# VMEM-BENCH: ATTRIBUTING LONG-RANGE MEMORY FAILURES IN CAUSAL LONG-VIDEO GENERATION

**Anonymous authors**
Paper under double-blind review

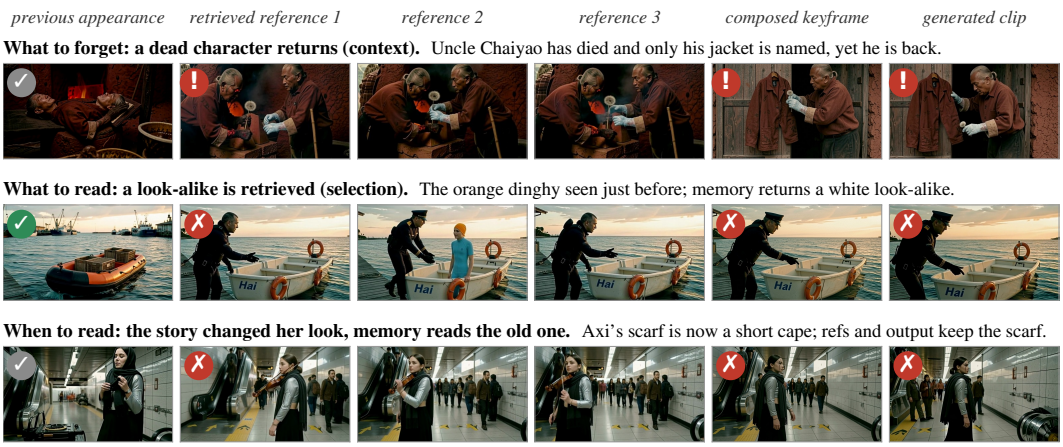


Figure 1: **Similarity decides what memory reads, and it cannot tell the past a story needs from the past that merely resembles it.** One Track-B rollout per decision. Top: a dead character returns because his jacket is named. Middle: a white look-alike replaces the orange dinghy just seen. Bottom: the instruction says Axi's scarf is now a short cape, yet memory returns her old look and the clip copies it (the no-memory rollout, not shown, draws the cape; a description-based judge counts her present). Badges: required entity correctly shown (check) or not (cross); removed entity present (!).

## ABSTRACT

Causal long-video generators increasingly treat memory as retrieval: to let characters and objects return after they leave the screen, they store past segments as frames, latents, or KV blocks and read back whatever best matches the current instruction or query. To test this bet, we propose **VMem-Bench**, a benchmark that scores memory both before and after generation and attributes each failure to the memory decision behind it. **Track-A** replays 92 real films (16,548 segments) and scores the references a memory hands to the generator before the next segment is revealed, with no generator involved. **Track-B** generates 30 long stories (2,824 segments) with one shared generator, run with retrieval memories and with two controls, no memory and each needed entity's previous appearance (*last-seen*), and scores whether required entities appear and look as before and whether removed ones stay out. Given the entities' names, retrieval finds most of the needed past, and its references keep returning entities closer to how they looked before at every gap. Its failures lie in decisions that similarity does not make: after the story changes an entity's look, it still hands over the old look, and the entity is drawn as described less often; it hands over more than is needed, costing precision; and it brings removed entities back, most of all when the prompt still names them. Last-seen references, which are told which entities are needed, gain comparable recall at lower precision and avoidance costs, yet they too hand over pre-change looks and fail when the instruction points to an earlier moment. Most inspected failures were settled before generation, in what memory stored or retrieved. Retrieval fails because it reads what resembles the instruction, not what is still true in the story; this suggests that it needs control that decides from the story when to read, what to read, and what to forget.

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

Video generators are moving from single clips to long, incremental rollouts: they emit a story segment by segment, take new instructions along the way, and never regenerate the past (Huang et al., 2025; Yang et al., 2025; Sand.ai et al., 2025; Yuan et al., 2026; Meng et al., 2026). Over such a rollout a character leaves and returns, a prop changes state, and some things must *not* come back, such as a character who has died. Because many generators condition only on a short recent window, recent systems add a long-range memory, and they build it as retrieval: store the past, then read back what best matches the current instruction or query. The store holds frames (Yu et al., 2025), latents (Hu et al., 2026b), KV blocks (Ji et al., 2025; Liu et al., 2026), or object slots (Dou et al., 2026); the read ranks them by relevance to the current instruction or query; and most never forget, or evict only the least relevant entries once a budget fills (Table 11). None of those we review tracks from the story whether an entity still exists, or whether a segment needs memory at all.

Such a memory can bring back a character who has died. Why, and what else does it get wrong? The final video cannot say at which step a failure arose. Each released system pairs its memory with its own generator, so a difference in the output mixes what the memory selected with how that generator renders it, and video benchmarks (Huang et al., 2023; Shi et al., 2026; He et al., 2026) score only the output. Fig. 1 shows why this matters: memory brings back a character who has died, retrieves a look-alike boat, and hands over a character's old look after the story has changed it. In each row the generator renders what memory handed over, and a score that checks the clip against the instruction's description even counts the last as correct.

**Scoring memory before and after generation.** A memory turns the past into a small set of references for the next segment:

$$O_{<n} \xrightarrow{\text{write}} \mathcal{M}_n \xrightarrow{\text{read with intent } g_n} \widehat{\mathcal{C}}_n \xrightarrow{\text{generate}} x_n. \tag{1}$$

Past observations $O_{<n}$ are written into a memory $\mathcal{M}_n$; the current instruction $g_n$ reads a finite context $\widehat{\mathcal{C}}_n$ from it; the generator turns that context into the next video $x_n$. Memory has done its job once $\widehat{\mathcal{C}}_n$ is fixed. VMem-Bench therefore measures it at two points. **Track-A** scores $\widehat{\mathcal{C}}_n$ itself on real films whose next segment is revealed only after scoring, so no generator is involved. **Track-B** lets the system generate its own story and scores the video $x_n$: whether required entities appear, whether returning ones look the same as before, and whether removed ones stay out. There, a **shared-generator testbed** runs four retrieval memories with one generator and one prompt stream, next to two controls: **no memory**, whose difference estimates what memory adds, and **last-seen references**, each needed entity's previous appearance, which separates what conditioning on references costs from what choosing them by similarity costs. A stage-by-stage human inspection then locates where Track-B failures arise.

**Attributing the failures.** Given the entities' names, retrieval finds most of the needed past, and its references keep returning entities closer to how they looked before at every gap; its failures lie in decisions that similarity does not make. *When to read:* a stored view helps only while it still shows how the entity should look; after the story changes an entity's look, retrieval still hands over the old one, and the entity is drawn as the instruction describes less often. *What to read:* on Track-A, the memory that finds the most past evidence also hands over the most wrong references; after generation, retrieval costs two to three times the precision that correct references cost. *What to forget:* retrieval brings removed entities back, most when the prompt still names them. *What was written:* a third of closed-loop misses come from evidence memory held but did not retrieve, and nearly a third from a wrong earlier rendering that the loop had stored. A rule that is told which entities are needed, the last-seen control, gains comparable recall at lower costs, but it too hands over pre-change looks and fails when the instruction points to an earlier moment. This suggests that similarity retrieval needs *memory control*: deciding, as the story unfolds, when to read, what to read, and what to forget.

**Contributions.**

- We propose VMem-Bench, a benchmark that scores long-range memory both before generation (Track-A: 92 films, 16,548 segments, 2,803 entities) and after it (Track-B: 30 stories, 2,824 segments, nine memory stresses), for presence, identity, and exclusion, with causal access, frozen gold, and context-budget accounting.

- We isolate what memory adds and what similarity costs, using a shared-generator testbed with no-memory and last-seen controls and paired confidence intervals, and locate where closed-loop failures arise through a stage-by-stage human attribution.
- We show that similarity retrieval keeps returning entities closer to how they looked before but hands over stale looks, costs precision, and brings removed entities back, and that recall-versus-gap curves and final-video scores hide all of this.
- We relate each observed failure to a decision that similarity does not make, namely when to read, what to read, what to write, and what to forget, and identify memory control as the open problem for causal long-video generation.

## 2 Related Work

**Causal long-video generation and memory.** Modern video models increasingly support long, incremental, and interactive generation (Yang et al., 2025; Yuan et al., 2026; Meng et al., 2026; Team, 2026). Their histories are carried through recent windows, frame/latent/KV caches (Zhang et al., 2025; Ji et al., 2025; Yi et al., 2026; Dou et al., 2026), retrieval over observed video (Yu et al., 2025; Hu et al., 2026b; Zhao et al., 2026), or entity-cued KV archives (Liu et al., 2026). These families keep very different internal states, so their native diagnostics cannot be compared; what they share is the context they hand to the generator and the video that results, which is what VMem-Bench scores (Table 10, App. A.4.1). They also share a read rule, relevance to the current instruction or query, even with object slots (Dou et al., 2026) or an entity-ID index (Liu et al., 2026); any eviction drops the least relevant entries, not those the story has retired. VideoMemory (Zhou et al., 2026) goes further, with a memory agent that keeps explicit character, prop, and background records and updates their states after each shot. It is, however, a script-to-shot pipeline: a storyboard agent plans the whole script before any shot is rendered, so every decision is made with the full plot in view. In causal generation, where the story arrives one instruction at a time, we are not aware of a memory that decides when to read and what to forget; our results suggest this is where the open problem lies.

**Memory evaluation beyond video.** LongMemEval (Wu et al., 2025a) and MemoryAgentBench (Hu et al., 2026c) separate a memory's indexing, retrieval, updating, and forgetting from merely having a long context, and agent, embodied, and egocentric work treats visual history as a budgeted resource (Wu et al., 2026; Yadav et al., 2025; Hu et al., 2026a; Wang et al., 2026; Li et al., 2026; Bei et al., 2026). RAG evaluation faces the same attribution problem as ours; RAGChecker (Ru et al., 2024) scores retrieval and generation separately and finds that better retrieval can add distracting context. VMem-Bench applies this component view to video, where the store is also rewritten from the system's own output.

**Video-generation benchmarks.** VBench (Huang et al., 2023) measures broad perceptual and semantic dimensions; MovieBench (Wu et al., 2025b) and MSVBench (Shi et al., 2026) extend evaluation to movie-level or multi-shot generation; EntityBench (He et al., 2026) explicitly tracks character/object/location consistency over recurrence gaps. These benchmarks judge generated outputs. VMem-Bench adds the missing measurement before generation and a no-memory control after it, which together show what memory contributes; our results show that recall-versus-gap curves, the usual summary, change shape with the mechanism and appear even without memory.

## 3 VMem-Bench: Scoring Memory Before and After Generation

VMem-Bench measures memory at the two points of Eq. 1: the context it hands to the generator (Track-A) and the video that results (Track-B). Fig. 2a shows one step of each. Take segment 15 of the film *Spring* (Fig. 11): the instruction asks for a girl, her staff, a nest, and a puppy last seen five segments earlier. Track-A asks whether the references a memory returns show the puppy; Track-B asks whether the puppy appears in the generated clip and is the same puppy as before; the no-memory control asks whether it would have appeared anyway, and the last-seen control what the puppy's own previous appearance would have done.

### 3.1 Causal Evaluation Contract

Both tracks process a story one segment at a time under one contract and differ only in what is scored.

**Evaluation contract**

**Input.** At segment $n$ the system receives only the instruction $g_n$, and after scoring the observation of segment $n$.
**Memory.** It may use only a memory $\mathcal{M}_n$ built from earlier segments; future segments and all scoring gold stay hidden.
**Output.** A context $\widehat{\mathcal{C}}_n$ of reference images; its size is the context budget $B_n = |\widehat{\mathcal{C}}_n|$.
**Scored.** Entities and their states, never how a memory works inside (App. A.1.2).

**Track-A: what memory hands to the generator.** Real films. $\widehat{\mathcal{C}}_n$ is scored against the entities in the next real segment, which is shown only afterwards. No generator runs, so a miss reflects memory (writing, upkeep, or reading) or the judge. Instructions name the entities on screen, so Track-A scores the retrieval boundary under explicit entity-bearing intents.

**Track-B: what survives generation.** Closed loop. The system generates $x_n$ from $\widehat{\mathcal{C}}_n$, the video is scored, and memory is updated from the system's *own* output, as in real use; the shared-generator testbed and its two controls separate memory from generator (Sec. 4.1).

### 3.2 Data, Gold, and Memory Stresses

Track-A is built from BlenderOpenMovies and LSMDC (Table 2): 92 films, 16,548 segments, 44.91 hours, and 2,803 typed characters, props, and locations, with slices for reappearance, multi-instance, scene-return, and state-change events.

Track-B contains 30 human-authored stories expanded into 2,824 five-second segments (3.92 hours), with reappearances up to 77 segments apart. The instruction gives an entity's look only when it first appears or changes state; afterwards it names the entity, so a character's look must come from memory. Many props, however, carry their look in the name ("the two brass fog bells"), and most returns after 30 or more segments are either such props or characters that come back changed, with the new look spelled out. Each segment is tagged with the memory stresses it exercises (Fig. 2b): six ask the system to *remember* something and three to *keep something out* (Table 1).

Track-A gold comes from a staged pipeline: VLM proposals are normalized by deterministic contracts and stronger-model review triggers targeted repair. Humans fix the canonical entity roster, and every blocked unit plus a stratified audit sample of the rest passes blocking human review before a deterministic freeze (App. A.1.1).

