BRIDGES OVER HEAVY HITTERS: A CAUSAL-HITS THEORY AND COUNTERFACTUAL BENCHMARK FOR ATTENTION-GRAPH KV-CACHE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free KV-cache compression decides which past tokens a long-context language model may forget. Heavy-hitter methods retain tokens with large accumulated attention, whereas Stream-Influence scores tokens by centrality in the sparse causal attention graph. We show that Stream-Influence is a one-pass causal form of Kleinberg's HITS: a token's authority is the attention it receives from later tokens, weighted by their hub scores. Without normalisation, attention sinks make the recursion diverge, and we give a bounded alternative. On bridge-chain graphs, sparsified centrality recovers every bridge once the budget exceeds the chain length, while accumulated attention keeps only a prefix proportional to the budget, because causal mass makes early tokens heavy hitters regardless of content. A symmetry theorem then shows that if several structurally identical chains are present and only the question identifies the gold chain, every score depending on the question through a single hop assigns the same value to a gold bridge and its twin. This class includes H2O, SnapKV, StreamingLLM and Stream-Influence. BRIDGEBENCH instantiates both regimes with known bridges. Physical eviction on models from 3B to 14B confirms the obstruction: unseeded scores retain at most 8% of full-cache accuracy at a 10% budget and 21% at 20%. A beam-restricted walk seeded at the question over coreference heads recovers the gold chain, retaining 55% and 81% at these budgets against 10% and 25% for the best baseline. The same heads, transferred without re-calibration, keep 74% of the gold sentences on HotpotQA, where SnapKV keeps 12%.

1 INTRODUCTION

The key–value (KV) cache is the dominant memory cost of long-context inference, and *training-free* cache compression asks *which past tokens may the model forget?* and answers with a per-token score. *Heavy-hitter* methods (Zhang et al., 2023; Liu et al., 2023; Jo & Shin, 2024) keep the tokens with the largest accumulated attention; *recency* methods (Xiao et al., 2023) keep a sink prefix and a sliding window; observation-window methods such as SnapKV (Li et al., 2024) keep the tokens most attended by the last few queries. All three read the attention matrix as a collection of independent *magnitudes*.

A third view reads it as a *graph*. CentralKV (Wu et al., 2026) treats the sparse causal attention pattern as a directed acyclic graph over tokens and scores each token by an online centrality, Stream-Influence, arguing that multi-hop reasoning depends on *structural bridges*, tokens with low accumulated attention but high graph centrality, and reporting LongBench gains over H2O and SnapKV. That evidence is end-to-end: LongBench does not reveal which tokens matter, so it cannot say *why* a graph score helps. We use the recursion and defaults of Wu et al. (2026) unchanged; none of the theory, benchmark or method below appears there.

This paper. We supply the missing analysis, the missing controlled evidence and, because the evidence contradicts part of the folklore, the missing fix.

(1) What the score computes (§4.1–4.2). Stream-Influence is a one-pass, *causal* analogue of Kleinberg's HITS (Kleinberg, 1999) with max-aggregation: a token's *authority* is the attention it receives

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.

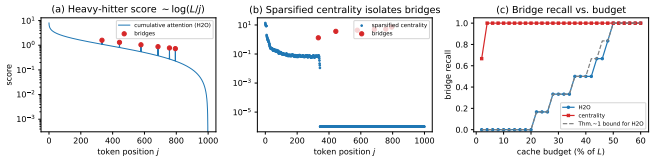

Figure 1: **Theorem 7 on a linear bridge chain** (L=1000, six bridges, c=0.5). (a) Accumulated attention decays as $\log(L/j)$, so content-free early tokens out-rank every bridge. (b) Sparsified centrality puts bridges 1–2 orders of magnitude above the background. (c) Bridge recall vs. budget; dashed: the bound $\#\{k : b_k \leq e^c \rho L\}/n$ on H2O's recall.

from later tokens, weighted by their *hub* scores. This factorises the score exactly as (sparsified, decayed accumulated attention) $\times$ (hub amplification) and exposes a defect: unnormalised, the recursion diverges geometrically under an attention sink, so that the "centrality" degenerates to the attention of the last few queries; a HITS-style normalisation bounds it.

