SCRIPTY: AN EXTERNAL SCRIPT SUPERVISOR FOR CONTINUITY IN STORY-LEVEL VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Story-level video generation runs a model clip by clip inside a harness that writes prompts and picks reference images, yet a detail set in one clip, such as a cut braid, often reverts in the next: retaining event history does not specify the current state, and correct text can still conflict with an outdated reference image. The harness must know which states still apply, and both the prompt and the reference images must show them. We present Scripty, a continuity module that works like a film set's script supervisor, attached via host-specific adapters without modifying host code or models. From frozen storyboards, asset lists, and reference images it builds a state record, each change citing its storyboard line, and it keeps each state until the storyboard changes it, injecting states into prompts and guiding reference-variant generation. Generated prompts and clips never serve as state evidence. We introduce StateHold, a benchmark that asks whether a later clip depicts the state the storyboard specified at that point.

On a fixed host across 42 stories, Scripty reaches 48.8% cross-unit fact pass rate versus 26.0% for native recap (+22.8 pp); adding reference-image variants improves this further, with some within-unit losses. These results support improved storyboard-state conformance but do not establish better viewer-perceived continuity or reliable state expiry.

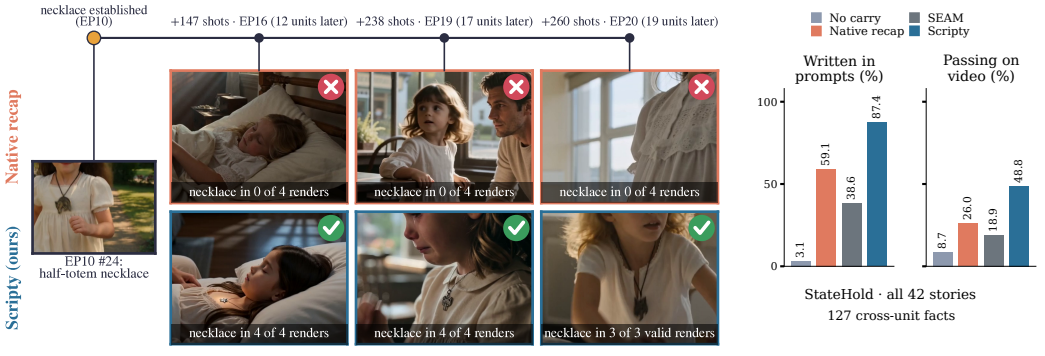

Figure 1: **One state, 260 shots later (illustrative).** Luna's necklace in a 25-episode SEAM-Bench drama, outside StateHold. At three checkpoints +147 to +260 shots after its establishment, paired-seed renders show it in 11/11 valid Scripty clips and 0/12 recap clips (0/6 with her chest in frame). Checkpoints were selected on Scripty's prompt and framing; frame selection ignores necklace presence. This tests coarse presence, not the half-totem shape, and is not representative long-gap evidence (Appendix D.15, §D.2). Right: pooled 42-story rates for states correctly written into prompts versus passing on video, measured by different judges. A separate supplementary video illustrates three StateHold cases.

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.

1 INTRODUCTION

Storyboard-specified states can silently revert in generated video: in our 42-story corpus, a boy's cut braid returns two clips later and a transferred pendant reappears on the giver. We study this in story-level pipelines that generate a narrative unit by unit, each unit comprising one or more planned shots. A video-generation pipeline, or host, wraps the video model in a harness that determines what information carries between units (Wu et al., 2025; Huang et al., 2026; Shi et al., 2025).

The typical harness (Figure 2) runs a planner, a writer and the video model. Recaps, visual memory (Zhang et al., 2025b), asset libraries and SEAM's entity-attribute memory (Liu et al., 2026a) carry information across units. This paper focuses on two requirements that remain open: the host must make the storyboard-specified state explicit for each unit, and must condition generation on reference images consistent with that state. We test whether a storyboard-grounded record of persistent state that drives both the prompt and the reference images raises cross-unit fact pass rates beyond the harness's native recap.

Scripty is an external continuity module that attaches to its host through an adapter, leaving host source code unchanged. From frozen upstream storyboards, entity and scene metadata, and reference images, Scripty builds a shot-indexed record of persistent state with storyboard evidence for each change (Figure 3). Generated prompts and clips never serve as state evidence. The record drives both model inputs: text constraints for the host writer (L0), and, when existing references do not depict a required state, host-generated reference variants of base images (L1).

StateHold fixes upstream storyboards and the initial reference-image library across conditions, comparing pipelines that differ in how they carry state, check prompts and adapt references. Paired visual questions test whether later clips show states established by the storyboard, with entity-presence gates and a within-unit guardrail. The full system scores 48.8% versus 26.0% cross-unit fact passes on the 42-story corpus (story sign test $p = 0.00032$); question banks were frozen by content hash before the reported runs. Text injection (L0) accounts for much of this measured gain. A text-layer audit finds 87.4% of states written into prompts but only 48.8% passing on video (Table 13).

The main host is in-house, so we also test adapters to two public hosts (ViMax, VideoMemory) and provide StateHold's frozen upstream inputs, storyboards and question banks; what the review materials include, and what is planned for release, are specified in the Reproducibility Statement.

We contribute: (i) Scripty, an external state module that uses one evidence-linked record of persistent state to guide text prompts and state-specific reference images in an existing host, attaching through host-specific adapters without host-source edits; (ii) StateHold, a fixed-upstream benchmark for later-state conformance, with complementary visual questions and separate cross-unit and within-

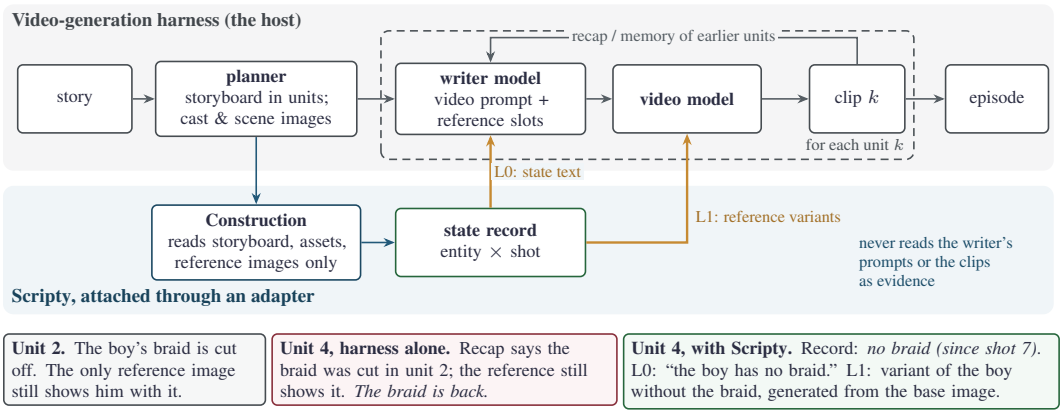

Figure 2: **A video-generation harness and where Scripty attaches.** The harness plans the story into units and, unit by unit, has a writer model turn the storyboard into a video prompt with reference images for the video model. Scripty builds a state record from the harness's upstream material and feeds the same record back through both channels the video model listens to: the prompt (L0) and the reference images (L1). Illustrative example.

unit evaluations; and (iii) empirical analysis: fixed-host comparisons and component studies that characterize gains in state conformance, prompt-video discrepancies and reference-variant trade-offs. Supplementary studies, including a viewing-preference experiment that does not establish a perceived continuity difference ($p = 0.375$), are reported in the experiments.

2 Related Work

Video pipelines and memory. Agentic pipelines coordinate story planning, shot design, assets and generation (Wu et al., 2025; Huang et al., 2026; Shi et al., 2025; Wang et al., 2026a; Xiao et al., 2026), and multi-shot generators produce several shots at once (Meng et al., 2026; Zhang et al., 2026). StoryMem uses visual memory across shots (Zhang et al., 2025b), and VideoMemory keeps an entity-centric memory of visual and semantic descriptors, updated after each shot (Zhou et al., 2026). These are natural hosts for an external state module; Scripty keeps a host's own memory in place and asks what explicit state adds beyond it.

Structured state and text-side continuity. FilmWorld and IAMFlow explore persistent world or identity representations (Zuo et al., 2026; Liu et al., 2026b). CineCrew makes cinematic and asset constraints explicit in a film description language, anchors clips with asset packs and storyboard keyframes, and uses a critic agent that checks generated clips and triggers refinement (Chen et al., 2026b). CineCrew and HoloCine are complete systems; Scripty is a module that could sit on top of such a system, builds its record only from upstream inputs, never from generated prompts or clips, and keeps a storyboard-evidenced state record that drives both the prompt and the reference images of a host without source edits. SEAM extracts per-shot entity attributes from the script, retrieves causally prior memory, filters it and rewrites prompts (Liu et al., 2026a); our same-host reimplementation is not an official SEAM-Bench result (Appendix G). CANVAS keeps explicit character, location and prop states for storyboard generation and updates its visual anchors from generated frames (Mondal et al., 2026). Our distinction is the combination of an upstream-only record and control of the host's conditioning images. CANVAS also updates images, but reads generated frames; text-side memory alone leaves a contradictory reference unchanged. Table 3 compares these mechanisms, and Appendix B covers further work.

Evaluating continuity. Benchmarks for story and multi-shot video cover consistency, quality and instruction following (Zhang et al., 2025a; Chen et al., 2026c; He et al., 2026; Yoon et al., 2026; Zhuang et al., 2026; Shi et al., 2026b;a); state-focused suites test whether states survive occlusion or segment boundaries (Chen et al., 2026a; Ma et al., 2026; Miao et al., 2026), and VBench scores single-clip quality (Huang et al., 2024). StateHold asks the narrower question of whether a later generation unit shows a state that the storyboard set up earlier, with complementary question pairs, presence gates, a within-unit guardrail and story-level statistics.

3 Method

The host calls Scripty once per story to build its record and once per unit for constraints (Figure 3). Model roles and released prompts are indexed in Table 30 and Appendix H.

3.1 Inputs: frozen upstream artifacts

Adapters read frozen storyboards, shot-to-unit groups, entity/scene metadata and reference images; missing required assets are supplied before generation. Arms share host storyboards and rendering assets; Construction-only extras are listed separately (Table 29). Generated prompts and clips never become state evidence. Prompt checking reads drafts (§3.5); reference selection reads slots and states (§3.6); neither changes the record. Adapters intercept writing requests or tool calls without host-source edits; L1 also needs image-to-image generation and reference control.

3.2 The state record

The record assigns a persistent-state view S_i to each shot i, with S_0 before the first shot. It retains accepted unit-level net changes, not every transient event; shot indexing does not make it a complete

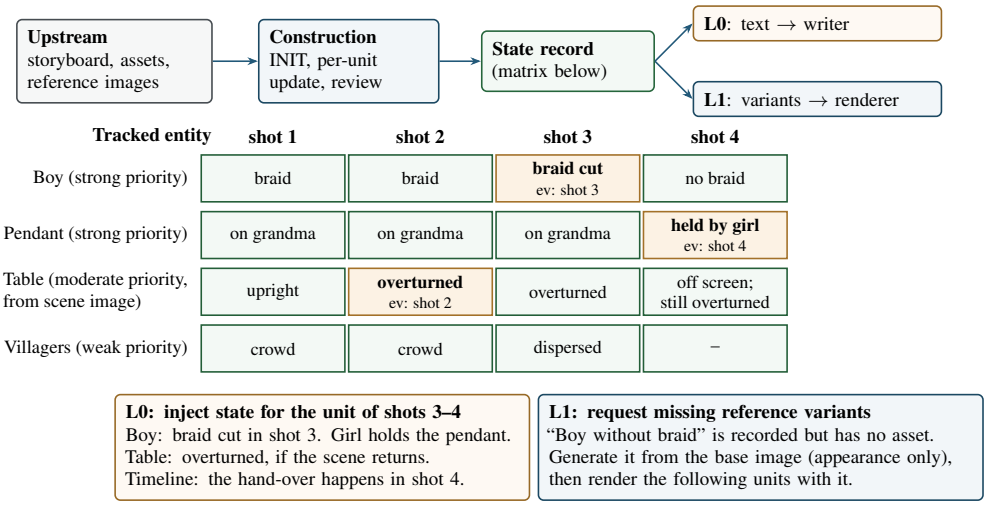

Figure 3: **From upstream material to both channels.** Construction builds the record from the host's storyboard, asset lists and reference images (including objects visible in scene images) and applies evidenced changes unit by unit, shot by shot. The record keeps one row per entity and one column per shot; each cell is the retained persistent-state view after the shot, not a complete transient-state history; shaded cells mark the shot whose storyboard line changes it. Being off screen does not end a state. The record feeds the host twice: text constraints for each unit (L0) and reference variants for states that no existing asset shows (L1). Illustrative example.

shot-by-shot world state. Read across shots, it is a matrix with one row per entity (Figure 3). For each entity, S_i stores whether it is on screen, where it is, which way it faces, its appearance (clothing, hair, injuries, what it carries) and its physical state (open, broken, lit), plus relations such as who holds what. Every value carries its provenance: the shot and storyboard line that established it, or, for some initial values, the asset list or scene image it came from.

Entities come from cast and asset lists, scene furnishings and images, and storyboard objects discovered by INIT. Strong-priority entities are story-critical and must be kept wherever they appear; moderate-priority entities are tracked normally, while weak-priority background groups receive one-line summaries (Appendix A).

3.3 Construction: initialize, then update unit by unit

Construction starts with INIT, which reads the inputs and builds S_0, the state already in force when the first shot begins. INIT sees all shots of the episode, so it can find every entity the story needs, but storyboard evidence for an initial value may cite only the first shot, and the result of an action inside the first shot belongs to S_1. A mechanical check enforces this and verifies each quote (Appendix A). Construction then proceeds unit by unit. For each unit a language model reads the unit's shots and the current state and proposes the unit's net changes, each tied to the shot and quote that support it, so a change that is undone within the same unit is usually not recorded and the per-shot states then omit it; a second model call checks that the changes are supported and that nothing the unit establishes was missed; and a deterministic executor applies them shot by shot, $S_i = \mathrm{Apply}(S_{i-1}, \Delta_i)$, where Δ_i is the accepted change set for shot i. A light switched on and off within one unit may leave no recorded change; splitting that unit can change extraction. The writer still sees the current storyboard.

The executor enforces identities, references and field types, so a model cannot break the record's structure; a wrong semantic reading can still pass review and enter it. Review and re-drawing are bounded (Appendix A); a failing unit stops dependent units instead of passing on a guess.

3.4 How long a state lasts

An appearance is a list of state attributes, each with its establishing shot. An attribute stays until the storyboard removes or excludes it; momentary in-unit actions are not recorded. This conservative rule can carry a transient look too long (§6).

3.5 Text injection (L0)

Before writing, the host receives an end-of-unit view: all entities visible during the unit, including mid-unit departures and carried objects, their appearance, position and physical state, and required off-screen entities. Strong-priority entities, header-declared entities and visible props are must-keep; low-confidence inferences are optional. The adapter inserts this text in a fixed prompt slot. To avoid showing a transferred pendant before its handover, the injection also supplies start/end states and storyboard lines at shots with supported changes. The host writer drafts the prompt; a checker verifies must-keep coverage and consistency with the record, allowing two rewrites before retaining the last draft and logging failure. It does not check timing, so changes can still appear early.

3.6 Reference variants (L1)

For each reference slot, a model separately reads the start- and end-of-unit state record and selects the default image, another registered version or a new-image option. Different choices or missing citations retain the preceding selection. Two new-image choices trigger a check that the cited depicted states already hold at the unit's start. Otherwise the aggregator selects a previously recorded, start-compatible *variant specification*: state text plus image-generation edits, possibly rendered for the first time now. With none available it retains the preceding image unchecked for conflicts. The host generates variants from base images using the filtered edits and renders the clips; this constrains intended state, not pixel correctness (Appendix A). Variants flagged by the image check are redrawn once and the second image is kept even if flagged again (Appendix D.12). L0 keeps the host references; L1 adds this reference-variant channel. Audited specifications respect the record's start-state timing, not necessarily image semantics (Appendix D.12).

4 StateHold: does the later clip keep the state?

What is compared. StateHold tests whether a later generation unit shows a visible state that the storyboard set up in an earlier one. The benchmark authors prepare and freeze the upstream for each story, namely the unit split, the storyboard, the asset lists and the reference images, and give it unchanged to every compared pipeline. Pipelines therefore differ only from the storyboard onward: how they write each unit's prompt, what state they carry and which reference images they render with (Figure 5 in Appendix C shows the scoring flow).

Facts and questions. Each fact names an entity, the state the storyboard establishes (a cut braid, a pendant in the girl's hand), the unit where it is established and the unit where it is checked. Questions are written from the storyboard alone, never from any pipeline's state or output. A *cross-unit* fact is checked in a later unit and measures persistence. A *within-unit* fact is checked in the same unit and serves as a guardrail: carrying more state across units must not break what each unit shows on its own. For a cross-unit fact the judge sees the establishing clip and the later clip, and presence gates first check, from a textual description of the entity or location, that it is recognizably on screen in each clip. The judge then answers two questions about the later state with opposite expected answers (*is the pendant in her hand?* yes; *is it not in her hand?* no), seeing only the clips and question text, never reference images. With $Q^{\pm}_{a,f}$ and $G_{a,g}$ indicating that arm a's answer to the question or gate is *correct*, a fact passes when all its gates and both questions are answered correctly:

$$Y_{a,f} = Q^{+}_{a,f} Q^{-}_{a,f} \prod_{g \in G_f} G_{a,g}, \qquad P_a = \frac{1}{|F|} \sum_{f \in F} Y_{a,f}. \tag{1}$$

Over cross-unit facts P_a is the primary score, CUC; over within-unit facts it is the WUC guardrail. "Cannot determine" counts as a failure and stays in the denominator. Failed cross-unit facts are split into four gate/question cases (Appendix D). The primary score does not ask whether the establishing

clip itself showed the state. The score therefore measures whether, with the entity recognizable in both clips, the later clip shows the state the storyboard requires. A state drawn late (wrong in the establishing clip, right in the later one) passes; Appendix D.5 adds an establishing-clip question for every cross-unit fact, and the blind human check (Appendix D.17) adds it on a sample. Every arm is judged once with the same judge, the same bank version and the same judge inputs, namely the clips and the question text (Appendix C.3).

Pools and review. StateHold contains 42 stories, grouped by source-story author into GPT-family (20) and other sources (22, including four public-domain adaptations). Both subsets contain shorter 5-unit stories and designed longer 6/8/10-unit stories; these are sequence lengths, not fixed state gaps. Each story has one source script. There are 127 cross-unit and 93 within-unit facts. The 18- and 24-story subsets, called FORMAL/EXPAND below, remain diagnostic strata. Facts receive mechanical checks and storyboard review; frozen-bank and sampled-human-review scope are documented in Appendix C.

Analysis. The primary comparison is the full system, Scripty L1, against the host's native recap on cross-unit facts. The main results table reports the full pool; the preregistered subset (the confirmatory reading) appears alongside the full pool in the registered-inference tables (Tables 8–9). Direction is summarized per story with a two-sided exact sign test, and uncertainty with a story-level bootstrap interval of the pass-rate difference; if the sign test is not significant we report direction and interval only. On within-unit facts, a difference significantly below -10 percentage points counts against Scripty; an interval above that bound is not a positive claim.

Secondary contrasts are uncorrected and their p values descriptive. The within-unit guardrail exposes gains bought at other states' expense; the length-matched control tests added prompt length (§5.3); and the joint score requires the establishing clip to show the state too (Appendix D.5).

The rule families have distinct validity roles (Table 5). Bounded retries can still leave void pairs (Table 4); they do not certify semantic correctness.

5 Experiments

5.1 Setup

The main host is ShortDrama, an in-house short-drama harness (Figure 2). Six arms share frozen upstream inputs: *no-carry* uses only the current unit's storyboard; *native recap* adds the host's summary; *SEAM* applies our entity-attribute-memory reimplementation to no-carry prompts, without the recap (Liu et al., 2026a); *SEAM+recap* applies the repair to recap prompts; *Scripty L0* injects state text; and *Scripty L1* also uses reference variants. L1 re-renders units with changed references under the same text and reuses other L0 clips and judgments. All arms use gpt-6-luna at high effort for writing; Scripty also uses it for Construction and prompt checking. All reported L1 results use the two-reading agreement rule and the pinned engine (Appendix F.1); models, implementation details and costs are in Appendix F and Table 25.

5.2 Main results

Table 1 reports the unified corpus. On the unified corpus the full system passes 62 of 127 cross-unit facts against 33 for native recap (+22.8 [+11.0, +34.2] pp; sign test $p = 0.00032$). L1 minus L0 is +9.4 [+1.6, +17.5] pp ($p = 0.0169$). Within-unit rates are 75/93 versus 73/93: +2.2 [-9.0, +13.3] pp, with the pooled interval above the -10 pp boundary. Tables 8 and 9 report the two strata separately. Against SEAM+recap the margin is +14.2 pp ($p = 0.02945$), an unadjusted secondary contrast (Appendix D.4); joint establishment-and-later rates are lower, 40/127 (31.5%) versus recap's 17/127 (13.4%) (Tables 15 and 16); on this stricter score Scripty L1 minus L0 is +4.7 pp (40/127 versus 34/127), about half the primary-metric increment. Long-gap gains remain unestablished (Appendix D.2). Three whole-chain draws retain L1's lead over recap, although only about half the cross-unit facts keep the same outcome (Appendix D.1). Removing gates or undecidable facts symmetrically preserves positive differences; worst-case resolutions of undecidable answers and non-registered subsets are less conclusive (Table 10). Both source-author subsets show positive

Table 1: **Unified StateHold results: 42 stories.** All arms share storyboards, rendering assets and the writer model. CUC measures conformance of the later clip to the storyboard-specified state (Joint CUC requires both clips); WUC is the within-unit guardrail. Paired differences are against native recap, with story-bootstrap 95% intervals. Note: SEAM+recap was added later and was not judged under the question-pair retry rule that the other five arms follow.

Method	CUC pass (%)	Joint CUC (%)	WUC pass (%)	Δ CUC [95% CI]	p
No-carry	11/127 (8.7)	9/127 (7.1)	74/93 (79.6)	-17.3 [-26.2, -9.4]	< 0.001
Native recap	33/127 (26.0)	17/127 (13.4)	73/93 (78.5)	–	–
SEAM (in-host)	24/127 (18.9)	18/127 (14.2)	75/93 (80.6)	-7.1 [-16.7, +1.6]	0.332
SEAM + recap	44/127 (34.6)	–	75/93 (80.6)	+8.7 [+0.0, +16.7]	0.019
Scripty L0	50/127 (39.4)	34/127 (26.8)	78/93 (83.9)	+13.4 [+2.2, +25.8]	0.064
Scripty L1	62/127 (48.8)	40/127 (31.5)	75/93 (80.6)	+22.8 [+11.0, +34.2]	< 0.001

CUC: 127 facts in 42 stories. WUC: 93 facts in 41 stories; one story has no WUC item. Undecidable answers count as not passed. Calls and checking budgets differ. Per-stratum results are in Table 6. Joint CUC requires the establishing clip to show the state as well (Tables 15–16); SEAM+recap has no establishing-clip judgments. Secondary p values are unadjusted.

L1–recap differences (Figure 6); Per-stratum results remain in Table 6. Baseline-only contrasts show that host memory already helps (Table 11).

5.3 Where the gain comes from

Table 2 varies one component with the other fixed; Table 7 retains the less conclusive per-stratum readings.

Is it prompt length? The length-matched history-summary control passes 22/127 (17.3%) cross-unit facts, 22.0 pp below L0 ($p = 0.0023$), although information access, checking and compute still differ. Removing check-driven rewrites changes cross-unit scores little but lowers within-unit rates by about 10 pp in each stratum (intervals include zero; Table 12). Information access still differs (Table 28).