### 3.3 Metrics

**Track-A.** A pinned Qwen3-VL-32B judge, blind to system identity, checks each reference against the frozen roster and the real next segment. At segment $n$, let $C_n$ be the entities that must be remembered (on screen, excluding first appearances) and $k_n$ the number of references. Precision is the fraction of the $k_n$ references that show a correct entity; Recall is the fraction of $C_n$ covered by at

Table 1: Track-B memory-stress families. Each rate is a micro-average over that family's opportunities.

| Family | Metric | Satisfied when |
|---|---|---|
| Long-gap reappearance | recall | A character or prop returns after at least 30 segments. |
| Look-alike | accuracy | The identity matches the intended member of a confusable pair. |
| State change | accuracy | A transformed entity appears in its new state. |
| Count / multi-instance | accuracy | The rendered count matches the authored count. |
| Avoidance | avoidance-OK | A removed, destroyed, or deceased entity is absent. |
| False friend | accuracy | A new look-alike is not collapsed into an existing identity. |
| State persistence | accuracy | A later segment keeps an earlier transformed state. |
| Indirect reference | avoidance-OK | A deprecated entity stays absent even if named in the prose. |
| Temporal reference | recall | A name-free time reference recovers the intended entity. |

Table 2: **VMem-Bench Track-A corpus statistics** (per dataset). Entities = characters + props + locations; hard-case columns count segments per bucket. BlenderOpenMovies is stylized animation, LSMDC live-action film clips; the two are analyzed separately throughout.

| | Corpus | | | | | | | Hard-case segments | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Movies | Segments | Entities | Char. | Prop. | Loc. | Dur. (h) | Re-app. | Multi-inst. | Scene-ret. | State-chg. |
| BlenderOpenMovies | 23 | 1,364 | 369 | 106 | 149 | 114 | 3.84 | 415 | 674 | 106 | 86 |
| LSMDC | 69 | 15,184 | 2,434 | 881 | 690 | 863 | 41.07 | 5,425 | 8,596 | 1,582 | 179 |
| Total | 92 | 16,548 | 2,803 | 987 | 839 | 977 | 44.91 | 5,840 | 9,270 | 1,688 | 265 |

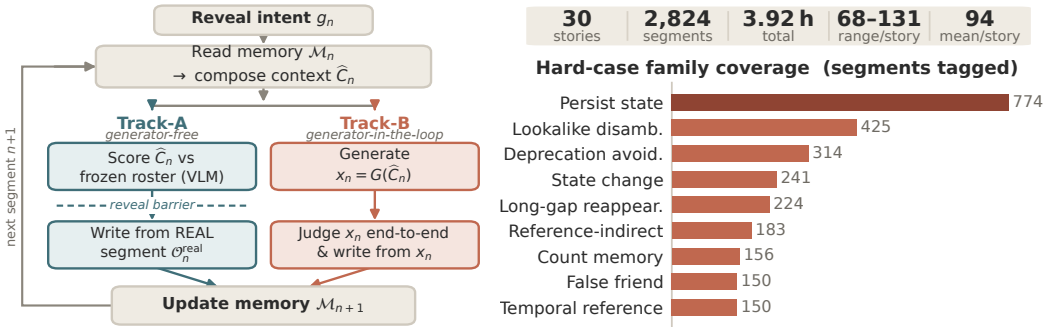


Figure 2: **(a) One step of VMem-Bench.** For each segment, memory is read with the instruction to compose a context $\widehat{C}_n$. **Track-A** scores that context against the entities of the real next segment *before* the segment is shown, then updates memory from the real segment; no generator is involved. **Track-B** generates $x_n{=}G(\widehat{C}_n)$, scores the video, and updates memory from the system's own output. **(b) Track-B scale.** Segments tagged with each of the nine memory stresses (a segment may carry several).

least one of them. Both, and their F1, are computed per segment and averaged over a film's segments, then over films (so the reported F1 is not the harmonic mean of the reported averages). Repetition is measured as $R_V$, the share of references that repeat an already covered view of the same entity, and $R_S$, their mean DINOv3 similarity; $\text{SelEff} = \text{Prec}\,(1 - R_V)$.

**Track-B.** The same judge sees each generated segment with a blinded list of entities that should appear, should not appear, look alike, or are decoys, each with its gold description, and marks which are present. An entity counts as present when the clip shows something that matches its description; the judge does not compare it with the entity's earlier renderings, so a fresh, description-matching stranger scores as a hit. Track-B therefore also scores *identity*: whether a returning entity looks the same as its previous appearance in the same rollout, measured as the DINOv3 similarity of the localized entity over returns detected in both clips, with ArcFace for faces, MegaLoc for locations, and a pairwise VLM judge as a check; thresholds are calibrated on control pairs before any condition is scored (App. A.1.4). Counts are pooled over all segments. Recall is the fraction of required character and prop appearances that are present; Precision is the fraction of present story entities that were required; F1 is their harmonic mean; Avoid-OK is the fraction of forbidden appearances that stay absent. Each memory stress uses its own rate. Differences to the no-memory control carry paired, story-clustered bootstrap 95% CIs, conditional on one rollout per story (App. A.1.3). Each condition builds its own history, so a paired difference compares whole rollouts on the same instructions.

## 4 Findings: Attributing Memory Failures

### 4.1 Setup

**Four retrieval memories and two controls.** The testbed stores every past segment's text and sampled frames and retrieves references for the next one in four ways: *Text-frame* ranks single frames by text similarity to the instruction; *Seg+uniform* ranks whole segments by text and samples their frames evenly; *Seg+DINO-key* ranks segments the same way but keeps visually diverse DINOv3 keyframes; *Seg+frame* re-ranks the frames of the top segments. On Track-A each returns up to $B{=}16$

Table 3: **Track-B: scoring the generated videos per memory stress** (30 stories per system, pooled), grouped by decision; all columns are higher-is-better. Top: four retrieval memories and two controls on one generator, no memory and last-seen references (each required entity's previous appearance in the same rollout, chosen with gold); controls are not ranked. Green (red) cells are significantly above (below) no memory, paired story-bootstrap 95% CI; uncolored cells are not significantly different. Bottom: released systems with their own generators, not comparable with the top block. Overall is pooled character/prop recall. Bold and underline mark the best and second best retrieval memory.

| | **When** | **What to read** | | | | **Write** | | **Forget** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| System | Long-gap rec. | Look-alike acc. | False-fr. acc. | Count acc. | Temporal rec. | State-chg acc. | Persist acc. | Avoid avoid. | Indirect avoid. | Overall↑ rec. |
| *Shared generator: retrieval memories and two controls* | | | | | | | | | | |
| Text-frame ret. | 0.843 | 0.947 | **0.947** | **0.842** | 0.813 | **0.991** | 0.789 | **0.885** | 0.694 | 0.623 |
| Seg+uniform ret. | **0.890** | 0.962 | 0.920 | 0.769 | **0.880** | **0.991** | 0.775 | 0.848 | **0.749** | 0.642 |
| Seg+DINO-key ret. | 0.862 | 0.969 | 0.920 | 0.763 | 0.873 | 0.983 | 0.788 | 0.848 | 0.694 | **0.684** |
| Seg+frame ret. | 0.829 | **0.978** | 0.907 | 0.818 | 0.853 | **0.991** | **0.796** | 0.843 | 0.617 | 0.671 |
| *No memory* ($K$=0) | *0.895* | *0.955* | *0.960* | *0.758* | *0.873* | *0.987* | *0.681* | *0.964* | *0.863* | *0.563* |
| *Last-seen refs* | *0.881* | *0.957* | *0.980* | *0.715* | *0.773* | *0.995* | *0.779* | *0.930* | *0.809* | *0.632* |
| *Released systems (own generators)* | | | | | | | | | | |
| MemFlow | 0.238 | 0.957 | 0.440 | 0.615 | 0.300 | 0.991 | 0.758 | 0.921 | 0.754 | 0.360 |
| LongLive-RAG | 0.386 | 0.981 | 0.740 | 0.718 | 0.327 | 0.994 | 0.712 | 0.983 | 0.929 | 0.284 |
| IAMFlow | 0.186 | 0.949 | 0.393 | 0.659 | 0.253 | 1.000 | 0.751 | 0.916 | 0.781 | 0.356 |

references. On Track-B each queries only its own generated history, returns $K$=3 references, and feeds one shared generator: FLUX.2 composes a keyframe from the instruction and the references (from the instruction alone when there are none), and Wan2.2-TI2V-5B turns it into the five-second clip that joins the history (App. A.4.3). All Track-B comparisons use this pipeline with one rollout per story and condition, and paired CIs are conditional on these rollouts. Two controls use the same generator and prompts. The *no-memory* control gets no references, so the difference to it is what memory adds. The *last-seen* control chooses by entity instead of by similarity: for each needed entity it hands over that entity's most recent appearance in the same rollout, chosen with the gold entity list but without any gold text or image (App. A.4.4).

**Released systems.** LongLive-RAG (Hu et al., 2026b) and MemFlow (Ji et al., 2025) run their released memory on Track-A; on Track-B they and IAMFlow (Liu et al., 2026) keep their own generators and are compared as whole systems (App. A.4.5, A.4.1).

### 4.2 When to Read: A Stored View Helps Until the Story Changes It

Against the gap since an entity was last seen, the retrieval memories recall more at long gaps (0.65–0.71 at 1–4 segments, 0.83–0.89 at 30 or more), but the same generator without memory rises even more steeply, from 0.57 to 0.90 (App. A.3): a returning entity is named in the instruction and, after a long absence, usually described. What a read adds depends on what the stored view still shows (Fig. 3). *While the look is not in the text,* it supplies it: a character named only by name and seen within four segments is drawn 51% of the time without memory, 61–70% with retrieval, and 62% with last-seen references, every paired CI above zero; a changed state that must persist rises from 0.681 to 0.775–0.796. *At any gap,* it keeps the entity the same: the DINOv3 similarity between an entity's previous and current appearance rises from 0.34 without memory to 0.49–0.57 with references, and from 0.52 to 0.63–0.73 for returns after 30 or more segments (mostly props), where presence as described does not change; the entity is detected on both sides in 88–90% of returns in every condition, so these means cover nearly the same returns (App. A.1.4); without memory the generator draws a stand-in that fits the description. *Once the story changes the look,* the stored view is stale: characters that return changed after 30 or more segments, with the new look in the instruction, are drawn as described 97% of the time without memory; all four retrieval memories lower this (−0.02 to −0.13; two CIs below zero), and the returns move toward the old appearance (similarity 0.20 to 0.36–0.39). Because the look changes at the return, every stored view of the

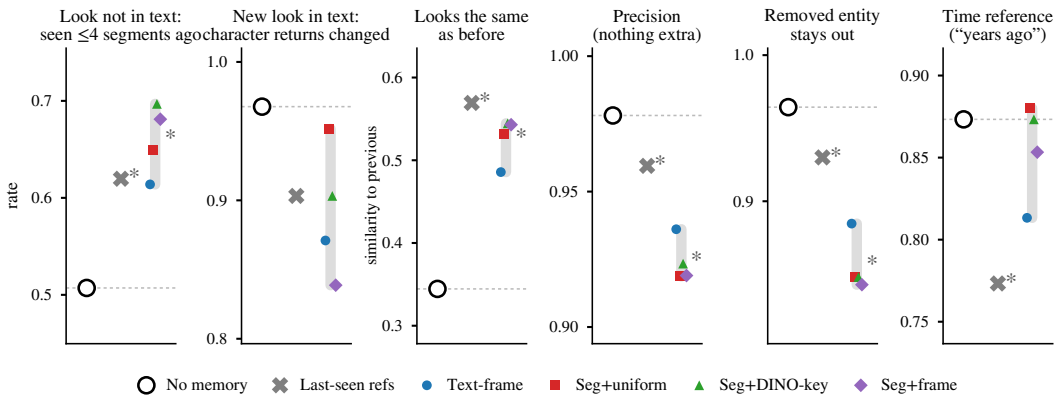


Figure 3: **Track-B: what references do, on one generator.** Each panel compares the same generator with no memory (open circle), with last-seen references (grey cross: each needed entity's previous appearance in the same rollout), and with the four similarity retrieval memories (colored). References help a character named only by name and seen within four segments; when the instruction spells out a new look, stale ones hurt. Returning entities look more like their previous appearance (DINOv3 similarity) with any references. References also cost precision and let removed entities back, and similarity retrieval raises both costs; last-seen references, which show the entity as it looks now, hurt instructions about an earlier moment. A star marks a difference to no memory whose paired 95% CI excludes zero (for retrieval: all four).

Table 4: **Track-A: scoring the references each memory hands to the generator**, at each memory's own budget (Refs = references per segment). Precision, Recall, and F1 say how many references are right and how many needed entities they cover; $R_V$ and $R_S$ measure repetition among the references (lower is better; App. A.2). Best is **bold**, second best underlined.

| System | BlenderOpenMovies | | | | | | | LSMDC | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Prec↑ | Rec↑ | F1↑ | SelEff↑ | $R_V$↓ | $R_S$↓ | Refs↓ | Prec↑ | Rec↑ | F1↑ | SelEff↑ | $R_V$↓ | $R_S$↓ | Refs↓ |
| *Retrieval memories (testbed)* | | | | | | | | | | | | | | |
| Text-frame ret. | **0.692** | 0.887 | **0.817** | 0.323 | 0.492 | 0.634 | 15.47 | **0.630** | 0.802 | **0.670** | 0.330 | 0.404 | 0.543 | 15.94 |
| Seg+frame ret. | 0.667 | 0.869 | 0.792 | 0.294 | 0.525 | 0.637 | 15.47 | 0.580 | 0.772 | 0.630 | 0.285 | 0.433 | 0.573 | 15.93 |
| Seg+uniform ret. | 0.662 | 0.883 | 0.796 | 0.297 | 0.515 | 0.650 | 15.47 | 0.566 | 0.769 | 0.622 | 0.264 | 0.454 | 0.605 | 15.92 |
| Seg+DINO-key ret. | 0.579 | **0.923** | 0.752 | **0.431** | **0.225** | **0.488** | 13.80 | 0.484 | **0.849** | 0.591 | **0.393** | 0.143 | **0.393** | 14.82 |
| *Released memories* | | | | | | | | | | | | | | |
| LongLive-RAG | 0.619 | 0.648 | 0.647 | 0.325 | 0.433 | 0.773 | **5.98** | 0.447 | 0.506 | 0.441 | 0.267 | 0.283 | 0.665 | **6.00** |
| MemFlow | 0.321 | 0.597 | 0.422 | 0.194 | 0.294 | 0.625 | 15.94 | 0.240 | 0.530 | 0.302 | 0.185 | **0.133** | 0.424 | 16.00 |

character predates the change; last-seen references hand one over as well and move the return further (0.44). A memory that does not track the story's state, whether it chooses by similarity or by entity, hands over a look the story has left behind.