(2) When heavy hitters must fail, and when *any* unseeded score must (§4.3–4.4). On bridge-chain graphs, sparsified centrality recovers every bridge at any budget above the chain length, whereas accumulated attention retains a bridge at position b only if $b \leq e^c \rho L$: causal mass accumulation makes *early* tokens heavy hitters regardless of content (Fig. 1). But we also prove a symmetry obstruction: when the context contains several structurally identical chains and only the question singles one out, every score whose dependence on the question is one-hop, including H2O, StreamingLLM, SnapKV *and* Stream-Influence, scores the gold bridges and their distractor twins identically. The question must be propagated.

(3) BRIDGEBENCH and a counterfactual protocol (§5). BRIDGEBENCH-SYN provides synthetic attention graphs with known important sets; BRIDGEBENCH-REAL procedurally generates documents of m structurally identical entity chains rendered as shuffled facts with a k-hop question about one of them, so that bridges, evidence and answer are known exactly and cannot be recalled from world knowledge. *Physically pruning* the KV cache of five open LLMs (3B–14B; a 32B model in Appendix J) confirms both theorems: every unseeded score retains bridges at the random rate and at most 8% (21%) of full-cache accuracy at a 10% (20%) budget.

(4) Seeded sparse propagation (§5.1). The obstruction says what a working score must do. Real models contain *coreference heads* whose attention from a sentence lands on the earlier sentence sharing an entity with it. On the chunk graph of these heads, a walk seeded at the question and pruned to a small beam at every hop, the seeded counterpart of Stream-Influence's max-style path extension, visits the gold chain and little else. It retains 55% of full-cache accuracy at a 10% document budget and 81% at 20%, where the best baseline retains 10% and 25%; head selection, propagation and beam sparsity are each necessary and the gain is flat in context length. On HotpotQA and MuSiQue the heads transfer unchanged: seeded chunk selection keeps 74% of HotpotQA's gold sentences at a 10% budget (SnapKV: 12%; H2O and Stream-Influence: 0%), while propagation is unnecessary because those questions name every hop, the case the obstruction theorem exempts; on RULER variable tracking, which it does not, propagation pays again (78% vs. 49% retention at a 2% budget).

2 RELATED WORK

Importance-based KV eviction. H2O (Zhang et al., 2023) and Scissorhands (Liu et al., 2023) keep tokens with the largest accumulated attention plus a recency window; A2SF (Jo & Shin, 2024) adds forgetting, TOVA (Oren et al., 2024) uses only the latest query, StreamingLLM (Xiao et al., 2023) keeps sinks and a sliding window. SnapKV (Li et al., 2024) and PyramidKV (Cai et al., 2024) score prefill tokens by a window of trailing queries; PyramidKV, AdaKV (Feng et al., 2024) and DuoAttention (Xiao et al., 2024) vary budgets across layers or heads, FastGen (Ge et al., 2024) profiles heads, R-KV (Cai et al., 2025) adds a redundancy term, and ChunkKV (Liu et al., 2025) evicts semantic chunks. All treat attention as per-token magnitudes: a token's score never depends on *who* attends to it. Retrieval heads (Wu et al., 2025) and RazorAttention (Tang et al., 2025) identify heads that copy from context; our coreference-head calibration is in this spirit.

Graph views of attention. Attention rollout and flow (Abnar & Zuidema, 2020) compose attention across layers for attribution; induction heads (Olsson et al., 2022) are an explicit two-hop circuit. CentralKV (Wu et al., 2026) first used within-layer centrality for eviction; GraphKV (Li et al., 2025) concurrently propagates over a KV-similarity graph. Neither characterises what the score computes nor tests the bridge hypothesis in a controlled setting. Katz (Katz, 1953), PageRank (Page et al., 1999) and HITS (Kleinberg, 1999) iterate to a fixed point on static graphs; the causal DAG admits a one-pass computation, and the seeded walk of §5.1 is a sparse, personalised variant.

Evaluation. LongBench (Bai et al., 2024), RULER (Hsieh et al., 2024) and multi-hop QA (Yang et al., 2018; Trivedi et al., 2022; Ho et al., 2020) measure end-to-end quality without exposing which tokens are necessary, and the lost-in-the-middle effect (Liu et al., 2024) confounds position with importance. BRIDGEBENCH controls both and measures the answer after the cache is actually pruned; HotpotQA's supporting facts and MuSiQue's decompositions give natural-text analogues.