Record and text channel. A six-story audit covers 89.6% of an independently extracted change inventory, with omissions and stale injections; this measures extraction agreement, not absolute accuracy or shot timing (Table 17, Appendix D.6). At the prompt level Scripty writes the tested state in 111/127 cases (87.4%), versus 75/127 (59.1%) for recap: +28.3 [+19.4, +38.2] pp ($p < 0.001$). Conditional video rates are descriptive because the written sets differ across arms (Appendix D.3). Correct prompt coverage (87.4%) coexists with a lower video pass rate (48.8%); the different instruments and evaluation defects prevent a causal partition (Appendix D.3, Table 14, Figure 7).

Image channel. Single-draw L1 minus L0 is +9.4 pp on cross-unit facts ($p = 0.0169$) and -3.2 [-7.7, $+1.1$] pp within units. With fixed texts and images, three draws per side give $+12.9$ [$+5.0$, $+20.8$] pp on affected cross-unit facts, but -15.8 [-30.3, -3.3] on FORMAL's affected within-unit facts (Appendix D.10). Exploratory splits associate FORMAL's within-unit loss with failed image audits; L1–L0 is positive on facts marked as needing a variant and negative elsewhere (Appendix D.11). Individual flips do not identify causes (Appendix D.9).

Scene-image objects. Adding objects from scene images raises cross-unit rates by 7.1 pp at L0 and 8.7 pp at L1, with both intervals spanning zero; the interaction is small (+1.6 pp). A DeepSeek-V4.1-Flash Construction-and-writing chain retains a lead over recap on EXPAND ($+22.5$ pp, $p = 0.013$).

5.4 External hosts and a text-level check

We test L0 on the same three dramas in ViMax (Huang et al., 2026) and VideoMemory (Zhou et al., 2026): EP01–05 (414 planned shots per arm, 43 judged facts) and EP01–03 (72 shots, 33 facts), respectively. Host-specific proxies/wrappers preserve source code; Construction alone reads an ex-

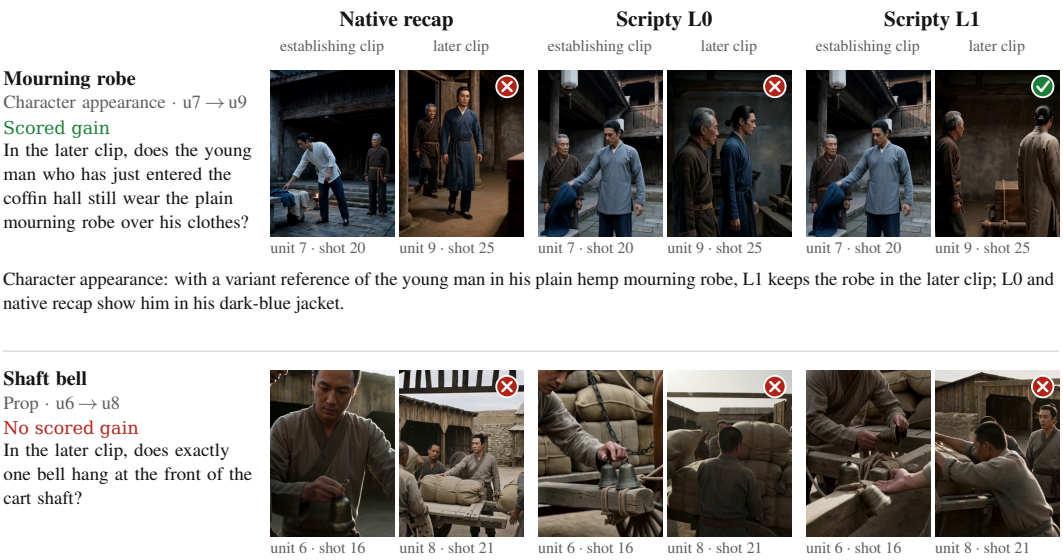

Figure 4: **Reference variants: scored gains and failures.** Establishing and later clips under recap, L0 and L1. Top: a mourning-robe variant helps retain the state. Bottom: no bell variant is generated and all arms fail. Cases follow the recorded selection rule; the full set includes a failed variant (Figure 8). The accompanying video is a separate curated selection of successful cases.

Table 2: **Component changes on the unified corpus.** One component changes while the other is fixed; pass-rate differences in pp with story-bootstrap 95% intervals. Secondary tests are descriptive. †DeepSeek whole-chain comparison with recap covers only the 24 longer stories.

Change	Δ CUC	Δ WUC	CUC p
Add L1 to L0	+9.4 [+1.6, +17.5]	-3.2 [-7.7, +1.1]	0.017
Scene-image objects (L0)	+7.1 [-4.5, +18.4]	+10.8 [+1.2, +20.7]	0.345
Scene-image objects (L1)	+8.7 [-2.3, +19.2]	+9.7 [-3.2, +22.6]	0.036
DeepSeek chain†	+22.5 [+10.0, +36.1]	+4.9 [-7.0, +20.0]	0.013

tra full-drama asset directory; backends are paired within each host, and the short-drama prompt checker is not transferred. ViMax L0 improves cross-unit passes (7/13 vs. 1/13; per-question $p = 0.031$), but loses expiry controls (3/9 vs. 6/9); this is a small integration demonstration (Table 18). An exploratory ViMax L1 run on one drama gives no cross-unit gain and worse expiry controls (Appendix G.5). VideoMemory passes 5/9 versus 1/9 cross-unit facts ($p = 0.125$), with expiry controls 4/5 each. Its backend substitutions and two one-arm keyframe fills are disclosed in Appendix G.6. These are integration demonstrations, not information-matched mechanism effects; different banks and backends preclude host rankings. On SEAM-Bench's storyboards (Liu et al., 2026a), we compare written prompts only. On SEAM-Bench's text level, after removing ledger items the shared storyboard does not support, the frozen engine retains more of the asserted state than our SEAM reimplementation (41.1% vs. 14.2%) but also carries more expired state (8.9% vs. 0.0%) (Table 23, Appendix E). A lifetime-aware variant outside the main system retains a permanent bracelet across a time jump in one drama, demonstrating behaviour rather than general efficacy (Table 24). A 20-episode visual demonstration both remembers a chip and over-retains mud (Appendix D.15). Generic video-quality metrics change little, while CLIP-T and some identity measures fall (Appendix D.13); comics carry scored pass rates from the same bank and judge as the video line (Appendix D.14; Figure 10).

5.5 JUDGE VALIDITY AND COST

Is the judge reliable? A retry-provenance audit found first-round reasons stored with later retry answers (Appendix D.19). Frozen scores are retained under this audit (Appendix D.7). Human calibration covers the baseline arms, not Scripty (Table 21). Rejudging under the same judge and clips preserves the direction of the main contrasts (Table 20, Table 22); same-family judges limit this evidence (Appendix D.18). Alternative vote policies preserve direction more than significance (Appendix D.11).

Do people agree? In a blind six-fact appearance check, L1 passes 4/6 later-state panels (one undecided), L0 3/6 and recap 2/6; human L0-to-L1 flips agree with the judge on 4/5 decided facts (Appendix D.17). An independent viewing-preference study (36 ratings on 12 stories) finds no significant difference, leaning towards the native recap (continuity: 1 vs. 4 stories, 6 ties, 1 without majority, $p = 0.375$; Appendix D.16). This study does not establish a perceptual benefit; it rates whole sequences, whereas StateHold checks named states.

What does it cost? The full system makes 11–15 times as many model calls per unit as the recap; with one final render per unit in each configuration, its total cost per story is about 1.9–2.6 times the recap's (Appendix F, Table 27). These are per-configuration costs; the research run additionally re-rendered changed units. Summed call durations are not end-to-end latency (Table 26).

6 CONCLUSION AND LIMITATIONS

Scripty uses one upstream state record to drive text and reference images without editing host source. Explicit state injection improves measured cross-unit fact pass rates on StateHold. Reference variants add a positive aggregate increment over L0, but their value remains conditional, with within-unit costs in repeated draws. These results support further study of explicit state, not a general improvement in viewer-perceived continuity or a clean causal separation of information, checking and image generation.

Limitations. Clustered simulations indicate low power for moderate gains in the smaller analysis subsets (Appendix D.11). The linear state chain does not support flashback branches; these are future work (Appendix A.8). Fixed-host effects do not rank whole systems. Construction may misread a state; transient appearances may persist; end-of-unit injection and a timing-blind prompt check may introduce changes early; variants may drift in identity; and fallback references may retain expired states (Appendix D.12). StateHold uses author-prepared storyboards and primarily checks later clips (Appendix D.5 adds establishment). Retry-reason provenance remains an unresolved measurement limit (Appendix D.19); judge sensitivity and limited human evidence remain (Table 20, Appendix D.16).

REPRODUCIBILITY STATEMENT

Models, versions, settings and costs are listed in Appendix F, which also separates three levels of reproduction: recomputing statistics, re-judging clips and regenerating clips. The review supplement contains: the frozen Scripty engine code and every prompt used by the reported runs (including the judge prompt and the scoring code); both frozen question banks with their content hashes; the per-question judge responses behind every reported arm; an index of the main-table judged clips with per-file SHA-256 hashes; an offline bundle that recomputes the reported tables with no model call and no network access; and the anonymous human-study responses. It does not contain the in-house short-drama host or any SDP product code; the full video and reference-image library (only the index, its hashes and one curated example video ship). Appendix D.19 records measurement limitations. Original media, consent records and historical diagnostics are preserved separately; retired diagnostic claims are explicitly marked. On acceptance we plan to release the remaining permitted benchmark materials and documented adapters. The in-house short-drama host's interfaces and backend requirements are described here.

THE USE OF LARGE LANGUAGE MODELS

Writing assistance. LLM-based tools assisted manuscript drafting. All generated text was reviewed and edited by the authors, who remain responsible for every claim.

Research ideation and execution. Code and experiment orchestration were written with AI coding assistants and reviewed by the authors against tests and run records. Judging uses Gemini models compared with small human-rater samples (Appendix D.18). State-record checks and historical loss diagnoses are AI-assisted. Unresolved causal and judge-error attributions are not claimed.

Synthetic dataset generation. Source stories and question banks were drafted by several model families, cross-reviewed by other models and checked by the authors (Appendix C). Generated media themselves constitute synthetic corpora.

Experimental roles are separate: source/anchor authoring, Construction, semantic review, prompt writing, SEAM extraction/repair, image generation and judging. Their actual versions are disclosed rather than summarized as one experiment model. Automated review is not human annotation, and the authors remain responsible for the final claims and retained limitations.

ETHICS STATEMENT

The corpus contains fictional narratives and generated media. Retained source material, reference assets and derived annotations follow their documented use and release conditions. Corpus exclusions are recorded with their reasons; excluded material is not silently replaced by a favourable new sample. SEAM-Bench source data is used for evaluation and is not redistributed with our generated corpus. Actual human review roles and blinding are stated in the evaluation protocol. The fact-level human study involved six volunteer raters who gave in-page consent, could withdraw by anonymous ID and received no fixed payment; anonymous answers are kept on the research group's server, visible only to the authors, and deleted three years after publication. The study was determined to be exempt from ethics review by our institution. The final blind check and the viewing-preference study use four non-author raters who volunteered without payment and can withdraw by anonymous ID; their consent page covers release. Only the responses of raters who agree are released, in anonymized form (rater IDs only; no names, contact details or timestamps; for research use only) and renumbered R1–Rk. The mapping from IDs to persons and the responses of raters who do not agree stay on the research group's server, visible only to the authors, and are deleted three years after publication. Before publication a rater may ask by ID to withdraw; the responses are deleted within seven days and the results recomputed. Aggregate results include all raters. All four raters consented to release, so every response is released. This study was likewise determined to be exempt from ethics review by our institution; only the anonymized responses of consenting raters are released. No private product-user data enters the study. Scripty supplies state and reference guidance; the host's generators retain their own safeguards.

REFERENCES

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, Weiyue Li, Alan Yuille, Paul Pu Liang, and Yilun Du. MemoBench: Benchmarking world modeling in dynamically changing environments. In *ECCV*, 2026a. arXiv:2606.27537.

Jiaben Chen, Sixun Dong, Qinhong Zhou, Raine Ma, Zhiyang Dou, Wojciech Matusik, and Chuang Gan. Better call CineCrew: Consistent ultra-long narrative-to-film generation. In *ECCV*, 2026b. arXiv:2609.07720.

Jiamin Chen, Qianben Chen, Jiawen Zhang, Yidi Wu, Yuchen Li, Xiaokun Zhang, Wangchunshu Zhou, and Chen Ma. DirectorBench: Diagnosing long-form video generation with personalized multi-agent evaluation. *arXiv preprint arXiv:2605.30090*, 2026c.

Ruozhen He, Meng Wei, Ziyan Yang, and Vicente Ordonez. EntityBench: Towards entity-consistent long-range multi-shot video generation. *arXiv preprint arXiv:2605.15199*, 2026.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. ViMax: Agentic video generation. *arXiv preprint arXiv:2606.07649*, 2026. Code: HKUDS/ViMax.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, et al. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. arXiv:2311.17982.

Yixuan Lai, Tianjia Shao, Weijia Dou, Siyu Zhu, and Jingdong Wang. GroundShot: Visually consistent multi-shot long video generation via entity-grounded shot scheduling. *arXiv preprint arXiv:2606.20799*, 2026.

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, et al. Anim-Director: A large multimodal model powered agent for controllable animation video generation. In *SIGGRAPH Asia*, 2024. arXiv:2408.09787.

Jiaqi Liu, Maolin Ran, Xiaoyang Lu, Jian Wang, Weiwen Liu, Jianghao Lin, Yong Yu, and Weinan Zhang. SEAM: Shot entity-attribute memory for consistent short-drama generation at scale. *arXiv preprint arXiv:2608.22725*, 2026a. Releases SEAM-Bench (data: Jackyqq/SEAM-Bench on Hugging Face).

Jinzhuo Liu, Jiangning Zhang, Wencan Jiang, Yabiao Wang, Dingkang Liang, Zhucun Xue, Ran Yi, and Yong Liu. Advancing narrative long video generation via training-free identity-aware memory. *arXiv preprint arXiv:2605.18733*, 2026b. Code: Eddie0521/IAMFlow.

Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. VideoStudio: Generating consistent-content and multi-scene videos. In *ECCV*, 2024. arXiv:2401.01256.

Jinpeng Lu, Dexu Zhu, Haoyuan Shi, Linghan Cai, Guo Tang, Yinda Chen, Jie Cao, Duyu Tang, Yi Zhang, Yong Dai, and Xiaozhu Ju. Current world models lack a persistent state core. *arXiv preprint arXiv:2606.20545*, 2026.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of sight, out of mind? Evaluating state evolution in video world models. In *ECCV*, 2026. arXiv:2603.13215.

Logan Mann, Abdur Rahman, Mohammad Saifullah, Taaha Kazi, and Vasu Sharma. NarrativeWorldBench: A frontier-saturated benchmark and a latent world model for long-horizon co-creative audio drama. In *ICML Workshop on High-dimensional Learning Dynamics (HiLD)*, 2026. arXiv:2606.17391.

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. In *CVPR*, 2026. arXiv:2510.20822.

Yingmao Miao, Pengfei Zhang, Chaoran Xu, Meng Yu, Jing Tang, Xiangxiang Chu, Chao Shen, and Chenhao Lin. Do video generators track the world across segments? A benchmark and method for world-state reasoning in video continuation. *arXiv preprint arXiv:2609.03673*, 2026.

Ishani Mondal, Yiwen Song, Mihir Parmar, Palash Goyal, Jordan Boyd-Graber, Tomas Pfister, and Yale Song. CANVAS: Continuity-aware narratives via visual agentic storyboarding. In *EMNLP (to appear)*, 2026. arXiv:2604.13452.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. AniMaker: Multi-agent animated storytelling with MCTS-driven clip generation. In *SIGGRAPH Asia 2025 Conference Papers*, 2025. doi: 10.1145/3757377.3764009. arXiv:2506.10540.

Haoyuan Shi, Mingtao Chen, Shuo Jiang, Ziyan Chen, Xuyi Sheng, Yiming Liu, Ying Zhang, Miao Wang, Jianxiang Lu, Fanyang Lu, Songyuanyi Lu, Xiele Wu, Zhichao Hu, Yuhong Liu, and Richeng Xuan. DramaChain Bench: An end-to-end benchmark for short-drama generation. *arXiv preprint arXiv:2609.00646*, 2026a. Technical report.

Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. MSVBench: Towards human-level evaluation of multi-shot video generation. In *Findings of ACL*, 2026b. arXiv:2602.23969.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

Bo Wang, Haoyang Huang, Zhiying Lu, Fengyuan Liu, Guoqing Ma, Jianlong Yuan, Yuan Zhang, Nan Duan, and Daxin Jiang. StoryAnchors: Generating consistent multi-scene story frames for long-form narratives. *arXiv preprint arXiv:2505.08350*, 2025.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. MAViS: A multi-agent framework for long-sequence video storytelling. In *Proceedings of EACL*, 2026a. arXiv:2508.08487.

Yuji Wang, Yuheng Chen, Teng Hu, Ran Yi, Yijia Hong, Han Feng, Weijian Cao, Chengjie Wang, Lizhuang Ma, and Jiangning Zhang. PersonaShot: Benchmarking person-centric narrative continuity in multi-shot video generation. *arXiv preprint arXiv:2608.16717*, 2026b.

Zun Wang, Jialu Li, Han Lin, Jaehong Yoon, and Mohit Bansal. DreamRunner: Fine-grained compositional story-to-video generation with retrieval-augmented motion adaptation. In *AAAI*, 2026c. arXiv:2411.16657.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent CoT planning. *arXiv preprint arXiv:2503.07314*, 2025. Code: showlab/MovieAgent. The MoviePrompts set ships with this paper.

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain Cinema: Towards short movie generation. In *ICLR*, 2026. arXiv:2507.18634.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, et al. FilmAgent: Automating virtual film production through a multi-agent collaborative framework. In *SIGGRAPH Asia Technical Communications*, 2024. arXiv:2501.12909.

Daewon Yoon, Hyeongseok Lee, Wonsik Shin, Sangyu Han, and Nojun Kwak. MSG Score: Automated video verification for reliable multi-scene generation. In *AAAI Workshop on Creative Video Modeling (CVM)*, 2026. arXiv:2411.19121.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. ST-Bench: the story-video benchmark released with StoryMem, 2025a. Released with arXiv:2512.19539; code: Kevin-thu/StoryMem.

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. Code: Kevin-thu/StoryMem.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. STAGE: Storyboard-anchored generation for cinematic multi-shot narrative. In *CVPR*, 2026. arXiv:2512.12372.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, et al. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, et al. VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2412.02259*, 2024.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-cong Chen. VideoMemory: Toward consistent video generation via memory integration. *arXiv preprint arXiv:2601.03655*, 2026. Code: EnVision-Research/VideoMemory.

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. In *CVPR*, 2026. arXiv:2505.24862.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. *arXiv preprint arXiv:2401.09414*, 2024.

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.

A State Record and Construction Details

This appendix describes the evaluated engine. Where the two pools were run with different Construction versions, the difference is stated in Table 25.

A.1 Record format

For every shot i the record holds a state S_i. Each tracked entity has an instance with five fields: presence (on screen or not), position, orientation, appearance and physical state. Appearance is a list of state attributes; each state attribute carries an identifier, the shot that established it, and the storyboard line that supports it. Relations record who holds what, and the engine derives three kinds of holding from positions with storyboard evidence: attached or worn, held in the hand, and carried by a vehicle or animal. Instances are tracked entities with strong or moderate priority; the two tiers have the same fields and differ in which fields are required and in whether the writer must keep them. Crowds and other background groups are weak-priority groups, stored as a one-line group summary that later units can only rewrite, not split into entities.

A.2 Where entities come from

The entity list has five sources: (i) the host's character, creature and prop tables; (ii) props that the asset descriptions mention without an image, admitted only if the episode's storyboard names them; (iii) one fixed-furnishing entity per scene whose description gives it a narrative state; (iv) objects visible in the scene reference images, named with the storyboard's wording where possible and excluding anachronistic, unclear or carried objects; and (v) objects that INIT finds in the storyboard but not in any image, which start as "not yet present" and appear only when a later shot supports them. When an object from (iv) looks like an asset from (i)–(iii), the pair is flagged; the injection shows it once, under the asset's identifier, and a value read from the scene image never overrides a value from the storyboard or the asset list.

A.3 Construction procedure

INIT. INIT receives all shots of the episode, the asset lists and descriptions, the scene descriptions and the reference images, and returns S_0. It may look ahead to register entities, but any storyboard evidence for an initial value must come from the first shot; a mechanical check rejects evidence from later shots before semantic review. The INIT reviewer sees the scene images. A rejection of a value that was read from a scene image is logged but does not force a new INIT, because Construction cannot change what the image shows.

Per-unit update. For each unit a single model call returns the changes for all its shots, each tied to a shot and a storyboard line. Only net changes are recorded; momentary actions inside the unit are not. The engine applies the changes shot by shot to produce S_i. At a scene change the engine marks entities as off screen unless the storyboard keeps them, without touching their position or state, and an attached or vehicle-carried object follows its holder's presence. A semantic reviewer checks each unit once per round against the storyboard: are the changes supported, and did the unit establish anything that is missing? A unit gets at most six review rounds, stops early when the same field is rejected in three consecutive rounds, and can be re-drawn at most three times; INIT likewise. A patch round rewrites only the fields the reviewer named. A unit that still fails stops its dependent units.

How long an appearance lasts. A state attribute stays until the storyboard removes it, either with evidence in the current unit or by stating a new state that excludes it. A time jump ends attachments tied to a character's role but does not end appearance state attributes.

A.4 What the injection contains

For each unit, the injection starts from the state at its last shot and lists entities visible anywhere in the unit with their current fields, entities that leave midway or remain at the same place off screen, scene objects, carried objects, the scene layout, and entities that must stay off screen. Eligible

double registrations of a census asset and a scene object are treated as one object after provenance and placement checks; ambiguous or diverging pairs remain separate. Its temporal section contains the state record boundaries and source storyboard visual text for shots with a relevant state-boundary difference, or an explicitly marked event source. It omits ordinary unchanged shots. These rows do not determine within-shot order or automatically bind the source text to an entity or visible action. Strong-priority entities, entities declared in the unit header and every prop on screen are marked must-keep; low-confidence inferred details are optional. A prompt checker reads the writer's prompt against the must-keep list and returns it at most twice; after that the last draft is used and the miss is logged. Values of scene-image objects never touched by the storyboard are not checked.

A.5 Reference variants (L1)

For each unit and each writer-attached reference slot mapped to a host base family, a model separately reads the start- and end-of-unit state attributes in the record and chooses from the writer's default image, any other described version in the same host-registered family, and a new-image option. The candidate set always includes the default and new-image options; some slots have no other registered version. Agreement on a non-default choice requires valid cited state attributes on both sides. Different choices or missing citations retain the preceding selection. For NEW/NEW, the aggregator checks every state attribute cited by the filtered end-of-unit drawing edits against all state attributes in the start-of-unit record. If one arose within the unit, it reuses the most covering previously recorded start-compatible variant specification; if there is none, it retains the preceding reference without checking it for conflicts. A variant specification recorded from an earlier unit's end state can be rendered for the first time now, so it should not be called an already registered image. The image instruction comprises a fixed opening, a fixed ban on lettering and modern objects, and the model's cited edits after mechanical filtering of actions, positions, story events, transient looks, fragments, lettering, invisible items and unsupported quantity wording. Filtering all end edits changes that side's choice to default before aggregation; it does not by itself guarantee that the final reference is default. The edits are not deterministically deduplicated. The host's image model generates a variant from the base image; if the base is unavailable, generation fails without a text-only substitute. Generated variants are checked for lettering and modern objects and redrawn once if flagged. The second image is retained and its check result recorded even if it is flagged again; a flagged result is not a quality-pass guarantee. Selection runs under a guard that refuses to open anchor tables, question banks or judgments.