**Do the released memories work?** Their own scores cannot say: with their own generators, the released systems recall 0.284–0.360, below our shared generator with no memory at all (0.563). Per-clip VBench quality (Huang et al., 2023) cannot tell either: it differs by at most 0.19 standardized units across the retrieval memories and rates the no-memory rollout *higher* on within-clip consistency (App. A.4.6).

### 4.3 What to Read: Similarity Hands Over More Than Is Needed

Track-A scores the references themselves (Table 4, Fig. 4). The retrieval memories return 13.8–15.9 references per segment, LongLive-RAG about six, and MemFlow about sixteen. Seg+DINO-key covers the most required entities (Recall 0.923 on BlenderOpenMovies, 0.849 on LSMDC) because it picks visually diverse keyframes, but it also has the lowest Precision among the retrieval memories (0.579/0.484): diversity brings in views that show none of the needed entities (Fig. 11). Text-frame

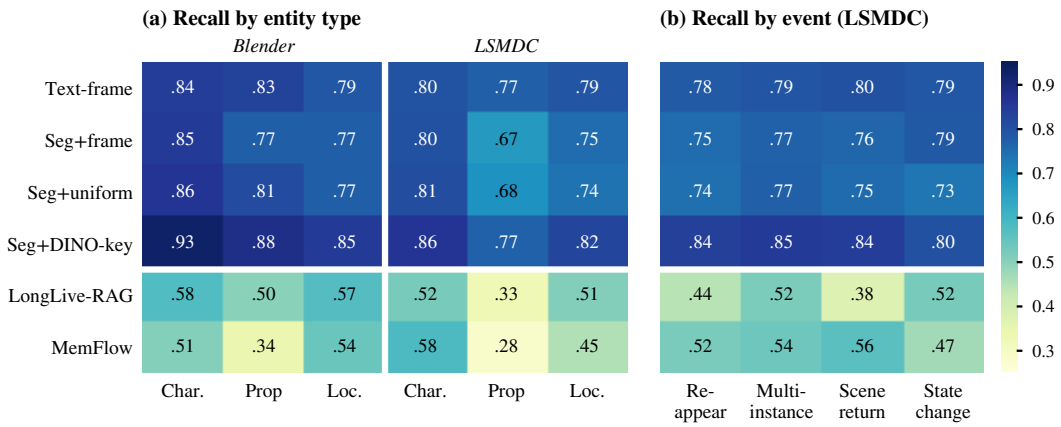


Figure 4: **What is hard on Track-A.** Continuity recall of each memory, **(a)** by the kind of entity that must be remembered and **(b)** by the event that makes a LSMDC segment hard (lighter means lower recall). Props and returning scenes are hardest, especially for the released memories (below the white line); diverse views help least for state changes, which no earlier view shows. Seg+DINO-key and Text-frame rank first and second on both corpora except Blender multi-instance.

covers slightly less (0.887/0.802) with the best F1 (0.817/0.670), and it leads at every budget (App. A.2): among the tested memories, the one that finds the most, Seg+DINO-key, also hands over the most wrong views; after generation it again recalls most, and Text-frame is again the most precise (Table 3).

After generation, the last-seen control separates what any reference costs from what similarity adds. With last-seen references, Precision falls from 0.978 to 0.960, and the fall grows with the number of references ($-0.015$ with one, $-0.026$ with two, an uncertain $-0.024$ on the 149 segments with three; App. A.3). Similarity retrieval, which always hands over three, costs $-0.04$ to $-0.06$, and all four retrieval memories fall significantly below the last-seen control; the nearest match can also be a look-alike (Fig. 1, middle row). Choosing the right entity is not enough when the instruction points to a particular moment. For time references ("the girl, years ago"), last-seen references show the entity as it looks now and lower recall from 0.873 to 0.773 (CI $[-0.15, -0.05]$), and Text-frame retrieval also falls (0.813): a memory has to choose not only which entity to show but when.

### 4.4 What to Forget: Retrieval Brings Removed Entities Back

When an entity must stay out, memory hurts, and similarity is most of the reason. Against no memory, Avoid-OK for destroyed or dead entities drops from 0.964 to 0.843–0.885 with retrieval, every paired CI below zero, and from 0.863 to 0.617–0.749 when the prompt still mentions the entity. Last-seen references, chosen to show only needed entities, cost less than half as much ($0.964 \rightarrow 0.930$): a past frame can also show an entity that has since left the story. A retriever ranks the dead character by how much he resembles the prompt, not by whether he is still in the story, so naming his jacket brings him back (Fig. 1, top row); in all 26 sampled cases, the forbidden entity was already in the retrieved references (Sec. 4.5). With two references, last-seen references still cost about half as much as retrieval's three (App. A.3).

### 4.5 What to Write: Most Failures Are Settled Before Generation

A missed entity may never have been rendered correctly before (writing), not retrieved (selection), dropped by the generator (synthesis), or missed by the judge; the testbed records every step to tell them apart. We inspected 224 blinded units from the four retrieval memories, drawn from judge-flagged failures by failure type and gap, with judge-confirmed controls and hidden repeats (App. A.4.7).

Of the required-entity failures (Fig. 5; inclusion-weighted with 95% CIs, raw counts 48, 38, 11, and 50 of 147), selection accounts for 36% [27, 45], writing for 29% [22, 37], synthesis for 7% [3, 13],

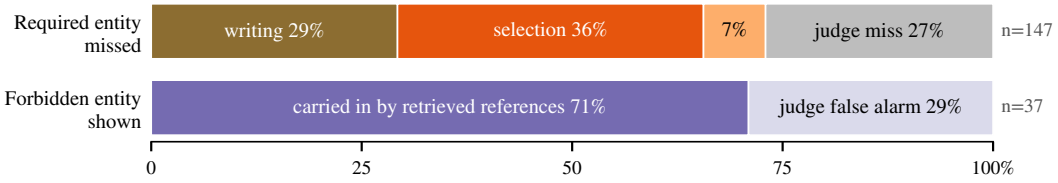


Figure 5: **Track-B: where a failure arises**, from a stage-by-stage human inspection of the four retrieval memories (pooled, inclusion-weighted). A missed required entity was already wrong in its previous appearance (writing), present there but not retrieved (selection), retrieved but lost in the keyframe (synthesis), or rendered and missed by the judge. A forbidden entity on screen was carried in by the retrieved references or is a judge false alarm. Per-stratum shares are in Fig. 15b.

Table 5: **From findings to design.** Each decision, the Track-B evidence that similarity does not make it, and a design hypothesis that VMem-Bench can test.

| Decision | Evidence | Hypothesis to test |
|---|---|---|
| When to read | References lift a name-only character seen within four segments ($+0.11$ to $+0.19$), not one whose look is in the text | Read only when the text lacks the look |
| Which moment | Latest views are stale for time references ($-0.10$) and for characters returning changed ($-0.10$, $-0.13$) | Index by entity, time, and state |
| How much | Precision falls with each reference; retrieval's three cost two to three times as much | Hand over few, precise references |
| What to forget | References let removed entities back (Avoid-OK $-0.08$ to $-0.12$) | Track who is still in the story |
| What to write | 29% of sampled misses start from a stored wrong rendering | Check a rendering before storing it |

and judge misses for 27% [18, 36]. Selection is the part Track-A exposes; writing, a wrong generated appearance that memory faithfully stored, is an error only the closed loop exposes. Together, 65% of flagged misses, and 90% of those confirmed as real, are settled before the generator runs. Of 37 resolved forbidden renders, none was introduced by the generator alone: 26 were visible in the retrieved references (labeled context), and 11 were judge false alarms (Sec. 4.4).

### 4.6 Toward Memory Control

Each observed failure points to a decision that similarity does not make (Table 5); we conjecture that retrieval needs a control layer that tracks which entities exist, in what state, and which moment the instruction means. Its simplest form is an entity table that records who is still in the story, how each entity looks now, and when it was seen, and chooses references by entity, state, and time rather than by resemblance; last-seen references are its crudest version. Our evidence motivates such control, and VMem-Bench can measure it.

## 5 Conclusion

VMem-Bench scores long-range memory before generation, on the references it hands over, and after it, against no memory and last-seen references. Measured this way, similarity retrieval finds most of the needed past and keeps returning entities closer to how they looked, but it hands over stale looks and more than is needed and brings removed entities back, because it reads what resembles the instruction rather than what is still true in the story; a rule told which entities are needed cuts these costs yet misses changed looks and earlier moments. A new memory should be scored the same way: a recall-versus-gap curve or a final-video score alone cannot show whether memory helped, what it cost, or at which decision it failed, and VMem-Bench is built to show all three.

## USE OF LARGE LANGUAGE MODELS

Generative AI tools were used for literature retrieval and discovery; research ideation and execution, including feedback on methodology/experiments, software/code editing, and interpretation of already-computed results; and drafting and polishing portions of the manuscript, figures, tables, and LaTeX. They were not used to generate the reported Track-A/Track-B measurements, the Track-B semantic source stories, or mathematical claims/proofs. Track-B source stories are human-authored; scoring gold and SUT prompts are deterministic transformations of that source layer. The authors independently checked benchmark definitions, protocols, code changes, numerical results, citations, and scientific claims and take full responsibility for the paper.

## ETHICS STATEMENT

Track-A uses publicly released research video (BlenderOpenMovies open movies and LSMDC movie clips) under their existing licenses; raw film files are not redistributed. Track-B stories are fictional benchmark prompts rather than depictions of real people. The supplementary human annotation was done independently of gold construction and system implementation and asked only whether fictional entities are visible in generated clips; no annotator identity or other personal information is recorded. We do not see a specific dual-use risk beyond that of the underlying video generators.

## REPRODUCIBILITY STATEMENT

Frozen gold, scoring contracts, adapter entry points, causal guards, and the numbers in the main tables are specified in the appendix and released benchmark artifacts. Code, configuration, dataset construction scripts, scorer settings, and evaluation contracts will be released so both tracks can be reproduced from the same frozen protocol. Implementation notes that would de-anonymize the submission are omitted from the anonymous manuscript and restored for camera-ready.

## REFERENCES

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40750–40784, 2026.

Weijia Dou, Hui Li, Jiahao Cui, Lei Zhou, Jingdong Wang, and Siyu Zhu. Slotmemory: Object-centric kv memory for streaming long-video generation. *arXiv preprint arXiv:2605.31033*, 2026.

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.

Junjun Hu, Xinda Xue, Botao Ren, Minghua Luo, Jintao Chen, Haochen Bai, Liangliang You, and Mu Xu. Astranav-memory: Contexts compression for long memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8097–8109, 2026a.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. Longlive-rag: A general retrieval-augmented framework for long video generation. *arXiv preprint arXiv:2606.02553*, 2026b.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026c.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Aiden Yiliu Li, Nels Numan, and Anthony Steed. Visual agentic memory: Enabling online long video understanding via online indexing, hierarchical memory, and agentic retrieval. *arXiv preprint arXiv:2605.16481*, 2026.

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*, 2026.

Yihao Meng, Zichen Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Yue Yu, Hanlin Wang, Haobo Li, Jiapeng Zhu, Yanhong Zeng, Xing Zhu, Yujun Shen, Qifeng Chen, and Huamin Qu. Causalcine: Real-time autoregressive generation for multi-shot video narratives. *arXiv preprint arXiv:2605.12496*, 2026.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities. `https://www.minimax.io/blog/minimax-h3`, 2026. Accessed: 2026-09-20.

Qwen Team. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

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 the Association for Computational Linguistics: ACL 2026*, 2026.

Catnip AI Team. Mainecoon: Pursuing a real-time audio-visual social world model. Technical report, Catnip AI, 2026. URL `https://mainecoon.tech/`.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. Egomemory: A comprehensive memory benchmark for long-horizon egocentric video understanding. *arXiv preprint*, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025a.

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28984–28994, 2025b.

Yong Wu, YanZhao Zheng, TianZe Xu, ZhenTao Zhang, YuanQiang Yu, JiHuai Zhu, Chao Ma, BinBin Lin, BaoHua Dong, HangCheng Zhu, RuoHui Huang, and Gang Yu. ContextBudget: Budget-aware context management for long-horizon search agents. *arXiv preprint arXiv:2604.01664*, 2026.

Karmesh Yadav, Yusuf Ali, Gunshi Gupta, Yarin Gal, and Zsolt Kira. Findingdory: A benchmark to evaluate memory in embodied agents. *arXiv preprint arXiv:2506.15635*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. Worldkv: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025.

Lin Zhao, Yushu Wu, Yifan Gong, Yanzhi Wang, and Pu Zhao. Omnimem: Scalable and adaptive memory retrieval for long video generation. *arXiv preprint arXiv:2605.30519*, 2026.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Y. 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.

