MEMORY OUTWEIGHS CONTROL: A CAUSAL AUDIT OF CACHED HISTORY IN AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive video generators are sold on the promise that a viewer can steer the story as it unfolds. We audited whether the steering arrives. On two autoregressive video diffusion backbones we measured how far a mid-rollout prompt switch moves the continuation, in two regimes: with no cached history, and with twelve frames of it. With no history the prompt fully determines the video — a switch to an unrelated scene leaves the two continuations near-orthogonal in feature space (1.00 cosine distance). With twelve frames cached, the same switch retains 0.98% of that effect on one backbone and 0.15% on the other. The backbone that keeps *less* is the one built for prompt switching, whose KV-recache re-encodes the entire prefix under the new prompt on every switch and did so here. Ten seeds each; the no-history regime doubles as the positive control.

That measurement reframes the object we set out to formalize. We define the *minimal sufficient visual memory* at a branch point through a causal information bottleneck, and we can define it — but on current systems the decision it is defined against does not reach the continuation, so the specification is not wrong so much as without anywhere to land.

A distance between generations could be blind to a switch that preserves texture, so we checked with a quantity that cannot be: CLIP alignment to the prompt text. On both backbones, with no history the continuation follows the newly chosen prompt in 10/10 seeds and with twelve frames cached in 0/10, scoring 0.369 against the prompt the viewer left and 0.089 against the one they chose. Two backbones and two metrics that share no blind spot put the surviving fraction between 0.15% and 0.41% for that switch. Over a grid of 4 switches — from an alternative action inside the established scene to an unrelated one — at 5 prefix lengths, the median across 128 measurements is 10.1% and the largest single one 44.9%; 0 exceed half. Most of the steering is absorbed, not all of it.

Where a switch is lost entirely, one frame accounts for it. Adding the single oldest cached frame — an emergent attention sink — to a nine-frame window reproduces the full-history collapse, while adding two other old frames does not; a ten-frame history containing that slot suppresses the switch more than an eleven-frame one without it. The effect is indifferent to what the frame depicts — one lifted from an unrelated history works as well — but not to whether it is a real frame, since zeroing the slot restores the switch.

The audit also finds structure in the memory itself. Pruning by measured relevance matches the full cache down to 50% of it, while recency ordering leaves the noise floor at the *first* frame removed. The same two frames at opposite ends carry the effect through opposite mechanisms, so a necessity criterion defined by overwriting — ours, and the natural one — certifies a memory that is unusable (14.45× the floor), and the obvious repair (2.47×) still cannot certify sufficiency: the remaining frames are individually dispensable and jointly required. The audit and every run record are released.

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



Figure 1: **What 0.98% surviving looks like.** One branch point, one seed, three frames per clip. *Top:* the prefix the viewer was watching when the choice appeared. *Middle:* the switch with the cache hidden — unopposed, the choice replaces the scene outright. *Bottom:* the same switch with the cache visible: subject and place survive intact, and what the chosen prompt achieves is a leak into the background of the last two frames that never displaces what is in front of it. That leak is why the surviving fraction is small rather than zero.

1 Introduction

Interactive narrative video is sold on a simple promise: a viewer picks one of several continuations at a decision point and the story follows. Keeping the story coherent across those branches requires *memory*, and recent long-video and world-model systems supply it generously — a growing key–value cache, or a bank of past frames to retrieve from (Yu et al., 2025; Yang et al., 2026; Huang et al., 2025). Carrying all of it is expensive: cache and attention both grow with generated frames, and in a branching product, where continuations are often produced speculatively, that bill is paid once per branch and compounds down the tree.

We set out to ask how little of that history is causally required once the viewer's choice is fixed — the **minimal sufficient visual memory** (MSVM) — and to identify it by intervention rather than by heuristic.

This paper. We give a causal definition of MSVM — predictive sufficiency for the post-decision generation distribution, plus causal necessity under interventions that hold the decision fixed (Section 3) — and an audit that measures it on a frozen generator.

Why the decision must be clamped. Sufficiency alone is fragile: the decision correlates with the past, so a memory can look sufficient by proxying which branch the viewer takes rather than by carrying the appearance information that keeps branches consistent. Intervening while holding a fixed cuts the back-door $M \to a \to Y$. The criterion pair is not ours — Zhou et al. (2026) introduced it for keyframe selection in video understanding — but a viewer decision has no analogue there.

What we found instead. Before any of that can be measured, the decision has to reach the continuation. It does not. On both backbones we audited, a mid-rollout prompt switch to an unrelated scene retains around one percent of the effect it has when no history is cached — and the backbone that retains least is the one built for prompt switching. The formalization is therefore stated here for a regime current systems do not yet provide, and the empirical contribution of this paper is an audit of the memory itself, with the decision held fixed.

Contributions.

1. A measurement of how much viewer control survives a cached history: 0.98% and 0.15% of a prompt switch on two autoregressive video backbones, against 1.00 with no history — the latter doubling as the positive control that the conditioning does reach the generator (Section 5).
2. An *audit* for the memory itself: per-frame cache ablation that preserves rotary positions, three classes of $\mathrm{do}(\cdot)$ intervention, and set-level divergences scored against a null estimated at the same sample size and in the same paired form as the effect.
3. *Findings* from that audit: where a switch is lost entirely, a single cached frame accounts for it, and any genuine frame in that slot will do; a sufficiency knee at 50% retention; and evidence that *our own* overwrite-based necessity criterion — and the obvious repair — cannot certify a sufficient memory.
4. A causal information-bottleneck *formalization* of minimal sufficient visual memory (Section 3), stated for a decision-conditioned regime that we report is not yet available.

Work on context in video generation runs in two directions, neither of which asks our question: making context longer and more complete (Yu et al., 2025; Chen et al., 2026; Oshima et al., 2025), and shrinking the cache by correlational scoring (Ji et al., 2026; Luo et al., 2026). Asking what the continuation is causally entitled to — and measuring it against a noise floor — is what we add; Section 2 makes the boundary precise.

2 Related Work

Work on context in long and interactive video generation has, over the past year, split into two directions that are both orthogonal to ours. One makes the context *longer and more complete*; the other makes the cache *smaller* by heuristic scoring. Neither asks which part of the past is causally required once the viewer's decision is fixed.

Making context longer and more complete. Autoregressive-diffusion generators (Yin et al., 2025; Huang et al., 2025) produce frames against a rolling KV cache and LongLive (Yang et al., 2026) adds frame-level prompt switching. On top of these, Context as Memory (Yu et al., 2025) retrieves past frames, Context Forcing (Chen et al., 2026) removes the short-context teacher bottleneck, WorldPack (Oshima et al., 2025) packs more history into a fixed budget, and open interactive world models (Zhao et al., 2026; Valevski et al., 2025; Bruce et al., 2024; Chen et al., 2024) accumulate state for long rollouts — the shared principle being *maximal* memory. We ask the dual question.

Making the cache smaller. Token eviction and cache compression for language — H2O (Zhang et al., 2023), Scissorhands (Liu et al., 2023), attention sinks (Xiao et al., 2024), SnapKV (Li et al., 2024), adaptive per-head policies (Ge et al., 2024) — has now been carried to autoregressive video: Forcing-KV (Ji et al., 2026) prunes static and dynamic heads differently on exactly the backbones we study, and Future Forcing (Luo et al., 2026) scores cached tokens against a predicted future query. These are strong efficiency baselines and, without exception, correlational: the criterion is what attention has pointed at or is predicted to. Notably, StreamingLLM's prescription — keep the earliest tokens as sinks plus a recent window — is the structure our audit recovers from measurement in Section 5. It is also, independently, what interactive video systems have converged on by hand: LongLive (Yang et al., 2026) ships a twelve-frame attention window with the first three frames pinned as a sink. Our measurement and their design agree on a structure that the recency-scored compression policies carried over from language do not preserve.

Causal information bottleneck: sufficiency and necessity. The criterion pair we build on is not ours. ReaSon (Zhou et al., 2026) defines keyframes for video *understanding* as those satisfying both predictive sufficiency and causal necessity, using a learned policy to select a sufficient frame subset and counterfactual interventions to test necessity. We adopt that pair and change three things, each forced by moving from understanding to branching generation. (i) The target is a *generative distribution* over the continuation rather than an answer to a question, so sufficiency is a divergence between generation distributions (Def. 1) rather than accuracy preservation. (ii) The intervention acts on the *generator's own context and KV cache* rather than on the input frames of a frozen discriminative model. (iii) Most importantly, our setting contains a variable that has no analogue

in keyframe selection — the viewer's decision a — and it must be *clamped* during intervention. Without clamping, a memory unit that merely predicts which branch the viewer will take is credited with causing the continuation through the back-door path $M \to a \to Y$ and is wrongly retained. Clamping is what would make the necessity test measure causation of the continuation rather than anticipation of the choice — on a backbone where the decision reaches the continuation at all, which Section 5 reports ours does not.

Information bottleneck. The IB (Tishby et al., 2000) formalizes the compression–relevance trade-off, generalizing the minimal sufficient statistic (Cover & Thomas, 2006), and has shaped how deep representations are analyzed (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017; Alemi et al., 2017; Achille & Soatto, 2018; Federici et al., 2020). Ours is the sufficiency objective with a distribution over future video as the target; since estimating that in pixel space is hopeless we use sample-based divergences (Gretton et al., 2012; Székely & Rizzo, 2013), and variational estimators (Belghazi et al., 2018; van den Oord et al., 2018; Poole et al., 2019) where a bound is wanted.