A.6 Comparison with SEAM and CANVAS

No single row is new on its own: SEAM already extracts structured state from the script as a training-free prompt-layer stage, and CANVAS already plans per-shot appearance and object states and uses them to pick visual anchors. What Scripty adds is the combination tested here: a record built only from upstream material with storyboard evidence for every value, which is never updated from the host's outputs, and which drives both the prompt text and state-specific reference variants inside an unchanged host. The experiments isolate the image side (L0 versus L1 and the re-rendering control) but not evidence binding by itself.

A.7 Known limitations of the evaluated engine

Rejected values read from scene images are logged rather than corrected. Objects that appear only in the storyboard depend on INIT reporting them, and some are missed. Names derived from scene images can differ between episodes, so cross-episode carry of such objects can miss. A recorded "in someone's hand" position is often not updated when the storyboard moves the object, and hand-held objects do not follow their holder's presence, which leaves some held props off screen in the prompt. Registering scene-image objects roughly doubles the size of the injected state. Storyboard objects still marked "not yet present" can be listed among the scene objects. The L0 reference image shows the state before an event, so a correct prompt can still be overruled by the image; L1 addresses this from the following unit.

Table 3: Mechanism comparison with the two closest systems, as described in their papers (Liu et al., 2026a; Mondal et al., 2026).

	SEAM	CANVAS	Scripty
Task and output	Rewritten shot prompts for downstream generators	End-to-end storyboard (keyframe image) generation; keyframes also fed to a video model in its appendix	State constraints and reference images for an existing video host
State source	Per-shot visual state extracted from the episode script	Global continuity plan by an LLM from shot descriptions	Host storyboard, asset lists and reference images only
Evidence per value	None attached	None attached	Provenance for every value (shot and storyboard line; asset list or scene image for some initial values)
Update	Memory graph over shots; retrieval of causally prior nodes, then a selective filter	Plan per shot; memory anchors updated after every generated frame	Proposed net change per unit, reviewed, applied shot by shot by a deterministic executor
Ending a state	Causal ordering only	No explicit rule	Until the storyboard removes or excludes it; no lifetime classes
Reference images	Not changed (text only)	Anchors extracted from generated frames	Base, a host-authored version, or a variant generated from the base for the recorded state
Reads generated output	No	Yes, to update anchors	No (open loop)
Integration	Training-free prompt-layer stage	Complete pipeline	Adapter into the host's prompt template and reference slots

A.8 Planned extension: flashback branches

The evaluated engine keeps a single linear chain of states per episode. A flashback breaks this chain: a shot set years earlier, or in a place the story has not shown, does not continue the current state. The planned extension re-runs INIT at such a shot and starts a new root; shots inside the flashback continue that root, and when the story returns to the present the next state continues the last mainline shot. The record then becomes a forest, and a flashback can show the boy with his braid without undoing the cut in the present. No reported number depends on this extension: none of the StateHold storyboards contains a flashback, and the SEAM-Bench run drops flashback shots before Construction.

B Further Related Work

Pipelines. MovieAgent decomposes a synopsis into scenes and shots with a hierarchical chain of thought, renders keyframes with a region-controlled image model and animates them (Wu et al., 2025). ViMax issues one model call per shot and threads identity through a registry of character portraits (Huang et al., 2026); it is our second host through an adapter (Appendix G). AniMaker plans each shot as duration, character set and background, keeps character and background banks, and selects clips by search (Shi et al., 2025). StoryMem conditions each shot on a purely visual memory of earlier keyframes (Zhang et al., 2025b), and VideoMemory keeps an entity-centric memory bank of characters, props and backgrounds (Zhou et al., 2026). MAViS assigns a per-shot design of several fields once, at design time, without carrying or updating them as the plot changes (Wang et al., 2026a). Captain Cinema generates keyframes for a whole short film and then interpolates them (Xiao et al., 2026); HoloCine generates multi-shot narratives holistically in one model (Meng et al., 2026); STAGE anchors multi-shot generation on storyboard frames (Zhang et al., 2026). CineCrew states

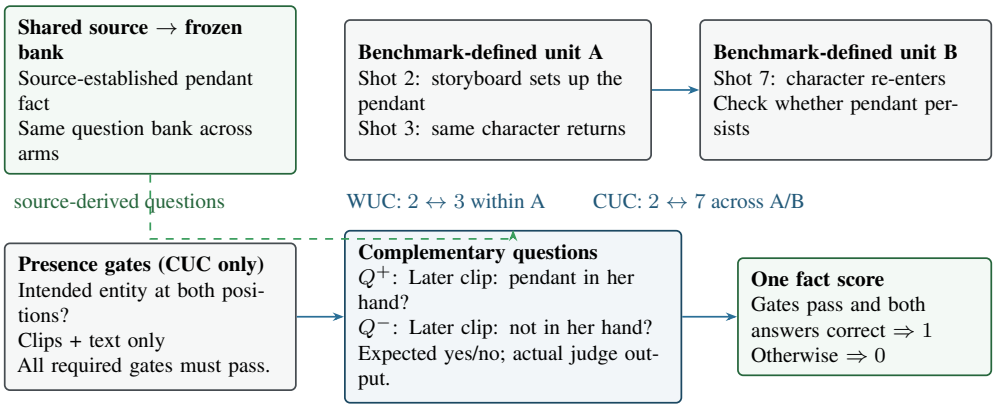

Figure 5: **StateHold separates shared truth from generated evidence.** Illustrative example. Unit boundaries come from the benchmark, independently of API call counts. The visual judge sees only clips and question text, with no reference images, including for presence gates; it sees neither the storyboard nor Scripty state. Unknown answers score zero in the fixed denominator. WUC has no presence gate; inference clusters by story.

cinematic and asset constraints in a film description language, anchors clips with asset packs and storyboard keyframes, and refines them with a critic agent that reads generated clips (Chen et al., 2026b). Further pipelines make related trade-offs between planning and asset reuse (Xu et al., 2024; Long et al., 2024; Zhuang et al., 2024; Li et al., 2024; Zheng et al., 2024; Chen et al., 2025; Wang et al., 2026c; 2025; Lai et al., 2026). None of these keeps an evidence-bound record of state changes from the storyboard that drives both the prompt and the reference images of a host it does not modify, which is the combination studied here (Table 3).

Persistent world and identity representations. FilmWorld formalizes novel-to-film generation as world modelling over a persistent world state with character, location and prop families, verifies rendered frames against its plan and regenerates (Zuo et al., 2026). IAMFlow maintains a persistent identity registry and writes attributes back from rendered frames (Liu et al., 2026b). Both read generated output and are coupled to their own pipelines; they are orthogonal to, and could be combined with, a state layer that works from upstream material only.

Benchmarks. MemoBench asks whether a state survives an absence produced by camera motion within one clip (Chen et al., 2026a); STEVO-Bench and WRBench report that video world models do not keep persistent state (Ma et al., 2026; Lu et al., 2026); StateBench tests whether generators keep visible, occluded and transitioning states across segments and proposes an entity-state agent (Miao et al., 2026). MSVBench scores state shift and persistence on multi-shot video with judge-drawn anchors and rubrics (Shi et al., 2026b); EntityBench fixes a per-shot entity schedule before generation and registers presence rather than change (He et al., 2026); DirectorBench scores scripts with explicit consistency requirements on a 1–5 rubric (Chen et al., 2026c); ViStoryBench evaluates story visualization (Zhuang et al., 2026); MSG Score compares key attributes across shots with visual questions (Yoon et al., 2026); PersonaShot scores character continuity on real footage (Wang et al., 2026b); DramaChain Bench scores the full short-drama pipeline with human annotators (Shi et al., 2026a); NarrativeWorldBench tests cross-episode state in audio drama (Mann et al., 2026). SEAM-Bench provides cross-episode short-drama scripts and text-layer judging (Liu et al., 2026a); we use its data under our own protocol (Appendix E). The VBench family measures single-clip quality and within-clip consistency on independent prompts (Huang et al., 2024; Zheng et al., 2025; Sun et al., 2024), which does not reach states that must hold across generation units. StateHold differs from these suites in fixing, before generation, which storyboard state each later unit must show, and in scoring it with paired questions and presence gates at the level of stories.

C StateHold Protocol and Analysis Contract

C.1 Bank construction and eligibility

Question construction uses upstream source and anchors common to the compared arms. Each fact identifies the established state, entity, source evidence, establishing/checking units, span class and any eligibility restriction. Its positive and negative questions are intended to negate the same proposition; semantic validity still requires checking. Presence gates are tied to the intended entity or location through a fixed textual description rather than being inferred from a treatment's output. Salience and visibility eligibility must be determined from the protocol, not from whether a system happened to render a detail.

Each fact belongs to CUC or WUC by generation-unit boundaries. A long gap is metadata rather than an independent sample. Execution/transition questions are excluded from a persistence pool. Registration of a contrast through `bears_on` remains separate from general schema membership. Empty registered pools are undefined and may not be reported as zero. Shared source reduces some confounding but cannot exclude treatment interactions with schema-aligned authoring. The sensitivity table reports the non-registered facts, the complement of the registered subset; this checks whether the gain is confined to the registered facts, not whether the bank favours the facts Scripty's schema can express.

C.2 Review and freeze

Mechanical validity, model review and human review are separate stages. The released bank includes the review procedure, the disposition of each edited or killed item, and its source hash. AI review is not counted as a second human rater.

Final bank: FORMAL (18 stories, 99 facts; freeze sha256 920e7ce4…) and EXPAND (24 stories, 121 facts; freeze sha256 0b6951ca…), frozen on 23 Sep 2026; digests identify the unanonymised manifests, and the shipped copies differ only by path neutralisation as documented in the supplementary package. Actual review: The banks underwent checks of source references, intended question-pair negation, unit binding and eligibility; a later audit found violations; the disposition of each edited or killed item ships with the released bank. Written review rules required that the positive and negative question share the same scope, gates be phrased as visible checks, jargon and personal names be removed, and where a change inside the checking unit would contradict the fact, the time qualifier be removed or only the shots before that change be judged; facts whose check shot restated the answer, or whose new state the storyboard never showed, were removed. A read-only model drafted each proposal, a second read-only model checked it independently, and the authors approved the changes. Drafting and checking used pooled quality counts only, never per-arm results or clips. No change altered an expected answer, entity, establishing unit or checking unit. Edits/kills: 89 (FORMAL) + 118 (EXPAND) reworded facts/0 (FORMAL) + 10 (EXPAND). Timing: Both banks were frozen on 23 Sep 2026. Scoring uses each arm's first judging pass; this is distinct from the retry rule inside one pass, under which a question pair whose two answers disagree is re-asked for up to three rounds (§C.3). The judge prompt digest is 22292fcc throughout, and the SEAM-Bench text check uses the same frozen engine (fc921a028) as the video runs.

Evaluator-information exposure. A read-only audit of the generation chain looked for evaluator-authored information reaching generation. In the final runs, the writing schedule that the host's adaptation plan had derived from the design-time anchor tables and the anchor labels in the asset census were removed for every arm, and every arm draws its reference images from the same frozen host asset library; the Scripty L1 selection runs with a guard that refuses to open anchor tables, question banks or judgments. INIT may look at all shots of an episode but may cite only the first as evidence.

C.3 Judging, cache identity and missing observations

The judge receives only the bound media and the question text; no reference image is provided, including for presence gates. It does not receive the source answer or Scripty state. The question pair is scored only if both answers are correct; "cannot determine" remains in the denominator as

zero. Within a judging pass, a fact's two questions are retried only when their answers are not mutually consistent, that is, both yes, both no, or either "cannot determine". Both questions are then judged again; if the second answers are consistent they are used, otherwise a third round is run and each question takes the majority of its three answers (no majority counts as "cannot determine"). A pair still inconsistent after three rounds is void: it scores zero and is listed separately. Consistent answers that give the wrong state are never retried; gate questions, which are also asked as positive and negative pairs, follow the same rule. Over ten layers (five arms across both pools; fact and gate pairs; the sixth main-table arm, SEAM+recap, was never judged under this rule), 290 of 2,340 pairs (12.4%) triggered a retry; 133 were resolved in the second round, 7 in the third, and 150 (6.4% of all pairs) were void. Per-arm counts are in Table 4. Within one arm's judging layers, a pair whose input is unchanged keeps its original answer instead of being drawn again, the rule by which L1 keeps the L0 rows of untouched units; each baseline arm is judged on its own fresh pass. API/backend

Table 4: Judge pair retries per arm on the main runs (excludes SEAM+recap, which was not judged under the retry rule): question pairs (fact and gate) / pairs retried / resolved in the second round / resolved in the third round / void.

Arm	FORMAL	EXPAND
No-carry	191 / 21 / 13 / 0 / 8	277 / 31 / 15 / 0 / 16
Native recap	191 / 24 / 12 / 0 / 12	277 / 36 / 15 / 1 / 20
SEAM (in-host)	191 / 25 / 10 / 0 / 15	277 / 32 / 16 / 0 / 16
Scripty L0	191 / 19 / 6 / 1 / 12	277 / 41 / 19 / 2 / 20
Scripty L1	191 / 22 / 8 / 0 / 14	277 / 39 / 19 / 3 / 17

changes cannot silently reuse a cell from another instrument.

Cache identity binds model/backend, prompt/version, question text and ordered media hashes. A compatible cached answer is not a fresh replicate. Every arm is scored on its first judging pass on the frozen banks, so all arms judged under the retry rule (five of six) follow the same judging-pass and retry policy; this does not by itself make the judge's noise identical across arms. Over the three baseline arms on EXPAND (80 cross-unit and 41 within-unit facts each), agreement between the first and a second pass is 90.0–97.6%; this measures repeatability, not accuracy. Human agreement is reported separately: performed as a separate fact-level study, not as part of bank review. Six non-author raters independently judged 40 sampled facts (10 CUC and 10 WUC per pool, 34 stories) on no-carry and native-recap clips, blind to method, two raters per fact; all 80 primary case reviews and all 24 third ratings on disagreeing cases were returned, and 37 facts are resolved for both arms. Pre-adjudication agreement between raters is 70.6% over 320 question-by-arm items. Against the adjudicated human reference the automated judge agrees on 62 of 75 fact-arm panels (82.7%, Cohen's kappa 0.66); it fails 11 of the 44 panels humans passed, 10 of them on the WUC guardrail, so it is stricter than humans there, and it passes 2 of the 31 panels humans did not pass, both panels that humans could not judge. Boundaries: raters and judge answered the 17 Sep question wording, not the frozen wording scored in the main table (against the frozen FORMAL and EXPAND banks, 34 of 40 facts match a bank fact, 19 FORMAL and 15 EXPAND; C015 is stricter in the frozen bank, which asks whether the jar stays the same jar rather than whether it is still a wine jar; C038 was removed from EXPAND because answering the guardrail item required cross-unit memory; and C031, C032, C039 and C040 were never in the bank). The frozen wording itself has not been re-rated by humans; only the two baseline arms were shown; and the sample is small. The study supports that the items are human-judgeable and shows the direction of judge error; it is not evidence about Scripty's effect. For a run failure, the ledger states the intended source range, accepted prefix, released output and missing downstream stages. An accepted-only score cannot be called full-coverage accuracy. Report actual-output and full-denominator/sensitivity readings under the frozen missing-data policy, rather than dropping failures after seeing their sign.

C.4 Fact-level human study of the judge

A seed fixed before any human answer drew 10 CUC and 10 WUC facts per pool, sampling stories first and then facts, with inclusion probabilities retained; each fact is shown for the no-carry and native-recap arms only. Each fact went to two of six non-author raters (native Chinese speakers, friends of an author unconnected to the research, with no fixed payment), who saw the two arms

as anonymous panels with seeded left/right placement, re-encoded from the original renders under the frozen time windows, with an author-reviewed Chinese translation of the question text. Raters answered gate and fact questions with yes, no or cannot judge, the last reserved for missing visual evidence. A disagreement went to a third non-author rater who had not seen the case and did not see earlier answers; each question was decided by majority, a three-way split stays unresolved, and the automated judge never acts as a rater. Fact outcomes reuse the frozen scoring rule (all required gates and both fact questions correct). Judge answers for the same cases were produced under the same presentation and compared case by case. We report pre-adjudication agreement, human-reference coverage, false-pass and false-fail rates and method direction (Table 21; summary in the judging subsection above); results are in-sample and conditional on this rater panel. The study validates the instrument on the rated items; it is not a preference study and not evidence about Scripty. Human fact outcomes, passed over resolved facts, no-carry vs native recap: FORMAL CUC 0/9 vs 3/10 and WUC 10/10 vs 9/10; EXPAND CUC 0/8 vs 2/9 and WUC 10/10 vs 10/10 (in EXPAND CUC, 2 no-carry and 3 native-recap non-passes are panels the raters marked as not visible). The native recap minus no-carry difference in pass proportion is paired over facts resolved for both arms, so it is not the difference of the two ratios: FORMAL CUC +0.33 [0.00, 0.67] over 9 paired facts and WUC -0.10 [-0.30, 0.00] over 10; EXPAND CUC +0.13 [0.00, 0.38] over 8 and WUC 0.00 over 10 (story-level bootstrap, 2000 draws, conditional on this rater panel). Without C021, C005 and C029 (37 facts): FORMAL CUC 0/8 vs 2/9, paired +0.25 [0.00, 0.63] over 8; EXPAND CUC 0/6 vs 2/7, paired +0.17 [0.00, 0.50] over 6; the guardrail strata are unchanged, and pre-adjudication agreement is 71.5% over 284 items. Source review of the sampled items: An author's source review after collection, which changed no collected answer, passed the rated materials for all 80 panels (time windows correct, no identity cues) and the source for 39 of 40 facts. It flagged C021 (the farm tool named in the source is unclear, and its positive and negative questions are not of the same kind), C005 (the break may lie on a stretch of the path that is never filmed) and C029 (the gate region is too vague), and it judged the wording unclear for 28 of 40 facts (FORMAL: the phrase 'same checking time'; EXPAND: positive and negative questions worded differently and a boilerplate scope sentence). All 40 facts are kept as collected. In this small sample the flagged facts show no more rater disagreement than the others (question-level pre-adjudication agreement 73.4% on the 28 flagged facts, 59.4% on the other 12); this does not show that the wording had no effect. The rated wording predates the released banks, whose current wording has not been re-rated by humans.

C.5 State-record audit protocol

For a few stories per pool, a source inventory of state changes is compiled from the storyboards by a reader who does not see Scripty's record: each entry names the entity, the category (appearance, held or carried object, presence and position, scene object), the new state and the shot. A recorded change matches an inventory entry when the same entity reaches the same state within the same unit; precision is the share of recorded changes that an entry supports, and recall the share of entries that the record contains (Table 17). The audit therefore measures unit-level coverage of changes, not whether the record places a change at the right shot. The record is meant to keep the changes that matter for later units; a change undone within the same unit is usually not recorded. An omission found here is charged to Construction once and not counted again downstream.

C.6 Statistical unit and reading rules

Pass percentages are pooled fact counts, while uncertainty is computed over complete stories. The focal difference is Scripty L1 minus native recap on the registered cross-unit facts, with full-pool results beside it. A two-sided exact sign test over stories excludes ties and reports them separately. Bootstrap intervals recompute the same pooled difference after each resample of whole stories (2000 resamples, seed 20260915).

The WUC floor is -10 pp against the native recap and no-carry. An interval crossing the floor does not establish non-inferiority, and a non-significant difference from zero is not a substitute for that test. Adverse point estimates are shown even when precision is poor. Significance language applies only to the declared focal contrast; span bins, other hosts, exploratory SEAM metrics and mechanism probes retain their stated descriptive scope.

C.7 RULE FAMILIES AND WHAT EACH CAN ESTABLISH

Table 5 retains the checks needed to interpret StateHold without reproducing the chronological development of the instrument. Structural and semantic checks have different powers: a program can verify that both question strings exist, while logical complementarity and visual decidability still require semantic review. A report must identify the checks actually exercised by the frozen instrument.

Table 5: StateHold rule families, their responsibility and their evidentiary limit.

Rule family	Required check	Limit of that check
Source provenance	Establishment, checking interval and source evidence exist	A valid address does not establish entailment
Unit binding	Bound shots exist and pool assignment follows unit boundaries	Does not establish that the state is visible
Pair structure	Positive and negative halves exist, are distinct and share a fact	Logical complementarity needs semantic review
Persistence scope	The item asks about retained state, not execution of an action	Plausible persistence still needs a supported interval
Carried-state eligibility	Silent interval and any invalidating context are inspected	Source silence is not unlimited validity
Visual decidability	Target attribute and exposure can support a media-only answer	Fine detail or audio evidence can invalidate an item
Presence gates	Intended entity or location is fixed by a textual description	A gate failure is not a factual-state judgment
Registration	Contrast membership comes from explicit anchors	Registration is not inferred from treatment success
Duplication and quotas	Duplicates identified; availability and shortfalls retained	Quotas must not manufacture unsupported facts
Freeze identity	Bank, anchor, media and instrument identities are bound	Hash agreement does not establish semantic correctness
Disposition history	Kills, edits and exclusions retain reasons and exposure timing	Post-readout changes remain post-readout
Judgment consistency	Complementary answers and retries follow a fixed rule	Repeatability does not establish human agreement

Specific quota and salience settings belong to the frozen instrument and accompanying bank, not a universal promise that every story supplies the same number of facts. An unavailable fact is reported as a shortfall; it is not repaired by inspecting which output would make a convenient test. A registered subset can legitimately contain fewer contributing stories than the full bank, and that difference is printed beside the rates.

C.8 A WORKED FACT AND ITS FAILURE MODES

Consider a constructed example in which a source establishes a cup on a table at the end of unit A, and later source leaves that placement unchanged during a visible return in unit C. The intended fact is the cup's placement, not whether the earlier placing action was well performed. The judge sees both bound clips; the positive question asks whether, in the later clip, the cup is on the table, and the negative asks whether it is not. Their expected answers are yes and no respectively, under the same exposure and identity conditions.

The two questions must cover the same scope. Asking whether the cup remains in C and whether Mara moved it during B would introduce an action question and would not be a logical complement. If the cup cannot be identified in C, its gate fails; if it is identified but the target location cannot be decided, the response is "cannot determine". Both produce zero under the primary score, but their reasons differ. If the source itself never makes the target visually answerable, that is an item-validity problem to be recorded under the review/disposition procedure, not simply another system error.

The same reasoning applies to WUC, except its gates are absent under the declared protocol. A source-supported state change between the two queried moments makes a persistence premise inap-

propriate; it must not be counted as an error merely because a system represents that change. This is also why an NLI contradiction indicator cannot substitute for the fact protocol.

C.9 Pooled differences and story-level uncertainty

Let F_s be the eligible facts of story s for one declared pool and contrast, and let $Y_{a,f}$ be the binary score of Equation 1. The reported pooled difference is

$$\widehat{\Delta} = 100\,\frac{\sum_s \sum_{f\in F_s}(Y_{A,f} - Y_{B,f})}{\sum_s |F_s|}, \qquad d_s = \frac{\sum_{f\in F_s}(Y_{A,f} - Y_{B,f})}{|F_s|}. \tag{2}$$

$\widehat{\Delta}$ is in percentage points. The sign test uses the nonzero signs of d_s; if F_s is empty, d_s is undefined and the story contributes no sign. Zero differences are ties, not missing stories. A story-level bootstrap samples stories and recomputes Equation 2, keeping the paired arms and all sampled facts together. It must not silently substitute a mean of per-story percentages for the pooled estimand. The exact bootstrap settings and any secondary signed-rank handling are recorded with the analysis.

Large numbers of facts from one generated story do not create additional independent replicates. Likewise, a longer accepted prefix supplies more observations but not a new independent run. Per-host and unequal-writer strata remain separate. When the final bank follows trial exposure, the report retains that limitation regardless of the nominal p-value or interval obtained from these computations.