# A Appendix

## A.1 Benchmark Construction, Gold, and Scoring

### A.1.1 Benchmark Construction Pipeline (S1–S4)

Track-A gold is built in four stages followed by a deterministic freeze. Model outputs are replaceable proposals; stable identifiers, prompt completeness, and frozen gold are fixed by deterministic contracts and human adjudication. **S1** drafts bounded segments, present entities, first appearances, and finite-ontology state events; a human-confirmed canonical roster is the naming authority. **S2** deterministically normalizes schema, identifiers, timestamps, and duplicates and runs structural linting. **S3** reviews each segment with Qwen3-VL-8B (Qwen Team, 2025); low-confidence/conflicting presence claims are rechecked and confirmed entities trigger deterministic prompt repair. Each unit receives Pass, Warn, Block, or RetryableError. **S4** sends every Block plus a stratified audit sample of the remaining units to blocking human review. Accepted decisions re-run the full deterministic validator. The final freeze writes the gold registry and layout hash; a changed layout/exclusion set creates a new data version. Every stage writes a derived artifact and never mutates the original draft. Fig. 6 summarizes both tracks' construction, and Fig. 7 shows what the Track-A stages change in practice.

### A.1.2 Anatomy of a Frozen Gold Instance

A frozen Track-A movie is an ordered sequence of prompt-complete intents with harness-side typed entity records, lifecycle/state events, required memory records, and forbidden/deprecated records. Only the current intent and, after reveal, the real segment are visible to the system. Typed gold identities, persistent gold records, required targets, and forbidden targets remain scoring-only. Entity records use {`character`, `prop`, `location`} with canonical names and aliases. Track-A's reappearance, multi-instance, scene-return, and state-change labels are diagnostic tags over this same frozen structure. Fig. 8 shows one frozen film.

A Track-B story is authored rather than annotated: a human writes the entities, their states, look-alike pairs, and scenes, and the scoring gold and prompt stream are derived from that skeleton. Fig. 9 shows how one story places every memory stress on its timeline.

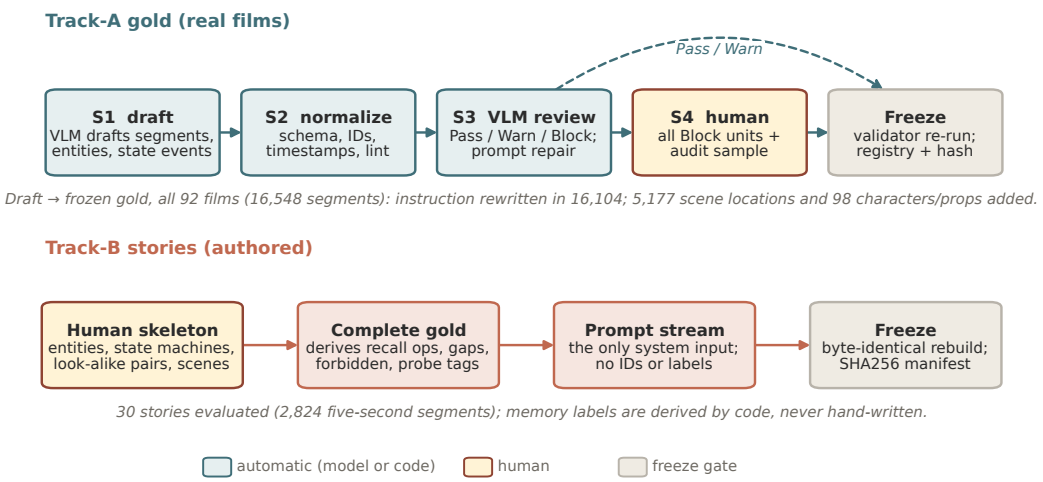


Figure 6: **How the two tracks are built and frozen.** Track-A gold starts from a VLM draft of each real film; deterministic normalization and a VLM review route every blocked unit, plus a stratified sample of the rest, to blocking human review before the full validator re-runs and the gold is frozen. Track-B stories are authored: a human writes only the semantic skeleton, code derives every memory label and the prompt stream, and a story is frozen only if both rebuild byte-identically from the skeleton.

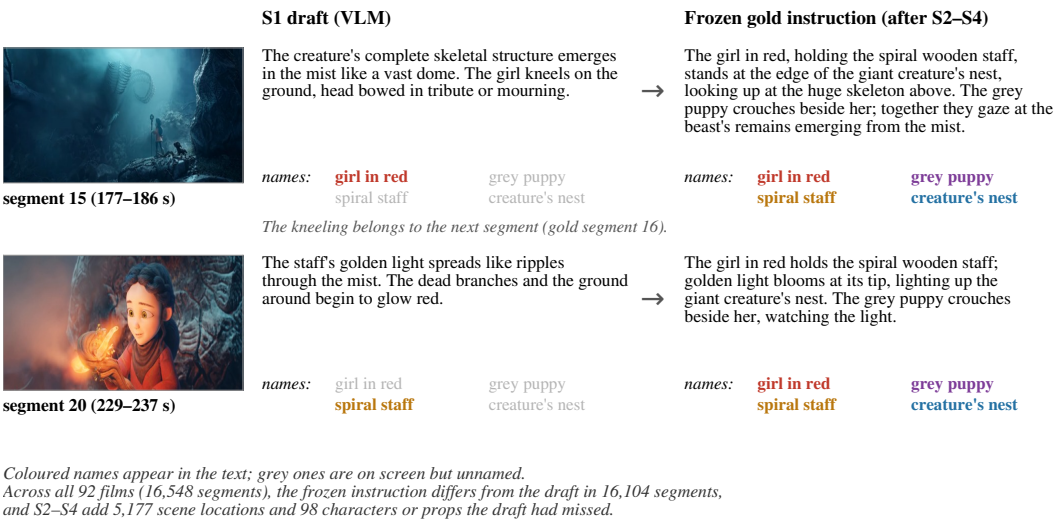


Figure 7: **From S1 draft to frozen gold.** In two *Spring* segments, the VLM draft describes the action but leaves entities that are on screen unnamed, and the first draft also borrows an action from the next segment. The frozen instruction names every present entity by its roster name, so a memory can be asked for each of them. The footer counts the same kind of change over the whole corpus.

### A.1.3 Benchmark Scoring Details

**Track-A metrics.** Precision, Recall, F1, SelEff, $R_V$, and $R_S$ are defined in Sec. 3.3. The judge scores each reference independently against the frozen roster and the real next segment. Redundancy then uses multimembership groups, so a reference that covers several gold entities contributes to each of those groups rather than to an arbitrary first label. $R_V$ is repeated-view redundancy inside those groups, and $R_S$ is the pair-weighted mean DINOv3 cosine among crops of the same entity. Headline Precision, Recall, and F1 are not rewritten by that grouping. Dataset numbers are unweighted means over a film's segments, then over films.

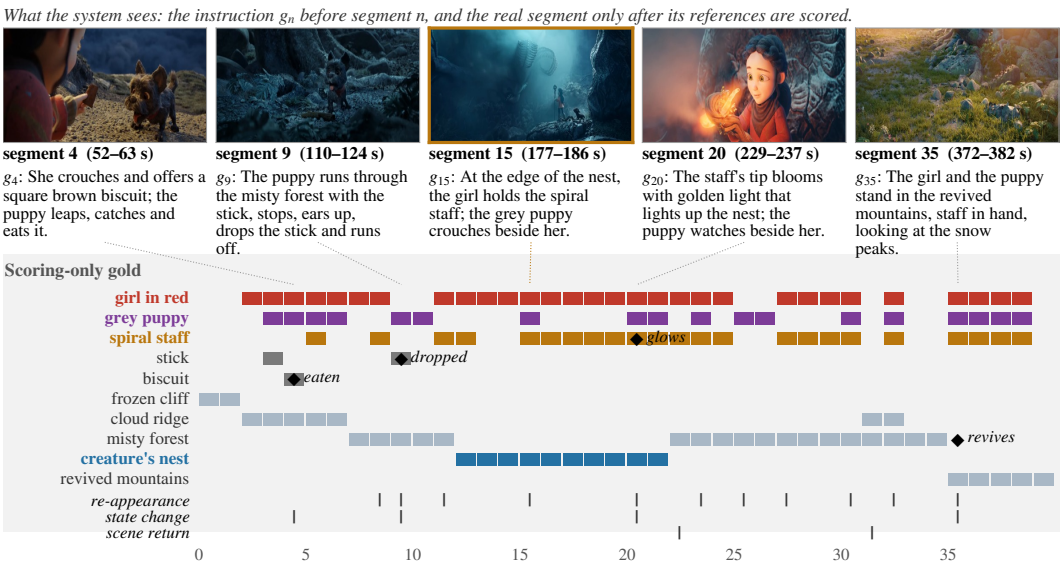


Figure 8: **Anatomy of a Track-A film** (*Spring*, Blender Studio, CC BY 4.0). Top: five story segments with the instruction $g_n$ the system receives (English excerpts of the Chinese gold). Bottom, on grey: scoring-only gold for ten of the film's eleven entities over segments 0–39 (the rest are titles and credits): where each entity is on screen, its state events, and the per-segment diagnostic tags. The grey puppy leaves after segment 10 and returns in segment 15, the case shown in Fig. 11.

**Causal guard and budget sweeps.** Every reference carries source time and is materialized at that recalled time. Any item at or after the current segment start is dropped. Fig. 12 truncates each system's own ranked output at increasing caps $K = 1, \ldots, 16$; it is a diagnostic budget sweep, not per-$K$ retuning.

**Track-B metrics.** Each family $f$ defines an opportunity set $\mathcal{O}_f$ and predicate $\mathrm{sat}_f(o)$ with $p_f = |\mathcal{O}_f|^{-1} \sum_{o \in \mathcal{O}_f} \mathrm{sat}_f(o)$. Must-remember families use recall, identity/state/count families use accuracy, and must-not-render families use avoidance-OK.

Family definitions and native metrics are in Table 1.

Track-B uses one deterministic judge call per segment ($k_{\text{votes}}$=1, temperature 0). The blinded roster mixes present, forbidden/deprecated, look-alike, and deterministic decoy entities. Decoy FPR / abstention for MemFlow, IAMFlow, and LongLive-RAG are $0.050/0.057$, $0.054/0.049$, and $0.016/0.132$, respectively; the last pair illustrates why precision-like metrics must be read with coverage. When the judge returns no parsable answer for a segment, every entity on its list counts as abstained: required entities are missed and forbidden ones absent. This happens for 87, 82, 99, and 112 of the 2,824 segments of Text-frame, Seg+uniform, Seg+DINO-key, and Seg+frame, and for 58 of the no-memory control (2–4%), so it slightly lowers the retrieval memories' recall relative to the control. Fig. 10 shows judge verdicts next to human labels on real units, including the two kinds of miss that make Track-B recall conservative.

**Uncertainty of memory effects.** Every difference between a retrieval memory and the no-memory control is given with a paired, story-clustered bootstrap: the 30 stories are resampled with replacement 2,000 times (seed 20260925), the same draw is applied to both conditions, and counts are pooled within the draw before taking the rate. The 95% interval is the 2.5th–97.5th percentile of the difference; all values are in `tables/experiments/memory_gain_bootstrap.json`. Each condition is one closed-loop rollout per story, so these intervals capture story-level variation, not rollout-to-rollout variation.

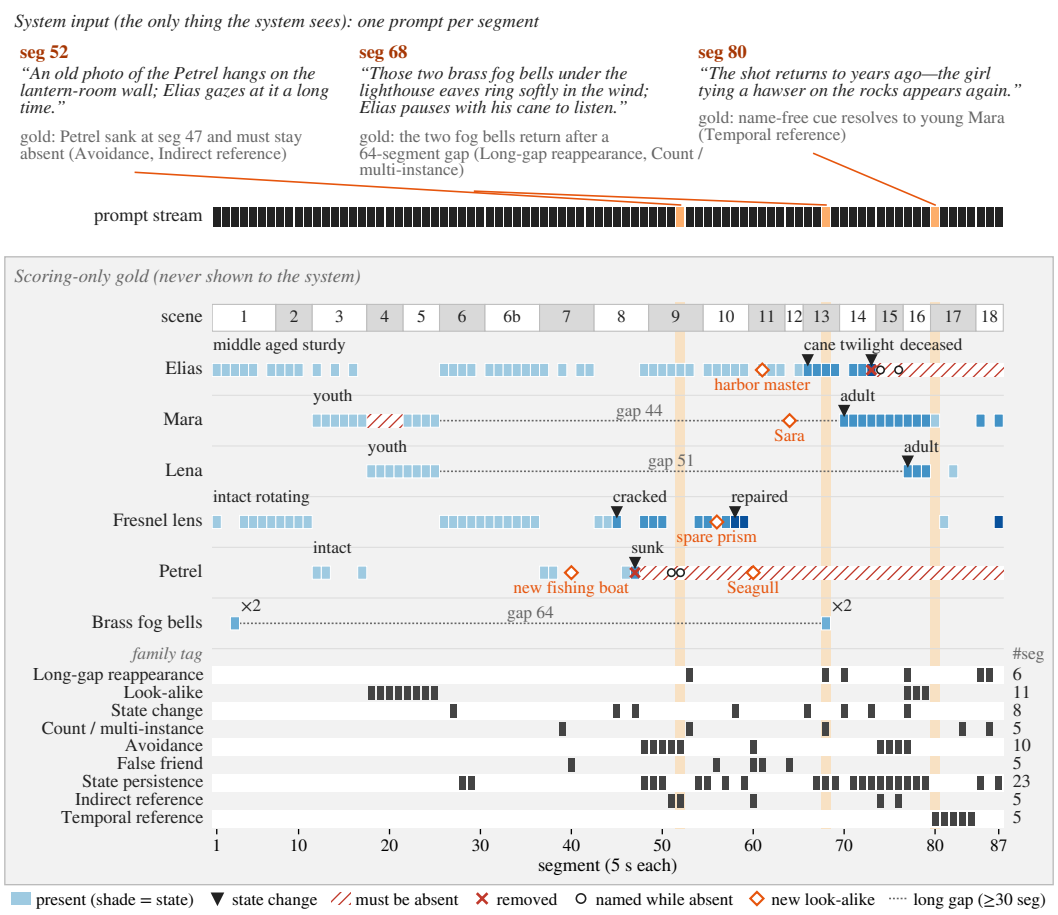


Figure 9: **How a Track-B script plants memory stresses.** Story `0001_lighthouse_keeper` (87 five-second segments, 19 scenes) drawn as a timeline. The system receives only the prompt stream (top; three prompts quoted verbatim); everything in the grey panel is scoring-only gold. Each lane is one of the story's 19 entities: shade gives its authored state, hatching marks segments where it must be absent (after it is destroyed or dies, or while only its look-alike twin is on screen), open circles mark prompts that still name a removed entity, and diamonds mark newly introduced look-alikes; the bottom rows tick the segments tagged with each family of Table 1, with per-family counts at right.