Controlled studies and measurement of memory. Echo-Memory (King et al., 2026) holds backbone, optimizer, and metrics fixed while swapping the memory mechanism — raw retention, compression, spatial summaries, recurrence — and reports that replay fidelity is not a sufficient proxy for whether a world was remembered — the warning that decides where our sufficiency criterion is stated (Section 3). VBench (Huang et al., 2024) supplies the consistency measures we report, and MBench (Zhang et al., 2026) decomposes memory capability into entity, environment, and causal consistency — their "causal" axis meaning narrative plausibility, ours a shift under $\mathrm{do}(\cdot)$ with the decision clamped; the two should not be conflated. Echo-Memory compares memory *mechanisms*; we quantify, for one mechanism, how much of it the continuation actually uses.

Editing internal state to establish what a component causes is standard practice in causal inference and causal representation learning (Pearl, 2009; Peters et al., 2017; Schölkopf et al., 2021); our protocol specializes it to the memory of a video generator (Ho et al., 2022; Ha & Schmidhuber, 2018), where a "unit" is a KV block or latent frame and the "effect" is a distribution over continuations.

3 Problem Formulation

We state the criterion the audit is built on and keep it short: the decision variable it is defined around does not, on the systems we audited, reach the continuation (Section 5), so the definitions below are stated for a regime those systems are advertised as providing rather than one we could exercise. Derivations, the information-bottleneck relaxation, and the estimator details are in Appendices A–D.

Setup. At a decision point the story has produced a visual history H — past frames plus whatever internal state the generator exposes (KV entries, latent frames, a retrieval bank). The viewer selects a decision a from the branches offered, and a frozen generator samples a continuation $Y \sim p_\theta(Y \mid H, a)$. A retention map gives $M = \phi(H)$ to condition on instead. We want the smallest M that is indistinguishable from H, for every a.

Definition 1 (Decision-invariant predictive sufficiency)**.** $M = \phi(H)$ is ϵ-*sufficient* if for every $a \in \mathcal{A}$

$$\mathbb{E}_H \left[D\big(p_\theta(Y \mid H, a) \,\big\|\, p_\theta(Y \mid M, a)\big) \right] \;\le\; \epsilon. \tag{1}$$

This is the information-bottleneck relevance term conditioned on the decision and required to hold across decisions — the sense in which sufficiency here is *decision-invariant*. It is stated on the generation distribution rather than on a consistency score, because replay fidelity is not a sufficient proxy for what a model remembered (King et al., 2026); consistency enters only as an estimator.

Sufficiency alone permits redundancy, and worse, permits a memory that is sufficient only because it predicts *which* branch the viewer takes. The second criterion rules both out by intervention, with the decision clamped so the back-door $M \to a \to Y$ is cut.

Definition 2 (Causal necessity)**.** For a partition $M = (M_S, M_{\bar{S}})$, the component $M_{\bar{S}}$ is *causally necessary at level* η if

$$\mathbb{E}_H \left[D\big(p_\theta(Y \mid M, a) \,\big\|\, p_\theta(Y \mid M_S, \mathrm{do}(M_{\bar{S}} \leftarrow \tilde{m}), a)\big) \right] \;>\; \eta \tag{2}$$

for interventions $\tilde{m}$ in a specified class, with a clamped. M is *minimal* when every nonempty component is necessary, and the MSVM is the Ω-cheapest ϵ-sufficient M that is also minimal (Eq. (4), Appendix B).

A load-bearing assumption, which we later measure to be false. Def. 2 decides necessity by *substitution*: it asks what happens when a unit's contents are replaced, which presumes a unit contributes through its contents. A unit can instead contribute by occupying a position — attention has to land somewhere — in which case substitution is nearly free while deletion is not, and the criterion declares the unit unnecessary exactly when it matters most. We keep it as stated because it is the natural formalization and the one a reader would write down, and Section 5 shows it failing on the single most load-bearing unit of a real generator's cache — the same unit that, we also find there, is what absorbs the viewer's choice.

4 THE AUDIT INSTRUMENT (DIVM)

DIVM is the instrument this paper measures with, not a system it proposes to deploy. It operationalizes Section 3 on a *frozen* pretrained interactive video generator in two stages: (1) a sample-based estimator of the information-bottleneck objective that scores candidate memory subsets, and (2) a counterfactual-intervention protocol that tests causal necessity. Both are exercised by every measurement in Section 5; no generator weights are trained and nothing is learned, which is what keeps the audit reproducible on a single GPU. A third stage — turning the audit into an adaptive retention policy — is what the framework was originally for; Section 5 reports the measurement that removes its premise, and we return to it in Section 6 rather than presenting an unvalidated policy here.

4.1 MEMORY UNITS AND CANDIDATE SETS

We expose the generator's history H as a set of *memory units* $\{h_1, \ldots, h_T\}$—the natural retention granularity of the backbone: per-frame KV blocks for an autoregressive-diffusion transformer, or latent frames / stored reference frames for a retrieval-based generator. A retention map ϕ selects a subset (optionally with per-unit compression). The search space is the set of subsets; DIVM scores units and retains a prefix under a footprint budget.

4.2 STAGE 1: ESTIMATING THE IB OBJECTIVE

For a candidate memory M we estimate the two terms of the information-bottleneck objective (Eq. (3), Appendix A) by surrogate. Because Y is high-dimensional video we never do density estimation in pixel space: the relevance term is the divergence of Eq. (1) between continuations sampled from the full-history and candidate-memory generators under matched seeds, measured in a perceptual embedding and averaged over $a \in \mathcal{A}$; the compression term is the footprint $\Omega(M)$, a monotone proxy.

A per-unit *relevance score* s_t is obtained by leave-one-unit-out ablation: the increase in the divergence surrogate when unit h_t is removed from an otherwise-full memory. Units are then ranked by s_t. (Attention-based importance, as in KV-compression baselines, is the cheaper correlational alternative this scoring is meant to be tested against; that comparison is E2, which Section 5 explains we could not run.)

4.3 STAGE 2: COUNTERFACTUAL NECESSITY CERTIFICATION

Ranking by relevance is still correlational. Stage 2 applies Def. 2: for a retained set M_S, we *clamp the decision* a and intervene on the complement, $\mathrm{do}(M_{\bar{S}} \leftarrow \tilde{m})$, drawing $\tilde{m}$ from (a) a different history (identity/scene swap), (b) a noised version of the unit, or (c) a null/masked value. We measure the induced shift in the generation distribution (Eq. (2)). Two uses:

1. **Prune spurious units:** a unit whose removal changed the relevance surrogate (Stage 1) but whose *clamped-decision* intervention produces shift $\leq \eta$ is proxying the decision, not causing the continuation; it is dropped.

2. **Certify minimality:** the final retained set is accepted only if every unit is causally necessary at level η, i.e. no unit can be overwritten without a $> \eta$ distributional shift.

The protocol is summarized as Algorithm 1 in Appendix C.

Cost. Both stages run at inference on a frozen generator: Stage 1 is $O(T)$ parallelizable generation passes per decision point and Stage 2 a bounded number of intervention passes on the retained set, which keeps the audit within a single consumer GPU.

5 EXPERIMENTS

This section reports what the audit measured, in the order the measurements forced. We ask first whether the viewer's decision reaches the continuation at all — it does not, on either backbone — which is what turns the rest into an audit of memory with the decision held fixed. Every numerical cell is emitted from a machine-written results file, and no quantity appears here that a recorded run did not produce. Protocol, compute accounting and known pitfalls are in Appendices E–M.

5.1 SETUP

Backbones. We require a generator that produces at least two plausible continuations at a decision point and exposes *manipulable* memory units — per-frame KV blocks, latent frames, or retrieved reference frames that can be individually ablated, overwritten under $\mathrm{do}(\cdot)$, and generated from with the decision clamped; one that folds its history into a fixed-length recurrent state offers no ablation handle and is excluded rather than approximated. Candidates are Self Forcing (Huang et al., 2025), LongLive (Yang et al., 2026), and minWM (Zhao et al., 2026). Everything runs at inference on one consumer GPU; no generator weights are trained.

Evaluation sequences. Each sequence is a scripted prompt pair — a shared prefix and a continuation — generating a shared history and a branch point with $|\mathcal{A}| \geq 2$. All sequences reported here are authored by us; no product or user data is used (Appendix G). A prefix of twelve latent frames is generated, its cache snapshotted, and continuations of nine frames produced from that snapshot under each retention plan.

Metrics. Divergences between conditions are set-level energy distances on DINOv2 features; consistency is VBench subject consistency (Huang et al., 2024) against the shared prefix, a read-out rather than the definition (Section 2); necessity is the intervention-induced shift of Eq. (2) with the decision clamped.

5.2 E0: BACKBONE QUALIFICATION

Gate: a backbone must expose ablatable units at a branch point, or the question cannot be asked of it. Self Forcing qualifies and is the backbone audited in detail below: 21 addressable frames, a 5.62 GB cache, 12.3 s per branch, and full-history subject consistency 0.914. LongLive qualifies on the same criteria; it carries the attenuation measurement in its shipped configuration and the memory audit in the one where that audit is definable (Appendix J).

Self Forcing passes on structure: latent frame t occupies a fixed token range in a flat per-block KV buffer with no local-attention window. Rotary positions are written into the keys before caching, so a unit must be dropped by *gathering* the retained ranges — compaction would measure a positional perturbation instead of a memory ablation — and gathering with nothing dropped reproduces the ungathered rollout exactly (Appendix F).