C.10 What sensitivity analyses diagnose

Gate-free scoring asks how much the reading depends on identity recognition; it does not establish a causal effect after controlling for gates. Abstention sensitivity asks how a specified alternative treatment of indeterminate answers affects the result. Killed-item restoration is meaningful only when those items can still be scored under a defensible counterfactual; an invalid premise should not be restored merely to obtain a larger table. Schema stratification examines whether question design favours facts represented by the system. Full-range accounting includes failed or unprocessed portions under the declared missing-data policy. These analyses answer different objections and cannot substitute for one another. The existing sensitivity table provides their result and denominator slots. They are reported as sensitivity readings, including the worst-case abstention imputation; none of them is a precondition for stating the primary contrast.

D Full Results

The main table combines 42 stories; FORMAL/EXPAND below identify source-and-length strata, not separate primary corpora.

D.1 Inference, sensitivity and controls

Tables 8 and 9 give, for each pool, the preregistered subset and the full pool with per-story signs, exact sign-test p values and the within-unit guardrail. Table 10 repeats the focal contrast under gate-free scoring, two treatments of undecidable answers and the non-registered subset. Table 11 compares the baselines with each other, and Table 12 holds the length-matched history-summary control. The summary is about as long as Scripty's injected state (median ratio 0.90 / 0.94 for FORMAL / EXPAND, interquartile ranges 0.83–0.98 / 0.87–1.02; 80 of 90 and 174 of 192 units within 0.8–1.2), and the writer's first prompt has essentially the same length in both arms (median ratio 0.998 / 1.002). Both arms have the same cap of two check-driven rewrites, but the summary's checker is more lenient than Scripty's prompt checker, so the control actually checked and rewrote less: no, one and two rewrites in 30, 34 and 26 units against 16, 22 and 52 for Scripty (FORMAL), and 100, 65 and 27 against 43, 50 and 99 (EXPAND). The call budget is therefore not matched: the control makes about half as many model calls (6.4 / 5.7 per unit against 11.8 / 11.5). The sources also differ by design, and the summarizer's are a subset of Scripty's: Scripty's INIT sees every shot of the episode to find the entities (it may cite only the first shot), whereas the summarizer sees the storyboard only up to the current unit, because a free-text summary has no mechanical check against writing later events; it receives the reference-image names but not the images. The bootstrap

Table 6: Video results with the same writer model across arms; call and checking budgets differ. CUC, WUC and Δ CUC are on the full pools (L1 minus recap: sign test $p = 0.013$ / 0.019). The last column repeats the primary comparison on the preregistered subset only (20 / 29 facts).

Method	Stories	CUC (%)	WUC (%)	Δ CUC [95% CI]	Subset Δ, p
FORMAL stratum					
No-carry	18	4.3	75.0	-27.7 [-40.5, -15.6]	–
Native recap	18	31.9	73.1	N/A (reference arm)	–
SEAM (in-host)	18	19.1	78.8	-12.8 [-31.9, +4.3]	–
SEAM + recap	18	48.9	75.0	+17.0 [+2.2, +31.1]	–
Scripty L0 (text only)	18	51.1	82.7	+19.1 [-2.3, +39.5]	–
Scripty L1	18	63.8	76.9	+31.9 [+14.0, +48.9]	+35.0 [0.0, +66.7], $p = 0.146$
EXPAND stratum					
No-carry	24	11.2	85.4	-11.2 [-23.9, -1.2]	–
Native recap	24	22.5	85.4	N/A (reference arm)	–
SEAM (in-host)	24	18.8	82.9	-3.8 [-14.3, +5.3]	–
SEAM + recap	24	26.2	87.8	+3.8 [−8.8, +13.1]	–
Scripty L0 (text only)	24	32.5	85.4	+10.0 [-2.3, +23.4]	–
Scripty L1	24	40.0	85.4	+17.5 [+2.6, +31.6]	+31.0 [+11.8, +50.0], $p = 0.012$

CUC / WUC: pass rates on 47 / 52 facts over 18 stories (FORMAL) and 80 / 41 over 24 (EXPAND). No fact was lost to a failed render; uncertain or void answers count as not passed. Δ CUC: against the native recap in the same pool, pp, story-level bootstrap intervals; sign tests and the preregistered subset are in Tables 8 and 9. L1 is the full system; L0 adds text constraints and keeps the host's references. Scripty arms also receive the native recap; SEAM repairs the no-carry prompts and receives no recap. SEAM + recap is an additional control; neither L1 contrast against it is significant (Appendix D.4).

Table 7: Where the gain comes from. Each row adds one component with the other held fixed and gives the change in pass rate, in percentage points with story-level bootstrap 95% intervals. ‡Whole chain on DeepSeek-V4.1-Flash, Scripty L1 minus the native recap on the same model (EXPAND only).

	FORMAL		EXPAND	
Change	Δ CUC	Δ WUC	Δ CUC	Δ WUC
Add L1 to L0	+12.8 [0.0, +26.1]	-5.8 [-11.5, 0.0]	+7.5 [-2.4, +17.3]	0.0 [-7.3, +7.3]
Scene-image objects (L0)	+14.9 [0.0, +30.6]	+13.5 [+1.9, +26.4]	+2.5 [-12.3, +18.2]	+7.3 [-6.2, +22.5]
Scene-image objects (L1)	+10.6 [-4.8, +24.4]	+13.5 [-3.9, +30.2]	+7.5 [-7.1, +22.5]	+4.9 [-12.8, +23.1]
DeepSeek chain‡	–	–	+22.5 [+10.0, +36.1]	+4.9 [-7.0, +20.0]

interval and the sign test answer different questions: the first bounds the pooled fact-level difference, the second asks whether more stories move one way than the other, so on FORMAL the interval can exclude zero while the sign test is not significant. A second control removes the prompt check from Scripty L0 on FORMAL: the state record and the injected text are byte-identical to L0, but no writer check runs and no check-driven rewrite happens. Cross-unit it passes 46.8%, -4.3 [-16.7, $+9.8$] from L0 ($p = 0.45$) and $+2.1$ [-12.2, $+16.7$] from the repeat run of L0, so the change lies within the spread of two repeat runs (51.1% and 44.7%); 70.2% of its facts have the same outcome as L0, against 59.6% between the two repeats. It still leads no-carry ($+42.6$ [$+29.8$, $+55.1$], $p < 0.001$), SEAM ($+27.7$ [$+8.5$, $+44.9$], $p = 0.012$) and the native recap ($+14.9$ [0.0, $+29.5$], $p = 0.18$). Within-unit it drops to 73.1% (-9.6 [-22.2, $+3.9$], $p = 0.34$), consistent with the check's role of making the writer include the unit's own props and states. Denominators per pool and arm: every arm is scored on all 47 cross-unit and 52 within-unit facts of the 18 FORMAL stories and on all 80 cross-unit and 41 within-unit facts of the 24 EXPAND stories (23 of which have within-unit facts); no fact is dropped for a missing clip.

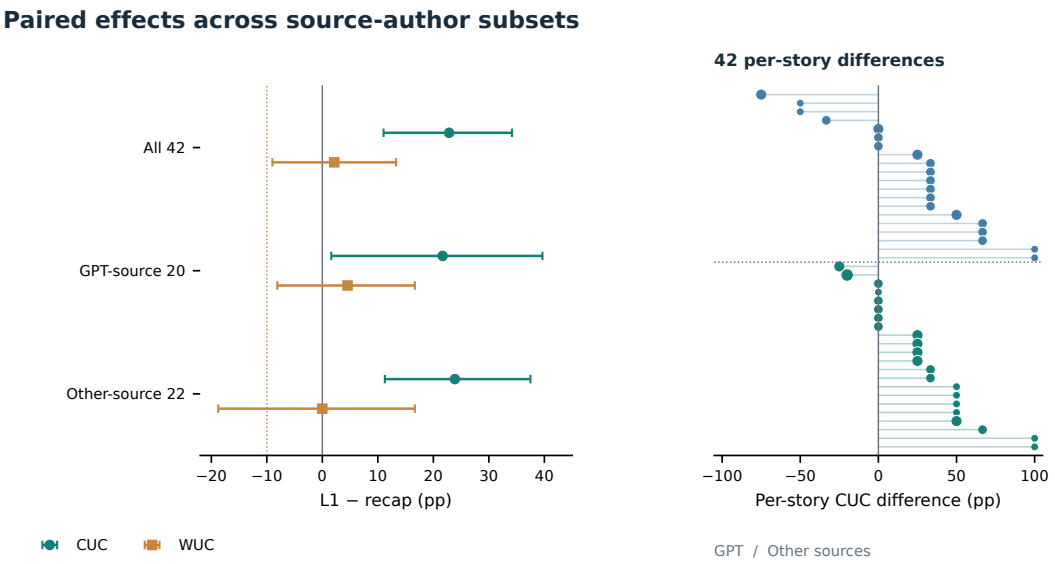

Figure 6: **Unified result and source-author subsets.** Fact-weighted L1 minus native-recap differences with whole-story bootstrap 95% intervals. The right panel shows all 42 per-story CUC differences, grouped by source author; marker area scales with fact count. Source grouping is observational. WUC has 93 facts in 41 stories. The dashed line is zero; the dotted line is the -10 pp WUC guardrail.

Ablation pass rates. Pass rates CUC / WUC (%), FORMAL / EXPAND: full Scripty 63.8 / 40.0, 76.9 / 85.4; L0 51.1 / 32.5, 82.7 / 85.4; no scene-image objects with L0 36.2 / 30.0, 69.2 / 78.0; with L1 53.2 / 32.5, 63.5 / 80.5; DeepSeek chain (EXPAND only) 47.5, 92.7. Differences are in Table 2.

Table 8: Primary inference on FORMAL: Scripty L1 (A) against each comparator (B), on the pre-registered subset and the full pool.

Comparator / reading	Pool	Stories	Facts	A (%)	B (%)	Δ (pp)	95% CI
Native recap / registered	CUC	16	20	75.0	40.0	+35.0	[0.0, +66.7]
Native recap / full	CUC	18	47	63.8	31.9	+31.9	[+14.0, +48.9]
No-carry / full	CUC	18	47	63.8	4.3	+59.6	[+47.8, +71.4]
Native recap / guardrail	WUC	18	52	76.9	73.1	+3.8	[-12.0, +19.2]
No-carry / guardrail	WUC	18	52	76.9	75.0	+1.9	[-14.3, +17.0]

Focal CUC signs (registered / full): 9/4/3 / 12/4/2 (positive/tied/negative; ties excluded from the sign test); two-sided exact sign p: 0.146 / 0.013. WUC lower bound: -10 pp.

Table 9: Primary inference on EXPAND, with the same structure as Table 8: Scripty L1 (A) against each comparator (B).

Comparator / reading	Pool	Stories	Facts	A (%)	B (%)	Δ (pp)	95% CI
Native recap / registered	CUC	18	29	62.1	31.0	+31.0	[+11.8, +50.0]
Native recap / full	CUC	24	80	40.0	22.5	+17.5	[+2.6, +31.6]
No-carry / full	CUC	24	80	40.0	11.2	+28.8	[+17.1, +41.2]
Native recap / guardrail	WUC	23	41	85.4	85.4	0.0	[-14.9, +13.6]
No-carry / guardrail	WUC	23	41	85.4	85.4	0.0	[-15.0, +14.0]

Focal CUC signs (registered / full): 10/7/1 / 15/5/4 (positive/tied/negative; ties excluded from the sign test); two-sided exact sign p: 0.012 / 0.019. WUC lower bound: -10 pp.

Table 10: Sensitivity of the primary contrast (Scripty L1 minus native recap, cross-unit facts), per pool, with the definitions fixed in Appendix C.

Reading	Stories / facts	CUC Δ (pp)	Status / interval
FORMAL			
Gate-free	18 stories / 47 facts	+34.0	computed; 95% CI [+17.8, +47.8]; signs 14/2/2, sign p = 0.004
Symmetric removal of undecidable facts	17 stories / 36 facts	+38.9	symmetric removal (uncertain/void facts of either arm dropped from both); 95% CI [+19.0, +58.8]; signs 12/3/2, sign p = 0.013. Symmetric not-passed scoring is the main reading (+31.9). Bounds, not sensitivity readings: comparator's own uncertain/void facts counted as passed +19.1 [0.0, +36.2]; every such fact resolved against Scripty +6.4 [-19.1, +27.1]
Killed items restored (if defensible)	6 killed FORMAL facts	N/A	not performed: killed items were never judged on the scored videos and their premises were adjudicated invalid
Non-registered facts	17 stories / 27 facts	+29.6	non-registered facts; 95% CI [+7.7, +52.0]; signs 7/9/1, sign p = 0.070; registered subset +35.0 [0.0, +66.7] over 20 facts in 16 stories
Intended full-range / failure accounting	18 stories / 47 facts	+31.9	identical to full: all 18 stories complete in both arms, 0 missing-fact rows, 0 generation-failure facts; 95% CI [+14.0, +48.9]; signs 12/4/2, sign p = 0.013
EXPAND			
Gate-free	24 stories / 80 facts	+13.8	computed; 95% CI [0.0, +27.4]; signs 13/7/4, sign p = 0.049
Symmetric removal of undecidable facts	23 stories / 56 facts	+21.4	symmetric removal (uncertain/void facts of either arm dropped from both); 95% CI [+3.8, +37.0]; signs 14/5/4, sign p = 0.031. Symmetric not-passed scoring is the main reading (+17.5). Bounds, not sensitivity readings: comparator's own uncertain/void facts counted as passed -2.5 [-21.0, +13.9]; every such fact resolved against Scripty -15.0 [-36.0, +4.0]
Killed items restored (if defensible)	10 facts killed in the blind .1 to .2 revision (5 CUC)	N/A	not performed: killed items were removed before the frozen .2 bank was judged; no first-pass answers exist for them
Non-registered facts	24 stories / 51 facts	+9.8	non-registered facts; 95% CI [-10.0, +30.0]; signs 8/12/4, sign p = 0.388; registered subset +31.0 [+11.8, +50.0] over 29 facts in 18 stories
Intended full-range / failure accounting	24 stories / 80 facts	+17.5	identical to full: all 24 stories complete in both arms, 0 missing-fact rows, 0 generation-failure facts; 95% CI [+2.6, +31.6]; signs 15/5/4, sign p = 0.019

Table 11: Supplementary full-pool contrasts, by pool. All arms use matched writers; SEAM means our reimplementation in the same host. Story-level bootstrap 95% intervals.

Comparison / stratum	Stories	CUC Δ (pp)	WUC Δ (pp)
FORMAL			
Native recap vs. no-carry, matched writers	18	+27.7 [+15.6, +40.5]	-1.9 [-11.5, +8.0]
SEAM reimpl. vs. no-carry, matched writers	18	+14.9 [+2.4, +27.1]	+3.8 [-9.4, +17.3]
Scripty L1 vs. SEAM reimpl., matched writers	18	+44.7 [+26.5, +62.2]	-1.9 [-19.6, +13.7]
EXPAND			
Native recap vs. no-carry, matched writers	24	+11.2 [+1.2, +23.9]	0.0 [-9.5, +10.3]
SEAM reimpl. vs. no-carry, matched writers	24	+7.5 [+1.2, +14.9]	-2.4 [-10.0, +5.7]
Scripty L1 vs. SEAM reimpl., matched writers	24	+21.2 [+7.8, +35.4]	+2.4 [-14.6, +17.5]

Table 12: Length-matched history-summary control and a no-prompt-check control. The history-summary control replaces Scripty's injected state with a model-written summary of the storyboard up to the current unit, of similar length, followed by a consistency check with the same cap on rewrites; writer, host recap, reference images, renderer and judge are unchanged. The no-prompt-check control keeps Scripty L0's state and injected text and removes the prompt check. Differences are in percentage points with story-level bootstrap 95% intervals; p is the per-story exact sign test.

	FORMAL		EXPAND	
Condition	CUC (%)	WUC (%)	CUC (%)	WUC (%)
Native recap	31.9	73.1	22.5	85.4
History summary + check	21.3	71.2	15.0	85.4
Scripty L0	51.1	82.7	32.5	85.4
L0 without prompt check	46.8	73.1	38.8	75.6
Scripty L1 (full)	63.8	76.9	40.0	85.4
Summary − Scripty L0	-29.8 [-50.0, -7.3]	-11.5 [-26.4, +3.8]	-17.5 [-28.4, -6.9]	+0.0 [-16.7, +15.8]
p	0.118	0.508	0.013	1.000
Summary − native recap	-10.6 [-26.9, +7.0]	-1.9 [-13.5, +9.6]	-7.5 [-17.7, +2.4]	+0.0 [-16.3, +14.0]
p	0.344	1.000	0.388	1.000

Scope: FORMAL (18 stories, 47 CUC / 52 WUC facts) and EXPAND (24 stories, 80 CUC / 41 WUC facts), first judging pass; every bank-referenced unit of the control rendered and judged (18 of 18 and 24 of 24 stories, no missing facts). Also: history-summary control minus no-carry, FORMAL: CUC +17.0 [+6.4, +29.3] ($p = 0.016$), WUC -3.8 [-15.7, +11.1] ($p = 0.508$); EXPAND: CUC +3.8 [-5.1, +12.2] ($p = 0.581$), WUC +0.0 [-11.9, +11.6] ($p = 1.000$). Intervals are story-level bootstrap 95% intervals (2000 resamples) of the pooled difference; p is the two-sided exact per-story sign test.

Why cross-unit facts fail. We split each failed cross-unit fact into four cases: the entity is missing from the establishing clip, the entity is not identified in the later clip, both gates pass but the later clip shows the wrong state (a later-state mismatch), or an answer is undecidable. For Scripty L1, on FORMAL, 17 of 47 facts fail: 53% later-state mismatch, 12% entity not identified later, 0% entity absent from the establishing clip, 35% undecidable; on EXPAND, 48 of 80 facts fail: 67% later-state mismatch, 2% entity not identified later, 4% entity absent from the establishing clip, 27% undecidable. For Scripty L0, on FORMAL, 23 of 47 facts fail: 65% later-state mismatch, 9% entity not identified later, 0% entity absent from the establishing clip, 26% undecidable; on EXPAND, 54 of 80 facts fail: 65% later-state mismatch, 0% entity not identified later, 4% entity absent from the establishing clip, 31% undecidable. For the three baselines, later-state mismatches make up 78–89% of their failures on FORMAL and 71–83% on EXPAND. Presence-gate failures are rare in every arm, so most failures are wrong states in the later clip with the entity present in both clips. Because no question checks the establishing clip's state, these cannot be split into states never drawn and states drawn and then lost.

Repeating the whole chain. We ran the whole chain three times on the same stories (Construction, writer, renderer and judge all drawn again; for L1 the variant selection and images as well); the first run is the main table. Cross-unit pass rates of the three runs: FORMAL L0: 51.1%, 44.7% and 53.2% (lead over the native recap +19.1 pp, $p = 0.227$, +12.8 pp, $p = 0.388$, +21.3 pp, $p = 0.039$); 45% of cross-unit and 62% of within-unit facts have the same outcome in all three runs; averaging the three runs per fact, L0 minus the native recap is +17.7 [+1.5, +32.7] pp ($p = 0.180$); FORMAL L1: 63.8%, 63.8% and 57.4% (lead over the native recap +31.9 pp, $p = 0.013$, +31.9 pp, $p = 0.021$, +25.5 pp, $p = 0.021$); 49% of cross-unit and 60% of within-unit facts have the same outcome in all three runs; averaging the three runs per fact, L1 minus the native recap is +29.8 [+14.4, +45.5] pp ($p = 0.021$); EXPAND L0: 32.5%, 36.2% and 38.8% (lead over the native recap +10.0 pp, $p = 0.267$, +13.8 pp, $p = 0.267$, +16.2 pp, $p = 0.118$); 42% of cross-unit and 68% of within-unit facts have the same outcome in all three runs; averaging the three runs per fact, L0 minus the native recap is +13.3 [+2.4, +23.4] pp ($p = 0.012$); EXPAND L1: 40.0%, 45.0% and 43.8% (lead over the native recap +17.5 pp, $p = 0.019$, +22.5 pp, $p = 0.004$, +21.2 pp, $p = 0.021$); 51% of cross-unit and 71% of within-unit facts have the same outcome in all three runs; averaging the three runs per fact, L1 minus the native recap is +20.4 [+7.8, +32.1] pp ($p = 0.011$). On FORMAL, the L1-minus-L0 increment across the three draws is +12.7, +19.1 and +4.2 pp. The between-run standard deviation of the cross-unit pass rate is 2.6–4.4 pp. The pooled rates and the lead of L1 over the recap are stable across runs, but single facts flip often. The experiment does not separate Construction from downstream randomness, and three runs give only a rough estimate of run-to-run spread.

D.2 Exploratory: increment by unit gap

Exploratory, not preregistered: cross-unit facts grouped by the number of units between the establishing and the checked unit (pp, story-level bootstrap interval for L1 minus recap; no significance claims, the buckets are small and unbalanced across stories). On FORMAL: 1–2 units apart (41 facts): native recap 13, L0 21, L1 26 passed; L0 minus recap +19.5, L1 minus recap +31.7 [+13.2, +50.0], L1 minus L0 +12.2; 3–4 units apart (6 facts): native recap 2, L0 3, L1 4 passed; L0 minus recap +16.7, L1 minus recap +33.3 [0.0, +66.7], L1 minus L0 +16.7. On EXPAND: 1–2 units apart (17 facts): native recap 3, L0 6, L1 10 passed; L0 minus recap +17.6, L1 minus recap +41.2 [+20.0, +64.3], L1 minus L0 +23.5; 3–4 units apart (35 facts): native recap 10, L0 14, L1 18 passed; L0 minus recap +11.4, L1 minus recap +22.9 [+3.3, +41.2], L1 minus L0 +11.4; 5–8 units apart (28 facts): native recap 5, L0 6, L1 4 passed; L0 minus recap +3.6, L1 minus recap -3.6 [-23.3, +17.4], L1 minus L0 -7.1. Comparing the text-only L0 with L1 separates a decline already present on the text side from one introduced by the image side.

D.3 Text-layer diagnostic

A text-layer judge from a third model family (DeepSeek-V4.1-Flash) reads blinded check-unit prompts against evaluator-defined states, without storyboards, reference images or arm names. Across 42 stories and 127 facts, Scripty writes the correct state in 111/127 (87.4%) prompts versus 75/127 (59.1%) for recap: +28.3 [+19.4, +38.2] pp ($p < 0.001$). Once written correctly, video pass rates are 40.5% for L0, 53.2% for L1, 37.3% for recap and 32.7% for SEAM. These conditional sets differ, so the differences do not identify causal stages. Table 13 reports pooled counts; the source-stratum readings are retained separately. In 197 prompts scored by DeepSeek and Gemini, category agreement is 97.0% ($\kappa = 0.94$); this is a cross-judge check, not human ground truth.

Table 13: Text-layer diagnostic on all 42 stories and 127 cross-unit facts. Correct/wrong/absent state mentions are scored independently of videos; conditional video rates describe different fact sets across arms. L1 uses the same text as L0.

Arm	Correct count	Correct / wrong / absent (%)	Video pass if correct (%)
No-carry	4/127	3.1 / 7.9 / 89.0	0.0
Native recap	75/127	59.1 / 3.9 / 37.0	37.3
SEAM (in-host)	49/127	38.6 / 9.4 / 52.0	32.7
History summary + check	47/127	37.0 / 4.7 / 58.3	31.9
Scripty L0	111/127	87.4 / 0.8 / 11.8	40.5
Scripty L1	111/127	87.4 / 0.8 / 11.8	53.2

Table 14: Text-layer diagnostic: share of cross-unit check prompts that write the tested new state correctly / wrongly / not at all (%), and the share of correctly written facts that pass on video (%). Scripty L1 writes the same text as L0 and differs only in the reference images.