### A.1.4 Track-B Identity Consistency

The presence judge checks each entity against its gold description only, so it cannot tell a returning character from a description-matching stranger. Two versioned scorers in VMem-Bench measure identity directly; the full protocol, prompts, thresholds, and model installer are in the benchmark's scoring documentation.

**Units.** From the frozen gold, one unit per return of a character or prop that was absent in the previous segment but seen before (1,130 units), plus a seeded 20% sample of returns from the previous segment (583) and the characters that return changed after 30 or more segments (61). The previous appearance is the latest earlier segment whose cast lists the entity; both appearances come from the same rollout of the same condition.

**VLM pairwise judge (`trackB-identity-1.0`).** For each appearance, three frames (15, 50, and 85% of the clip) side by side. The judge receives both strips, the entity's name, kind, and gold description at each appearance (with a note when a state change is expected), is told to ignore camera, pose, lighting, and background, and answers in JSON whether the entity is visible in each strip and, if in both, whether it is the same individual (face, hair, build, clothing; shape, color, material,

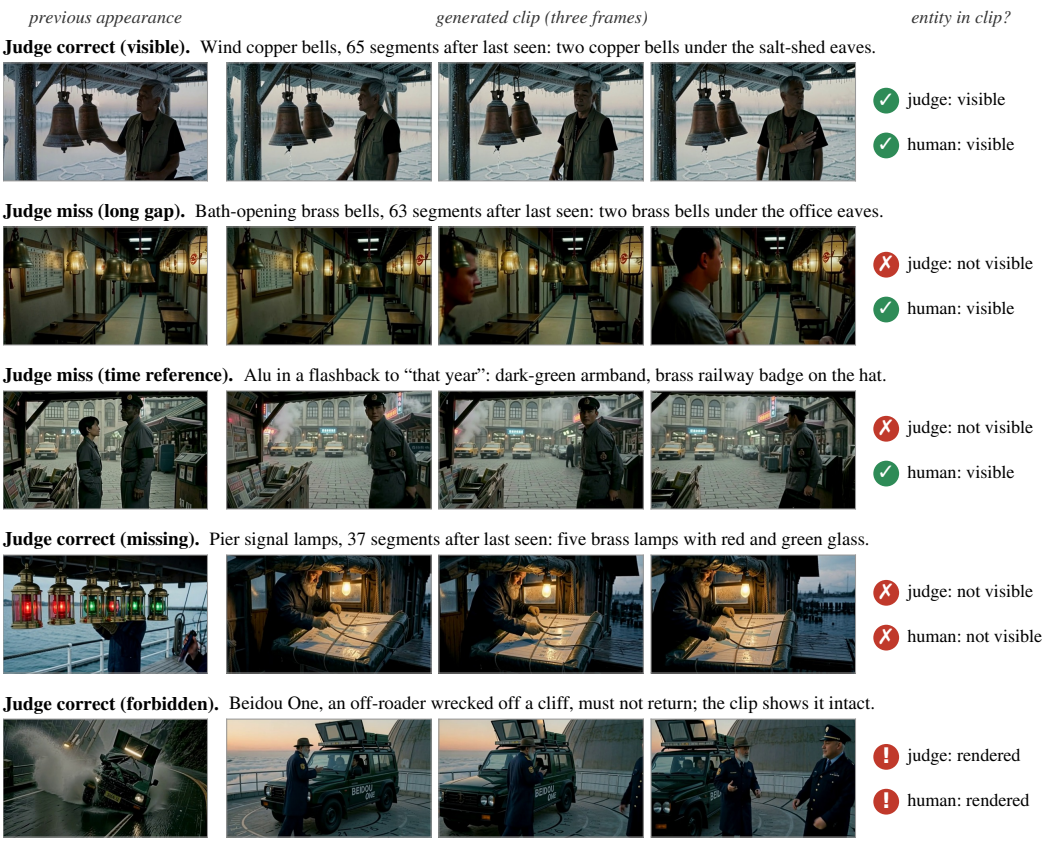


Figure 10: **Track-B judge verdicts checked against human labels on real units.** Each row shows the entity's previous appearance in the same rollout, three frames of the generated clip, the verdict of the pinned Qwen3-VL-32B judge, and the human label for the clip. In rows 1, 4 and 5 the judge is right about a required entity being present or absent and about a forbidden entity being rendered; 28 of 32 judge-confirmed controls were confirmed. In rows 2 and 3 the judge reports a required entity as missing although it is clearly rendered; such misses make up 27% of sampled required-entity failures and concentrate at long gaps and time references, so Track-B recall there is conservative.

markings for props). The condition is hidden. The scorer also reports the share of units where the entity is visible in both clips and judged the same, but this rate depends on how well the *previous* appearance was rendered, which memory affects through the earlier segments of the rollout: for characters returning changed after 30 or more segments, the current clip shows the character in almost every condition (0.98–1.00) and all visible pairs are called the same, and the whole difference between conditions (0.11 vs. 0.44–0.56) comes from the old appearance. We therefore report the same-individual rate among returns visible in both clips, which compares the two appearances only.

**Embeddings (`trackB-embed-1.0`).** GroundingDINO localizes the entity in each frame from its name and description; the crop is embedded with DINOv3 and, for characters, the largest face in the crop with ArcFace. Locations use the whole frame through MegaLoc, a place-recognition model, with whole-frame DINOv3 as a secondary check; whole-frame similarity is not used for characters or props, since a memory that copies a reference's composition would score high without the entity being the same. The similarity of a unit is the highest cosine over frame pairs detected on both sides; we report its mean over units detected on both sides.

**Calibration and uncertainty.** Positive controls pair two disjoint frame sets of one clip; negative controls pair two different entities of the same kind and story (Table 6). The embeddings separate them almost perfectly, the VLM judge less so: it rarely calls two visible look-alikes different, so its same-individual rate is optimistic. DINOv3 crop similarity is therefore the primary identity measure

Table 6: **Calibration of the identity-consistency metrics.** Positive controls pair two disjoint frame sets of the same clip; negative controls pair two different entities of the same kind and story. For the embeddings, the threshold is the Youden-optimal cut on these controls, fixed before any condition was scored. The VLM judge answers the same question twice with 0.993–1.000 exact agreement.

| Metric | pos./neg. pairs | AUC | threshold | pos. median | neg. median |
|---|---|---|---|---|---|
| DINOv3 crop | 385/389 | 1.000 | 0.67 | 0.96 | 0.19 |
| ArcFace face | 201/157 | 0.997 | 0.35 | 0.84 | 0.11 |
| MegaLoc scene | 200/200 | 1.000 | 0.66 | 0.89 | 0.07 |
| DINOv3 scene | 200/200 | 1.000 | 0.84 | 0.98 | 0.17 |
| VLM pairwise judge | same when both visible: 0.98 (pos.), 0.78 (neg.); neg. not consistent: 0.93 | | | | |

Table 7: **Track-B identity consistency.** No-memory value and each condition minus no memory (bold: paired story-bootstrap 95% CI excludes zero). Embeddings: mean cosine similarity between the previous and current appearance (DINOv3 on the localized entity; MegaLoc on the whole frame for locations), over returns detected on both sides. VLM: share judged the same individual among returns visible in both clips (lenient, Table 6). Returns are character and prop re-appearances after at least one segment away; returns changed are characters whose look changed after 30 or more segments, for which a higher similarity means the return looks more like its old look.

| Stratum | $n$ | None | Text | Uniform | DINO-key | Frame | Last-seen |
|---|---|---|---|---|---|---|---|
| *DINOv3 crop similarity (trackB-embed-1.0)* | | | | | | | |
| All returns | 1130 | 0.34 | **+0.14** | **+0.19** | **+0.20** | **+0.20** | **+0.22** |
| Characters | 486 | 0.27 | **+0.11** | **+0.17** | **+0.19** | **+0.17** | **+0.24** |
| Props | 644 | 0.41 | **+0.16** | **+0.20** | **+0.21** | **+0.22** | **+0.22** |
| Gap 1–4 | 688 | 0.30 | **+0.14** | **+0.19** | **+0.21** | **+0.20** | **+0.26** |
| Gap 5–29 | 293 | 0.35 | **+0.17** | **+0.16** | **+0.19** | **+0.21** | **+0.16** |
| Gap $\geq$30 | 149 | 0.52 | **+0.11** | **+0.21** | **+0.17** | **+0.17** | **+0.18** |
| Returns changed, $\geq$30 | 61 | 0.20 | **+0.17** | **+0.19** | **+0.18** | **+0.17** | **+0.24** |
| *VLM pairwise judge: same individual when visible in both (trackB-identity-1.0)* | | | | | | | |
| All returns | 1130 | 0.85 | **+0.09** | **+0.11** | **+0.12** | **+0.12** | **+0.11** |
| Characters | 486 | 0.77 | **+0.12** | **+0.16** | **+0.18** | **+0.19** | **+0.15** |
| Props | 644 | 0.88 | **+0.09** | **+0.09** | **+0.10** | **+0.10** | **+0.10** |
| Gap 1–4 | 688 | 0.84 | **+0.09** | **+0.10** | **+0.12** | **+0.14** | **+0.11** |
| Gap 5–29 | 293 | 0.87 | **+0.09** | **+0.11** | **+0.11** | **+0.09** | **+0.10** |
| Gap $\geq$30 | 149 | 0.86 | **+0.10** | **+0.13** | **+0.13** | **+0.14** | **+0.11** |
| *ArcFace face similarity, characters* | | | | | | | |
| Characters | 486 | 0.13 | +0.02 | **+0.06** | **+0.06** | **+0.04** | **+0.09** |
| Returns changed, $\geq$30 | 61 | 0.06 | **+0.04** | **+0.07** | **+0.07** | **+0.07** | **+0.06** |
| *MegaLoc scene similarity, locations* | | | | | | | |
| Locations | 425 | 0.11 | **+0.14** | **+0.16** | **+0.17** | **+0.19** | **+0.09** |

and the VLM judge a check. Three frames can miss a brief appearance. Differences to no memory use the same paired story bootstrap as all Track-B results (Table 7).

### A.2 Additional Track-A Results

**Repetition.** The repetition columns of Table 4 separate two kinds of redundancy. Seg+DINO-key has the best selection efficiency ($0.431/0.393$), the lowest appearance similarity $R_S$ ($0.488/0.393$), and near-lowest view repetition $R_V$ ($0.225/0.143$; MemFlow $0.133$ on LSMDC), yet it is the least precise retrieval memory: diverse views are not the same as correct views. Fig. 11 shows both effects on one segment.

**Budget.** Fig. 12 lets each memory hand over more of its own top-ranked references: F1 rises with shrinking returns, and the ordering of the memories holds at every budget.

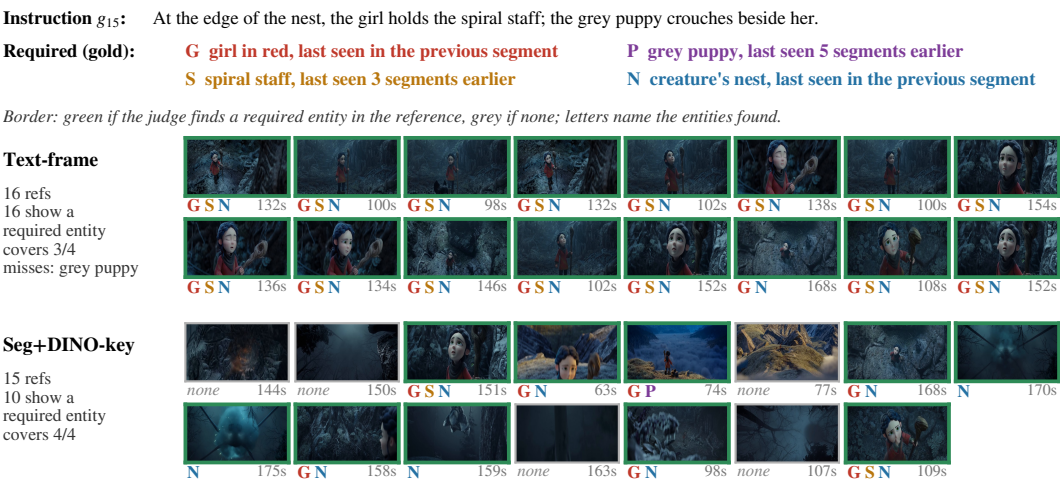


Figure 11: **Every reference two memories hand over for *Spring* segment 15.** Text-frame returns 16 near-identical views of the girl in the forest and never the puppy, last seen five segments earlier. Seg+DINO-key samples diverse views, finds the puppy on the cloud ridge (74 s), and covers all four entities, but five of its 15 references show none of them. Letters are the automatic judge's labels; it also tags some misty-forest frames before 144 s as the nest, which does not change which references count as correct here.