3 PRELIMINARIES: THE CAUSAL ATTENTION GRAPH AND STREAM-INFLUENCE

Setting. A decoder-only Transformer processes tokens $x_1, \dots, x_L$ and stores a key–value pair per token, layer and head. Under a budget of $B < L$ entries, a *training-free eviction policy* keeps a set $\mathcal{C}_t \subseteq \{1, \dots, t\}$, $|\mathcal{C}_t| \le B$, dropping the position with the lowest *importance score*. As is standard, a *system window* of the first w_{sys} tokens (attention sinks) and a *local window* of the last w_{loc} tokens are always retained; policies differ only in how the remaining slots are ranked. $A_{t,j} \ge 0$ is the causal-softmax attention from query t to key $j < t$, averaged over the heads and layers a policy observes.

Baseline scores. H2O (Zhang et al., 2023) ranks by *accumulated attention* $H_j(t) = \sum_{t' \le t} A_{t',j}$; SnapKV (Li et al., 2024) by the maximum (pooled) attention from the last w queries; StreamingLLM (Xiao et al., 2023) by recency.

The causal attention graph. CentralKV (Wu et al., 2026) instead reads A as a weighted directed acyclic graph $G_t = (V_t, E_t)$: nodes are tokens, and there is an edge $t' \to j$ with weight $A_{t',j}$ whenever j is among the top-K attention targets of t',

$$\mathcal{N}(t') = \arg\operatorname*{top}_K \{A_{t',j} : j < t'\}, \qquad E_t = \{(t', j) : t' \le t,\ j \in \mathcal{N}(t')\}. \tag{1}$$

Sparsification is what makes the graph meaningful (a dense softmax row connects every earlier token); edges point backwards in time, so G_t is revealed in reverse topological order, one node per step.

Stream-Influence. The score of node j at time t, $S_j(t)$, is defined by the coupled recursion (Wu et al., 2026)

$$S_j(t) = \gamma\, S_j(t-1) + A_{t,j}\, \hat{S}(t)\, \mathbb{1}[j \in \mathcal{N}(t)]\,, \tag{2}$$

$$\hat{S}(t) = 1 + \lambda \max_{v \in \mathcal{N}(t)} S_v(t-1), \tag{3}$$

with decay $\gamma \in (0, 1)$ and recursion weight $\lambda \ge 0$; $S_t(t) = 0$ for the newest token. Unrolling equation 2,

$$S_j(t) = \sum_{t' \le t:\ j \in \mathcal{N}(t')} \gamma^{\,t-t'} A_{t',j}\, \hat{S}(t'), \tag{4}$$

so a token's score is the decayed attention it received over its *sparse* in-edges, each weighted by the $\hat{S}(t')$ of the sending query, i.e. the best influence path arriving at t', which j inherits by extending it. Storing $\tilde{S}_j = S_j/\gamma^t$ makes the update touch only K nodes (Appendix L). *Notation:* $B = \rho L$ is the budget, K the out-degree, $M(t) = \max_j S_j(t)$, and $\bar{A}_{t,j} = A_{t,j}\mathbb{1}[j \in \mathcal{N}(t)]$ the sparsified attention.

4 THEORY: STREAM-INFLUENCE AS CAUSAL HITS

Proofs are in Appendix A.

4.1 STREAM-INFLUENCE IS ONE-PASS CAUSAL HITS WITH MAX-AGGREGATION

HITS (Kleinberg, 1999) assigns every node an *authority* $a_j = \sum_{i \to j} h_i$ and a *hub* score $h_i = \sum_{i \to j} a_j$, solved by iterating to a fixed point with normalisation after each step. Define, for the causal attention graph, the authority of token j as $a_j(t) := S_j(t)$ and the hub score of query t as $h_t := \hat{S}(t)$.

Proposition 1 (Causal HITS)**.** *Stream-Influence is the unique solution of the system*

$$a_j(t) = \sum_{t' \le t} \gamma^{\,t-t'} \bar{A}_{t',j}\, h_{t'}, \qquad h_t = 1 + \lambda \max_{v \in \mathcal{N}(t)} a_v(t-1), \tag{5}$$

which differs from HITS in three ways: (i) the hub aggregator is $\max$ *rather than* $\sum$*; (ii) every query carries a unit source term, so* $h_t \ge 1$ *and authority is at least sparsified, decayed accumulated attention; (iii) the graph is a DAG revealed in reverse topological order, so the system is lower-triangular in time and is solved* exactly *by a single forward pass with* $O(K)$ *work per step, with no power iteration.*

With $\max$ in (i), authority propagates along a chain $b_1 \leftarrow b_2 \leftarrow \cdots$ multiplicatively by λ per hop instead of being averaged with a query's other targets, a tropical Katz centrality (Katz, 1953).