	FORMAL (47 facts)		EXPAND (80 facts)	
Arm	Correct / wrong / absent	Pass if correct	Correct / wrong / absent	Pass if correct
No-carry	8.5 / 10.6 / 80.9	0.0	0.0 / 6.2 / 93.8	–
Native recap	57.4 / 4.3 / 38.3	48.1	60.0 / 3.8 / 36.2	31.2
SEAM (in-host)	36.2 / 10.6 / 53.2	47.1	40.0 / 8.8 / 51.2	25.0
History summary + check	42.6 / 4.3 / 53.2	45.0	33.8 / 5.0 / 61.2	22.2
Scripty L0	93.6 / 0.0 / 6.4	54.5	83.8 / 1.2 / 15.0	31.3
Scripty L1	93.6 / 0.0 / 6.4	68.2	83.8 / 1.2 / 15.0	43.3

Further arms and a second judge. With the same rule (whether the prompt sent to the video model writes the anchor's new state), Scripty L0 writes the state in 90.0% of EXPAND cross-unit prompts in the DeepSeek chain against 45.0% for the native recap on the same backbone ($+45.0$ [$+31.6$, $+57.9$] pp, $p = 0.0001$); because DeepSeek writes these prompts, a Gemini judge scores them, and DeepSeek as a cross-check gives $+42.5$ (agreement 96.2%, $\kappa = 0.93$). Without the prompt check, L0 writes 81.2% on EXPAND, 2.5 pp below sample A ($p = 0.77$) and 21.2 pp above the recap ($p = 0.013$), so the check hardly changes whether the state reaches the prompt. SEAM on top of the recap writes 72.3% and 71.2%, above the recap by $+14.9$ and $+11.2$ but below Scripty L0 by 21.3 and 12.5 pp. A third Construction sample writes 89.4% and 90.0% ($+31.9$, $p = 0.023$, and $+30.0$, $p < 0.0001$, over the recap), within 6 pp of samples A and B. Re-scoring all existing arms with a Gemini judge leaves the main contrast at $+35.6$ and $+22.8$ pp (from $+36.2$ and $+23.8$), with 96.5% agreement between the two judges. On ViMax, where no anchor table exists and the old and new states are derived mechanically from the frozen banks' paired questions, Scripty writes the carried state for 10 of 14 cross-unit facts against 4 for native ViMax, but also writes the expired state in 7 of 9 expiry controls against 0, so the failure to forget already appears at the prompt level, as on SEAM-Bench; these counts are too few for a test. Thirty-three Gemini outputs that failed to parse three times are excluded from all Gemini rates; DeepSeek had no such failures.

D.4 SEAM ON TOP OF THE NATIVE RECAP

This arm applies the same SEAM repairer with the same settings to the native-recap writer's prompts. Entity-state extraction and the memory graph depend only on the shared storyboard and are reused from the SEAM arm; filtering and rewriting are drawn anew; rendering, judge and scoring rule are those of the main table. Cross-unit pass rates are 48.9% (23/47) on FORMAL and 26.2% (21/80) on EXPAND, against 19.1% and 18.8% for SEAM on the no-carry prompts. Scripty L1 leads by +14.9 [-6.5, +34.0] pp (stories 9/5/4, $p = 0.27$) and +13.8 [-2.4, +31.4] pp (13/6/5, $p = 0.096$), so neither difference is significant; Scripty L0 and this arm are close (L0 minus it +2.1 and +6.2 pp, both not significant). Against the recap alone the arm gains +17.0 pp on FORMAL ($p = 0.065$) and +3.8 pp on EXPAND ($p = 0.29$). Within-unit rates do not differ clearly between these arms. The repairer injects into fewer units than on the no-carry prompts (40 of 72 and 64 of 168), since the recap drafts already carry some appearance and held objects. One source of noise: on FORMAL, 47 facts sit in units the repairer left unchanged, whose clips are byte-identical to the native recap's, yet re-judging flipped 3 cross-unit facts, all in this arm's favour. Scoring those facts with the recap's judge rows instead, reported only here, gives 42.6% and a Scripty L1 lead of +21.3 [-2.0, +42.6] pp ($p = 0.15$). The SEAM arm on the no-carry prompts was scored the same way as the main table, so the main comparisons are consistent. These stratum readings are retained alongside the combined main-table comparison.

D.5 ESTABLISHING AND LATER STATES TOGETHER

The primary score checks only the later clip, so a state that the establishing clip never showed but the later clip does would still pass. We therefore added one mechanically generated question per cross-unit fact on the establishing clip, asked of the window from the start of the change shot to the end of the establishing unit (only the last shot when the change is in the last shot); the question

bank and the window rule were fixed and hashed before judging. The judge, prompt and first-pass rule are those of the main table, and no clip was re-rendered. With both clips required, Scripty L1 still leads the native recap by +29.8 [+13.0, +45.5] pp ($p = 0.021$) on FORMAL and +11.2 [+3.6, +18.9] pp ($p = 0.022$) on EXPAND, no-carry by +42.6 [+30.4, +53.5] pp ($p < 0.001$) and +13.8 [+4.9, +22.5] pp ($p = 0.013$), and SEAM by +36.2 [+22.0, +50.0] pp ($p = 0.001$) and +6.2 [-2.4, +15.8] pp ($p = 0.388$) (Table 15). The arms differ little in establishing the state, which the establishing unit's storyboard states for every arm (+12.8 and -7.5 pp against the recap, both not significant); the baselines lose mostly by showing the state when it is established and losing it later. Conversely, 8 of L1's 30 later-clip passes on FORMAL and 14 of 32 on EXPAND are not shown in the establishing window, so the primary score overstates how often L1 carries a state that was drawn correctly in the first place. One reading goes against L1: on EXPAND its establishing rate is below L0's (-7.5 [-14.1, -2.2] pp, stories 0/19/5, $p = 0.062$), so some re-rendered establishing units show the state less well than the original L0 clip. As a diagnostic only, because each arm conditions on a different set of facts, the later pass rate among facts established correctly is 64.7% against 28.6% for the recap on FORMAL and 48.6% against 20.9% on EXPAND. The lead does

Table 15: Establishing and later states together (cross-unit facts, fixed denominators: 47 on FORMAL, 80 on EXPAND). Est.: the establishing clip shows the new state in the window after the change; Later: the primary StateHold pass; Joint: both. Counts: both / establishing only / later only / neither.

Arm	Est. (%)	Later (%)	Joint (%)	Counts
FORMAL				
No-carry	72.3	4.3	4.3	2/32/0/13
Native recap	59.6	31.9	17.0	8/20/7/12
SEAM (in-host)	70.2	19.1	10.6	5/28/4/10
Scripty L0	68.1	51.1	42.6	20/12/4/11
Scripty L1	72.3	63.8	46.8	22/12/8/5
EXPAND				
No-carry	50.0	11.2	8.8	7/33/2/38
Native recap	53.8	22.5	11.2	9/34/9/28
SEAM (in-host)	52.5	18.8	16.2	13/29/2/36
Scripty L0	53.8	32.5	17.5	14/29/12/25
Scripty L1	46.2	40.0	22.5	18/19/14/29

not hinge on the window rule or on a single run (Table 16). Of the 726 arm–question windows, 416 start at a detected cut, 120 are the whole unit (change in the first shot) and 190 use equal-length shot boundaries because the detector found more or fewer strong cuts than planned. Judging from the end of the change shot instead, judging the whole establishing clip, keeping only facts that need no estimated boundary, or using the second whole-chain run (Appendix D.1) leaves every contrast positive. Against the native recap on FORMAL the difference stays between +23 and +32 pp but is not significant from the end of the change shot or on the whole clip ($p = 0.065$) or on the measured-cut subset ($p = 0.180$, 25 facts); against SEAM on EXPAND it ranges from +6.2 to +15.2 pp and is significant only in the second run and on the measured-cut subset.

Table 16: Robustness of the joint score (Scripty L1 minus each arm, pp; story-level sign-test p in parentheses; NC: no-carry). Main: window from the start of the change shot; Second run: same window on the second whole-chain run; After change: window from the end of the change shot; Whole clip: the whole establishing clip; †main window, only facts whose establishing window needed no estimated shot boundary in both arms (25–49 facts per contrast).

	FORMAL			EXPAND		
Readout	Recap	SEAM	NC	Recap	SEAM	NC
Main	+29.8 (.021)	+36.2 (.001)	+42.6 (<.001)	+11.2 (.022)	+6.2 (.388)	+13.8 (.013)
Second run	+31.9 (.006)	+38.3 (<.001)	+44.7 (<.001)	+16.2 (.039)	+11.2 (.039)	+18.8 (.013)
After change	+23.4 (.065)	+31.9 (.003)	+36.2 (<.001)	+8.8 (.022)	+7.5 (.227)	+13.8 (.013)
Measured†	+24.0 (.180)	+40.0 (.004)	+33.3 (.008)	+14.3 (.039)	+15.2 (.031)	+23.3 (.004)
Whole clip	+27.7 (.065)	+36.2 (.002)	+42.6 (.001)	+15.0 (.049)	+11.2 (.146)	+17.5 (.035)

D.6 STATE-RECORD AUDIT AND LOSS ATTRIBUTION

The record covers 89.6% of an independent inventory of storyboard changes on 6 stories (Appendix D.6). Table 17 compares the state record with an inventory compiled without seeing it. Earlier loss-stage assignments are now unresolved (Appendix D.19). Figure 7 illustrates two recorded chains. The record contains 89.6% of the 144 changes in the independent inventory, and at least 81.8% in every category. Strict precision is low (44.9%) mainly because the two sides count at different granularity: the record splits one hand-over into a change for the object and for each person involved, and it also logs small moves within a scene that the inventory does not list. Of the 332 recorded changes, 149 match the inventory, 140 more are backed by the storyboard text although the inventory does not list them, 37 are not real state changes (a restatement of the same state or a momentary action), and 6 have no support, mostly background extras moved back and forth as one entity; counting the first two groups as supported gives 87.0%. Matching is by unit, so these rates describe unit-level coverage, not whether a change sits at the right shot. Counting (state, target unit) pairs, the injection delivered 24 of the 26 pairs that should apply (18 distinct states); 3 of 57 pairs that no longer applied (54 distinct states) were still delivered: an outfit change represented as two versions of one character, both listed on screen, and a lamp still lit and clothes still soaked after a time jump. Counted per holding interval (object, holder and interval, reported at the first affected shot), seven held-object cases (six intervals spanning 38 shot states, plus one unit-level exclusion of an object already counted) had a stale position (1) or were marked as not visible although their holder was (6); a further 5 were still recorded in hand after being put down. The inventory is not ground truth: two extractions with the same prompt agree at 89.4% (Dice), and a manual check of 98 matching decisions agreed with the automatic one 92.9% of the time, the automatic judge being slightly strict. The numbers are therefore agreement with an independent extraction on a small sample, not absolute accuracy. A record change matches an inventory change when both reach the same state within the same unit; the audit does not check the shot at which the change happens. Zero-model tests of the frozen engine (fixtures and recorded model outputs, no model calls; 555 tests pass and 2 are skipped, plus 55/55 for the L1 step) cover recording of state changes, shot placement that is invariant to how shots are split into units, exit and re-entry, visibility of carried items following their carrier, the release of character attachments after an in-unit time jump, and filtering of the new-image option when no drawable edit remains; they do not cover transfer of an item between holders or fact-level permanent-versus-temporary expiry within an episode. These tests show that the engine handles given model outputs correctly, not that the model records every event.

D.7 DISCORDANT FACTS AGAINST THE NATIVE RECAP

On the registered first pass, cross-unit outcomes (both pass / recap only / L1 only / both fail) are 11/4/19/13 on FORMAL and 11/7/21/41 on EXPAND; within-unit outcomes are 29/9/11/3 and 31/4/4/2. These score counts are unchanged. Earlier AI-assisted, single-reviewer dossiers followed records, prompts, references and sampled frames, but also used stored judge reasons. A retry-provenance audit found that a final answer could retain its first-round reason (Appendix D.19). Reason-only claims of judge error and the diagnosis-filtered effect estimates are therefore not claimed. The layer-count partition is preserved as a count partition only, not a validated causal breakdown. Each such claim remains unresolved unless the final contributing responses, scored media/time window and predicate-level visual evidence are jointly established; a reason/answer mismatch alone is insufficient. Retiring these attributions neither changes the original score nor assigns the failures to model variability.

D.8 VIMAX

Table 18 gives the counts across three ViMax dramas; the adapter and a separate exploratory L1 pass on one drama are described in Appendix G.

D.9 FROM L0 TO L1: WHICH FACTS FLIP

The final L1 changes 17 FORMAL and 24 EXPAND fact outcomes relative to L0, all in re-rendered units with a changed reference. FORMAL has 10 cross-unit gains, 4 cross-unit losses and 3 within-unit losses; EXPAND has 14 cross-unit gains, 8 cross-unit losses, one within-unit gain (`cuihuo f006`) and one within-unit loss. These are score transitions, not verified causal benefits or harms.

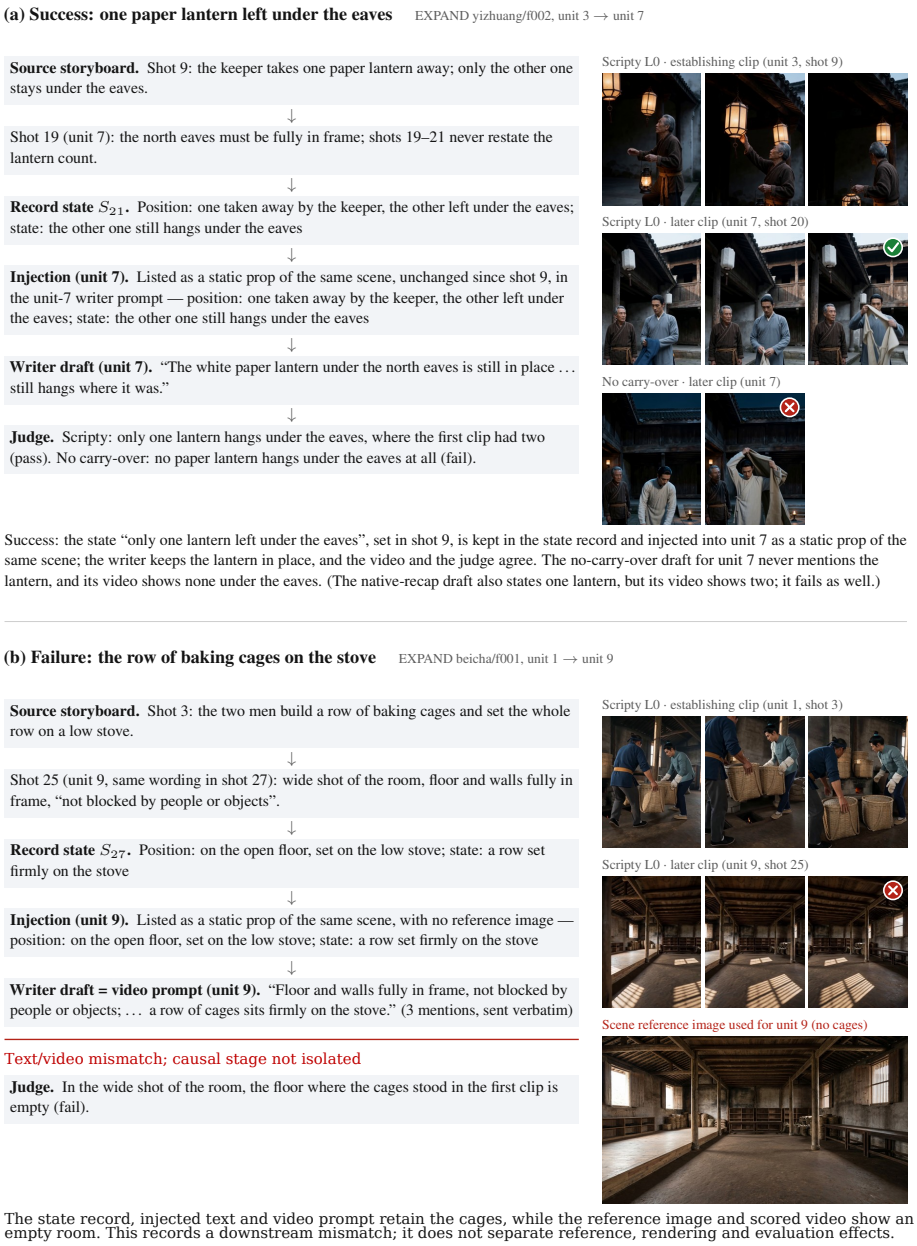

Figure 7: One successful and one failed fact followed from the storyboard through the state record, the injected text and the writer's prompt to frames of the rendered clips (EXPAND, Scripty L0; the success is contrasted with no-carry). In the failure, every text step keeps the cages, while the scene reference and video both show an empty room. This correspondence does not isolate the reference image as the cause.



Figure 8: All cases selected by a rule fixed in advance on EXPAND sample A: the establishing and later clips of each fact under the native recap, Scripty L0 and Scripty L1, including a case where L1 draws the variant wrong and one where it does not help. No scene-type fact went from fail to pass under L1 on this sample, so that slot is empty. Cases were picked by a rule written before selection; the full candidate set is in the supplementary material. Gain and loss labels refer to frozen scores, not independently verified visual truth. Selected frames do not establish whole-unit persistence; the yanshang curtain state at the unit's end remains visually uncertain.

Table 17: State-record audit. A source inventory of state changes was compiled from the storyboards without looking at Scripty's record; precision is the share of recorded changes that the inventory supports (the extra row also accepts support from the storyboard text), and recall the share of inventory changes that the record contains. The lower block checks, at the units where a state should or should not apply, what the injection actually delivered.

Category	Inventory changes	Inventory match (%)	Recall (%)
Appearance	11	81.2	81.8
Held and carried objects	40	40.4	92.5
Presence and position	58	52.3	91.4
Scene objects	35	35.1	85.7
All categories	144	44.9	89.6
supported by inventory or storyboard text		87.0	
Carry at the target unit			
Should carry: delivered	26		24 (92.3%)
Should not carry: wrongly delivered	57		3 (5.3%)
Held objects: stale position or not following holder			7 (1 stale, 6 not with holder)

Stories, inventory procedure and matching rule: six single-episode stories (EXPAND sample A: zhiyuan, dianmao, kejing; FORMAL: fengbo, miaoma, yedu), stratified and seeded, excluding stories used in debugging; the inventory was extracted from the host storyboard alone by gemini-3.1-pro-preview (a different model family from the GPT-family Construction model), and a change counts as matched when the same entity reaches the same state within the same unit, decided by lexical rules first and the same model as an independent judge otherwise.

Table 18: Scripty L0 in two external hosts. Native and augmented arms share host storyboards and rendering assets, with the same seed per shot and unchanged host source. Construction-only inputs are listed in Table 29. The dramas are shared, but coverage and banks differ: ViMax, three dramas (beyond-the-wall, his-toyboy, werewolf; EP01–05 each), 174, 103 and 137 shots per group; VideoMemory, three dramas (beyond-the-wall, his-toyboy, werewolf; EP01–03 each), 24 shots per drama and group. ViMax's per-question cross-unit sign test gives $p = 0.031$ (6 wins, 0 losses), while its story-level bootstrap interval touches zero; excluding facts where either arm fails a gate gives 5/11 vs. 1/11. VideoMemory gives $p = 0.125$ (4 wins, 0 losses). These are exploratory within-host comparisons.

	ViMax		VideoMemory	
Facts	Native	+ Scripty L0	Native	+ Scripty L0
Cross-unit	1/13	7/13	1/9	5/9
Within-unit	4/9	8/9	1/7	3/7
Expiry controls (state should no longer show)	6/9	3/9	4/5	4/5
Stated in the scene itself	10/12	9/12	12/12	12/12

Coverage, ViMax: 46 facts frozen, each drama's bank before its rendering; 3 (his-toyboy w01, beyond-the-wall f05 and w01) were dropped in both groups before judging because their unit videos stayed incomplete after the pre-set repair round, leaving 43 judged per group; shots rendered 172/174, 102/103 and 136/137 (native) and 166/174, 98/103 and 136/137 (Scripty). VideoMemory: 33 facts frozen before rendering, all scored; every gate passed in both groups.

The stored reason for `cuihuo f006` is not bound to its final retry answer; it cannot prove either a true visual improvement or a same-response judge contradiction. The earlier variant-matching notes remain exploratory. Same-request re-rendering also flips 16–26% of affected facts (Appendix D.10), making a single flip insufficient for causal attribution; it does not show that a known fraction of these flips is noise.

D.10 RE-RENDERING CONTROL

To measure the conditional reference-variant contrast across repeated video draws, we re-rendered every unit whose reference images L1 changes (FORMAL 44, EXPAND 73 units) two more times on each side from the identical request with an explicit seed that is the same on both sides for a given unit and draw (the platform echoed the sent seed for all 234 pairs), while the original main-run pair was rendered without a seed parameter and so carries independent platform-drawn seeds (L0 references against L1 variant images; text prompt and all other parameters byte-identical), giving three samples per side, and re-judged every fact and gate that touches these units with the same frozen judge and first-pass rule; the original pair reproduces the main-table L1 and L0 readings. On the 111 cross-unit facts that touch re-rendered units, persistence is higher with the variants in every sample ($+10.8$, $+15.3$ and $+12.6$ pp); averaged over samples the variant-minus-base difference is $+12.9$ pp [$+5.0$, $+20.8$] (26 of 42 stories higher, 7 lower; sign test $p = 0.001$), $+13.2$ [$+1.6$, $+24.8$] on FORMAL and $+12.7$ [$+1.9$, $+23.0$] on EXPAND, and $+13.9$ [$+6.4$, $+21.5$] without trimmed-window questions. Re-rendering the same request moves a side's pass rate by at most 5.4 pp but flips 16–26% of these facts, so a single-fact L0/L1 flip is insufficient for causal attribution; this control does not identify the fraction of flips caused by re-rendering noise. Within-unit consistency on the 33 affected facts is lower with the variants but not significantly (-8.1 [-19.0, $+1.1$]; FORMAL alone -15.8 [-30.3, -3.3], sign test $p = 0.125$, 19 facts), gate pass rates are unchanged ($+0.5$ pp), and the six video-quality metrics and CLIP-T differ between sides by at most 0.004; on EXPAND imaging quality ($+0.0039$) and CLIP-T (-0.0025) have intervals that exclude zero, but the magnitudes are negligible on a 0–1 scale.

D.11 EXPLORATORY FOLLOW-UP ANALYSES

These analyses reuse existing judge rows only (no new rendering or judging), with the main-table scorer, story-level bootstrap (2000 draws, seed 20260915) and exact per-story sign test, and report the two pools separately. Every split was chosen after the main results were known.

Where within-unit losses under L1 come from. A variant image counts as verified when the image judge of Appendix D.12 says the target state is drawn and finds no lettering or modern object. A re-rendered unit counts as verified when all of its variants are. On FORMAL, 26 of 44 re-rendered units are verified (38 of 55 images); on EXPAND, 45 of 73 (59 of 87). Almost all failures are states drawn only in part. Lettering or modern objects affect 1 and 8 units and one within-unit fact in each pool. In the main run, 3 FORMAL within-unit facts go from pass under L0 to fail under L1 (1 in a verified unit, 2 in unverified units). Same-side re-rendering flips such facts from pass to fail at 12.5%, so 2.0 such losses are expected from re-rendering alone. EXPAND has 1 such loss against 2.8 expected. In the re-rendering control (three draws per side), the FORMAL within-unit difference is -5.6 [-16.7, +5.6] pp in verified units (12 facts) and -33.3 [-51.9, -11.1] in unverified units (7 facts). On EXPAND the two values are 0.0 (8 facts) and +5.6 (6 facts). In a counterfactual, facts that touch an unverified unit take their L0 score. The FORMAL within-unit difference against the native recap then becomes +7.7 [-8.0, +22.6] (was +3.8 [-12.0, +19.2]), and against no-carry +5.8 [-11.3, +21.2]. The cross-unit margin over the recap becomes +23.4 [+4.3, +43.5] (sign test p=0.146; was +31.9, p=0.013). On EXPAND the within-unit reading does not change (0.0 [-14.9, +13.6]; L1 and L0 have the same within-unit pass rate there, with one fact lost and one gained), and the cross-unit margin is +18.8 [+6.0, +32.0] (p=0.013).