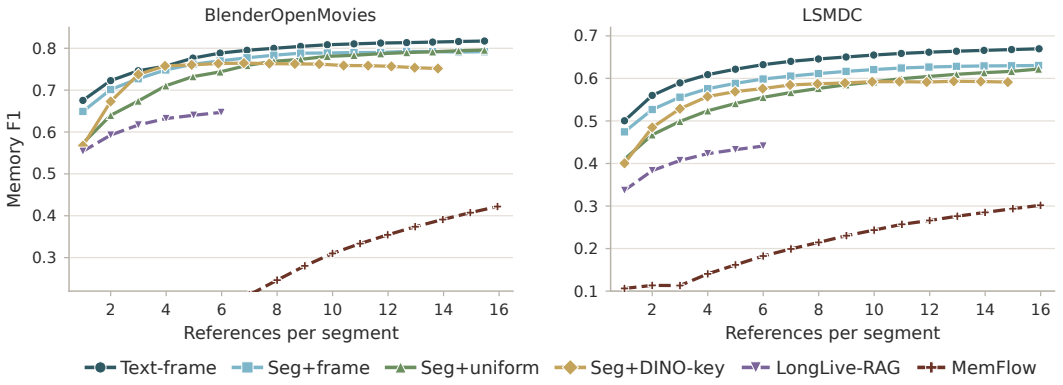


Figure 12: **Track-A: quality as a memory may hand over more references.** Each curve keeps a memory's own top-ranked references up to a growing cap. F1 keeps rising with the cap but with shrinking returns, and the ordering holds at every budget: Text-frame leads throughout, and the released memories stay far below. Left: BlenderOpenMovies; right: LSMDC.

**History length.** Fig. 13 tracks F1 as the archived film grows: retrieval stays nearly flat, while the released memories keep losing; since they evict or compress entries while the testbed re-ranks the full history, this decline mixes lossy storage with reading.

### A.3 Additional Track-B Results

**Recall against the gap.** Fig. 14 shows why a recall-versus-gap curve misleads: every condition, including no memory, recalls more at long gaps, because a returning entity is named in the instruction and, after a long absence, usually described (Table 9); only the difference to no memory isolates memory.

**Per-stress effects.** Fig. 15 maps every memory stress for no memory, last-seen references, and the four retrieval memories, and repeats the human attribution by stratum.

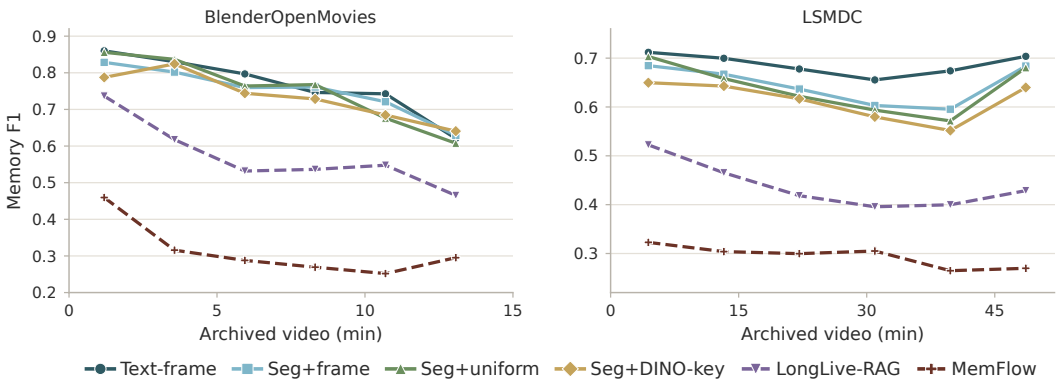


Figure 13: **Track-A memory F1 versus archived-video length** (same runs as Table 4). Retrieval remains comparatively stable as history grows; the released causal-memory paths lose more F1. The far bins contain few films. Left: BlenderOpenMovies; right: LSMDC.

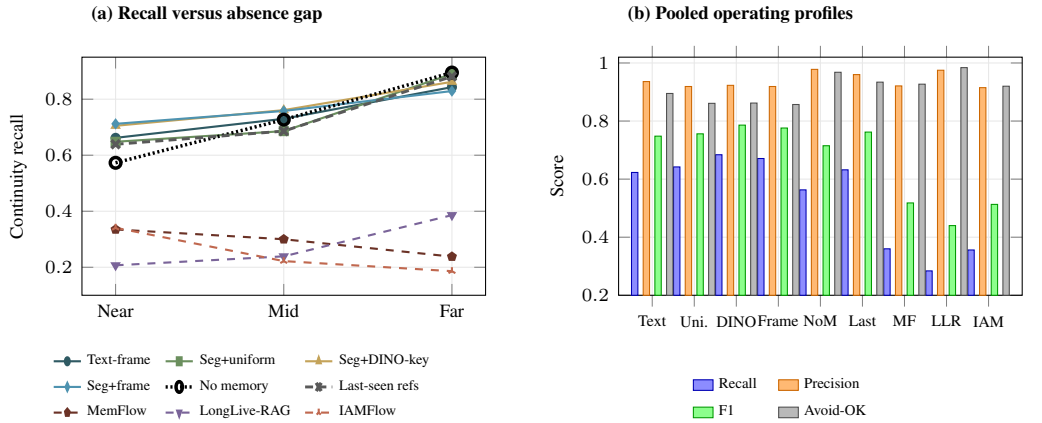


Figure 14: **Track-B generated-video scores** (same 30-story pool as Table 3). **(a)** Continuity recall by absence-gap bin; the dotted line is the same generator with no memory and the grey dashed line with last-seen references. Support is $N$=701/293/210 for near/mid/far. **(b)** Pooled Recall, Precision, F1, and Avoid-OK.

**Robustness.** Table 8 repeats the comparisons with no memory on segments whose judge output parsed in both arms, since an unparsable output counts every required entity as missed and retrieval produces more of them (82–112 segments against 58 for no memory and 71 for last-seen references). All main effects stay: gains at 1–4 segments and for persistent states, no presence gain at 30 or more segments, and lower precision and avoidance. Between 5 and 29 segments only Seg+frame gains significantly. The second part splits the last-seen control by how many references it used: both costs grow with the count, so part of retrieval's larger cost comes from handing over three references, but at two references last-seen still costs about half as much as retrieval.

**Entity kind and gap.** Table 9 splits recall by entity kind. The gap bins mix kinds: returns within four segments are mostly characters named only by name, while returns after 30 or more segments are mostly props whose names describe them and characters coming back changed with the new look in the instruction. Memory's gain is concentrated in the first group; in the second, references add nothing for props and hurt for changed characters.

Fig. 16 shows what the Track-B numbers look like in pixels: on the same prompt stream, retrieval restores a character seen one segment earlier that the no-memory generator replaces with a stranger, and it also returns a dead character whom the prompt names only through his plaque.

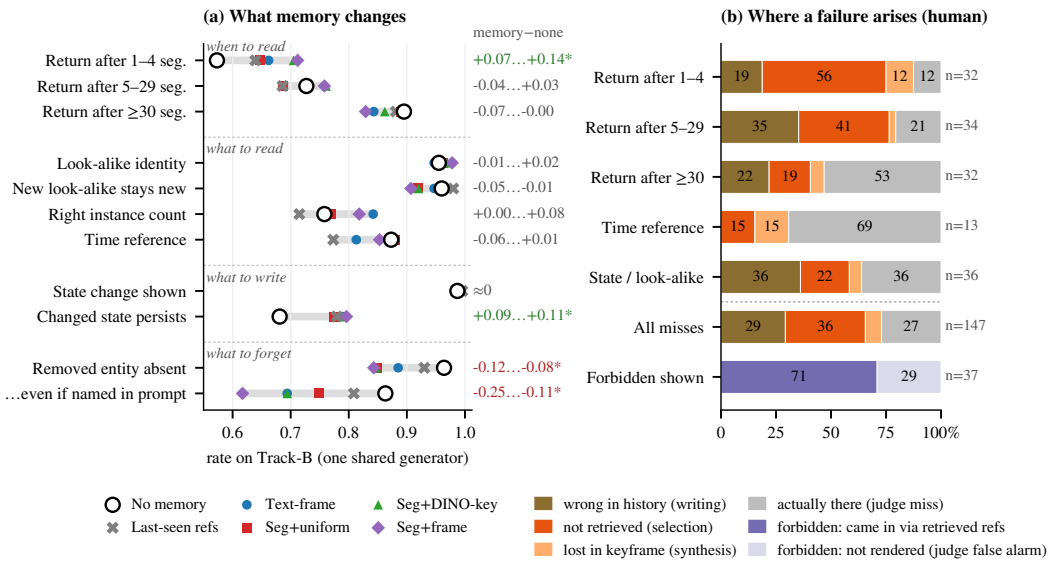


Figure 15: **What memory changes, and where failures arise. (a)** Track-B rate per memory stress for the shared generator with no memory (open circle) and with each retrieval memory (filled), with last-seen references as grey crosses, grouped by the decision each stress tests (Sec. 4); the gap rows mix entity kinds (Sec. 4.2). A marker right of the circle means memory helps; left means it hurts; the right column gives memory minus no memory (* when all four paired 95% CIs exclude zero). Memory helps for recent returns and persistent states and hurts where an entity must stay out. **(b)** Human attribution of Track-B failures (Sec. 4.5), by stratum: the share already wrong in the entity's previous appearance, present there but not retrieved, retrieved but lost in the keyframe, or actually rendered and missed by the judge. The last row splits judge-flagged forbidden renders into those carried in by the references (purple) and judge false alarms (light purple). The two aggregate rows are inclusion-weighted.

## A.4 Systems and Adapter Fidelity

### A.4.1 Adapter Principles

Released causal systems are driven through authors' released memory mechanisms and weights with only protocol-required substitutions; the four retrieval memories are ours and form the controlled testbed. All scored rows share frozen gold, causal guard, materialization, and scoring. No gold crop or gold identity is exposed to a system. On Track-A, self-generated observation is replaced by the revealed real segment; if a native write requires a model forward, we use one teacher-forced clean-context forward without denoising. Recalled items are mapped back to source frames through temporal identities and dropped if they violate the causal guard. Methods that require future story information or modalities absent from passive film are not assigned surrogate quantitative scores. Table 10 summarizes why the families' internal diagnostics are not comparable and which common observables VMem-Bench uses.

Table 11 lists, for each memory we cite or test, what it stores, how it decides what to read, and whether and how it forgets.

### A.4.2 Released Causal-Memory Systems

**LongLive-RAG.** Each revealed Track-A segment is VAE-encoded to Wan latents (Wan Team, 2025) with the released LongLive-RAG (Hu et al., 2026b) wrapper and published LatentAE. The released latent-memory rule performs cosine top-$k$ retrieval with `memory_size=6`, `recent_exclude=5`, and `sink_size=1`. Track-B uses the released generator-coupled system.

**MemFlow.** MemFlow (Ji et al., 2025) combines a recent window with a compressed long-term KV bank. We teacher-force each revealed Track-A segment through the released base+LoRA clean-