Clamping is structural here, and verified rather than assumed: overwriting half the history leaves the cross-attention fingerprint unchanged to the last bit, while switching the decision moves it by five orders of magnitude more (Appendix E).

Most of the steering does not arrive. We measure how much of a mid-rollout prompt switch survives a cached history, using a ratio that needs no threshold. For a fixed seed, switch the branch

prompt to an unrelated scene and record the distance between the two continuations twice: with the history hidden (d_{none}) and with it visible (d_{full}). Their ratio is the fraction of the switch that survives.

With no history the prompt fully determines the video: d_{none} is 1.00 on both backbones, i.e. near-orthogonal in feature space. That regime is also the positive control: it rules out a conditioning path that never reaches the generator, which a small effect alone cannot distinguish from a decision-inert model. With twelve frames cached, the same switch retains 0.98% of its effect on Self Forcing (0.51–13.4% over 10 seeds) and 0.15% on LongLive (0.03–0.59%); Table 1.

Does that hold for switches a viewer would plausibly make? One prompt pair at one prefix length is the evidence base that misled us twice here. So we ran the grid: 4 switches (an alternative action inside the established scene, a different act in the same place, a new scene, the unrelated control) $\times$ 5 prefix lengths $\times$ 8 seeds (Appendix K). Across all 128 measurements the switch retains a median of 10.1% (IQR 5.4–16.6%); the largest single measurement is 44.9% and 0 exceed half. The unrelated control at twelve frames returns 1.1%, replicating the headline above with disjoint seeds.

Two features of that grid are not what we expected. The collapse begins as a *cliff rather than a slope* — three cached frames already absorb most of the switch — and survival is *not* ordered by how far the switch reaches: at twelve frames the in-scene alternative action retains 0.6% over 20 seeds, *less* than either intermediate switch. Two of the four drop again once the *whole* history is visible, which the next paragraph is about. Appendix K has the per-cell table.

Or is the metric simply blind to the switch? A distance between two generations can be small either because the switch failed to reach the generator or because the embedding cannot see what it changed — a live worry, since DINOv2 features are dominated by texture and layout, which the cached history supplies to *both* branches. So we repeated the measurement on the same videos with a quantity that cannot share that blind spot: CLIP alignment to the prompt *text*. With no history the continuation aligns with the newly chosen prompt in 10/10 seeds; with twelve frames cached, in 0/10 — it scores 0.369 against the prompt the viewer *left* and 0.089 against the one they chose. LongLive splits the same way. Formed as the ratio above, the surviving fraction is 0.98% and 0.41% on Self Forcing, 0.15% and 0.33% on LongLive: two backbones, two metrics that share no blind spot, four numbers below half a percent.

Where the choice does collapse, one frame accounts for it. Two of the 4 switches lose almost everything once the whole history is visible and two do not, and for the two that do the collapse is not about how much history is present. Against the nine most recent frames they retain 7.5% and 16.2%; adding the next two oldest frames leaves that where it was (6.2%, 20.6%). Adding instead the *single* oldest frame — the emergent attention sink that the memory audit below finds dominant — takes them to 1.2% and 0.8% — the full-history value (1.1%, 1.0%). A ten-frame history containing the sink suppresses more than an eleven-frame one without it. What does the suppressing is that slot holding *some* genuine frame rather than that frame's contents: a frame lifted from an unrelated history works as well (0.7% against 0.8%), zeros do not (6.3%). For the two switches that never collapse the frame does nothing at all (Appendix I). We can say what causes the collapse where it happens, not which switches it happens to.

LongLive is the one worth pausing on: the backbone whose stated contribution is frame-level prompt switching retains 6.5$\times$ *less* of the switch than the one making no such claim, with its KV-recache running. We cannot say why from these measurements and do not speculate.

What follows for the framework. Definitions 1–2 are stated for a decision that reaches the continuation, and on both systems audited it does not, so the clamped-decision test has nothing to clamp against and the decision-conditioned experiments are not merely unrun but presently unrunnable — a fact about the systems rather than the formalization, which is not wrong so much as without a place to land. The rest of this section audits memory with the decision held fixed, which here costs nothing.

5.3 E1: Memory ablation — does the knee exist? (H1)

Claim tested: H1; the curves are Figure 2. The decision is held fixed throughout, which here costs nothing.

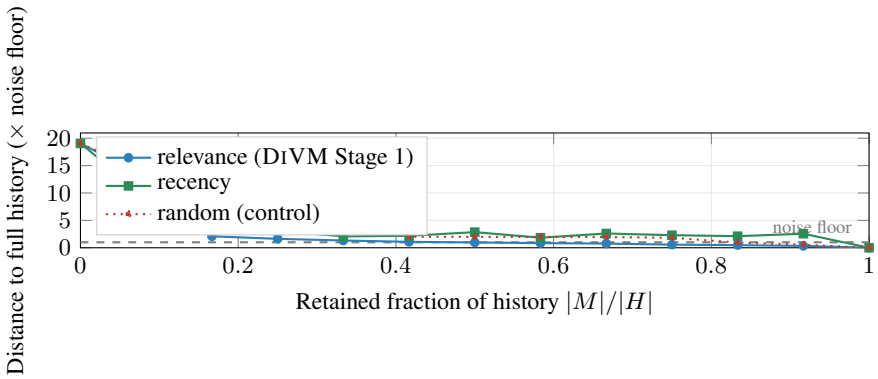


Figure 2: **E1 — how far the generation distribution moves as history is pruned, by pruning order.** Energy distance to the full-history condition in units of the same-condition split-seed noise floor; below 1 the pruned memory is indistinguishable from keeping everything. 6 seeds per point, 4 sequences. Relevance ordering stays under the floor to 50% retention and random to 83%; recency leaves it at the *first* unit removed (100%) — a one-unit gap, not a ranking result (see text).

We prune under three orders — recency, DIVM Stage-1 relevance, and random, the last a control rather than a strawman: if it tracks relevance, unit identity carries no information. Conditions are sets of rollouts, distances are energy distances between sets, and everything is reported against a same-condition split-seed null measured identically (Appendix D). The knee exists — the curve sits at the noise floor while a substantial part of the history is discarded, then leaves it — and what the three orders add is a mechanism for where each one leaves.

We summarise each order by its *sufficiency budget*: the smallest retained fraction from which the curve never leaves the measured noise floor — defined against the null rather than by eye, so it is not a budget chosen because it flatters us. Relevance holds to 50% retention on the averaged curve and random to 83%; recency never reaches the floor at all (100%), being outside the noise band from the *first* unit removed. The per-sequence budgets behind that average are not tightly clustered — relevance ranges from 25% to 67% across sequences — so the average is a summary, not a typical value. We draw no ranking from the gap between the first two, and the next paragraph gives the reason: the gap is a fact about one unit, not about the quality of two orderings. What is a result is recency's.

One unit explains all three, and it is the same one: the oldest retained frame carries a leave-one-out effect of 0.222 against a median of 0.049 across the others, and is dominant in 4/4 of the sequences, and among the top two in 4/4. Each order's curve departs from the floor at exactly the budget where that unit leaves its retained set. So we do not read this as "relevance beats recency" in general — random tracks relevance at high retention precisely because a random draw usually does not take the one unit that matters. What we can say is narrower: recency-shaped retention has a specific, reproducible liability, and it matters because scoring by recency or by attention statistics is what current KV-compression policies for autoregressive video do (Ji et al., 2026; Luo et al., 2026). The architecture designates no attention sink here (`sink_size` $= 0$), so the unit's dominance is emergent, mirroring the sink phenomenon documented for language models (Xiao et al., 2024).

The same structure on a second backbone. We repeated the audit on LongLive — not as it ships, but in the one configuration where its cache still holds the history to be interrogated at all, since as shipped it discards the prefix while the branch is generated. What follows is therefore about a second architecture operated in that regime, not about the shipped system; Appendix J reports the difference, what it costs, and the controls the ablation had to pass. The curves, the leave-one-out profile and the interventions all reappear there: relevance ordering holds to 58% retention against 50%, recency reaches the floor on neither, and one unit dominates by a similar margin. It is the oldest cached frame: among the top two in 4/4 of the LongLive sequences and 4/4 here, strictly first in 3/4 and 4/4, and standing 3.6× and 5.4× above the median of the other units. The one sequence where it is not strictly first loses by 0.5% of that sequence's noise floor — a tie between two co-dominant units, not

a counterexample. No sink is designated in that configuration either, so in both the dominant unit is one the model settled on rather than one an engineer pinned — though on LongLive it emerges only once the pinned sink is switched off — and the three-way intervention separates position from content there as it does here: where the dominant unit is the oldest frame, deleting and zeroing it cost $1.84\times$ and $1.72\times$ the floor against $0.64\times$ for overwriting it, and where it is the newest the three invert exactly as they do on this backbone.

A two-line rule is not worse. The measured profile — oldest frame dominant, newest few next, flat middle — is the shape of an existing prescription: earliest tokens as sinks plus a recent window (Xiao et al., 2024). Run against Stage-1 scoring on the first two sequences it ties on distance and picks the same units at 5/10 of the budgets, while a recent window *without* the sink clearly fails. Whatever Stage-1 buys, it is not a better retained set than a two-line rule.