Facts that need a variant. A cross-unit fact counts as needing a variant when L1's selector marked the tested entity's registered reference as contradicting the required state in the fact's later unit. The tested entity was matched to selector cells by reading the state text, before any outcome was computed. This covers 25 of 47 FORMAL facts and 43 of 80 EXPAND facts; 3 ambiguous facts per pool are excluded and reported separately. On the facts that need a variant, L1 minus the native recap is +36.0 [+14.3, +56.5] (p=0.021) and +25.6 [+6.0, +46.2] (p=0.049), and L1 minus L0 is +28.0 [+8.3, +47.8] (p=0.039) and +23.3 [+7.9, +40.0] (p=0.022). On the remaining facts, L1 minus L0 is -10.5 [-21.1, 0.0] and -8.8 [-19.4, 0.0], and L1 minus the recap is +31.6 [+5.0, +58.8] (p=0.125) and +8.8 [-8.8, +25.0]. Counting the ambiguous facts as needing a variant leaves these readings qualitatively unchanged.

One judging policy for every arm. L1 minus the native recap on cross-unit facts, FORMAL / EXPAND: First pass: +31.9 [+14.0, +48.9] (p=0.013) / +17.5 [+2.6, +31.6] (p=0.019); Second pass for every arm: +23.4 [+6.4, +41.2] (p=0.065) / +12.5 [−1.2, +26.4] (p=0.064); Per-fact mean of the two passes: +27.7 [+12.8, +40.9] (p=0.021) / +15.0 [+1.8, +28.6] (p=0.012); Pass only if both passes pass: +25.5 [+8.3, +41.7] (p=0.035) / +15.0 [+2.5, +26.9] (p=0.019); L1 first pass against the baselines' second pass: +27.7 [+9.5, +43.5] (p=0.035) / +13.8 [−1.2, +29.1] (p=0.041). The paired first-minus-second-pass difference of the margin is +8.5 [−12.5, +28.0] and +5.0 [−5.8, +15.5]. Uniform three-pass coverage is now complete for both pools; Appendix D.19 reports every policy without replacing the registered first-pass table. On the 40-fact EXPAND drift sample, the three-pass majority gives L0 minus the recap = +21.1 [−15.8, +53.3] (19 facts), identical to the first pass on the same facts.

Minimum detectable effect. FORMAL has 18 stories and 47 cross-unit facts (story-level bootstrap SE 8.7 pp); EXPAND has 24 stories and 80 facts (SE 7.6 pp). At 80% power and two-sided α = 0.05, the analytic minimum detectable effect (2.8 SE) is 24.4 and 21.2 pp, and it rises to 25.9 pp at 16 stories. We also ran a paired binary simulation with story clusters. It is calibrated to each pool's recap pass rate, the observed facts per story and a single-fact re-rendering flip rate of 16% or 26%, and its between-story effect heterogeneity is fitted to the observed SE. Under this model the minimum detectable effect is 32 / 26 pp (sign test / interval) on FORMAL and 28 / 22 pp on EXPAND. Detecting +15 pp with 80% power needs about 48–56 stories per pool for the interval and about 80 for the sign test. With 16 stories, power for +15 pp is 0.15–0.19 (sign test) and 0.37–0.40 (interval). With no between-story heterogeneity, the most favourable case, 16 stories give 0.25–0.47 and 0.47–0.70. At 16 stories, raising the facts per story to 8 lifts interval power to 0.50–0.62 but leaves sign-test power at 0.14–0.29.

D.12 GENERATED REFERENCE VARIANTS

On the variants used in the final runs (FORMAL: 55 images, 18 stories; EXPAND: 87 images, 24 stories) we apply an image judge and fix identity thresholds in advance from base images only (the 99th percentile of cross-story similarity between base images of the same type: character 0.92 on the head-shot panel, scene 0.64, prop 0.67). The target state is drawn and identity kept in 60.0% [45.1, 74.2] and 58.6% [50.6, 68.8] of the images; the state alone is drawn in 69.1% and 74.7% (at least partly in 98.2% and 100%), identity is kept in 89.1% and 81.6%, and 3.6% and 8.0% contain lettering or a modern object (intervals: story-level bootstrap). The repeat runs, the no-scene-object runs and the DeepSeek chain give joint rates of 55.7–64.0%. Re-judging a random 20% of the images reproduces the three-level verdict for 11 of 11 and 16 of 17. The identity measure is a threshold-based proxy whose below-threshold side was fully inspected: all 12 character images below threshold (FORMAL 5, EXPAND 7) showed no identity swap; the drops come from layout drift of the four-view sheet (4, the head-shot panel replaced by a full- or half-body view), target states that alter the head itself (3: a shaved queue, a shaved beard, a cheek scab) and target costume changes visible in the head-shot crop (5). For props and scenes the whole image is compared, and the state change is often the whole shape (bamboo poles tied into a frame, gold leaf scattered on the floor), which is why prop identity on EXPAND is only 55% (11 images).

Timing of the selected specifications. In seven historically audited L1 runs (A/B in each pool, one no-scene-image-object run per pool and EXPAND DeepSeek), selected variant specifications did not introduce cited facts before they held in Scripty's own start-of-unit record. This is a check of the selection plan, not independent storyboard annotation or generated-image semantics. Later sample-C runs have separate judge-key checks and were not part of this seven-run audit. Retained references can still carry expired state because the fallback does not check them for conflicts (Appendix A).

D.13 GENERAL VIDEO QUALITY AND SIDE EFFECTS

We aggregate the six VBench-style metrics (Huang et al., 2024) and CLIP-T over all 42 stories (unit means within stories, then equal story weights; 10,000 paired story-bootstrap draws). These are quality controls, not persistence scores. L1 minus native recap is: subject consistency +0.0030 [-0.0026, +0.0087]; background consistency +0.0049 [+0.0004, +0.0095]; motion smoothness +0.0006 [+0.0001, +0.0011]; temporal flickering +0.0007 [-0.0003, +0.0017]; imaging quality +0.0029 [-0.0001, +0.0061]; aesthetic quality +0.0021 [-0.0007, +0.0050]. CLIP-T against each unit's summary is -0.0125 [-0.0166, -0.0081]; this decrease is retained. Face-identity comparisons use only paired available observations: cref base A -0.0070 [-0.0202, +0.0054] (42 stories); cref base B -0.0067 [-0.0237, +0.0105] (30 stories); cself cross A -0.0136 [-0.0394, +0.0126] (42 stories); cself cross B -0.0677 [-0.1134, -0.0137] (19 stories). A/B denote appearances expected to stay unchanged/change, respectively. Source-stratum analyses and per-story values remain in the supplementary records. The aggregation does not establish causal attribution or a general perceptual improvement.

D.14 COMIC RENDERING

As a check outside video, the same storyboards were rendered as comic pages under all four arms (no-carry, native recap, SEAM, Scripty), with the same page-packing prompt, image model, unit split, and reference-image rules. The two writers ran separately, so their texts differ in more than the carried state (36–62% similarity). Units were picked by a rule written before any page was made, from the bank structure alone.

A scored 18-story four-arm run (FORMAL, 2026-09-17) was completed on the same frozen question bank, the same four arms, and the same v1 whole-page judge as the 18-story FORMAL video reading, not the 42-story main-table reading scored with windowed.v2; only the rendering layer differs. Table 19 compares pass rates side by side.

Cross-unit arm ordering is consistent across renderers: Scripty leads in both, with Scripty − no-carry deltas of +53.2 pp (comic, 95 % CI [+39.2, +66.0], signs 15/3/0) vs. +46.8 pp (video, CI [+31.1, +63.3]), and comparable margins over recap and SEAM (Table 19 footnote). However, per-fact three-state agreement between the two renderers is only 67.7% (268/396): roughly a third

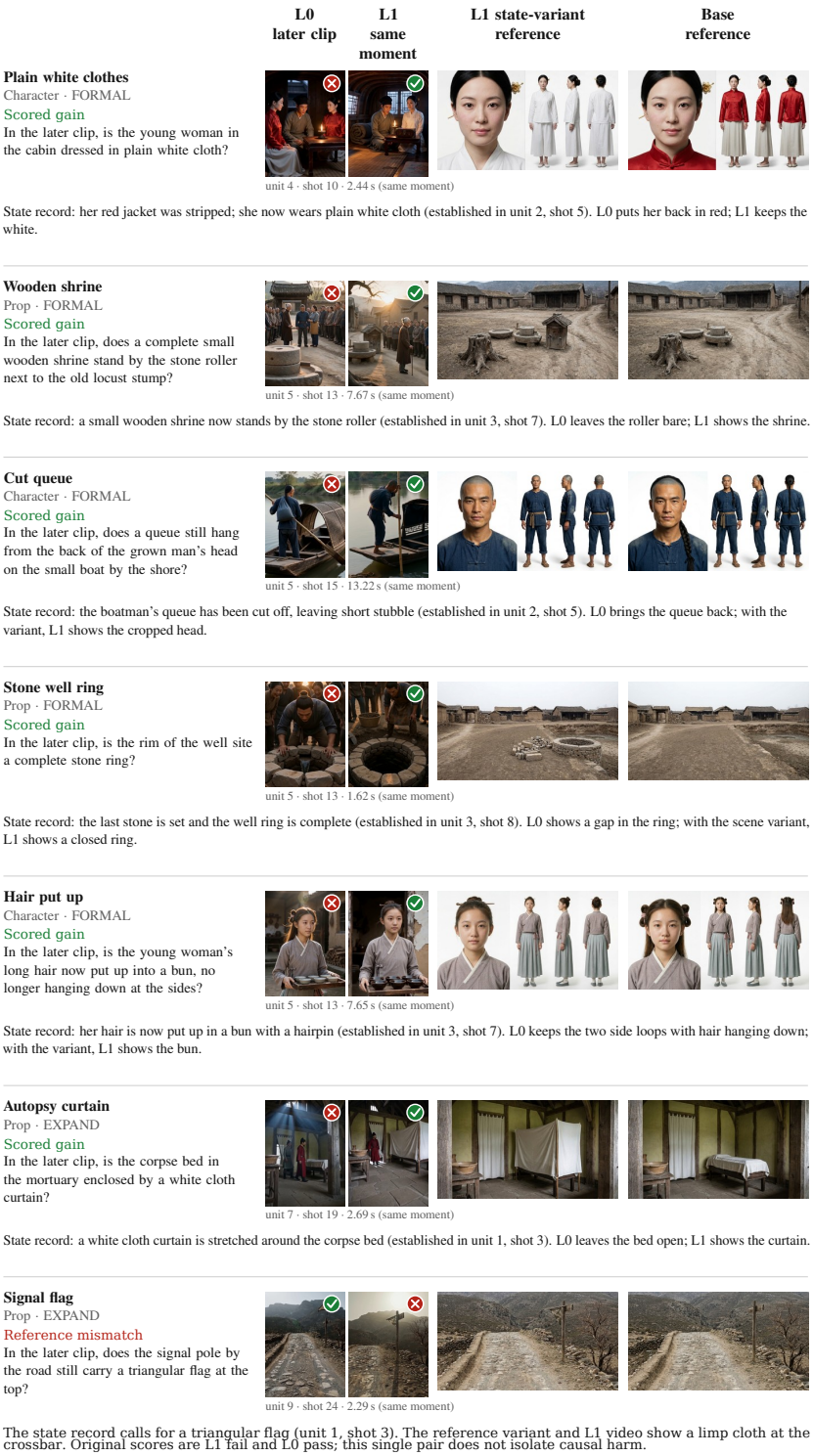

Figure 9: L1 at work: for each fact, the L0 later clip, the L1 clip at the same moment, the state-variant reference that L1 generated and the base reference it was generated from. The last row is a failure in which the variant draws the tested object wrongly; the original L0 score is retained, since an unbound stored reason cannot overturn it (Figure 8). Cases were picked by a rule written before selection; the full candidate set is in the supplementary material. Labels refer to frozen scores; the yanshang curtain example is not a verified whole-unit success.

Table 19: Four-arm comic vs. video comparison (18 stories, FORMAL0913). Pass rates are fact-weighted; cross-unit pool n=47, within-unit pool n=52. Both lines use the same frozen question bank, four arms, and judge; only the rendering layer differs. The comic side uses the v1 whole-page judge (`c25e2386`), not the windowed.v2 ruler of the main table.

Pool	Arm	Comic (%)	Video (%)
Cross-unit	No-carry	6.4	4.3
	Native recap	40.4	23.4
	SEAM	23.4	17.0
	Scripty	59.6	51.1
Within-unit	No-carry	82.7	84.6
	Native recap	76.9	65.4
	SEAM	80.8	80.8
	Scripty	82.7	71.2
Δ (pp)	Scripty − no-carry (CUC)	+53.2	+46.8
	Scripty − recap (CUC)	+19.1	+27.7
	Scripty − SEAM (CUC)	+36.2	+34.0
	Scripty − no-carry (WUC)	+0.0	−13.5
	Scripty − recap (WUC)	+5.8	+5.8

95 % CIs (comic / video). CUC: Δ no-carry [+39.2, +66.0] / [+31.1, +63.3]; Δ recap [+0.0, +38.0] / [+6.5, +47.1]; Δ SEAM [+19.2, +53.5] / [+17.8, +51.1]. WUC: Δ no-carry [−11.1, +11.1] / [−22.2, −3.8]; Δ recap [−9.8, +20.4] / [−6.0, +17.6]. Comic story-level signs (+/−/tie, 18 stories): CUC no-carry 15/3/0, recap 9/6/3, SEAM 12/5/1; WUC no-carry 4/10/4, recap 5/10/3. Per-fact three-state agreement between renderers: 67.7% overall (268/396); per arm: no-carry 71/99, recap 59/99, SEAM 69/99, Scripty 69/99.

of individual facts flip. The correct claim is therefore that arm ordering and cross-unit effect sizes are similar, not that the two renderers agree per fact.

A guardrail asymmetry appears in the within-unit pool: Scripty − no-carry is +0.0 pp under comics (flat; CI [−11.1, +11.1]) but −13.5 pp under video (CI [−22.2, −3.8], a significant drop whose CI excludes zero). We have no explanation for this divergence and flag it as an open question.

The comic side was scored with the v1 whole-page reading judge (digest `c25e2386`), not the windowed.v2 ruler used by the main table. These numbers are therefore reported as this appendix comparison only and are not merged into the main results.

D.15 Long-horizon demonstration

As a demonstration rather than a test, we ran one 20-episode SEAM-Bench drama through the native recap, no-carry and Scripty (Figure 11). The chip implanted in Lina's arm in episode 4 is written into 23 shots of episodes 6–20 by the Scripty writer and into none by the other two, and it is visible in an episode-14 clip ten episodes later. The same carrying keeps mud from an episode-1 rain night in 60 shots of episodes 5–20, visible in episode 20. A flashback to Lina at age three carries neither state in any group. The shot-selection rules, including one keyword-picked probe row that the figure labels as such, were fixed before any complete clip had been seen. A display-only rule fills the two probe rows where the native recap and no-carry have no matching shot with that arm's earliest unused Lina shot of the same episode; it changes no reading. All rules are released. The figure presents this long-horizon demonstration; the accompanying video illustration is a separately curated StateHold example.

Scope of the accompanying video. The supplementary MP4 shows three FORMAL cases: a teahouse wall repainted dark indigo, wet mud on a clay horse, and a crossed seal on a mill door. Each case opens with a context card (the anchor event and the state to keep), then plays the complete later unit in three synchronized columns: no-carry, Scripty L0 and Scripty L1. All footage is the original frozen experiment video, unmodified and not re-rendered. Selection was post-hoc and outcome-aware: candidate units were those where both Scripty arms already pass the frozen check, and each displayed clip passed two independent Gemini-3.1-Pro full-video reviews with consistent target-

Scripty v2.1, FORMAL pool · page 1 of 9

Native recap **Scripty (full, L1)**

dangao
Probe: Hair bun
Unit 3 → Unit 5 (this page)
Bank question (video wording, translated):
Comparing the two clips, in the second clip: does the younger adult woman in the courtyard have no hair bun at the back of her head (hair hanging loose, not coiled into a round bun)?



duanzhan
Probe: Collapsed plank road
Unit 2 → Unit 5 (this page)
Bank question (video wording, translated):
Comparing the two clips, in the second clip: is there a clear gap in the middle of the wooden plank road along the cliff?



Figure 10: Comic rendering of the same unit under the native recap (left) and Scripty L1 (right), FORMAL, two stories on this page, each with the probe, the units and the bank question translated from the video wording. This page illustrates two of the 18 stories rendered for this qualitative check.

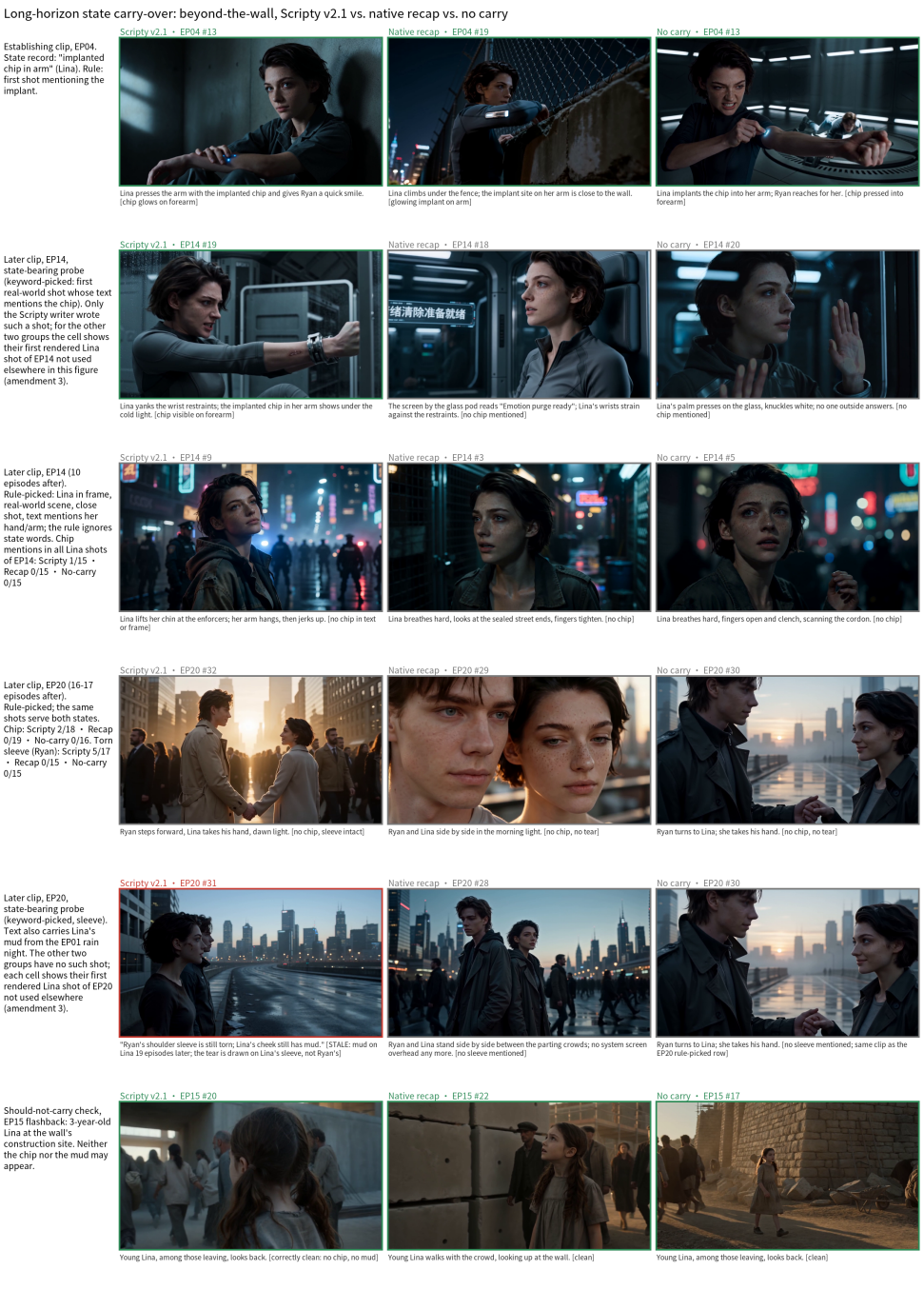

Figure 11: Long-horizon state carry-over on one 20-episode drama (native recap, no-carry, Scripty). Green: state carried and visible, or correctly absent; red: an expired state still visible; grey: not carried or not visible. Rows marked keyword-picked were chosen by a keyword rule added before any complete clip was seen; in those rows the native-recap and no-carry panels show the arm's earliest unused Lina shot of the same episode, which does not mention the state (a display-only rule). The on-screen text in the native-recap EP14 panel was generated by the video model. The same mechanism that remembers across many episodes also fails to forget.

state verdicts plus maintainer frame inspection; the end card states this selection rule. The video is not a representative evaluation of the overall gain, and minor generation artifacts noted by the reviews are disclosed on its end card. Scores, the paper's failure examples and original study-media correspondence are unchanged. Each judged clip appears in the shipped media index with its SHA-256, so a reader can trace any number in the paper to the clip it came from. The earlier single-case video, the longer exploratory video, the other comic pages and excluded human-study clips remain in the research archive, not in this submission's media selection; their omission is not a claim of quality or a replacement of the original observations.

Front-page example. Figure 1 uses *werewolf*, a 25-episode SEAM-Bench drama outside State-Hold, run through the native recap and Scripty as above. Luna's half wolf-totem necklace is established in episode 10. The three checkpoints are the shots, at least 147 storyboard shots later, where Scripty's writer mentions the necklace and frames Luna's chest: episodes 16, 19 and 20 (+147, +238 and +260 shots); Luna does not appear after episode 20. Each checkpoint was rendered four times per arm with paired seeds, under rules written and hashed before rendering. The necklace is visible in 11 of 11 valid Scripty renders and 0 of 12 native-recap renders; in 6 of the recap renders her chest is out of frame. One Scripty render was discarded under our child-appearance rule. Both arms' render prompts carry the same added age and clothing line. One frame per cell was picked by the clearest view of her chest, then the lowest seed, without regard to the necklace. The rendered pendant varies in shape and never shows the half totem of the record, and the episode-20 script itself mentions the necklace later in that episode. The left bar panel is the prompt-level text-layer reading (DeepSeek judge against the preregistered anchor table, Appendix D.3); the right panel is the cross-unit pass rate on the rendered video.

D.16 Viewing preference

Each page carries one mechanically generated cue from the shared storyboard, naming the scene and the characters on screen at the start of the window ("these clips start in the tavern courtyard; the characters at the start are ..."); no plot summary, props, costumes, states or events are given, and raters judge source fidelity against this cue and common sense, so that dimension measures consistency with the cue rather than fidelity to the full script. Windows were drawn by a fixed seed among those that fully contain at least one cross-unit fact of the frozen bank (both its establishing and its checked unit inside the three-unit window), taking first the stories not used in the earlier fact-level study; raters are not told which state to watch. The study therefore measures raters' overall sense of continuity over three consecutive units that span a cross-unit fact, not the persistence of a particular state, which the fact check below addresses. Eleven of the twelve stories, and three of the six fact-check stories, also appeared in that earlier study; no fact is repeated. The two parts are designed to take about 55 minutes per rater. Each story outcome is the majority of three ratings; ties and cannot-tell are excluded from the sign test. Continuity: Scripty L1 preferred on 1 story, the native recap on 4, 6 ties, 1 without majority (two-sided sign test $p = 0.375$). Cue consistency (asked as source fidelity): 1 / 5 / 5 / 1 ($p = 0.219$). Visual quality or overall: 2 / 4 / 1 / 5 ($p = 0.688$). At the rating level (36 ratings), continuity is 5 for L1, 14 for the recap and 17 equal. Leaving out any one rater keeps the direction (recap 1–3 stories, L1 0–1) and no reading becomes significant; no rater leans towards L1. These are 36 ratings on 12 stories, not 36 independent samples, and no significant difference does not show that the arms are equal. The viewing task rates three-unit windows without naming a target state. Its materials and question differ from the six-fact check; together they do not establish that a higher fact pass rate causes better perceived continuity. A null preference result is not equivalence and cannot be assigned to the host or renderer without further controls.