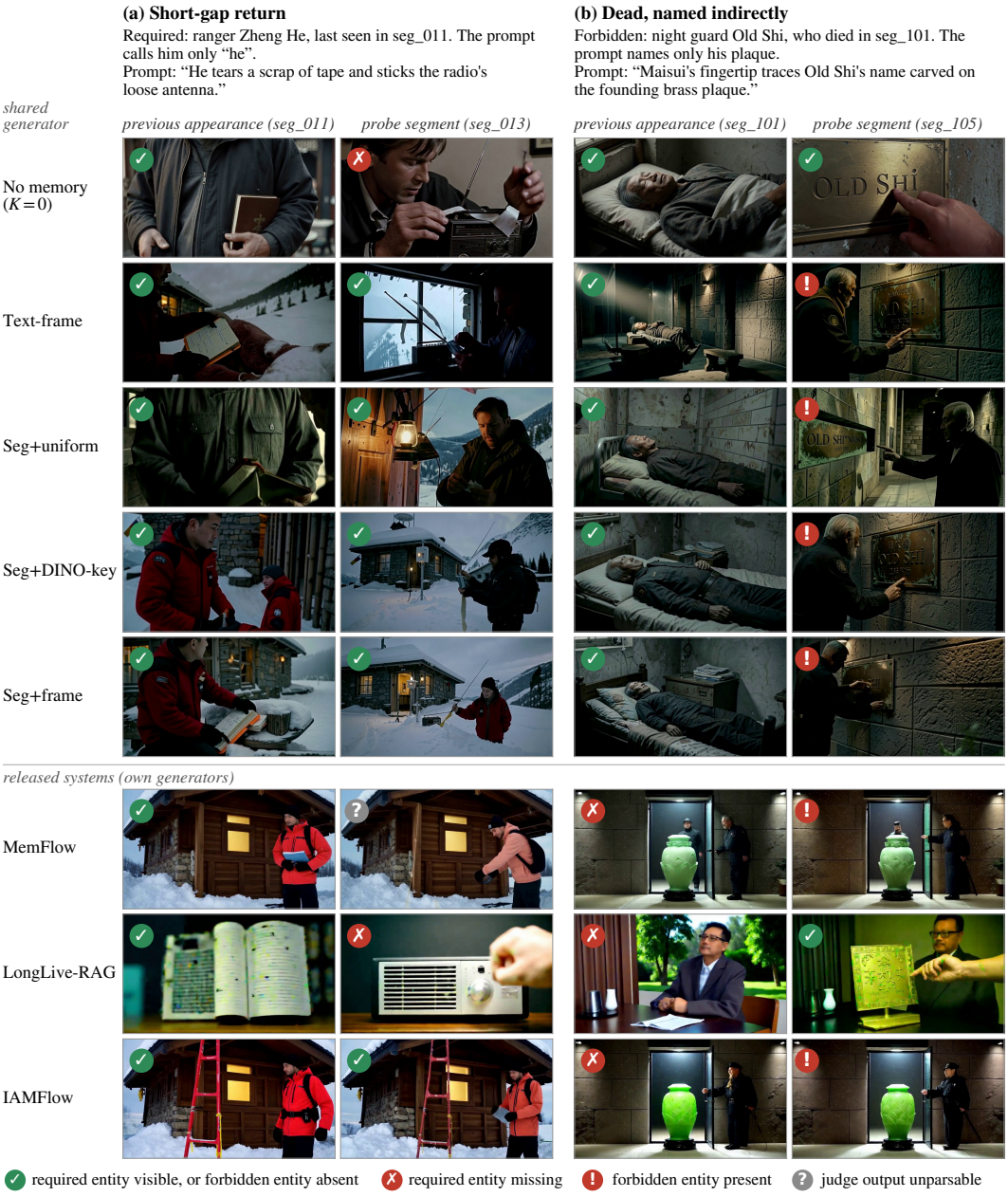


Figure 16: **Track-B rollouts on two memory probes: memory brings back a recently seen character, and also a dead one.** Rows are systems given the same prompt stream; columns show the entity's previous appearance and the probe segment (middle frame of each generated clip). In (a) the no-memory generator draws a stranger for "he" while all four retrieval memories restore the ranger; in (b) no memory keeps the dead guard out, but all four retrieval memories, MemFlow, and IAMFlow put him back at his plaque. Badges are the automatic judge's verdicts for that entity, and Track-B Recall and Avoid-OK pool such verdicts over all 2,824 segments; the released systems use their own generators.

Table 8: **Robustness of the Track-B memory effects.** Differences to no memory with paired story-bootstrap 95% CIs; bold when the CI excludes zero. (a) Dropping every segment on which either arm's judge output failed to parse leaves all main effects in place; only Seg+frame gains significantly at 5–29 segments. (b) The precision and avoidance costs of last-seen references grow with the number of references.

*(a) Minus no memory, segments with a parsable judge output in both arms*

| Condition | dropped | Recall | Precision | Avoid-OK | 1–4 | 5–29 | ≥30 |
|---|---|---|---|---|---|---|---|
| Text-frame | 123 | **+0.065** [+0.05, +0.09] | **-0.043** [-0.05, -0.03] | **-0.075** [-0.09, -0.06] | **+0.097** [+0.05, +0.14] | +0.000 [-0.04, +0.04] | -0.043 [-0.09, +0.01] |
| Seg+uniform | 119 | **+0.087** [+0.06, +0.11] | **-0.061** [-0.07, -0.05] | **-0.111** [-0.13, -0.09] | **+0.084** [+0.03, +0.13] | -0.022 [-0.07, +0.03] | +0.005 [-0.07, +0.08] |
| Seg+DINO-key | 140 | **+0.132** [+0.10, +0.16] | **-0.056** [-0.07, -0.04] | **-0.110** [-0.13, -0.09] | **+0.144** [+0.09, +0.20] | +0.036 [-0.01, +0.08] | -0.033 [-0.10, +0.03] |
| Seg+frame | 145 | **+0.123** [+0.10, +0.15] | **-0.060** [-0.07, -0.05] | **-0.115** [-0.13, -0.10] | **+0.154** [+0.11, +0.20] | **+0.068** [+0.03, +0.11] | -0.053 [-0.10, +0.01] |
| Last-seen refs | 106 | **+0.075** [+0.06, +0.09] | **-0.018** [-0.03, -0.01] | **-0.034** [-0.05, -0.02] | **+0.066** [+0.01, +0.12] | -0.026 [-0.07, +0.02] | -0.005 [-0.05, +0.04] |

*(b) Last-seen refs minus no memory, by number of references used*

| References | segments | Recall | Precision | Avoid-OK |
|---|---|---|---|---|
| 1 | 1803 | **+0.072** [+0.06, +0.09] | **-0.015** [-0.03, -0.00] | **-0.028** [-0.04, -0.01] |
| 2 | 842 | **+0.072** [+0.04, +0.10] | **-0.026** [-0.04, -0.01] | **-0.044** [-0.06, -0.02] |
| 3 | 149 | +0.017 [-0.05, +0.09] | -0.024 [-0.06, +0.02] | -0.039 [-0.09, +0.02] |

context commit path, record the local window and surviving bank blocks, and recover source frames via bit-exact fingerprints of retained KV blocks. The temporal RoPE table is extended to the film length with the same closed-form rotary embedding. Track-B uses the released generator-coupled path.

**IAMFlow.** IAMFlow (Liu et al., 2026) scores frames using the native diffusion transformer's KV cache. Its offline driver precomputes entities from the full prompt list; our causal adapter disables this future look-ahead and uses native per-prompt processing. Track-B uses the released generator-coupled stack. Track-A is reported only as a scalability diagnostic because the native per-frame KV archive exceeds host memory on long films.

### A.4.3 The Four Retrieval Memories

**Track-A retrieval definitions.** The adapter stores observed segment text and sampled video frames with temporal identities. **Text-frame** ranks historical frames by cross-modal similarity. **Seg+uniform** ranks segments by text similarity and uniformly samples frames. **Seg+DINO-key** uses the same segment ranking, 2-fps dense sampling, and greedy farthest-point DINOv3 diversity. **Seg+frame reranking** performs text coarse retrieval followed by frame-level cross-modal reranking with reciprocal rank fusion. Track-A reports their native cap $B{=}16$ and the budget sweep in Fig. 12.

**Audited Track-B generation loop.** Text-frame, Seg+uniform, and Seg+DINO-key use one frozen set of rollouts; Seg+frame uses a later rerun of the same retrieval and generation procedure. Each story starts with an empty retrieval store and an empty entity bank. At segment $n$, the authored prompt $g_n$ is used directly as the retrieval query. The history record for each prior segment stores

Table 9: **Track-B recall by entity kind and absence gap.** No-memory recall and each condition minus no memory (paired story-bootstrap 95% CI; bold when it excludes zero). Just seen means the entity was on screen in the previous segment. Characters are named only by name after their first appearance; 61 of the 62 character returns after 30 or more segments come back changed, with the new look in the instruction. Props often carry their look in the name. Text, Uniform, DINO-key, and Frame are the four retrieval memories.

| Kind | Gap (segments) | $n$ | None | Text | Uniform | DINO-key | Frame | Last-seen |
|---|---|---|---|---|---|---|---|---|
| Character | just seen | 2173 | 0.500 | **+0.101** [+0.07, +0.14] | **+0.147** [+0.11, +0.18] | **+0.192** [+0.16, +0.23] | **+0.172** [+0.13, +0.22] | **+0.125** [+0.10, +0.15] |
| | 1–4 | 430 | 0.544 | **+0.137** [+0.09, +0.19] | **+0.119** [+0.05, +0.18] | **+0.177** [+0.11, +0.25] | **+0.186** [+0.13, +0.24] | +0.051 [-0.01, +0.11] |
| | 5–29 | 63 | 0.746 | +0.000 [-0.11, +0.12] | -0.016 [-0.17, +0.12] | **+0.095** [+0.01, +0.19] | +0.064 [-0.05, +0.17] | -0.079 [-0.20, +0.05] |
| | ≥30 (returns changed) | 62 | 0.968 | **-0.097** [-0.20, -0.02] | -0.016 [-0.10, +0.05] | -0.065 [-0.16, +0.02] | **-0.129** [-0.25, -0.03] | -0.065 [-0.15, +0.00] |
| Prop | just seen | 1087 | 0.413 | **+0.034** [+0.01, +0.06] | +0.040 [+0.00, +0.08] | **+0.093** [+0.05, +0.13] | **+0.079** [+0.04, +0.11] | **+0.035** [+0.01, +0.06] |
| | 1–4 | 271 | 0.620 | +0.011 [-0.04, +0.06] | +0.004 [-0.05, +0.06] | **+0.059** [+0.00, +0.11] | **+0.063** [+0.01, +0.12] | **+0.089** [+0.03, +0.15] |
| | 5–29 | 230 | 0.722 | +0.004 [-0.05, +0.05] | -0.048 [-0.10, +0.00] | +0.017 [-0.03, +0.07] | +0.022 [-0.02, +0.07] | -0.030 [-0.08, +0.01] |
| | ≥30 | 148 | 0.865 | -0.034 [-0.10, +0.04] | +0.000 [-0.11, +0.11] | -0.020 [-0.11, +0.06] | -0.041 [-0.12, +0.05] | +0.007 [-0.05, +0.06] |

Table 10: Different internal memories, one common interface. The families expose different internal states and native diagnostics, so those quantities are not directly comparable and do not by themselves establish what reaches or survives generation. VMem-Bench maps them to common observables: boundary context $\widehat{\mathcal{C}}_n$ before synthesis (Track-A) and generated outcome $x_n$ after synthesis (Track-B).

| **Memory family** | **Internal interface / native diagnostic** | **What internal-only evaluation cannot establish** |
|---|---|---|
| Window / KV | frames, latents, or KV; recency / native attention; cache reach | Retention does not imply that intent-relevant evidence crosses the finite conditioning boundary; stale history may remain active. |
| Retrieval | indexed frames / clips; search + rank; Recall@$K$ / similarity | Larger $K$ can buy recall while increasing context spend or contamination; a good retrieval decision may still fail in synthesis. |
| Entity / state memory | persistent identity / state records; address, update, filter; record validity | Correct records do not guarantee that the right visual evidence is materialized under budget or realized correctly by the generator. |

that same authored prompt as both prompt text and description, together with frames decoded from that memory's own previously generated MP4; no real-film frame, gold crop, or other memory's trajectory enters the archive. Retrieval happens before write-back, with eligibility requiring the historical segment end time to be no later than the store clock. The resulting still images are then passed as composed references to a shared keyframe-first stack: FLUX.2 Klein 9B-KV FP8 composes either a text-only or multi-reference keyframe, and Wan2.2-TI2V-5B-Turbo expands it to a 121-frame segment (1280×704 render, delivered at 832×480). The generated segment is then appended to that memory's own store and becomes available to later segments. The retrieval reference budget is $K$=3; the audited code defaults are 1-fps uniform sampling, 2-fps dense sampling, top-5 text-ranked segments, three DINO keyframes per segment, and reciprocal-rank fusion with $k$=60 for Seg+frame

Table 11: **How causal video memories read and forget.** Every read matches the current instruction, latest frame, camera pose, or attention query by similarity or relevance, and forgetting is either absent, a budget eviction ranked by that same relevance, or age-based compression; no causal system evicts or withholds an entry because the story changed its state. Only IAMFlow and the non-causal VideoMemory keep explicit entity records, and both address them by the entities named in the current instruction or shot.

| System | Stored as | Read by | Forgets by | Causal |
|---|---|---|---|---|
| *Released / published memories* | | | | |
| FramePack (Zhang et al., 2025) | Frame latents, compressed by importance rank | All history, ordered by recency or similarity | Geometric compression; tail deleted or pooled | Partly (end-point plans) |
| MemFlow (Ji et al., 2025) | Local KV window + KV bank | Prompt-to-frame attention, top-$k$ | Budget; keeps top-$k$ by prompt relevance | Yes |
| IAMFlow (Liu et al., 2026) | Entity-ID registry + archived frame KV | Set cover of IDs named in prompt | None; archive grows, read capped | Yes |
| LongLive-RAG (Hu et al., 2026b) | Evicted latent KV + AE embedding | Cosine to latest latent, top-$K$ | None; every evicted frame kept | Yes |
| Context-as-Memory (Yu et al., 2025) | Raw generated frames | Camera FOV overlap with target view | None; read capped at $k$ | Yes |
| WorldKV (Yi et al., 2026) | Evicted KV chunks, pose-indexed | Camera/action pose distance, top-$k$ | No chunk eviction; redundant tokens pruned | Yes |
| OmniMem (Zhao et al., 2026) | Full-history KV blocks, GPU/CPU | Per-head attention top-$K$, local excluded | None; LRU only offloads to CPU | Yes |
| SlotMemory (Dou et al., 2026) | Slot-indexed KV tokens, fixed bank | Prompt-text + visual relevance, top-$k$ | Budget; evicts least prompt/visual relevance | Yes |
| VideoMemory (Zhou et al., 2026) | Entity ID, attribute text, reference image | LLM attribute match per planned shot | None; new states appended | No (full script) |
| *Our testbed probes (App. A.4.3)* | | | | |
| Text-frame | Segment text + timestamped frames | Instruction-to-frame cross-modal similarity | None; append-only, read capped | Yes |
| Seg+uniform | Segment text + timestamped frames | Segment text similarity; uniform frames | None; append-only, read capped | Yes |
| Seg+DINO-key | Segment text + timestamped frames | Segment text similarity; DINOv3 diverse frames | None; append-only, read capped | Yes |
| Seg+frame | Segment text + timestamped frames | Segment text, then frame rerank (RRF) | None; append-only, read capped | Yes |

reranking. The scorer is the same pinned Qwen3-VL-32B judge used for every Track-B row, applied to the frozen story gold.