Corollary 2 (Heavy-hitter factorisation)**.** *Let* $\bar{H}^{\gamma}_j(t) = \sum_{t' \le t} \gamma^{t-t'} \bar{A}_{t',j}$ *be sparsified, decayed accumulated attention. Then* $S_j(t) = \bar{H}^{\gamma}_j(t) \cdot \alpha_j(t)$ *with* $\alpha_j(t) = \sum_{t'} \gamma^{t-t'} \bar{A}_{t',j} h_{t'} / \bar{H}^{\gamma}_j(t) \in [1,\, 1 + \lambda \max_{t' \le t} M(t'-1)]$.

Corollary 2 separates the two departures from H2O: *sparsification* ($A \mapsto \bar{A}$, with decay) and *hub amplification* α_j, the in-edge-weighted mean hub score of the queries that read j. Setting $\lambda = 0$ gives $\alpha \equiv 1$, a decayed, top-K-sparsified heavy-hitter score, which is our key ablation.

4.2 BOUNDEDNESS: WHEN THE RECURSION DIVERGES, AND A NORMALISED FIX

Proposition 3 (Divergence under a persistent heavy hitter)**.** *Suppose some token* s *satisfies* $\bar{A}_{t,s} \ge a$ *for all* $t > t_0$ *(e.g., an attention sink that is always in the top-*K *set). If* $r := \gamma + \lambda a > 1$ *then* $S_s(t) \ge a \sum_{i < t - t_0} r^{\,i} = \Omega(r^{\,t})$*, and for every other token* $S_j(t) = \Theta\big(\sum_{t' \le t} \gamma^{t-t'} \bar{A}_{t',j}\, r^{\,t'}\big) \propto \sum_{t' \le t} \bar{A}_{t',j}\, (r/\gamma)^{t'}$.

Wu et al. (2026) note the sufficient condition $\gamma + \lambda a < 1$, i.e. $a < (1-\gamma)/\lambda = 0.004$ at the defaults, argue that it holds *on average*, and pin the sinks. The condition is on the *maximum* persistent weight, and it fails on ordinary prompts: on 1.3k-token BRIDGEBENCH-REAL prompts with Qwen2.5-3B the mean per-row maximum attention over evictable tokens is 0.08 (sinks excluded), twenty times the threshold; the largest score reaches 10^6 within the prompt, and with the sink included (r=1.25) fp32 overflows after $\sim$400 tokens (Appendix E). Once some token dominates, a query's weight grows geometrically with its *position*: the score drifts towards sparsified attention from the most recent queries: an observation-window heuristic, not a centrality.

HITS normalises after every iteration; the recursion does not. The repair is the HITS normalisation, applied online:

$$h_t = 1 + \lambda \, \frac{\max_{v \in \mathcal{N}(t)} a_v(t-1)}{M(t-1)}, \qquad M(t-1) = \max_j a_j(t-1). \tag{6}$$

Proposition 4 (Bounded normalised Stream-Influence)**.** *With equation 6,* $h_t \in [1, 1+\lambda]$ *and* $S_j(t) \le (1+\lambda)/(1-\gamma)$ *for all* j, t *and every* (γ, λ)*; Corollary 2 holds with* $\alpha_j(t) \in [1, 1+\lambda]$*, and the update remains* $O(K)$ *per step since* M *is a running maximum of the stored invariant.*

4.3 SEPARATION ON BRIDGE CHAINS

Definition 5 (Bridge-chain graph). *Fix L, a chain weight $c \in (0,1)$ and bridge positions $b_1 < \cdots < b_n < L$ with $n \ll L$. Every query t spreads its attention uniformly over its predecessors, $A_{t,j} = 1/(t-1)$, except that each bridge b_k (and the final query L) places weight c on its chain predecessor (b_{k-1}, resp. b_n; b_0 is the sink) and spreads $1-c$ uniformly.*

The uniform background contains no information, yet it is not neutral for accumulated attention:

Lemma 6 (Causal mass accumulation). *In the bridge-chain graph the dense accumulated attention of a non-bridge token is $H_j(L) = \ln(L/j) + O(1/j) + O(n/L)$, and of a bridge is $H_{b_k}(L) = \ln(L/b_k) + c + O(1/b_k) + O(n/L)$.*

Early tokens receive mass from every later query simply because there are more later queries.