This falsifies our own necessity criterion, and the obvious repair is not enough either. Def. 2 calls a unit unnecessary when overwriting it leaves the distribution in place. Applied per unit with η at the upper end of the floor's own confidence interval, that test keeps only the *newest* frame and discards the oldest, and the memory it certifies is unusable: distance $14.45\times$ the floor, subject consistency 0.521, close to keeping nothing at all.

The two interventions separate two kinds of necessity. The oldest frame has a removal effect of $3.66\times$ the floor and a replacement effect of $0.77\times$: it must be *present*, and its contents barely matter. The newest frame is the mirror image — removal $1.19\times$, replacement $4.95\times$: its contents matter and its slot can be covered by its neighbours. Neither is visible to a criterion that tests only one of the two.

Requiring a shift under *either* removal or replacement helps substantially — the certified memory moves from $14.45\times$ to $2.47\times$ the floor, consistency from 0.521 to 0.889 — and still does not yield a sufficient memory: every remaining unit falls below η on both tests, yet together they account for the residual gap. They are individually dispensable and jointly required. The problem is therefore the shape of the criterion rather than our choice of intervention, and it applies to the sufficiency-plus-necessity pair as usually stated (Zhou et al., 2026), not only to our instance.

5.4 What we did not run, and why

E2, E3 and E4 are specified and unexecuted, and one reason covers all three: each needs a decision that reaches the continuation, and the measurements above are that it does not. Appendix L states each design; we print no tables for them. Stated plainly, this measures a cache and not a choice, so the decision-invariance Sections 3–4 are built around remains a specification rather than a result.

Compute accounting. Totals are 15.8 GPU-hours on NVIDIA L40S-24Q (vGPU, 21.3 GB usable), peak 16.8 GB; every run logs backbone, seed, sample count, wall-clock and memory (Appendix M).

6 Discussion and Conclusion

Limitations. The audit spans two backbones but not equally: 4 scripted sequences on Self Forcing against 4 on LongLive, and the second is audited in a configuration it does not ship with, because as shipped its cache discards the prefix while the branch is generated (Appendix J). Histories are twelve frames throughout. Every threshold is a sample-based surrogate in an embedding space, no better than the measured floor (462 balanced splits).

The policy the framework was for. DIVM was designed to end in a predictor setting the retained budget per decision; we report none, because its premise — that the minimal sufficient memory varies *with the decision* — is what the measurements above remove on both backbones.

Conclusion. Interactive video generation is sold on the viewer steering the story; on the two systems we audited most of that steering does not arrive, by two metrics that share no blind spot. Where a switch is lost outright one cached frame accounts for it, and any genuine frame in that slot will do — the same unit the audit finds dominant on both backbones, settled on by each model rather than designed into it, and whose position-not-content character defeats per-unit necessity, ours included.

AI USE STATEMENT

Generative AI tools were used substantially in preparing this work, and we disclose that use in full rather than by category. They were used to draft and revise prose; to propose and refine hypotheses, including the attention-sink account of the dominant memory unit; to design experiments and, in particular, to design the controls and discriminating diagnostics on which the conclusions rest; to implement the evaluation harness, the backbone adapters, the intervention classes and the analysis scripts; to interpret results; to draft the formal statements of Section 3 and the proof sketch of Appendix A; to retrieve candidate references; and to conduct internal review rounds of the manuscript against a fixed rubric.

We have *not* used generative AI tools to generate, impute, extrapolate or adjust any experimental result. Every number in this paper is emitted by a script from a recorded run; the back-fill script refuses to write a value no run produced and aborts when two records disagree, and the build refuses to compile if any ratio quoted in units of the measured noise floor came from a record whose floor was estimated differently. Candidate references retrieved this way were each checked against the primary source before being cited, and several were discarded at that step.

We have reviewed all AI-assisted work. The authors take responsibility for the final content of this submission, including every claim, definition, interpretation and artifact produced with these tools.

ETHICS STATEMENT

This work studies how much visual context an already-trained generative video model must retain in order to stay consistent across branches. We do not train or release a new generative video model, and the method does not increase the fidelity of synthetic video: it removes context, and its best possible outcome is to leave the generated distribution unchanged while spending less memory.

We note the dual-use character of any efficiency result for generative media. Making branch generation cheaper makes it cheaper for benign and for malicious users alike, and interactive narrative video is a medium in which persuasive synthetic content can be produced at scale. We consider the marginal risk here small — the contribution is an analysis of context sufficiency rather than a capability increase — but we state it explicitly rather than leave it implicit. Relatedly, the consistency metrics we optimize for measure whether a character established *within a generated story* persists across branches; they do not measure, and this work does not pursue, transfer of a real person's identity into generated video.

All evaluation sequences reported in this paper are scripted by the authors: short prompt pairs describing a shared prefix and a continuation, containing no real people and no product data. We had considered drawing branch points from a deployed interactive narrative video product and describe that source in the appendix as future work; nothing in the results depends on it, and no such data was used.

REPRODUCIBILITY STATEMENT

All formal claims appear in Section 3; the full argument for Proposition 1, including the rate-to-footprint conversion and the cross-decision averaging, is given in Appendix A. The sample-based estimators for both information terms, with sample counts and variance control, are specified in Appendix D. The counterfactual intervention classes, the decision-clamping procedure for each backbone, and the calibration of the necessity threshold η are specified in Appendix E. Backbone versions, the definition of a memory unit per backbone, and the exact ablation interface are in Appendix F; construction of the branching evaluation sequences and the consistency-metric configuration are in Appendix G. Every reported run logs its seed, sample count, single-GPU wall-clock time, and peak memory (Appendix M), and every number in the paper is emitted from a machine-written results file rather than transcribed by hand. The audit itself — the backbone adapter, the position-preserving ablation, the intervention classes, the estimators, and the back-fill script that writes every number in this paper from those records — is included as supplementary material, together with the scripted evaluation sequences. That script is the only writer of numbers into this paper, it refuses to emit a value a run record does not carry, and it aborts rather than choose between two records that disagree

on the same quantity — a check added after one such collision put a default where a measurement belonged.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2018.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2020.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, XiTai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-KV: Hybrid KV cache compression for efficient autoregressive video diffusion models. *arXiv preprint arXiv:2605.09681*, 2026.

Wayne King, Zeyue Xue, Yuxuan Bian, Jie Huang, Haoran Li, Yaowei Li, Yaofeng Su, Yuming Li, Haoyu Wang, Shiyi Zhang, Songchun Zhang, Yuwei Niu, Sihan Xu, Junhao Zhuang, Haoyang Huang, and Nan Duan. Echo-Memory: A controlled study of memory in action world models. *arXiv preprint arXiv:2606.09803*, 2026.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiayi Luo, Qiyan Liu, Tengyang Wang, JunHao Liu, Jiayu Chen, Cong Wang, Hanxin Zhu, Chen Gao, Xiaobin Hu, Qingyun Sun, and Zhibo Chen. Future forcing: Future-aware training-free KV cache policy for autoregressive video generation. *arXiv preprint arXiv:2605.30083*, 2026.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. WorldPack: Dynamic frame compression for long-context video world modeling. *arXiv preprint arXiv:2512.02473*, 2025.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning (ICML)*, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, 2015.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations (ICLR)*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

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. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.22622.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.07772.

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. In *SIGGRAPH Asia*, 2025. arXiv:2506.03141.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, Hao Liu, Chen Li, Jing Lyu, and Yueqi Duan. MBench: A comprehensive benchmark on memory capability for video world models. *arXiv preprint arXiv:2606.00793*, 2026.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, Fan Bao, and Jun Zhu. minWM: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 2026.

Yuan Zhou, Litao Hua, Shilong Jin, Wentao Huang, and Haoran Duan. ReaSon: Reinforced causal search with information bottleneck for video understanding. In *AAAI Conference on Artificial Intelligence*, 2026. arXiv:2511.12530.

A Proof Details

We state the assumption and the proposition here rather than in the body, because nothing measured in this paper depends on either. They characterize the frontier the audit walks along; they do not identify which units realize it.

Assumption 1 (Rate–footprint compatibility)**.** There is $\bar{b} < \infty$ such that every retention map under consideration satisfies $\mathrm{I}(M; H \mid a) \leq \Omega(M)\,\bar{b}$, with $\bar{b}$ a bound on the conditional entropy of a single retained unit.

Assumption 1 is what lets a rate objective stand in for a unit count. It is tight when retained units are conditionally independent given a and loose otherwise, which is why Section 4 optimizes the footprint directly and uses the information form only for the characterization. We state it as an assumption because we can bound the two quantities against each other but cannot equate them.

Proposition 1 (IB relaxation)**.** *Let* $\mathrm{I}_{\mathrm{rel}}(\phi) = \mathrm{I}(M; Y \mid a)$ *with* $M = \phi(H)$*, averaged over* $a \in \mathcal{A}$*, and let* $\epsilon(\beta) = \mathrm{I}(Y; H \mid a) - \mathrm{I}_{\mathrm{rel}}(\phi_\beta)$ *for* ϕ_β *a minimizer of*

$$L_\beta(\phi) \;=\; \underbrace{\mathrm{I}(M; H \mid a)}_{\textit{compression}} \;-\; \beta \underbrace{\mathrm{I}_{\mathrm{rel}}(\phi)}_{\textit{decision-invariant relevance}} . \tag{3}$$

Then (a) *under Assumption 1 the constrained problem* (4) *is relaxed by* (3)*, and every solution of* (4) *is a minimizer of* L_β *for some* $\beta \geq 0$*;* (b) $\epsilon(\beta)$ *is non-increasing in* β*; and* (c) *if the supremum of* $\mathrm{I}_{\mathrm{rel}}$ *is attained, every limit point of* $\{\phi_\beta\}$ *as* $\beta \to \infty$ *is* 0*-sufficient.*

Part (b) needs no assumptions and follows from an exchange argument on two multipliers; part (a) needs Slater's condition, which the identity map supplies. Both are given in Appendix A.

