centering within each replicate. Ordinal alpha uses the three raw ratings per video.

D.6 Difficulty and Source-Paired Uncertainty

Table 14 reports all three FilmEval tiers, with five completed outputs per method and tier. Comparing methods within each tier shows that ReVi retains its lead as narrative complexity rises. Fidelity generally falls at higher difficulty: denser plots introduce more event dependencies that must survive compression, and a brief setup may explain a later action. The Paper Editor's causal selection targets this pressure. Visual grounding varies less systematically across tiers, consistent with the AV Planner's separate task of aligning narration and shots. Table 15 reports ReVi-minus-baseline quality differences and pointwise source-paired interval bounds. Higher average scores do not imply improvement on every source.

Table 14: Recap quality by narrative difficulty.

Tier	Method	RF ↑	NP ↑	AVG ↑	DE ↓	DP ↑
Easy	Direct LLM	76.81	67.84	64.57	199.06	0.0% (0/5)
Easy	ViMax	84.12	78.44	75.37	<u>91.92</u>	0.0% (0/5)
Easy	VideoClaw	81.99	80.89	<u>79.54</u>	128.44	0.0% (0/5)
Easy	MovieAgent	<u>84.91</u>	<u>81.15</u>	76.94	124.62	0.0% (0/5)
Easy	MM-StoryAgent	75.95	76.11	69.02	162.61	0.0% (0/5)
Easy	VGoT	78.91	77.49	73.92	148.39	0.0% (0/5)
Easy	**ReVi**	**96.14**	**93.34**	**89.39**	**2.01**	**100.0% (5/5)**
Medium	Direct LLM	73.30	69.16	60.46	189.25	0.0% (0/5)
Medium	ViMax	83.39	78.86	<u>81.78</u>	<u>103.05</u>	0.0% (0/5)
Medium	VideoClaw	80.40	<u>80.73</u>	80.57	157.32	0.0% (0/5)
Medium	MovieAgent	<u>84.04</u>	78.30	73.20	160.51	0.0% (0/5)
Medium	MM-StoryAgent	76.65	73.88	68.81	235.69	0.0% (0/5)
Medium	VGoT	78.93	76.09	72.86	159.35	0.0% (0/5)
Medium	**ReVi**	**90.97**	**90.12**	**86.74**	**2.34**	**100.0% (5/5)**
Hard	Direct LLM	73.28	62.54	59.84	291.78	0.0% (0/5)
Hard	ViMax	78.75	<u>78.02</u>	<u>78.85</u>	<u>131.16</u>	0.0% (0/5)
Hard	VideoClaw	80.94	75.99	71.79	149.97	0.0% (0/5)
Hard	MovieAgent	<u>82.22</u>	76.39	74.04	195.89	0.0% (0/5)
Hard	MM-StoryAgent	74.11	71.33	71.88	259.91	0.0% (0/5)
Hard	VGoT	74.59	75.03	71.17	182.85	0.0% (0/5)
Hard	**ReVi**	**87.64**	**87.15**	**89.88**	**2.75**	**100.0% (5/5)**

FilmEval results by difficulty; all five outputs per method and tier completed. Quality is on a 0–100 scale; duration statistics are percentages. Rankings are within each tier. Bold and underline mark the best and the second-best results, respectively; ties share a rank.

Table 15: Source-paired quality gains over baselines.

Baseline	RF	NP	AVG
	FilmEval		
Direct LLM	+17.12 [+14.52, +19.84]	+23.69 [+21.33, +25.94]	+27.04 [+23.45, +30.81]
ViMax	+9.50 [+6.87, +12.27]	+11.76 [+8.87, +14.70]	+10.00 [+7.34, +12.57]
VideoClaw	+10.47 [+8.85, +12.10]	+11.00 [+8.03, +14.17]	+11.37 [+8.21, +14.73]
MovieAgent	+7.86 [+5.28, +10.48]	+11.59 [+8.79, +14.23]	+13.94 [+10.06, +17.59]
MM-StoryAgent	+16.01 [+13.52, +18.50]	+16.43 [+13.06, +19.85]	+18.76 [+15.95, +21.68]
VGoT	+14.10 [+11.62, +16.44]	+14.00 [+11.58, +16.36]	+16.02 [+12.73, +19.26]
	Recap32		
Direct LLM	+21.74 [+19.58, +23.75]	+21.85 [+19.80, +23.72]	+19.63 [+13.76, +24.93]
ViMax	+12.39 [+8.95, +15.87]	+11.89 [+9.68, +14.40]	+8.82 [+3.94, +13.70]
MovieAgent	+9.70 [+6.94, +12.20]	+10.74 [+7.66, +13.96]	+10.58 [+6.36, +14.66]

Source-paired ReVi gains (points), with interval bounds on the second line of each cell. FilmEval resamples within difficulty; Recap32 resamples its 12 source IDs.

Duration and generation expenditure. Table 16 complements the quality comparisons with duration-error and generation-cost intervals.

Table 16: Duration-error and cost intervals.

Baseline	DE (pp)	Generation cost ($)
FilmEval		
Direct LLM	[-243.75, -194.99]	[-6.75, -5.45]
ViMax	[-121.66, -88.87]	[-3.44, -2.65]
VideoClaw	[-159.40, -120.99]	[-4.41, -3.10]
MovieAgent	[-167.09, -149.35]	[-4.75, -4.28]
MM-StoryAgent	[-224.66, -208.31]	[-5.90, -5.32]
VGoT	[-181.71, -135.14]	[-4.96, -3.36]
Recap32		
Direct LLM	[-270.60, -148.67]	[-9.90, -5.42]
ViMax	[-138.77, -103.24]	[-5.67, -3.90]
MovieAgent	[-200.50, -178.49]	[-7.99, -6.94]

Paired interval bounds for ReVi minus each baseline on duration error and generation cost. Quality differences are reported in Table 15.

Table 17: Intervals for inner-loop component ablations.

Comparison	RF	NP	AVG	DE (pp)
Inner Loop: ablated minus Full ReVi				
Causal Editing	[-6.37, -3.15]	[-2.50, -0.70]	[-1.94, 0.87]	[0.10, 0.53]
TTS Feedback	[-1.04, 0.67]	[-3.82, -1.80]	[-3.19, -1.02]	[6.13, 7.45]
Narrative Allocation	[-2.60, -1.45]	[-6.11, -4.60]	[-3.41, -1.89]	[1.69, 1.97]
Beat Planning	[-1.92, -0.24]	[-3.21, -1.36]	[-10.76, -7.10]	[-0.13, 0.82]
Repair Routing	[-6.73, -2.54]	[-3.38, -2.83]	[-3.36, -2.98]	[1.29, 1.87]

Paired interval bounds for inner-loop component ablations. Quality differences are in points; duration-error differences are in percentage points. Every comparison uses six source pairs.

Editing mechanisms. Table 17 collects paired quality and duration-error bounds for the editing ablations.