**Realization-layer scope.** The reported retrieval rollouts are keyframe-first: selected stills are composed into a single FLUX scene keyframe, and Wan receives that image as its first-frame condition. This is useful for holding a simple realization mechanism fixed, but it introduces a structural bias: memory that should emerge only later in the segment may be pulled into the first frame. The codebase also contains Wan2.2-I2V-A14B backends with SVI/Morphic LoRA variants, but the reported Track-B retrieval rows did *not* use those paths; they used Wan2.2-TI2V-5B-Turbo. A unified multimodal reference-conditioned generator such as MiniMax H3 (MiniMax, 2026), which accepts text, image, video, and audio context directly, is a natural future replacement because it may avoid collapsing multiple memory cues into one forced first-frame keyframe. When we designed the Track-B realization setup, H3 had not yet been open-sourced; the frozen evaluation therefore used the FLUX-to-Wan scaffold. We have not rerun the benchmark with H3 and make no claim that it improves VMem-Bench scores.

Table 12: IAMFlow host-memory growth observed during our Track-A scalability test. The limiting quantity is the number of archived KV frames, which scales with total observed video duration rather than simply with the number of benchmark segments. The predicted GiB column is the archived-KV tensor payload from shape alone; OOM rows exceed the fixed host-memory budget before a full-movie scored output can be produced.

| Movie | Segs. | Mean seg. | KV frames | Pred. GiB | Outcome |
|---|---|---|---|---|---|
| `0053_Rendezvous_mit_Joe_Black` | 221 | 7.3 s | 2071 | 555 | completed |
| `sita_sings_the_blues_part1` | 164 | 14.9 s | 3259 | 873 | OOM |
| `0013_Halloween` | 128 | 14.9 s | 2442 | 654 | completed |
| `1005_Signs` | 296 | 9.8 s | 3763 | 1008 | OOM |
| `0017_Pianist` | 235 | 13.8 s | 4287 | 1148 | OOM |

**Provenance audit and clean Seg+frame rerun.** The legacy Seg+frame rollout recorded 1,687 retrieval CUDA OOM failures across all 30 stories, so we do not use its scores. We instead reran all 30 stories from empty history with retrieval and generation resources isolated, changing no retrieval or generation semantics; the clean run completed with zero retrieval OOMs/failures. The testbed's production files are internally 0-indexed whereas Track-B gold is 1-indexed, so scoring uses the same $+1$ filename staging as the original three audited retrieval rows (`GT seg_k` maps to production `seg_{k-1}`). The corrected run scores all 2,824/2,824 segments. The release pins those snapshots rather than a moving source tree.

### A.4.4 Last-Seen Reference Control

For Track-B segment $n$, the control reads the gold cast and keeps the characters and props that must appear and have appeared before. It orders them by gap (longest first), keeps up to three, and for each takes the middle frame of this rollout's own clip of the segment in which the entity last appeared; duplicates are merged. The generator receives only these frames, never gold text, names, or images, and a guard admits only clips that this run produced strictly before segment $n$. Because the needed entity was usually on screen just before, 1,803 segments receive one reference, 842 two, 149 three, and 30 (the first segment of each story) none. The control shows what a memory gains by choosing the right entity at its latest moment; it is not an information upper bound, since it hands over fewer references than retrieval and its frame may itself be a wrong earlier rendering. A storage outage interrupted 26 of the 30 rollouts; each was rolled back to its first affected shot and regenerated, and all 30 pass a per-shot audit that the references match the rule above. On the four untouched rollouts, every last-seen effect has the same sign as on the 26 repaired ones (recall $+0.03$ vs. $+0.07$, precision $-0.047$ vs. $-0.015$, avoidance of destroyed or dead entities $-0.075$ vs. $-0.029$, persistence $+0.17$ vs. $+0.09$, time references $-0.10$ in both).

### A.4.5 IAMFlow Track-A Scalability

IAMFlow's native long-term archive copies per-frame K/V slices to host memory. For its Wan2.1-T2V-1.3B backbone, one archived frame costs approximately $30 \times 2 \times 1560 \times 1536 \times 2$ bytes $\approx 274$ MiB. In our driver, 15-second segments add about 5.3 GiB of anonymous host memory per segment; with an 800–900 GiB host-memory budget, the native carrier reaches a hard horizon of roughly three thousand archived KV frames. We do not substitute a bounded-KV variant because it changes the memory carrier used for recall.

### A.4.6 Per-Clip Quality Control

VMem-Bench measures memory, not visual quality; this control asks whether per-clip quality could explain the memory differences within the shared-generator testbed. Every one of the 2,824 Track-B clips of each condition is scored with VBench 0.1.5's own models and formulas (Huang et al., 2023): imaging quality (MUSIQ), aesthetic quality (LAION head on CLIP ViT-L/14), subject consistency (DINO), background consistency (CLIP), dynamic degree (RAFT), and motion smoothness (AMT). Image metrics use eight fixed frames per clip; runs are deterministic. Table 13 reports means with story-clustered bootstrap 95% CIs and the mean per-clip entity recall of the same clips. Across the four retrieval memories the largest standardized difference in any quality metric is $0.19$, and no metric

correlates with per-clip recall above $\rho$=0.19 in any condition. The no-memory control has lower recall but *higher* subject and background consistency (standardized differences 0.18–0.43): these metrics reward within-clip stability, which a rollout that never recalls earlier evidence can achieve. Released systems are omitted because they use different generators.

Table 13: Per-clip VBench-style quality on the shared-generator testbed (all 2,824 clips per condition). Each cell gives the mean with its story-clustered bootstrap 95% CI below; Clip rec. is the mean per-clip entity recall of the frozen judge on the same clips. Best in each column is **bold**, second best underlined.

| Condition | Imaging↑ | Aesthetic↑ | Subj. cons.↑ | Bg. cons.↑ | Dynamic↑ | Motion sm.↑ | Clip rec.↑ |
|---|---|---|---|---|---|---|---|
| Text-frame | 0.634 [0.622, 0.647] | 0.587 [0.576, 0.596] | 0.926 [0.924, 0.929] | 0.931 [0.929, 0.933] | 0.488 [0.456, 0.520] | **0.9934** [0.9932, 0.9937] | 0.661 [0.631, 0.691] |
| Seg+uniform | 0.650 [0.640, 0.662] | 0.588 [0.576, 0.599] | 0.929 [0.927, 0.931] | 0.931 [0.930, 0.933] | 0.453 [0.426, 0.480] | 0.9930 [0.9927, 0.9932] | 0.680 [0.651, 0.711] |
| Seg+DINO-key | **0.652** [0.642, 0.663] | **0.592** [0.580, 0.603] | 0.928 [0.926, 0.931] | 0.931 [0.929, 0.933] | 0.493 [0.465, 0.521] | 0.9931 [0.9928, 0.9933] | **0.731** [0.710, 0.751] |
| Seg+frame | **0.652** [0.640, 0.664] | **0.592** [0.581, 0.603] | 0.926 [0.924, 0.928] | 0.930 [0.929, 0.932] | 0.500 [0.469, 0.533] | 0.9932 [0.9930, 0.9934] | 0.727 [0.708, 0.748] |
| No memory ($K$=0) | 0.646 [0.639, 0.654] | 0.573 [0.567, 0.580] | **0.935** [0.933, 0.937] | **0.940** [0.939, 0.942] | **0.546** [0.525, 0.567] | 0.9932 [0.9930, 0.9934] | 0.553 [0.518, 0.588] |

### A.4.7 Supplementary Human Failure-Attribution Study

**Scope.** The study covers the four shared-generator retrieval rollouts, whose intermediates (references, FLUX keyframe, Wan clip) are all recorded; released systems keep memory inside the generator's KV or latent state, so only their final pixels could be inspected. Track-A needs no such study: no generator is in its scored path, and its gold is human-reviewed (App. A.1.1).

Table 14: **Where closed-loop failures arise**, by retrieval memory (%, inclusion-weighted). Left: judge-flagged failures of a required entity, by the stage at which it is lost: already wrong in its earlier appearance (Hist., writing), in the history but not retrieved (Select.), in the references but lost in the keyframe or clip (Synth.), or in fact rendered (Judge). Right: judge-flagged forbidden renders, by where the entity entered: carried in the references (Context), introduced by the generator (Gen.), or not rendered (Judge). Unresolved units excluded; pooled CIs in the main text.

| | Required entity missing or wrong | | | | | Forbidden entity rendered | | | |
|---|---|---|---|---|---|---|---|---|---|
| Memory | Hist. | Select. | Synth. | Judge | $n$ | Context | Gen. | Judge | $n$ |
| Text-frame | 45 | 28 | 1 | 27 | 36 | 60 | 0 | 40 | 10 |
| Seg+uniform | 34 | 35 | 9 | 22 | 38 | 100 | 0 | 0 | 9 |
| Seg+DINO-key | 13 | 44 | 18 | 25 | 35 | 56 | 0 | 44 | 9 |
| Seg+frame | 25 | 38 | 2 | 34 | 38 | 67 | 0 | 33 | 9 |
| Pooled | 29 | 36 | 7 | 27 | 147 | 71 | 0 | 29 | 37 |

**Frozen sample.** From the frozen judge's per-entity verdicts we form four unit types: a required recall entity absent from the judge's present set (*miss*; stratified by absence gap 1–4, 5–29, ≥30, and temporal reference), an entity present in a wrong persisted state (a state already shown earlier, so memory could supply it) or as the wrong look-alike member (*wrong*), a destroyed or deceased entity rendered (*forbidden*), and a required entity the judge found (*control*). Each retrieval memory contributes 56 units (28 miss, 10 wrong, 10 forbidden, 8 control; Table 15) drawn with seed 20260925; 10% of units reappear later as hidden repeats, and five unscored practice items come first. Segments without recorded references or keyframe are excluded before sampling. The manifest, media hashes, and inclusion probabilities were written before any label was collected.

Table 15: **Frozen human-study sample.** Units per retrieval memory and unit type, with the size of the judge-flagged pool each was drawn from in parentheses; the unequal draws are why shares are inclusion-weighted. Strata with too few candidates fall back to a neighbouring stratum (time references and look-alike errors are pooled with mid-gap misses and state errors). The 224 units are shown with 22 hidden repeats (246 scored items) after five unscored practice items, 251 items in all.

| | Miss (required entity absent) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Memory | gap 1–4 | 5–29 | ≥30 | time ref. | Wrong | Forbidden | Control | Units |
| Text-frame | 8 (140) | 8 (56) | 8 (20) | 4 (8) | 10 (109) | 10 (201) | 8 (596) | 56 |
| Seg+uniform | 8 (155) | 8 (63) | 8 (12) | 4 (6) | 10 (121) | 10 (265) | 8 (582) | 56 |
| Seg+DINO-key | 8 (116) | 10 (52) | 8 (18) | 2 (2) | 10 (117) | 10 (265) | 8 (639) | 56 |
| Seg+frame | 8 (109) | 8 (42) | 8 (20) | 4 (5) | 10 (111) | 10 (274) | 8 (632) | 56 |
| All | 32 (520) | 34 (213) | 32 (70) | 14 (21) | 40 (458) | 40 (1005) | 32 (2449) | 224 |

**Interface and rule.** The box below shows what each item displays, the question asked of every panel, and the deterministic rule that turns the answers into an attribution. The entity card gives the entity's gold appearance at that point of the story (its latest state).

**Human attribution item (one per unit)**

**Entity card.** *Old Shi (character): night patrolman in his fifties, gray crew cut, deep-gray security uniform and black leather shoes, patrol baton and key ring at the waist.* Required state or look-alike cue, if any, in the story's original Chinese.
**Four panels, left to right.** (1) the entity's earlier appearance in the same rollout; (2) the three references memory returned; (3) the composed keyframe; (4) the generated clip.
**Question per panel.** "Is the correct entity visible here?" (forbidden units: "Is this entity visible here?").
**Answer:** yes / no / unsure.
**Hidden from the annotator.** Memory identity, judge verdict, and unit type; required and forbidden items differ only in the question. Items from all memories and unit types are interleaved in a fixed seeded order; five practice items come first and hidden repeats appear in the second half.
**Attribution (deterministic, applied after labeling).** Required entity: clip shows it → judge miss; else references show it → synthesis; else earlier appearance shows it → selection; else → writing. Forbidden entity: clip does not show it → judge false alarm; else references carry it → context; else → generator. An "unsure" on a deciding panel leaves the unit unresolved.

**Results and reliability.** All 251 items were completed (median 11 s per item). The 22 hidden repeats, which measure how repeatable the labels are rather than agreement between annotators, received the same attribution in 20 cases; 28 of 32 judge-confirmed controls were confirmed, and 8 of 224 units were unresolved. Per-memory shares (Table 14) have intervals of about ±20 points, so the main text uses the pooled row. Judge misses concentrate in long-gap (17 of 32) and temporal-reference (9 of 14) failures, so Track-B recall in these families is conservative for the retrieval memories; the study did not sample the no-memory control, whose parse-failure rate is also lower (App. A.1.3).

### A.5 Limitations and Outlook

VMem-Bench diagnoses what similarity-based memory gets wrong; it does not prescribe how memory control should be built, and our claim that causal video memory is a control problem remains a conjecture until such a memory is built and measured. The benchmark covers story-driven generation, where instructions name characters, props, and places; interactive settings in which memory must follow user actions or camera control, as in world models, are a natural extension of the same two measurement points. As generators improve, new stresses can be added as families without changing the contract: the context a memory hands over and the video that results.