Theorem 7 (Separation). *Consider the bridge-chain graph with $K \geq 2$, pinned windows of size $w = w_{\text{sys}} + w_{\text{loc}} = o(L)$, and budget $B = \rho L$.*

(a) Sparsified centrality recovers every bridge. *For any $\lambda \geq 0$ and $\gamma \in (0,1]$ with $\gamma^L \geq \varepsilon > 0$, Stream-Influence (original or normalised) assigns each bridge a score $\geq c\varepsilon$ and each non-bridge, non-pinned token a score $\leq K(1+\lambda)/(L-1)$ under normalisation (resp. $O(K\hat{S}_{\max}/L)$). Hence for $L > K(1+\lambda)/(c\varepsilon) + w$ all n bridges rank above every background token and bridge recall is 1 at any budget $B \geq n + w$.*

(b) Accumulated attention loses deep bridges. *H2O retains bridge b_k iff the number of background tokens ranked above it, $\lfloor b_k e^{-c} \rfloor (1 + o(1))$, is below $B - w$, i.e. iff $b_k \leq e^c \rho L\,(1 + o(1))$. At budget $\rho < e^{-c}$ every bridge in the last $(1 - e^c\rho)$ fraction of the context is evicted, independently of L and of the model.*

Part (b) is tight on the synthetic graph (dashed line in Fig. 1c). Part (a) attributes bridge recovery to *sparsification*: a bridge is the only top-K target of its successor that carries non-background weight; Appendix A.7 shows (a) is robust to any background with per-row maximum $O(1/t)$ and that $\lambda > 0$ is needed to prefer a token read *only* by a bridge. Sparsification, however, separates bridges from *mass*, not from structurally identical tokens.

4.4 A SYMMETRY OBSTRUCTION, AND WHAT A SEEDED SCORE MUST DO

Definition 8 (Structurally equivalent distractors). *A context graph G has m* twin chains *if there are m disjoint vertex sets $C_1, \ldots, C_m$ and graph automorphisms π_i of $G \setminus \{q\}$ with $\pi_i(C_1) = C_i$, where q is the question (the final query tokens), and q's out-edges land in C_1 only.*

Call a score S *k-local in q* if S_v depends on the edges of q only through paths of length $\leq k$ from q to v. H2O, StreamingLLM and SnapKV are 1-local: a token's score changes with q only through the edge $q \to v$. So is Stream-Influence: q is the last node, and its hub score multiplies only its direct out-edges (Eq. 4).

Theorem 9 (Unseeded scores cannot select the chain). *Let G have m twin chains and let S be automorphism-invariant and 1-local in q. Then $S_v = S_{\pi_i(v)}$ for every $v \in C_1$ not adjacent to q and all i. Consequently, any policy ranking by S (ties broken independently of the chain) retains the non-anchor bridges of C_1 with the same probability as their twins, i.e. at most the fraction of the twin class that fits in the budget.*

The only asymmetry between the gold chain and its twins is the question's attention to the anchor, and a 1-local score cannot carry it further; BRIDGEBENCH-REAL (§5) instantiates exactly this structure.

Proposition 10 (Seeded sparse propagation recovers the chain). *Let the chunk graph $\bar{W}$ (Definition in §5.1) be such that, for every chunk u on the gold chain, the strongest symmetrised neighbour of u outside its own hop is the chain-adjacent chunk (*top-1 coreference*). Then the beam-1 walk seeded at q visits the h gold chunks in h hops and no other chunk, so any budget of $\geq h$ chunks retains the full chain. If top-b coreference holds only with probability p_b independently per link, the beam-b walk retains the chain with probability $\geq p_b^h$; dense propagation with leakage ε per hop gives the h-th chunk a weight of order $(1-\varepsilon)^h$ times the anchor's, so its retention decays with hop count while the beam walk's does not.*

Proposition 10 is the seeded analogue of Stream-Influence's max operator: sparse path extension keeps the signal on the chain, and it names the quantity to measure, the top-1 coreference accuracy of the heads used to build the graph, which our calibration step estimates.