D.17 Blind check on final outputs

The protocol (sha 77784a0a…, released with the study package) was fixed before release. Each page is judged by three of the four raters, assigned by a fixed rotation, and the order of its three panels is randomized. L0 and L1 share the establishing clip for three facts and the checked clip for one, so raters can see that those panels partly coincide. The L1 clips are exactly those the judge scored (media manifest sha 0e2eed9c…; judge cache-key check before sending: 364/364 rows match). Each of the 18 method panels (6 facts, native recap / Scripty L0 / L1) is judged by three of the four raters; each question takes the majority answer, and a question without a majority

Table 20: Judge robustness: Scripty L1 minus each baseline, cross-unit facts (pp, story-level bootstrap 95% intervals).

Pool	L1 minus	Pass 1 (primary)	Pass 2	Flash judge
FORMAL	No-carry	+59.6 [+47.8, +71.4]	+55.3 [+42.9, +68.9]	+66.0 [+52.3, +78.8]
	Native recap	+31.9 [+14.0, +48.9]	+23.4 [+6.5, +40.4]	+27.7 [+10.9, +43.8]
	SEAM	+44.7 [+26.5, +62.2]	+42.6 [+26.5, +60.0]	+46.8 [+27.1, +65.2]
EXPAND	No-carry	+28.8 [+17.1, +41.2]	+22.5 [+11.5, +34.2]	+40.0 [+31.6, +49.4]
	Native recap	+17.5 [+2.6, +31.6]	+12.5 [-1.2, +26.3]	+28.8 [+17.2, +41.3]
	SEAM	+21.2 [+7.8, +35.4]	+13.8 [+2.5, +25.6]	+27.5 [+15.1, +41.1]

Table 21: Agreement with independent fact-level human judgments. Human-reference coverage and judge error rates are reported separately, with frozen sampling and adjudication procedures. Human references and the judge answers compared with them use the 17 Sep question wording, not the frozen wording of the main table, and cover only the no-carry and native-recap arms; the table validates the instrument, not a method effect.

Pool	Fact-arm items	Agreement (%)	FPR (%)	FNR (%)	Human coverage (%)
CUC	35	91.4 (32/35)	6.7 (2/30)	20.0 (1/5)	87.5
WUC	40	75.0 (30/40)	0.0 (0/1)	25.6 (10/39)	100.0

FPR conditions on human-reference failures; FNR conditions on human-reference passes. Unknown human references stay visible in coverage. Sampling weights, qid agreement, inter-rater agreement, and story-level method-direction agreement: human reference = adjudicated consensus of two non-author raters per fact, with a third blind non-author rater on disagreement; 40 facts (10 CUC and 10 WUC per pool) drawn with seed 20260922 fixed before any answer, inclusion probabilities retained and results reported in-sample only; no-carry and native-recap arms only. Items are fact-arm panels with both a human reference and a judge answer; not counted are 4 panels left unresolved after the third rating (C003 and C026 no-carry, both C023 panels) and 1 panel without a judge answer (C021 no-carry). A human or judge 'cannot judge' is scored as not passed, as in the frozen scorer. Overall 62/75 (82.7%), Cohen's kappa 0.66; by pool FORMAL CUC 19/19 and WUC 17/20, EXPAND CUC 13/16 and WUC 13/20; question level 257 agree, 31 opposite, 8 human-only and 12 judge-only cannot-judge over 308 cells. Method direction, native recap minus no-carry, human vs judge on the same matched facts (paired, not a difference of the arm ratios): FORMAL CUC +33.3 vs +33.3 pp over 9 facts; FORMAL WUC -10.0 vs 0.0 pp over 10 facts; EXPAND CUC +14.3 vs +14.3 pp over 7 facts; EXPAND WUC 0.0 vs -10.0 pp over 10 facts. Sensitivity without C021, C005 and C029 (flagged in source review; 37 facts): 58/70 (82.9%), kappa 0.66, false passes 1/27, false fails 11/43. Judge answers are case-level runs on the same Chinese question text and AV24 clips as the raters, not the main-run judgments.

leaves the panel undecided. H-later (both later-clip gates met and the later-state question answered correctly) is the primary outcome: recap 2/6, L0 3/6, L1 4/5 decided, 4/6 with the undecided panel counted as not passed; by pool, FORMAL 1/3, 0/3, 2/3 and EXPAND 1/3, 3/3, 2/2. H-anchor (the state is shown in the establishing clip): 6/6, 5/6, 5/6. H-joint (both): 2/6, 2/6, 3/5. Human L0-to-L1 flips on H-later: 2 improvements (dushiniang f001, suxue f002), 0 regressions, 3 unchanged and 1 undecided; the judge's flips agree on 4 of the 5 decided facts, and of the 17 decided panels human and judge differ on one (dingzhang f003, L0), where the judge's paired answers contradict each other and one rater noted uncertainty about whether the sleeve was the same item. As preregistered, no overall agreement rate is reported. Raters agreed unanimously on 36/36 gate questions and 15/18 establishing-clip questions. For three facts L0 and L1 share the establishing clip and for one the checked clip, so their difference there comes from the other clip alone. With 6 facts, three per pool, these results are descriptive: no per-pool sign test can go below $p = 0.25$. Raters' free-text notes (not scored) mention errors outside Scripty's state mechanism: dialogue mixing Chinese and English comes from the host writer and occurs in every arm (about 3–6% of units); a character replaced by another in a final shot is a renderer error triggered by a shared-storyboard line that names a companion without framing him; reports of objects clipping or appearing from nowhere were not checked.

D.18 JUDGE VALIDITY

Cross-unit panels are dominated by human failures (30 of 35), so a judge that always answered "fail" would already agree on 85.7% of them; the informative numbers are the per-class error rates, and this study does not cover Scripty outputs or the final wording. The human reference is the adjudicated answer of two non-author raters per fact, with a third blind rater on disagreement. The study covers the no-carry and native-recap arms under an earlier question wording, so it checks the judge rather than any method; its cross-unit panels contain few human passes, so the per-class error rates matter more than overall agreement.

Second judge. We re-judged the five core arms (excluding SEAM+recap) with the low-cost Gemini-3.6-Flash, using the same judge code and prompt (digest 22292fcc) on exactly the clips the Pro judge scored; Flash judged once and is compared with Pro's first pass. For Scripty L1, per-question agreement with Pro is 92.7% on FORMAL and 89.4% on EXPAND (Cohen's κ 0.86 and 0.79). Flash is more lenient (L1 cross-unit pass rates 70.2% and 56.2% against 63.8% and 40.0%), but every cross-unit contrast of Scripty L1 against the three baselines keeps its direction and a story-level bootstrap interval above zero: against no-carry $+66.0$ and $+40.0$ pp (Pro $+59.6$, $+28.8$), against the native recap $+27.7$ [$+10.9$, $+43.8$] and $+28.8$ [$+17.2$, $+41.3$] (Pro $+31.9$, $+17.5$), and against SEAM $+46.8$ and $+27.5$ (Pro $+44.7$, $+21.2$). The one reading that changes is L1 minus L0 on FORMAL, whose lower bound sits at 0.0 under Pro ($+12.8$) and lies above zero under Flash ($+17.0$ [$+2.0$, $+31.1$]); no within-unit difference of L1 against a baseline is significant under either judge. Both judges are Gemini models, so this shows that the conclusions do not depend on the judge's size, not that a judge from another model family would agree.

Second judging pass. The registered inference uses the first judging pass. Re-judging EXPAND with the same judge and clips, Scripty L1 passes 38.8% of cross-unit facts on the second pass against 40.0% on the first (91.2% of facts keep their outcome); its margins become $+22.5$ [$+11.5$, $+34.2$] pp over no-carry, $+12.5$ [-1.2, $+26.3$] over the native recap and $+13.8$ [$+2.5$, $+25.6$] over SEAM, so the margin over the recap stays positive but its interval now includes zero. The repeat L1 run (sample B) goes from 45.0% to 38.8% (86.2% agreement), $+12.5$ [-2.5, $+26.8$] over the recap. On the second pass the within-unit differences against the baselines lie between -7.3 and -4.9 pp, all with intervals that include zero. To separate judge drift from the clips, we re-judged 40 cross-unit facts per arm a third time in one interleaved window. Scripty L0 (sample A) stays at 32.5%, 30.0% and 30.0% over the three passes, while the baselines' second and third passes agree with each other and sit slightly above their first (no-carry 10.0/15.0/12.5%, native recap 22.5/25.0/25.0%, SEAM 17.5/25.0/25.0%; pooled first-to-third-pass flips 7 up and 2 down, $p = 0.18$). The differences between passes are not significant and the sample is small. On FORMAL the second pass reproduces the main result: Scripty L1 goes from 63.8% to 59.6% (78.7% of facts keep their outcome), with margins of $+55.3$ [$+42.9$, $+68.9$] pp over no-carry, $+23.4$ [$+6.5$, $+40.4$] over the native recap and $+42.6$ [$+26.5$, $+60.0$] over SEAM; the repeat run (sample B) goes from 63.8% to 57.4% (89.4% agreement), with $+21.3$ [0.0, $+42.3$] over the recap. The FORMAL baselines move from 4.3%, 31.9% and 19.1% to 4.3%, 36.2% and 17.0% (no-carry, native recap, SEAM), so the upward shift seen on EXPAND does not recur, and within-unit differences stay within ± 6 pp and are not significant on both passes.

D.19 MEASUREMENT VALIDITY AND ATTRIBUTION PROVENANCE

Stored reasons and final answers. The frozen retry implementation can update an answer while retaining its first-round reason. A stored reason contradicting a final answer therefore cannot alone establish a wrong final label. Recorded answer sequences reproduce the scores; this provenance defect does not itself establish arithmetic corruption. The earlier development-pool diagnostic index binds 131 facts to 786 final fact-question rows, including 56 answers changed by retries. It is not an error-rate estimate or a new adjudication of all gates and pixels. Unclosed judge-error and causal-stage assignments remain withdrawn; attribution requires the final response, scored media and predicate/time-window evidence together.

Uniform three-pass coverage. The same frozen questions, clips, crop rule, judge model and prompt were used for all passes; only the pass/cache namespace differs. All main-table arms have

complete three-pass coverage in both pools. L1 minus recap under the mean of three is $+27.0$ [$+10.9$, $+41.8$] pp on FORMAL ($p = 0.0213$) and $+14.6$ [$+1.8$, $+27.2$] on EXPAND ($p = 0.0118$). FORMAL remains positive but not significant by the story sign test under pass 3 alone ($p = 0.2101$) or majority of three ($p = 0.0574$). Table 22 reports pass 3, majority of three and mean of three, including the within-unit guardrail. These appendix checks do not replace the registered first-pass main table. A positive pooled interval is not a significant story sign test, and no policy is selected for its favourable result.

Table 22: Uniform three-pass judging of the same main-table videos: L1 minus native recap (pp). All three policies are reported; the registered first-pass main table is unchanged. A positive bootstrap interval does not imply a significant story sign test.

Pool	Policy	CUC [95% CI]	Sign p	WUC [95% CI]	Sign p
FORMAL	Pass 3	+25.5 [+4.3, +45.8]	0.2101	+9.6 [−5.9, +25.0]	0.7539
FORMAL	Majority of 3	+29.8 [+10.4, +47.9]	0.0574	+3.8 [−10.0, +17.3]	1.0000
FORMAL	Mean of 3	+27.0 [+10.9, +41.8]	0.0213	+6.4 [−5.4, +18.5]	0.6291
EXPAND	Pass 3	+13.8 [+1.2, +25.6]	0.0490	−4.9 [−17.9, +8.2]	0.7266
EXPAND	Majority of 3	+17.5 [+2.6, +31.6]	0.0192	−2.4 [−14.7, +9.8]	1.0000
EXPAND	Mean of 3	+14.6 [+1.8, +27.2]	0.0118	−4.1 [−16.3, +8.1]	0.7744

E TEXT-LEVEL CHECK ON SEAM-BENCH

SEAM-Bench supplies serial-drama storyboards for a comparison of written prompts only; no video is rendered. Scripty, an episode-local writer without carry and our SEAM reimplementation receive the same storyboards. Current-episode source, carried state and character descriptions are kept separate, so information from later episodes cannot enter through the state. Recall and staleness are counted over a fixed inventory of source-defined assertions, compiled without looking at any arm's output; undecidable cells stay in the denominator.

After the ledger validity check below, Scripty's text carries the asserted state in 41.1% of on-screen cells, against 14.2% for our SEAM reimplementation and 9.6% for the episode-local no-carry writer; on shared cells it is better than SEAM in 67 and worse in 2 (Wilcoxon $p < 0.001$). Expiry is a known weakness: Scripty keeps an expired state in 8.9% of cells, against 0.0% for SEAM and no-carry, better than SEAM in 10 shared cells and worse in 14 ($p = 0.097$). Before the check the rates were 36.5%, 13.3% and 8.8% (paired 69 vs. 13) and 8.8%, 0.6% and 0.0% (paired 14 vs. 18). The paired comparison scores each cell $+1$ (updated state written), 0 (silent) or -1 (stale state written), whereas the stale rate counts only -1 cells over entity-present cells, so an arm that writes more states can both win paired cells and have a higher stale rate. The Wilcoxon tests pair assertion cells, which are correlated within each of only three dramas; we therefore read their p values (< 0.001 for recall, 0.097 for staleness) as descriptive rather than as evidence across stories. The staleness gap is a real cost of keeping state until the storyboard removes it. A five-dimension blind reading is corroboration only, because the judge identified Scripty's side in 50 of 65 pairs; in addition, only Scripty's writer passes a timeline gate that marks flashback shots in the scene column (11 episodes carry the mark, no control episode does), which can reveal its side, whereas the ledger judging, which asks whether a state is present, is little affected; against SEAM the differences are $+5.5$ in character-state continuity, $+4.8$ in scene consistency, $+13.5$ in prop continuity, $+4.2$ in composition generability and $+2.3$ in intent fit. On the state of an earlier engine version, swapping only Scripty's writer for claude-sonnet-5, deepseek-v4-pro or kimi-k2.6 gives the same direction on SEAM's own probes, and for claude-sonnet-5 on the blind reading. Limits: the assertion inventory and expiry table are ours and have had no human pass; shot alignment is our reconstruction; and the comparison reimplements a partial protocol rather than the official one. Protocol details: SEAM-Bench, three dramas (werewolf 25, his-toyboy 23 and beyond-the-wall 20 episodes; 68 episodes, 65 with carried state). Scripty reads the frozen no-carry upstream storyboard (the same writer tier) and runs the frozen v2.1 engine (fc921a028) used in the video experiments, with scene-image objects on, and with Construction, review and writer on gpt-5.6-terra at high effort (through a Codex subscription, with the gateway gpt-5.6-terra as fallback); all 65 carried episodes converged. The SEAM reimplementation and the episode-local no-carry writer are the frozen 11 Sep outputs, re-judged in this round. Recall and stale

Table 23: External source-grounded text evaluation. Fixed-source-inventory recall, stale-state rate, and source faithfulness retain separate definitions and denominators.

Method	Stories	Source recall (%)	Stale (%)	Faithfulness	Coverage (%)
No-carry (episode-local)	3 (68 ep.)	9.6	0.0	N/A	100
SEAM reconstruction	3 (68 ep.)	14.2	0.0	N/A	100
Scripty	3 (68 ep.)	41.1	8.9	N/A	100

Dataset, model chain, source independence, and reconstructed instrument: SEAM-Bench, three dramas (werewolf 25, his-toyboy 23 and beyond-the-wall 20 episodes; 68 episodes, 65 with carried state). Scripty reads the frozen no-carry upstream storyboard (the same writer tier) and runs the frozen v2.1 engine (fc921a028) used in the video experiments, with scene-image objects on, and with Construction, review and writer on gpt-5.6-terra at high effort (through a Codex subscription, with the gateway gpt-5.6-terra as fallback); all 65 carried episodes converged. The SEAM reimplementation and the episode-local no-carry writer are the frozen 11 Sep outputs, re-judged in this round. Recall and stale rates are over a fixed inventory of source-defined state and expiry cells (381 and 239, of which 299 and 160 remain after the ledger validity check reported with the table), counted over cells where the entity is on screen, judged by gemini-3.1-pro-preview; coverage is the share of the inventory scored. The source-faithfulness (NLI) axis is near-tautological for post-hoc repair and is not reported. This is our reimplementation of a partial protocol, not an official SEAM-Bench score. These numbers are not official SEAM scores unless the exact official protocol has been verified and reproduced.

rates are over a fixed inventory of source-defined state and expiry cells (381 and 239, of which 299 and 160 remain after the ledger validity check reported with the table), counted over cells where the entity is on screen, judged by gemini-3.1-pro-preview; coverage is the share of the inventory scored. The source-faithfulness (NLI) axis is near-tautological for post-hoc repair and is not reported. This is our reimplementation of a partial protocol, not an official SEAM-Bench score.

Ledger validity check. Because the ledger was generated by a model from the scripts without independent human annotation, we checked every ledger cell before recomputing both sides. The removal rules were written and hashed (sha256 3ddb0f6e…) before any verdict file was opened. A cell is removed when (R1) the shared upstream storyboard never establishes the state, or has not yet established it in that episode, including, on the staleness side, a termination the storyboard never shows (rule R1s, added while the check was being executed); (R2) a one-off action or momentary reaction is treated as a state lasting across episodes; (R3) the entity appears in that episode only in flashback, memory or photo shots; (R4) the storyboard explicitly changes the state before that episode, or re-creates it through a new event; or (R5) the staleness episode belongs to the same night or continuous scene, so whether the state should have ended cannot be decided. The inputs were only the ledger entry and the upstream storyboard, never any arm's output or verdict. Each of the 72 assertions was examined once by DeepSeek, which had to cite verbatim upstream shots (all 49 citations matched); the author spot-checked 219 cells, agreed on 78.5% and changed 7 assertions. The check removed 161 of 620 cells (82 of 381 on the recall side, 79 of 239 on the staleness side, 67 of them under R1s), and both sides were recomputed from the existing verdicts without re-judging. Using DeepSeek's verdicts without the author's changes gives the same conclusions (recall 40.2% vs. 13.8%, staleness 8.0% vs. 0.0%). The removal is not neutral between arms, and it favours Scripty: the rule categories were motivated by an earlier Scripty loss attribution that had seen outcomes; 11 of the 13 recall cells where SEAM beat Scripty are removed (in those cells Scripty had followed the storyboard's changed appearance and was scored against the ledger's outdated wording); Scripty's recall rises from 36.5% to 41.1% while the other arms rise by about 0.8 pp; and because the upstream storyboard is the frozen no-carry arm's own text, rules judged against it are not independent of that arm. The check is AI-assisted and author spot-checked, not an independent human review. The error attribution below was made on the ledger before this check.

Where Scripty's text-level errors arise. AI-assisted diagnostic under author review, not an independent annotation. We attributed every Scripty cell that failed the recall face (183 of 288 on-screen cells) and every expiry cell in question (the 15 stale hits and 3 cells where the paired comparison favoured SEAM without a stale hit) to the earliest stage at which the tested state was lost or went wrong: upstream storyboard or inventory item, state record, cross-episode injection, writer, judge. Evidence was first located mechanically by author-chosen keywords traced through the storyboards, the per-episode state seeds, the injected text and the writer output; the remaining cells went to a

single assistant model (DeepSeek, from a family distinct from the writers and the judge), which returned a stage label with verbatim quotes; the author then reviewed all cells of the rare labels, a sample of the common label and all 18 expiry cells. Agreement was 43/59 on the recall check and 12/15 on stale hits, and the author changed 34 of the 183 recall labels. Most recall misses arise in the state record (103), 86 of them carried objects that are recorded only in the transient action field, where the next action overwrites them, or never recorded; 64 trace to the item or the storyboard (states never depicted upstream, momentary actions treated as lasting states, entities seen only in flashbacks); the writer, injection and judge account for 11, 3 and 2. For staleness, 12 of Scripty's 15 hits come from the record keeping a state it should have dropped: an explicit time skip ("days later", "a month later") does not expire earlier states, a newly added garment does not replace the old one on the same body part, and a character who has left keeps its state unchanged. These 12 hits sit on six state chains, one of which (Kaius's wound and bloodied shirt carried from episode 3 to episode 9) accounts for 6; the other 3 are ambiguous items, and injection, writer and judge account for none. Within one episode, as on StateHold, a state attribute stays until the storyboard removes it (§3.4); across episodes, carried appearance passes through each episode's initialization, and there the v2.1 engine has no lifetime mechanism for appearance state attributes (lifetime classification was deliberately deferred): whether a carried appearance ends is left to the model at each episode's initialization, which clears all carried appearance after an explicit time jump (dropping permanent accessories) and otherwise keeps transient details indefinitely. The same engine serves StateHold and ViMax. A lifetime-aware expiry is future work. The general limitation, that the record carries a state forward but can lack a trigger that ends it, is noted in §6. SEAM's low stale rate goes with writing few cross-episode states at all (recall 13.3%); we note this only as an observation, because the SEAM arm was not audited in the same way and rates after removing item-level errors are not compared across arms. The judge verdicts were not re-run.

E.1 Exploratory: adding an appearance-lifetime mechanism

This demonstration uses one SEAM-Bench drama (*his-toyboy*, episodes 1–10) and a variant of the engine that is not the paper's main system; every main-text and main-table result uses v2.1. The variant stores each character's appearance as separate state attributes in five fields (attire, worn items, marks, transient states, hair), each with a lifetime class: *transient* facts last within a scene, *durable* facts can outlast a day but not necessarily a long gap, and *permanent* facts persist unless the script changes them. Time jumps are recorded as per-shot events that must quote a scene heading, a caption or on-screen text. At a jump, transient facts expire and permanent facts are kept; for each durable fact a model decides whether it survives the stated gap, a semantic reviewer checks the decision, and uncertain facts are kept. Expired facts leave a tombstone so that a later episode cannot silently restore them. All other settings match the main SEAM-Bench run (same sealed storyboards, writer, Construction and review models, episodes chained from episode 1), and the same ledger judge and frozen ledger score it; the SEAM and no-carry arms reuse their verdicts for the same cells. On paired

Table 24: Appearance lifetimes on *his-toyboy*, episodes 2–10 (ledger cells with the entity present).

Arm	Presence	Expiry (original)	Expiry (checked ledger)
Scripty with appearance lifetimes	8/17 (47.1%)	0/12	0/7
Scripty (main version, v2.1)	1/18 (5.6%)	0/11	0/7
SEAM (re-judged)	1/18 (5.6%)	0/13	0/7
No-carry	0/17 (0.0%)	0/13	0/7

presence cells the lifetime variant is better than v2.1 in 7 cells, worse in none and tied in 10. All 7 wins come from two assertions: the protagonist's bracelet, judged present in every episode from 4 to 10 after the four-year jump (v2.1 keeps it only in episode 4), and a bite mark in episode 3. The bite mark is dropped at the four-year jump; where the storyboard later mentions an "old bite mark", the judge labels it updated rather than stale. No arm has a stale expiry cell in this range. Running Construction on episode 4 a second time from the same episode-3 state with all caches cleared again detects the jump, keeps the bracelet as a permanent item and drops the bite mark. With one drama, one writer sample, 17 presence cells and a gain concentrated in the one assertion the mechanism targets, this demonstrates the intended behaviour rather than estimating a general improvement; the other two dramas were not run with this variant.