The two steps are recorded below, and we are explicit about which is an assumption rather than a theorem.

Strong duality (part (a)). Conditional mutual information $\mathrm{I}(M; H \mid a)$ is convex in the encoder $p(M \mid H)$ for fixed marginals, and the sufficiency deficit $\mathrm{I}(Y; H \mid a) - \mathrm{I}(Y; M \mid a)$ is likewise convex; the constraint set is therefore convex. The identity map $\phi = \mathrm{id}$ attains deficit $0 < \delta$ for any $\delta > 0$, so a strictly feasible point exists and Slater's condition holds. Averaging the per-decision constraints over the branch distribution before forming the Lagrangian is what makes the resulting relevance term *decision-invariant*; if some a has zero probability its constraint is vacuous and the corresponding branch is excluded from $\mathcal{A}$.

The rate-to-footprint step is an assumption, not a result (Assumption 1). Ω counts retained units and $\mathrm{I}(M; H \mid a)$ measures rate in nats. For a map that keeps units verbatim they are related by $\mathrm{I}(M; H \mid a) \leq \Omega(M)\,\bar{b}$ with $\bar{b}$ bounding a single unit's conditional entropy, and the bound is tight exactly when retained units are conditionally independent given a. Cached video frames are strongly dependent, so the regime of interest is the loose one. We did not find a two-sided bound and do not claim one; this is why Stage 1 optimizes the footprint directly rather than a rate.

Monotonicity (part (b)). Proved in the main text by the standard exchange argument on two multipliers; it requires no assumption and no regularity beyond the existence of minimizers.

Honest status. Parts (a) and (b) are complete. The bridge between rate and footprint is stated as Assumption 1 and used only where labelled. Nothing in the empirical sections depends on the proposition: the audit measures footprints and divergences directly.

B Formulation Details

Consistency functional. Cross-branch consistency is a functional $\mathcal{C}(Y; H)$ scoring how well a continuation preserves the identities and setting established in H. Sufficiency (Def. 1) is stated at the level of the generation distribution, which implies preservation of any such bounded functional; the specific consistency scores we report are estimators, not the definition.

Algorithm 1 DIVM memory identification (per decision point)

Require: frozen generator p_θ; history units $\{h_t\}$; decisions $\mathcal{A}$; budget/thresholds ϵ, η

1: // Stage 1: rank by leave-one-out relevance
2: **for** each unit h_t **do**
3: $\quad s_t \leftarrow$ divergence increase when removing h_t, avg. over $a \in \mathcal{A}$
4: **end for**
5: $M \leftarrow$ smallest top-s_t prefix s.t. full-vs-M divergence $\leq \epsilon$ ▷ sufficiency
6: // Stage 2: certify causal necessity (clamp a)
7: **for** each retained unit $u \in M$ **do**
8: $\quad$ shift $\leftarrow \mathbb{E}_a D\big(p_\theta(\cdot \mid M, a) \,\|\, p_\theta(\cdot \mid M \setminus u, \mathrm{do}(u \leftarrow \tilde{m}), a)\big)$
9: $\quad$ **if** shift $\leq \eta$ **then** drop u from M ▷ spurious / redundant
10: $\quad$ **end if**
11: **end for**
12: **return** certified $M \approx$ MSVM

Mutual-information form. Equivalently, sufficiency asks that M retain the information H carries about Y given the decision, $\mathrm{I}(Y; H \mid a) - \mathrm{I}(Y; M \mid a) \leq \delta(\epsilon)$.

The optimization. The MSVM at a decision point is the retention map

$$\phi^\star \in \arg\min_{\phi} \ \Omega\big(\phi(H)\big) \quad \text{s.t.} \quad \phi(H) \text{ is } \epsilon\text{-sufficient}, \tag{4}$$

with Ω a memory cost (retained KV entries, latent frames, or bits), whose solution additionally satisfies causal minimality (Def. 2). Problem (4) admits a Lagrangian information-bottleneck relaxation trading a sufficiency gap $\epsilon(\beta)$ against retained rate, monotone in the multiplier; that is Proposition 1 below. It is stated here rather than in the body because *no empirical result in this paper depends on it* — the audit measures footprints and divergences directly — and because what the relaxation cannot say is *why* a unit is necessary, which is the whole subject of Def. 2.

Clamping. Clamping a blocks the back-door through the decision, so a component that is useful only because it predicts which branch the viewer took — rather than because it carries scene information the continuation renders — shows no intervention effect and is correctly dropped. On the backbones audited here clamping is structural and costs nothing, which is itself the finding of Section 5.

C IDENTIFICATION ALGORITHM

Algorithm 1 states the identification procedure as specified in Section 4. We print it unmodified, including the Stage-2 pruning step that Section 5 shows to be wrong for sink-like units, so that what was specified and what was measured can be compared directly. A corrected version would require a shift under *either* removal or replacement in place of line 7's replacement-only test. Section 5 measures that variant — it moves the certified memory from 14.45× the floor to 2.47× — and still does not obtain a sufficient memory, so we report it as a diagnosis of the criterion rather than presenting it as the method.

D ESTIMATION OF THE INFORMATION TERMS

Both terms of Eq. (3) are estimated from samples; neither is obtained by density estimation in pixel space.

Relevance. For a candidate memory M we draw n continuations from $p_\theta(\cdot \mid H, a)$ and n from $p_\theta(\cdot \mid M, a)$ under *matched seeds*, embed each continuation, and estimate the divergence of Eq. (1) in embedding space with an energy distance (Székely & Rizzo, 2013) and, as a cross-check, a kernel two-sample statistic (Gretton et al., 2012). Matched seeds matter: without them the sampling noise of the generator swamps the effect of removing a single unit. Where a variational bound is preferred we use the standard contrastive estimators (Belghazi et al., 2018; van den Oord et al., 2018; Poole et al., 2019), whose bias characteristics are documented.

Compression. The footprint $\Omega(M)$ — retained units, or bits after per-unit quantization — serves as a monotone proxy, for the reason given in Appendix A(ii).

Per-unit relevance score. s_t is the increase in the divergence surrogate when unit h_t is removed from an otherwise-full memory, averaged over $a \in \mathcal{A}$. Leave-one-out over T units costs $O(T)$ generation passes per decision point; the passes are independent and run concurrently.

Sample counts and variance. Consistency estimates on generated video are noisy and the effect of removing one unit among many is small. We fix the number of sequences, seeds per sequence, and continuations per condition in advance, and report a standard error for every cell rather than a bare point value.

The noise floor is measured, not assumed, and measured in the same shape as the effect. Every reported distance is accompanied by a null obtained by splitting one condition's own seeds and taking the energy distance between the halves — but the split has to produce the same sample sizes the effect is computed at. Earlier versions of two of these experiments ran the full-history condition over 6 seeds and split them three against three, while every effect was a six-against-six comparison; the null was therefore estimated at half the effect's sample size, read high, and understated every ratio built on it. The defect was found and fixed in the pruning curves first, and survived in the intervention diagnostic afterwards, where it produced the same intervention on the same unit quoted at two different multiples of "the floor" in two sections of an earlier draft. Both now sample the full-history condition twice as deeply and take the null over all 462 balanced six-against-six splits of it, reported with a confidence interval; a check in the build refuses to compile if any record feeding a floor-normalised number in this paper used a different estimator. (We do not quote how large the understatement was: the affected records were regenerated by the corrected runs, so the figure would have to come from something other than a run record.) Nothing below that floor is called an effect. The floor is not a formality — an earlier version of this measurement compared two individual rollouts by root-mean-square distance in latent space, and every comparison, from dropping a single unit to dropping the entire history, landed at the same value as simply changing the seed. That statistic reports whether the trajectory diverged, which is nearly binary here, not whether the distribution moved, and any curve built on it is flat by construction. Measuring in a perceptual embedding and between sets rather than between samples is what recovered the signal.

E Counterfactual Intervention Protocol

Intervention classes. $\tilde{m}$ is drawn from three families: (a) *cross-history swap* — the corresponding unit from a different sequence with a different subject and setting; (b) *noised unit* — the original unit perturbed at a calibrated magnitude; (c) *null* — a masked or zeroed unit. Class (a) is the informative one and also the most fragile: the donor must be near enough in distribution that the generator does not collapse outright, since a collapse registers as a large shift for the wrong reason. Donors are matched on clip length and shot scale and screened by a distribution-check pass before use.

Clamping the decision. Clamping means the decision token, embedding, or conditioning signal is *held at its observed value* while the context is overwritten — not merely that the same branch label is passed alongside a re-derived conditioning vector. For each qualified backbone we document where the decision enters the computation and verify the clamp by intervening on the decision alone and confirming that the generation distribution moves.

On Self Forcing the clamp is structural. The decision is a text prompt; it is encoded once by the frozen text encoder and its projected keys and values are held in a cross-attention cache that is separate from the self-attention key/value buffer. Nothing in that path reads the history, so overwriting history cannot re-derive the conditioning. We confirmed this empirically rather than by reading: after nulling half the history units, the cross-attention key/value fingerprint across all 30 blocks was identical to machine precision, while switching the decision moved it by five orders of magnitude more. The second measurement matters as much as the first — without it, an unchanged fingerprint would be consistent with a conditioning path that is simply inert.