F CONFIGURATION, COST AND REPRODUCIBILITY

FORMAL uses a fixed five-unit scheme and EXPAND 6-, 8- or 10-unit schemes (pools in §4).

F.1 IDENTITY OF THE REPORTED SYSTEM

Table 25 lists the engine version, models, renderer, judge, banks and settings of the runs reported in the main text. Construction, checking, writing and judging are listed as separate roles even where they use the same model family.

Version equivalence. We rebuilt, with the v2.1 code and no model calls, what each run should have sent and compared it with what it did send. Re-rendering every state record reproduces the injected text byte for byte (FORMAL 90 of 90 units, FORMAL without scene-image objects 90 of 90, EXPAND 192 of 192). Rebuilding the writer requests reproduced all 232 FORMAL and 474 EXPAND logged requests. Their final per-unit chains contain 90/192 first drafts, 126/248 prompt-checker-feedback rewrites and 8/22 format- or packing-gate rewrites (224/462 requests). Another 2/3 first drafts and 6/9 gate-feedback rewrites belong to superseded whole-unit chains: recorded replies discarded on restart, not failed calls in this ledger. The separate FORMAL no-scene-image-object audit covers 207 requests. First drafts match byte for byte. For final prompt-checker rewrites the fixed feedback frame and embedded previous draft also match; for gate rewrites only the prefix before the gate message was compared. Every L1 reference-selection request (FORMAL 452, EXPAND 606, FORMAL without scene-image objects 448) and every variant-image instruction is reproduced as well. The FORMAL records were built at a commit whose source differs from v2.1 only by an injection option that changes none of the 90 FORMAL units, so the 56 units written before the switch to the v2.1 copy received the same requests; the commit of the no-scene-object arms differs from v2.1 only by the scene-object switch. For the EXPAND no-scene-object arms the state records (564 of 564 files), injected text (192 of 192 units) and first-draft writer requests (192 of 192) match v2.1. For each final-L1 arm we also re-derived the full chain without model calls: replaying the stored selection answers with the committed code reproduces every per-unit reference decision (FORMAL 452 of 452, EXPAND 606 of 606); every planned reference image is the asset registered and sent in the video request; every clip (78 and 148 units) was produced by exactly that request, with reused clips traced to the run that submitted an identical request; and every judge cache key recomputes from the question text and those clip bytes (382 of 382 and 554 of 554 rows); across all seven final L1 runs, 3,818 selection decisions and 3,362 judge keys were re-derived with no mismatch. The seven historical runs are FORMAL A/B, EXPAND A/B, one no-scene-image-object run per pool and the EXPAND DeepSeek chain. Later sample-C L1 runs have separate judge-key checks; they are outside these 3,818 selection decisions and 3,362 judge keys.

F.2 COST

Table 26 separates three kinds of cost. Construction runs once per story, offline, before any clip is written. Deployment adds per-unit work: the prompt check with at most two rewrites, the L1 reference selection, and the generated variant images; the deterministic query and projection themselves make no model call. Failed and superseded attempts during development are reported separately and are not part of the deployment cost. Scripty's deployment overhead is extra model calls: about 34 (FORMAL) / 52 (EXPAND) Construction calls per story, and per generation unit 2.3 / 2.2 prompt-checker calls and 1.5 / 1.4 rewrites beyond the first draft. Per generation unit, Scripty makes 2.49 / 2.41 writer calls (1 first draft + 1.49 / 1.41 rewrites) and 2.34 / 2.23 prompt-checker calls, i.e. 4.83 / 4.64 model calls on the writer side, against 1.11 / 1.26 writer calls and no prompt-checker calls for both no-carry and native recap; Scripty therefore uses 2.2x / 1.9x the no-carry writer calls (4.3x / 3.7x counting prompt-checker calls). Counts are over the final attempt chain of each of the 90 / 192 units; whole-unit re-runs after a process restart are excluded for every arm; priced at public gateway rates this is roughly CNY 32-63 / CNY 55-110 per story for Construction and CNY 26-35 / CNY 38-51 for the extra writer work (lower bound: cached input free; FORMAL Construction tokens are estimated because its gateway log kept no usage). Deployed directly, the re-rendered units replace their L0 renders at the same length, so L1 adds no rendering cost; relative to L0 it adds about CNY 5.4–6.6 / 5.9–6.3 per story in model calls at gateway prices and CNY 4.0 / 4.7 in variant images, about CNY 9.4–10.6 / 10.6–11.0 in total (Table 27). The L1 row of Table 26 (655 / 831

calls) counts the reference-selection, image-generation and image-check calls the final variants depend on, including failed attempts. Construction and writer costs are usage priced at public gateway rates; L1 re-rendering is billed from the render receipts. L0 rendering (about CNY 62 / 88 per story) is paid by every arm and is not counted. Table 27 sets this cost against quality. Without the prompt check, L0 makes about as many writer calls per unit as the native recap and still passes 46.8% and 38.8% of cross-unit facts against the recap's 31.9% and 22.5%, so most of the gain does not come from extra writer calls; the cost that remains is Construction, which runs once per story. Because rendering dominates, the full system's total cost per story is about 1.9–2.6 times the recap's, not the 11–15 times of its model-call count.

F.3 What can be reproduced

We distinguish three levels. (i) Recomputing the reported statistics from judge records: the offline bundle reproduces the reported tables with no model call and no network access. (ii) Recomputing the reported tables offline from the shipped judge records is supported, but re-judging individual clips is not, because the media files themselves are not shipped. (iii) Full regeneration of the study (re-rendering the video, re-judging from the media, or regenerating clips) additionally requires the host pipeline and its models, which are not shipped. What is released for each level is listed in the Reproducibility Statement.

G Host Interfaces and Fair Comparisons

G.1 What each arm receives

Table 28 lists what each arm's writer can read in the short-drama host. All arms share the frozen upstream (unit split, storyboard, asset lists and reference images), the writer model and the renderer; they differ in the carried information, in any prompt check and repair step (Scripty's prompt checker, SEAM's repair model), and, for Scripty L1, in the reference images.

SEAM reimplementation. SEAM is our reimplementation of Section 4.2 of the original paper, applied as a post-hoc repairer to the no-carry arm's prompts, whose writer receives neither a recap nor state; the same repair applied on top of the native recap is reported separately (Appendix D.4). Because StateHold stories are single episodes, retrieval is confined to earlier shots of the same episode (latest prior state of each shared entity with edge weight at least 0.3, falling back to the last three shots), so the cross-episode memory that the original paper credits with most of its gain is not exercised. There is no top-k or token budget beyond the model's selective filter; a failed rewrite is retried once before mechanical injection; the repair model is the writers' model. The original paper gives no numeric hyperparameters; ours were fixed before any StateHold reading and not tuned. SEAM injects into 64% (FORMAL) and 51% (EXPAND) of the repairable units, against 27.4% in the original report, partly because we repair per unit; units without an injection reuse the no-carry clip. Neither SEAM nor SEAM-Bench releases code, prompts or a rubric, so the implementation and the text-level axes are ours. A sibling implementation that shares the filter and rewrite components but retrieves across episodes reproduces the original operating point on SEAM-Bench (68 episodes, five backbones): NLI probe 0.343→0.111 against the reported 0.308→0.131, and continuity recall 0.91–1.00 against 0.946; the StateHold arm retrieves only within an episode. Our implementations, prompts, rubric and ledger are provided in the supplementary material.

G.2 Shared upstream and arm-private runs

The upstream material is prepared once per story and copied unchanged to every arm before the arms fork. Each arm then writes in its own workspace, so one arm's prompts, state or recap cannot become another arm's input. The actual writer requests are kept, so the exposure of each arm can be checked from what was sent rather than from directory names.

G.3 Reference images

All arms draw reference images from the same frozen host asset library, which also holds a few state versions that the host authored from the shared storyboard before any run and that every arm's

prompt step may select (Table 25). The L1 variants are generated from Scripty's own record with the host's image model and are part of the L1 treatment; they are never shared with other arms. The record of each unit's request keeps the available images, the chosen image and whether a variant was used, so a retained or lost state can be traced to what the renderer actually received. L1 reuses L0 wherever the reference images do not change: units whose images are unchanged keep their L0 clip and judge rows. Units with a swapped image are re-rendered with a request that is identical to the L0 request except for the swapped images (text prompt and all other request fields byte-identical, verified without model calls on the submitted requests, FORMAL 44/44, EXPAND 73/73); of the 78 / 148 units that carry judged facts, these 44 / 73 are re-rendered in the final L1. Only facts and gates that involve a re-rendered unit are re-judged (FORMAL 120 of 191 question pairs, EXPAND 162 of 277), with the same judge and pass-1 rule; all other rows are carried over unchanged from L0.

G.4 Integration contract

Table 29 lists the inputs and hook points. ViMax requires a writer-request proxy and VideoMemory two runtime tool wrappers; both use operator-prepared Construction assets. Engineering time was not measured. Preserving host source does not imply adapter-free setup, nor identical information access: shared host storyboards and rendering assets are distinguished from the additional inputs read only by Construction.

Additional Construction inputs. The old Scripty production bundle is distinct from the host-generated rendering assets shared by the arms. Its reference files are byte-identical to an earlier production: 122 name-card placeholders and five *beyond-the-wall* scene images. Those five match the earlier scene pool; the surviving index calls them Stage 0 products, but earlier generation requests or source frames were not recovered. The scene directory lists locations and first-appearance episodes beyond the tested episode. The inspected files did not directly show the evaluated later wounds, bite marks, bracelet or chip, but this is not proof of equal information or complete source independence. Native arms do not read the additional bundle. EP01–03 VideoMemory receipts bind its original prepared copy, not the later EP04–05 additions. No information-matched causal effect is claimed for these external comparisons.

G.5 ViMax

ViMax plans a story into scenes and shots and renders them with keyframe references (Huang et al., 2026). Our adapter converts ViMax's storyboard and cast into Scripty's input format and passes the state to ViMax's prompt-writing model through a proxy, so that the ViMax code is unchanged. In the paired comparison, the Scripty arm and native ViMax receive byte-identical storyboards and cast lists, and the only difference in each request is the injected state block; this is checked for every shot. Construction additionally reads a pre-existing Scripty v10 production asset/reference bundle, including a name card for one missing scene. The v10 label identifies these assets, not the v2.1 engine used here; they are operator-prepared inputs, not outputs of ViMax extraction. The shared upstream storyboard was generated with gpt-5.6-terra, while both arms use gpt-6-luna as the refining writer; because both arms receive the same storyboard, this mismatch does not affect the comparison between them. The reported ViMax comparison uses L0 only. An exploratory L1 pass on beyond-the-wall showed no cross-unit gain (0 of 4 facts in both arms) and worse expiry controls (0 of 3 against 3 of 3 for native ViMax), because the reference variants inherit the stale wet-and-muddy appearance; within-unit facts passed 3 of 3 against 1 of 3. Two facts were dropped from both arms in this pass because ViMax's own image-selection model returned empty content for all eight shots of one scene, apparently a content filter. With this few facts the result is not interpretable beyond that.

G.6 VideoMemory

VideoMemory (Zhou et al., 2026) plans each episode into a storyboard and an entity memory and renders keyframes and clips with its own agents. We use the official implementation (commit a4b1545) without source changes, checked before every run by commit, clean working tree and a hash over all host source files, and call the host's own graphs in its order with its default settings (8 shots per

episode). Each episode's storyboard, memory bank and asset images are generated once and shared byte-identically by both groups. Construction separately reads the pre-existing Scripty v10 roster and scene assets used in the ViMax test, with operator-supplied scene-name cards; these are not native VideoMemory artifacts. Scripty v2.1 builds its record from the sealed host storyboard, chaining EP01 to EP03 under the engine's bounded re-roll contract. At run time the host's two visualization tools are bound to line-by-line copies whose only change is a string appended to the keyframe and video sub-agents' messages: empty for the native group, so its messages are byte-identical to the host's, and for Scripty a per-shot block with the state before and after each shot. An audit of all logged requests found no block in the native group and exactly one per request in the Scripty group. The host's configured services were unavailable, so three backends were replaced identically in both groups, changing only the client inside each host tool: the host LLM (gpt-5.2) by gpt-6-luna, image generation (nano-banana) by the image tool of the same model family, and video (Sora 2, 8 s, 1280×720) by Seedance 2.0 at 480p and 5 s, with the same seed per shot in both groups. Keyframes that the video provider flagged as a possible real person were resubmitted once through its asset channel, in both groups. In one Scripty-group episode the image provider's safety filter rejected two keyframes; both were made with the host keyframe agent's own prompt and references, one by a fresh session of the same backend and one by gemini-3-pro-image, and returned through the host's image tool; the second shot is in no judged unit, and dropping the one fact anchored on the first gives 4/8 vs. 1/8 cross-unit. The banks (33 facts: 9 cross-unit, 7 within-unit, 5 expiry controls, 12 in-scene controls) were written from the scripts and sealed storyboards and frozen before rendering; the judge and first-pass rule are those of the main table. All 144 shots rendered on the first pass. In Table 18, the four cross-unit wins are marks the script keeps across units (a bite mark, an arm wound, a facial cut), and expiry and in-scene controls are identical between groups. With 9 cross-unit facts the difference is not significant; like ViMax, this is an integration demonstration.

G.7 Failures and interventions

We count intended stories and units, the constructed range, released prompts, rendered clips and judged facts separately, because a stop in Construction, a writer failure and a renderer refusal happen at different stages. Facts in a unit that failed to render stay in the denominator as not passed (exception: on the ViMax portability host, three facts whose unit videos remained incomplete after the pre-set repair round were dropped from both groups before judging; see Table 18). In the development-pool runs summarized here, the video platform rejected four clip submissions, all in Scripty arms: the EXPAND L0 clip of one unit and the FORMAL no-scene-object clip of another (content filter, one rejection each), and the FORMAL L1 clip of a third unit (two consecutive failures: content filter, then an image-format error on the same bytes). Each was resubmitted with a byte-identical request and the first successful clip was kept. No baseline clip was refused in that tally. Separately, three FORMAL baseline drafts (one no-carry, two native recap) referenced ten or eleven images, above the renderer's limit of nine, and were rejected at submission; they were rewritten under a nine-reference writer gate (for the native recap, that unit and the later units of the story), re-packed and re-rendered, and the rewritten clips were used. In L1, two variant images were rejected at asset registration and regenerated once.

H Model Roles and Prompts

Table 30 lists every language- or image-model role, what it sees and what it may produce. The full prompt texts, schemas and model settings are released with the code; the deterministic query and projection steps make no model call. Reproduction levels are described in Appendix F.3.

Table 25: Provenance of the evaluated treatment. Identities refer to the artifacts actually consumed and scored, not merely to directory names or intended configuration.

Component	Frozen identity / evidence
Treatment commit / run	Scripty v2.1, execution copy fc921a028 (clean) with SI_INITVAL=1, PACK_HEADER_FILL=1, QTY_PRESENT_ONLY=1; EXPAND sample A built on it; FORMAL records built at d8cd2dfd6, which differs from the v2.1 copy only by the QTY option; re-rendering all 90 FORMAL injections with either code gives identical text, so the 56 units drafted before the switch received the same requests (Appendix F.1); noSO ablation on 9be0ff015.
Source / roster / split	Storyboards and host inputs from the clean run roots (FORMAL 18 stories, EXPAND 24), evaluation terms stripped (LK6 v4 scan: 710/1059 files, 0 hits); every arm reads the same clean copy; EXPAND main table uses Construction sample A.
Bank / unit map / windows	FORMAL (freeze sha 920e7ce49c2e…, 18 stories); EXPAND (freeze sha 0b6951ca4747…, 24 stories); windows: scene-cut trim for facts flagged needs_trimmed_window, else whole unit clips. Digest note: see bank.id.
Construction / review / writer	All LLM roles gpt-6-luna, effort high: INIT image reading, Construction, semantic/INIT reviewer, tier audit (FORMAL gateway gpt-6-luna; EXPAND 19 stories Codex + 5 gateway), writer, prompt checker, baseline writers, SEAM repair (Codex CLI), L1 reference selection (Codex or gateway per story).
Image / video backend	ARK doubao-seedance-2-0-260128, 720p, 9:16, 24 fps, audio on, no watermark, per-unit duration from the host pack, platform-drawn seed; L1 variants gemini-3-pro-image-preview (img2img), checked by gemini-3.1-pro-preview.
Judge / prompt / settings	gemini-3.1-pro-preview via gateway, judge prompt fingerprint 22292fcc, pair retries 2 (five arms judged under this rule; SEAM+recap was not), pass 1 only, frozen scorer 6f12e5e0.
Shared reference library	All arms draw reference images from the same frozen host asset library, which besides base images contains a small number of state versions authored by the host before any run from the shared storyboard (e.g. the mule without its pack frame); each arm's own host prompt step may select them (native recap 18 of 226 units, no-carry 10 of 226, Scripty L0 11 of 226, SEAM 5 of 112). Only the Scripty L1 arm adds variants generated during the run from base images (img2img), checked and registered, and used only by that arm.
Draws / seeds / retries	Construction re-rolls FORMAL 1 unit / 0 INIT, EXPAND 5 unit / 0 INIT (caps 3/3); prompt checker $\leq$2 rewrites; renderer seed platform-drawn, one draw per unit; four platform rejections in three Scripty units, each resubmitted with a byte-identical request (Appendix G); judge pair retries up to 2 extra rounds within pass 1.
Human protocol / exclusions	Fact-level study of the judge with six non-author raters on the no-carry and native-recap arms under an earlier question wording (Appendix D.18)

Each pool and modality retains its own evaluation identity. Failed or incomplete coverage is disclosed in the corresponding result table.

Table 26: Cost of the Scripty arm by role, including work in failed attempts. Shared study costs (base images, rendering, judging) are not attributed to Scripty.

Role	Calls	Input tokens	Output tokens	Call time (sum)	Failed-run share
Construction proposals	424 / 802	9.5M* / 22.1M*	0.37M* / 0.79M*	3.2 h / 8.8 h	0 / 6.7%
Review / correction	189 / 445	26.5M* / 62.3M*	0.31M* / 0.73M*	1.4 h / 6.7 h	0 / 6.7%
Query / injection	0 / 0	0 / 0	0 / 0	N/A	N/A
Writer overhead	345 / 699	19.3M / 37.0M	0.92M / 1.80M	6.5 h / 13.5 h	2.3% / 2.0%
Host asset supply	N/A	N/A	N/A	N/A	N/A
L1 asset generation	655 / 831	3.3M / 3.7M	0.36M / 0.56M	5.8 h / 7.6 h	15.7% / 7.6%
Host rendering	N/A	N/A	N/A	N/A	N/A
Evaluation	N/A	N/A	N/A	N/A	N/A

Aggregation basis, concurrency, hardware/backends, reusable setup, missing records, and full end-to-end latency: Cells give FORMAL / EXPAND for the Scripty arm of the reported runs. Construction proposals and review include INIT; writer overhead counts prompt-checker calls and rewrites beyond each unit's first draft; L1 counts the reference-selection, image-generation and image-check calls the final variants depend on, including failed attempts (the variant image instruction is part of the reference-selection reply, not a separate call); variants generated and checked during the two-reading-agreement-rule fix but later discarded (82 / 89 calls) are not included. Wall time is summed call time; roles ran in parallel across stories. *Includes estimated tokens where the gateway log kept no usage (all FORMAL Construction tokens). Base reference images, rendering of the first clip of every arm, and judging are shared study costs and are not attributed to Scripty. Per-role wall times overlap and are not additive; N/A marks measures the backend did not report.

Table 27: Quality against cost per configuration. Calls per generation unit count the final attempt chain (writer / prompt check / L1 reference selection), plus Construction calls amortized over units. LLM cost per story is the gateway-price equivalent (lower bound: cached input free; upper: full price); writer and prompt-checker calls ran on a subscription. Rendering is billed from the render receipts and is the same for every configuration, because L1 replaces L0 renders rather than adding to them.

Configuration	CUC (%)	Calls / unit	+ Constr. = all	LLM (CNY)	Images	Total (CNY)
FORMAL						
Native recap	31.9	1.11 / 0 / 0	$+0.00 = 1.11$	13.6–14.1	–	75.3–75.8
L0, no prompt check	46.8	1.10 / 0 / 0	$+6.81 = 7.91$	46.5–78.6	–	108.2–140.3
L0	51.1	2.49 / 2.34 / 0	$+6.81 = 11.64$	73.8–115.5	–	135.5–177.2
L1	63.8	2.49 / 2.34 / 4.91	$+6.81 = 16.55$	79.2–122.1	4.0	144.9–187.8
EXPAND						
Native recap	22.5	1.26 / 0 / 0	$+0.00 = 1.26$	20.4–22.0	–	108.4–110.0
L0, no prompt check	38.8	1.09 / 0 / 0	$+6.49 = 7.58$	75.6–131.5	–	163.6–219.5
L0	32.5	2.41 / 2.23 / 0	$+6.49 = 11.13$	113.4–185.0	–	201.4–273.0
L1	40.0	2.41 / 2.23 / 3.09	$+6.49 = 14.22$	119.3–191.3	4.7	212.0–284.0

Rendering per story: CNY 61.7 (FORMAL) and 88.0 (EXPAND), included in the total.

Table 28: Inputs available to each method in the short-drama host.

Arm	Current input	Carried information	Reference images
No-carry	Current unit's storyboard	None	Host library
Native recap	Same	Host's own text recap of earlier units	Host library
SEAM (in-host)	Same	No recap; the no-carry prompt is repaired with retrieved and filtered entity-attribute memory	Host library
Scripty L0	Same	Host recap plus the injected state	Host library
Scripty L1	Same	Same as L0	Host library, or variants generated from Scripty's record

Table 29: Integration contract for the three video hosts.

	Short-drama host (in-house)	ViMax	VideoMemory
Inputs read	Storyboard in units with scene notes, cast and asset lists, scene descriptions, reference images	ViMax storyboard and cast; external roster and scene assets for Construction	Host storyboard and memory; external roster and scene metadata for Construction
Prompt entry	Fixed slot in the host's prompt template; the host's writer writes the prompt; Scripty's checker returns it at most twice	State block added to each request to ViMax's prompt-writing model through a proxy	State block appended to the keyframe and video sub-agents' messages by tool wrappers
Reference entry (L1)	Per-slot choice among host images and variants made by the host's image model from the base image	Not used in the reported comparison (one exploratory pass)	Not used
Channels reported	L0 and L1	L0 only	L0 only
Unchanged	Host code, writer, renderer, shared upstream	ViMax code; storyboard and cast byte-identical across arms	Host code (hash-checked); storyboard, memory and assets byte-identical
Audit	Actual writer requests and chosen references kept	Requests differ only in the state block, checked per shot	Every logged request checked for the block

Table 30: Model roles, their inputs and their outputs.

Role	Sees	Produces
INIT	All shots, asset lists and descriptions, scene descriptions, reference images	Entity list and S_0, citing only the first shot
Per-unit update	The unit's shots and the current state	Changes per shot, each with a shot and quote
Semantic reviewer	Storyboard, proposed changes and resulting states (scene images for INIT)	Unsupported changes and omissions to fix
Host writer	Current unit's storyboard, host recap, injected state	The unit's video prompt and attached reference slots
Prompt checker	Must-keep list and the written prompt	Missing states; at most two rewrites
L1 reference chooser	Reference slot candidates and the state at unit start and end	Default, existing version or new image, with the changed attributes
Image model	Base image and the filtered appearance instruction	The variant image
Fact judge	Clips and question text only; no reference image, storyboard or state	Answers to gates and questions
SEAM extractor and repairer	Storyboard, draft prompt and retrieved memory	Entity-attribute memory and the repaired prompt