Calibrating η. The threshold η separates "no effect" from "an effect" and must not be chosen so that the result comes out. We fix it from a null distribution measured before the main runs: intervene

on units already known to be irrelevant (padding positions, pre-prefix context) and place η at a fixed quantile of the resulting shift distribution, chosen and recorded before any E2 run. **No η was ever calibrated, because E2 was not run.** We note that the sink result of Section 5 makes the threshold secondary: the dominant unit fails the overwrite test at *any* η above the noise floor, so the criterion needs revising before its threshold is worth tuning.

F Backbone and Memory-Unit Details

For each candidate backbone we record the version and checkpoint; what constitutes one memory unit; the interface through which a unit is read, ablated, and overwritten; the maximum branch horizon; and whether any part of the history is folded into a fixed-length state that cannot be addressed. Candidates are Self Forcing (Huang et al., 2025), LongLive (Yang et al., 2026), and minWM (Zhao et al., 2026); the first two qualify and are used in Section 5, minWM was not audited.

Self Forcing (qualified). The DMD checkpoint, exponential-moving-average weights, four denoising steps, three latent frames generated per block, at 480×832. History lives in one flat key/value buffer per transformer block, 30 blocks of shape $[B, 32760, 12, 128]$ each for keys and values. A latent frame occupies exactly 1560 consecutive tokens, so frame t is the token range $[1560t, 1560(t{+}1))$ in every block, and the buffer holds 21 frames — the unit and the horizon both follow from that arithmetic rather than from a choice of ours. Local attention is disabled ($\texttt{local_attn_size} = -1$), so a query attends to the entire retained history; no attention sink is designated ($\texttt{sink_size} = 0$), which matters for interpreting the Stage-1 scores in Section 5. A unit is read, ablated, or overwritten by indexing its token range in all 30 blocks. The full cache is 5.62 GB.

Ablation is by gathering, not by compaction. Rotary position embeddings are applied to keys before they enter the cache, so removing a unit by sliding later tokens down would change every retained token's effective position. We instead pass attention only the retained token ranges, leaving positions untouched. We verify the indexing by gathering with nothing dropped: the result matches the ungathered rollout to the last bit.

G Evaluation Sequence Construction

Scripted branch points. Each scripted sequence fixes a shared prefix and offers branches differing along exactly one axis — appearance-preserving versus scene-reconfiguring — so that the stratification of Appendix H holds by construction rather than by post-hoc labelling.

Product decision points (not used). A deployed interactive narrative video product would supply realistic branch statistics that scripted sequences cannot, and we had planned to draw on one. *No product data was used in any result reported here*; every sequence is scripted. We record the intent because the decision-conditioned experiments would need such a source, and because a reader should know which parts of the design are aspirational.

Consistency-metric configuration. VBench (Huang et al., 2024) sub-metric versions, embedding models, and window settings are pinned and reported, since consistency scores are sensitive to these choices. Where Zhang et al. (2026) supplies scenarios matching our branch structure we adopt them rather than invent equivalents.

H Stratification by Decision Type and Scene (E4)

H2 asserts that the minimal sufficient memory is not a constant. E4 would test this by stratifying identified MSVM size by decision type and scene complexity. It shares E2's precondition and therefore its fate: stratifying by decision type presupposes that the decision changes the continuation, and Section 5 measures that on both backbones it does not. A stratification computed here would differ across strata only by sampling noise. We state the design and report no numbers for it. Note what this costs the paper: without E4 there is no evidence that an *adaptive* budget beats a constant one, so a constant budget is the recommendation our own measurements support, and the Stage-3 predictor of Section 4 is specified rather than justified.

I WHICH FRAME TAKES THE CHOICE

The grid left the behaviour at twelve frames unexplained, and twelve frames is the only visibility set in which the branch can see frame 0. This experiment separates those two facts. Four visibility sets are run on the same 12 seeds and the same prompts, each surviving fraction normalised by the same no-history cell:

visible set	same scene, diff. action	same place, diff. act	new scene	unrelated scene
9 most recent	7.5%	11.0%	18.5%	16.2%
11 most recent	6.2%	9.9%	11.5%	20.6%
9 most recent + *frame 0*	1.2%	12.6%	13.2%	0.8%
all 12	1.1%	13.2%	11.3%	1.0%

2 of the 4 switches collapse under the full history, in the sense that their twelve-frame value is less than half their nine-frame one. For 2 of those 2, one frame accounts for the whole collapse: the ten-frame set that includes frame 0 reproduces the twelve-frame value, while the eleven-frame set that excludes it stays with the nine-frame one. Read the middle two rows against each other — a *smaller* history suppresses the switch by more, provided the frame it adds is that one.

What this does not say. It does not say the sink suppresses steering in general. The two switches that do not collapse are unaffected by that frame — for the same-place switch, adding it *raises* the surviving fraction from 11.0% to 12.6%. So the finding is conditional: where a collapse occurs, it is a one-frame effect rather than a quantity-of-history effect; which switches collapse is not something these data determine, and we do not have an account of it.

An earlier version of this analysis reported the sink account holding for three of the four switches. That came from a test asking only whether the ten-frame value sat closer to the twelve-frame value than to the nine-frame one, which is satisfied by noise whenever all four conditions are equal — as they are for the same-place switch, where adding the sink frame raises rather than lowers the surviving fraction. A criterion for "one frame explains this collapse" has to require that a collapse exists first. The corrected criterion is the one stated above and is applied in the released run record.

Position or content? Frame 0 is three things at once: the attention sink, the oldest cached frame, and the frame in which the scene was established. The table above does not separate them, so we filled that slot four ways against the same nine-frame window, over 10 seeds:

slot 0 holds	same scene, diff. action	unrelated scene
nothing — not in the cache at all	6.2%	14.6%
its own frame (the establishing shot)	0.8%	0.7%
a frame from an unrelated history	0.7%	0.9%
its own frame plus noise at its own scale	5.0%	3.7%
zeros — addressed but carrying nothing	6.3%	8.9%

Two readings die here and a third survives. It is *not* the establishing shot: a frame lifted from an unrelated history suppresses the switch exactly as well as the real one (0.7% against 0.8%, 0.9% against 0.7%). Nor is it the bare position: a slot that is addressed but zeroed suppresses almost nothing (6.3% and 8.9%, against 6.2% and 14.6% for a slot that is not there at all), and noise recovers most of the switch too. What absorbs the viewer's choice is the sink slot holding *some genuine cached frame*; *which* frame is immaterial.

That is the same dissociation the necessity interventions of Section 5 find for this unit in the memory sense, from the other side — removing it costs 3.66× the noise floor while overwriting its contents costs only 0.77× — and the two together say something narrower than either alone: the generator needs a plausible frame at that slot, and is indifferent to what it depicts.

Table 1: **How much of a mid-rollout prompt switch survives a cached history.** For a fixed seed the branch prompt is switched to an unrelated scene and the two continuations compared, once with the history hidden (d_{none}) and once with twelve frames cached; attenuation is their ratio. 10 seeds, median (range). $d_{\text{none}} \approx 1$ is near-orthogonal: with no history the prompt fully determines the video, which rules out a mis-wired conditioning path as the explanation. LongLive's KV-recache re-encodes the cached prefix under the new prompt at every switch and did so here.

Backbone	d_{none}	Attenuation	Prompt switching
Self Forcing (Huang et al., 2025)	1.00	**0.98%** (0.51–13.4%)	not claimed
LongLive (Yang et al., 2026)	1.00	**0.15%** (0.03–0.59%)	headline feature, with KV-recache

J A Designed Sink Is Not the Same Object

Self Forcing's sink is emergent: the architecture designates none (`sink_size` $= 0$) and the audit finds one anyway. LongLive's is a design parameter — it ships pinning the first three frames alongside a twelve-frame attention window — so it can be dialled at inference without instrumentation, and we did, over 5 seeds, measuring each configuration's distance from the shipped one.

Two controls first, because the first version of this run passed neither and we nearly reported its output. Re-running the shipped configuration against itself gives exactly zero, as it must for a deterministic sampler. Cutting the attention budget to three frames — which leaves the sink and *no* recent context — gives 0.395, so the knob reaches the attention path. An earlier version set the policy on the attention modules only; the pipeline overwrites that from its own configuration on every call, and every configuration returned a distance of zero from the shipped one. The abort on a dead knob-control is now in the runner.

At the shipped budget of twelve frames, removing the three sink frames moves the generation by 0.005; removing three *recent* frames instead moves it by 0.002. The ratio, 2.4×, points the same way as Self Forcing — the sink end of the cache is worth more per frame than the recent end — but the honest observation is that both numbers are two orders of magnitude below what happens when the recent window is taken away (0.395). On this backbone the sink is a small effect.

It can even be a negative one. At a six-frame budget the shipped policy spends three of those frames on the sink and three on recent context, and drops further from the reference than a six-frame policy with no sink at all does. Frames pinned at the front are frames not spent at the back, and at a tight budget the back is worth more.

We therefore do not claim the Self Forcing result replicates here. A sink the model produced for itself and a sink an engineer pinned in place occupy the same slot and need not do the same work, and these measurements are consistent with their not doing it. What the two backbones do agree on is the direction: the oldest end of the cache is not interchangeable with the newest, and a retention policy that scores by recency alone will misplace it.

Porting the memory audit, and what it costs. The obvious completion of this paper is to repeat the pruning curves and the interventions on LongLive. Doing so is not a matter of pointing the same code at a different checkpoint, because as shipped the object the audit is defined on is not there. LongLive's cache is not a record of the history but a rolling buffer — 3 pinned sink frames followed by a window of the most recent 12 — and when the budget overflows it physically discards the oldest tokens rather than declining to attend to them. A leave-one-out over the prefix would be asking which frames the continuation needs while the generator removes them on its own schedule.

We measured that schedule rather than deriving it, by recording the eviction events the model performs during a branch continuation and reconstructing which absolute prefix frames remain addressable at each step. The instrumentation is observation-only and we checked that: generation with tracing on is bit-identical to generation with it off. Over a 12-frame prefix and a nine-frame continuation the prefix is evicted in blocks as the branch is generated, and what remains when the continuation ends is frames 0,1,2 — 3 of 12, and exactly the pinned sink. The other 9 are gone. That is a fact worth having on its own: a system marketed on long interactive context is, at the moment the viewer's choice is being rendered, conditioning almost entirely on 3 frames chosen by position rather than by content.

The configuration in which the audit is defined. Both obstacles — the eviction, and the fact that attention receives `cat(sink, window)` so a token index no longer names an absolute frame — are gated on the same two parameters, and both vanish at `sink_size` $= 0$ with global attention. Nothing is evicted, and attention receives one contiguous range beginning at token zero, so frame t is tokens $[1560t, 1560(t{+}1))$ exactly as on Self Forcing. We audit LongLive there. That is off its shipped setting and its adapter was trained with the sink and window in place, so the question is whether what we would be auditing is still the generator: over 5 seeds the full-cache configuration performs 0 evictions against 450 for the shipped one, holds its opening slightly better (0.962 against 0.958), and sits 0.0002 from the shipped output.

That last number is easy to over-read, so we will say what it does and does not license. It shows that this change of internal regime leaves the generated video essentially where it was — same output, different machinery. It does *not* show that the machinery is the same, and the audit is a measurement of machinery. What we changed is exactly the structure the audit looks at: no eviction, no pinned sink, global attention. So the correct reading of everything below is that the phenomenon reappears on a second architecture operated in this regime, not that the shipped LongLive has been verified. The finding reported earlier in this appendix about the shipped configuration — that its designed sink is a small effect there, and that we do not claim the Self Forcing result replicates in it — stands unchanged and is not in tension with this: the two sentences are about two different configurations, and we report the attenuation results of Section 5 in the shipped one and the audit in this one rather than letting them be read as one measurement.

One further consequence deserves stating outright, because it is the sharpest thing this configuration change costs. The sink the audit finds on LongLive is emergent — the architecture designates none here — but it emerges *after* the designed sink has been switched off. In the shipped model the first frames are pinned by construction, so what dominates there is a sink an engineer placed. When we say below that the dominant unit is one each model settled on rather than one designed into it, that is a statement about both generators in the configurations we audited them in; about the shipped LongLive it is not a statement at all.

Controls the ablation had to pass first. Three earlier attempts to instrument this backbone all completed, looked plausible, and were measuring something else, so the hook is checked against ground truth before it is allowed to produce a number. Every attention module reports `sink_size` $=$ 0 and global attention, and the cache is the full 21-frame buffer. Re-running a branch is bit-identical to itself. Forcing the gather path with nothing dropped is bit-identical to not gathering at all. Dropping exactly one history frame removes exactly 1 frame's worth of key tokens — the check that catches an index-mapping error directly, and the one an earlier version of this control failed because the probe was installed outside the filter and was reading k before any frame had been removed. Dropping the whole history moves the output substantially. Overwriting half the history leaves the cross-attention fingerprint unchanged to 0, so decision clamping is structural on this backbone as well.

What the audit finds there. Table 2 puts the two backbones side by side. Four of the audit's findings were re-measured on the second backbone and all four reappear — the pruning curves, the leave-one-out profile, which unit dominates it, and the three-way intervention below. The sufficiency budget under relevance ordering sits in the same range (58% against 50%), random ordering lands on the same number (83% against 83%), recency ordering never reaches the floor on either backbone (100% and 100%), and one unit dominates the leave-one-out profile by a similar margin (0.115 against a median of 0.032, versus 0.222 against 0.049). That unit is the oldest cached frame: it is among the top two in 4/4 of the LongLive sequences and 4/4 of the Self Forcing ones, and strictly first in 3/4 and 4/4. The one sequence where it is not strictly first is worth stating precisely rather than as a fraction: on beach the argmax goes to unit 4 instead, by 0.5% of that sequence's measured noise floor, with both units standing well above the rest. Two units are tied there and the counter has to pick one; reading that row as a sequence where the oldest frame does not matter would be reading a rounding decision as a result. It is also the sequence whose top two units separate least on *both* backbones — beach on Self Forcing and beach on LongLive — which is what one would expect if the tie is about the scene rather than about either generator. What was *not* re-measured there is the comparison against the hand-written sink-plus-window rule, so the claim that Stage-1 scoring buys nothing over two lines of prescription rests on Self Forcing alone.

Table 2: **The audit on a second backbone.** Sufficiency budgets are the smallest retained fraction from which the curve never leaves that backbone's own measured noise floor, so the two columns are comparable in floor units rather than in raw distance. Leave-one-out effects are the dominant unit against the median of the others. LongLive is audited at `sink_size` $= 0$ with global attention, the configuration in which the history is still there to interrogate (Appendix J); its attenuation numbers elsewhere in this paper come from its shipped configuration instead, and the two should not be pooled.

	Self Forcing 4 sequences	LongLive 4 sequences
sufficiency budget, relevance order	50%	58%
random (control)	83%	83%
recency	100%	100%
leave-one-out, dominant unit	0.222	0.115
median of the others	0.049	0.032
oldest frame among the top two	4/4	4/4
and strictly first	4/4	3/4
oldest frame ÷ median of others	5.4×	3.6×
measured noise floor	0.072	0.055

The margin is worth one further sentence. In this configuration LongLive's architecture designates no sink either — we set `sink_size` $= 0$ to make the audit definable — so the unit the profile concentrates on is emergent on both generators rather than pinned by either. Two models trained separately, on the same base, put their weight on the same slot when nothing tells them to.

The dissociation replicates too. A replication of the pruning curves is not by itself a replication of the position-not-content finding, so we ran the three-way intervention on LongLive as well, over 4 sequences. The prediction is a shape rather than a magnitude, and it is two-sided: a unit holding a *position* should cost about the same to delete as to zero and far more than to overwrite, while a unit carrying *content* should show the reverse.

On the 3 sequences where the diagnostic's dominant unit is the oldest frame, removing it costs 1.84× the floor, zeroing it 1.72×, and overwriting it with an unrelated history's frame only 0.64× — the position signature, and the same one Self Forcing gives for that unit over its own 4 sequences (2.57×, 2.23×, 0.65×), where the oldest frame is the dominant unit in 4 of 4. On the remaining 1 the dominant unit is the *newest* frame instead, and there the three arms invert: 1.24× and 1.18× against 7.33× for overwriting. That is not a failure of the account but the other half of it — Section 5 reports the same inversion for the newest frame on Self Forcing (1.19× removal against 4.95× replacement) — and it is why we group these by which unit the diagnostic found rather than pooling them: averaged across all 4 sequences the two signatures cancel almost exactly, and a pooled table would report no dissociation at all. The weakest unit stays *below the floor* under all three arms on both backbones (0.28–0.38× here, 0.31–0.52× there), so none of its three arms is an effect at all by the criterion this paper uses.

Both columns are produced by the same script with the same seed counts, so the comparison is between two backbones rather than between two protocols. They are not the same run as the deeper single-sequence certification of Section 5, which is why its number for this intervention (3.66× removal, 0.77× replacement) is not identical to the average here: that run measures one sequence at twice the sampling depth, and on that sequence this one gives 3.7× and 0.78×. The averages differ because they average different sequences, not because the two disagree about the intervention.

We are careful about what this does not cover. The second backbone is 4 sequences to the first's 4, and the two are audited in configurations that differ. Neither backbone's audit says anything about the decision axis, which on both is the thing that does not arrive.

K THE DOSE-RESPONSE GRID

Table 3 gives every cell behind Figure 3: 4 switches × 5 prefix lengths × 8 seeds, each entry the median and full range of the surviving fraction of the switch.

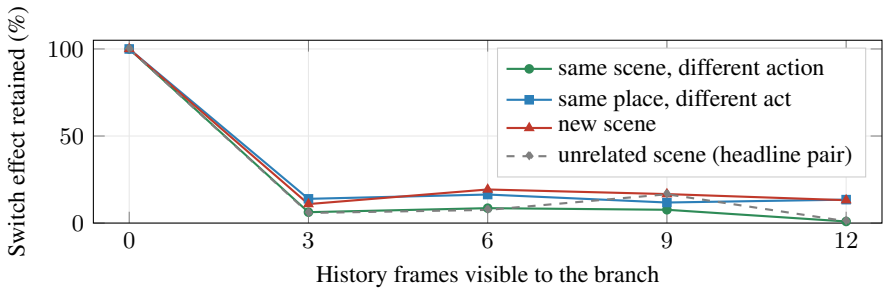


Figure 3: **Every switch we tried is absorbed, and how far it tried to reach does not save it.** Each curve is one prompt switch issued at the branch point, measured as the set-level distance between the switched and unswitched continuations, normalised by the same distance with no history visible — so every curve starts at 100% by construction, and the drop between zero and three frames is the result. The curves do not order by how far the switch departs from the established scene; the residual differences between them are discussed above and we do not account for them. 8 seeds per point.

Table 3: **How much of a prompt switch survives, cell by cell.** Median and full range over 8 seeds, as a percentage of the same switch's effect with no history visible. We give the range rather than an interquartile range because several cells are heavy-tailed, where a quartile falls on the jump and reports the convention rather than the data. Rows are ordered by how far the switch departs from the scene the prefix established. Every cell is below 44.9%; none of the 128 individual measurements exceeds half.

	history frames visible to the branch			
switch	3	6	9	12
same scene, different action	6.2 [2.8–16.2]	8.6 [4.8–21.8]	7.6 [4–23.5]	0.9 [0.2–15]
same place, different act	13.9 [8.9–24]	16.4 [3.9–29.8]	11.8 [5.1–27.2]	13.5 [5.4–17.5]
new scene	10.9 [4.4–18.9]	19.3 [8.7–24.4]	16.6 [7.7–44.8]	13.1 [3.7–43]
unrelated scene (headline)	5.8 [2.7–32.1]	7.6 [3–17.7]	16.5 [7.2–44.9]	1.1 [0.5–13.1]

What the residual variation does and does not support. The cells range from 0.9% to 19.3%, and we want to be explicit about which parts of that spread we are willing to read. Two statements are safe. Every cell is far below the no-history condition, and the largest *individual* measurement in the grid is 44.9%, with 4.7% of the 128 measurements above a quarter and 0 above a half — so the claim is that most of the switch is absorbed, not that none of it survives, and a reader should hold the tail in mind alongside the median. Second, the drop is essentially complete by three frames for the two intermediate switches, whose cells stay flat from three frames to twelve.

The pattern at twelve frames is real, and we do not fully explain it. At twelve frames the two *extreme* switches — the in-scene alternative action and the unrelated control — fall to 0.6% (IQR 0.3–1.1) and 0.8% (0.5–1.8), while the two intermediate ones stay at 12.2% (8.4–17.8) and 7.1% (3.7–13.4). Calling that noise would have been the convenient answer, so we added seeds instead: at 20 seeds per cell the two ends remain separated from the two middles with non-overlapping interquartile ranges. For those two switches the large drop also falls between nine and twelve frames rather than at three, which qualifies the cliff reading above rather than contradicting it — most of the absorption happens immediately, and for some switches a second step happens at the end.

Twelve frames is also the only condition in which the branch can see frame 0, which Section 5 identifies as an emergent attention sink. We state that coincidence because it is the obvious hypothesis and because the grid alone cannot test it: the sink is present at twelve frames for all four switches and only two collapse. We do not claim it as the mechanism.

Two estimators, one number. The attenuation of a single cell can be formed either seed-by-seed, $d_s(\text{history})/d_s(\text{none})$, or by dividing each seed's numerator by the median no-history distance. The first is paired and the second is not, and earlier in this project a mismatch of exactly this kind produced

two false negatives, so we checked rather than assumed: on the headline cell the two give 1.1% either way, because that switch's no-history distance barely varies across seeds (a factor of 1.0 between its smallest and largest). They separate only for the nearer switches, whose no-history distances vary by up to 1.8. The two forms are used in different places and we say which: Table 1 and its quoted ranges are paired, while the grid of Figure 3 and Table 3 divides by the median. The spread is also worth stating in the right shape. That cell is heavy-tailed rather than merely wide: 6 of its 8 seeds fall below two percent while the remaining two sit at 6.1% and 13.1%. An interquartile range lands on that jump and would report the quartile convention rather than the data, so we quote the median with the full range (0.51–13.4%) and say where the mass is. The switch is nearly always absorbed and occasionally is not; a summary that hides either half is the wrong summary.

L Experiments Not Yet Run

E2 and E3 are specified in Section 5 but have not been run, for the reason E0 established: on the qualified backbone the decision has no measurable effect on the continuation, so there is no decision-proxy unit for the clamped-decision test to catch and no back-door for the causal criterion to cut. Earlier drafts reproduced their tables here with every cell unmeasured; we have removed them. An empty table states nothing the paragraphs below do not state more precisely, and printing one invites the reader to treat a design as a result. The table and figure definitions remain in the source (`tab/e2_identification.tex`, `tab/e3_policy.tex`, `fig/pareto.tex`), unpopulated, so that the runs that would fill them have a fixed target.

Why E2 cannot be run on the qualified backbone. The crux evidence E2 is designed to produce is a shift-on-retained column in which attention importance keeps units below η while DIVM does not — a correlational criterion caught retaining a unit that only proxies the decision. Such a unit exists only if the decision influences the continuation, and E0 measured that it does not here: with any history visible, the branch prompt moves the generation distribution less than resampling the seed does, and so does a control prompt naming an entirely different scene. Running E2 anyway would compare two criteria on a causal graph in which the path $M \to a \to Y$ carries no effect; a null would say only that there was nothing to clamp. We would rather defer the experiment than publish an uninformative null dressed as evidence. If the two columns turn out indistinguishable on a backbone where the decision *does* reach the continuation, that is a real negative result for the causal criterion and we will report it as one.

M Compute Accounting

Every run records backbone, seed, number of sequences, number of continuations per condition, wall-clock time, and peak GPU memory, and every number in this paper is emitted from those records rather than transcribed by hand. Totals: 15.8 GPU-hours on NVIDIA L40S-24Q (vGPU, 21.3 GB usable), peak 16.8 GB.

N Further Limitations

Scope of what was actually measured. The two halves of the paper do not have the same reach, and conflating them would flatter the weaker one. The *attenuation* results — the headline, the semantic control, and the dose-response grid — span two backbones in their *shipped* configurations, 4 prompt switches, 5 prefix lengths, and up to 20 seeds per cell. The *memory audit* — the pruning curves, the leave-one-out profile, the intervention classes — also spans two backbones, but unevenly: 4 evaluation sequences on Self Forcing against 4 on LongLive, 6 seeds per condition on both, and the LongLive half in a *reconfigured* regime rather than the shipped one (Appendix J). The comparison against the hand-written sink-plus-window rule was run on Self Forcing only. All of it uses twelve-frame histories and nine-frame continuations at 480×832.

Within what we have run the leave-one-out profile is strikingly stable — the same dominant unit, the same second unit, the same flat middle, magnitudes within ten percent, and the same on both backbones — but a handful of scripted sequences that share one shape (a single subject in a coherent scene, near-static camera) is not generality. Whether the profile survives cuts, multiple subjects, or

a moving camera is untested. So is whether it holds on an architecture outside this family: both generators here are Wan-based, so a replication across them is weaker evidence than a replication across families would be.

The clamp has never been under load. The decision axis itself is measured, not assumed: Section 5 reports that on both backbones a switch retains under a percent of its no-history effect, by two metrics that share no blind spot. What that leaves untested is the machinery built on top of it. H2 and H3 are stated for a decision-conditioned memory, and neither backbone can exercise that condition, so what we measured is the memory side with the decision held fixed — the regime Def. 2 demands, but also the regime in which clamping costs nothing, so our clamp has never had to work. A backbone where the decision does reach the continuation could behave differently in ways we cannot anticipate from here, including ways that weaken the framework.

Cost model. The efficiency claim assumes the backbone's cost actually scales with retained units. This holds for KV-based autoregressive generators, where dropping a unit removes both its storage and its attention term, but retrieval-based generators pay retrieval and re-encoding costs that a smaller retained set does not necessarily reduce. The two families are therefore reported separately rather than pooled into a single saving figure.

Composition down the tree. MSVM is defined at a single decision point. A branching story has many, and the union of per-point minimal memories need not be minimal for the path — a unit dropped as unnecessary at point k may be required at point $k{+}2$ once an intervening decision has changed what the continuation must render. We do not claim compositional minimality and do not test it; the per-point result is the scope of the paper.

Consistency is not the only thing memory buys. We measure whether branches stay consistent with the shared past. A memory could be minimal for consistency and still be insufficient for narrative sense — a prop that must reappear because the plot demands it, not because it is visually established. Our criterion would discard it. That is a limitation of the target functional, not of the identification procedure, and it is why we report the consistency metric being optimized rather than claiming memory sufficiency in general.

O Additional Results

The per-sequence curves behind Figure 2, the full leave-one-out score vectors, the three-way intervention table, and the sink-plus-window comparison are in the released run records, one JSON per run, carrying the backbone, seeds, sample counts, wall-clock and peak memory for every condition. E2–E4 remain unrun for the reasons given in Appendix L; no table for them appears anywhere in this paper.

What we fixed in advance as reportable, and what became of it. Three outcomes were committed to before the runs, to be reported whichever way they fell. *(i) Evaluation sequences on which no knee exists.* Under recency ordering that is every sequence — none reaches the noise floor before full retention — and under relevance ordering it is none of them. *(ii) Units that pass Stage-1 relevance but fail Stage-2 necessity*, which would be the decision-proxy units clamping exists to catch. We cannot report their frequency: on both backbones the decision has no effect for such a unit to proxy, so the count is undefined rather than zero. *(iii) Any backbone on which* DiVM *fails to beat attention importance*, at the same prominence as one where it succeeds. That comparison is E2 and was not run, for the same reason.

Two of the three could not be honoured. We record that here rather than dropping the commitments quietly, since a pre-registration that is silently narrowed to the parts that worked is worth less than none.