5 BRIDGEBENCH AND SEEDED SPARSE PROPAGATION

BRIDGEBENCH is designed so that the tokens a policy *must* keep are known by construction and the structures of §4, bridge chains and twin chains, are present by design.

BRIDGEBENCH-SYN consists of five procedural families of row-stochastic causal attention matrices with ground-truth important sets: *linear bridge chain* (Definition 5), *multi-hop fork* (bridges also receive weak decoy attention), *needle-in-a-haystack*, *definition–use chain* (the heavy-hitter-friendly case) and *broad coverage*: a unit test for the *score*, independent of any model.

BRIDGEBENCH-REAL instances are lists of facts about $m{=}12$ fictional entity chains *company* $\rightarrow$ *founder* $\rightarrow$ *birthplace* $\rightarrow$ *province* $\rightarrow$ *capital*, rendered from four sentence templates, plus 12 fillers that mention chain entities without linking them; the 60 facts are shuffled and numbered. One chain is gold and a k-hop question ($k \in \{2, 3, 4\}$, "*In which province is the birthplace of the founder of Halcyon Instruments located?*") is appended with a one-shot example asking the model to spell out the chain (`founder: X; town: Y; province: Z`). Entity names are fictional, so answers cannot come from world knowledge. The instance exposes the *anchor* fact, the *bridge* entities, the k *evidence* facts, the answer, and $m{-}1$ twin chains indistinguishable from the gold chain except through the question (Definition 8). Prompts are $\sim$1.3k tokens.

Counterfactual protocol and metrics. We prefill with eager attention and harvest the attention matrix (mean over heads and layers for token-level policies; selected heads for the chunk graph). For each policy and budget we compute a keep set, *physically* delete the other entries from every layer of the KV cache, feed the final prompt token *after* pruning so that no full-context logits leak, and decode greedily; surviving keys keep their RoPE positions, so this is exactly eviction. The template (instruction, example, question) is pinned for every policy and the budget ρ applies to the fact list; four sink tokens and a 32-token local window are also pinned. Metrics: *EM* (the final field of the parsed chain contains the answer), reported relative to the same model's full-cache EM (*retention*); *chain integrity* (fraction of leading fields correct before the first error); *gold-chunk recall* (gold facts retained *entirely*); token-level *bridge*/*evidence recall*.

5.1 SEEDED SPARSE PROPAGATION

Theorem 9 says a working score must carry the question's signal several hops; Proposition 10 says it should do so sparsely over edges that encode coreference. This yields a three-step policy that we call *seeded sparse propagation* (SSP; Figure 2).

Chunks. The unit of eviction is a chunk (a fact line, a sentence or a fixed span), so that retained evidence is retained whole; q denotes the question chunk.

Coreference-head chunk graph. For heads $\mathcal{H}$, $W_{su} = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}\sum_{t\in s}\sum_{j\in u} A^{(h)}_{t,j}$ is the mass chunk s sends to chunk u ($W_{ss}{:=}0$). $\mathcal{H}$ is chosen *once per model* on three held-out synthetic instances: for every pair of facts sharing an entity we rank, per head, the earlier fact among all earlier facts by the mass the later fact sends to it, and keep the 32 heads with the highest mean reciprocal rank (best heads MRR$\approx$0.8, mean over all heads $\approx$0.37; Appendix D), as in retrieval-head calibration (Wu et al., 2025; Xiao et al., 2024); no question or gold label is used.

Seeded beam walk. Symmetrise $\bar{W} = W + W^\top$ (coreference is symmetric; the causal graph records only the later-to-earlier direction) and row-normalise to P. Start from $s^{(0)} \propto W_{q,\cdot}$, the question's own attention (alone, this is chunk-level SnapKV restricted to the coreference heads). For $h = 1, \ldots, H$: keep the b largest entries of $s^{(h-1)}$, zero the rest, step $s^{(h)} = P^\top s^{(h-1)}$, renormalise, and accumulate score $+{=}\ \lambda^h s^{(h)}$ ($H{=}4$, $b{=}2$, $\lambda{=}0.7$ throughout). A token's score is its chunk's score. The walk costs $O(HbS)$ for S chunks; the chunk graph is accumulated online from the same top-K edges Stream-Influence uses. SSP is a *seeded, chunk-level* instance of the causal-HITS view: the question is the sole hub with a source term and the beam plays the role of the max operator.

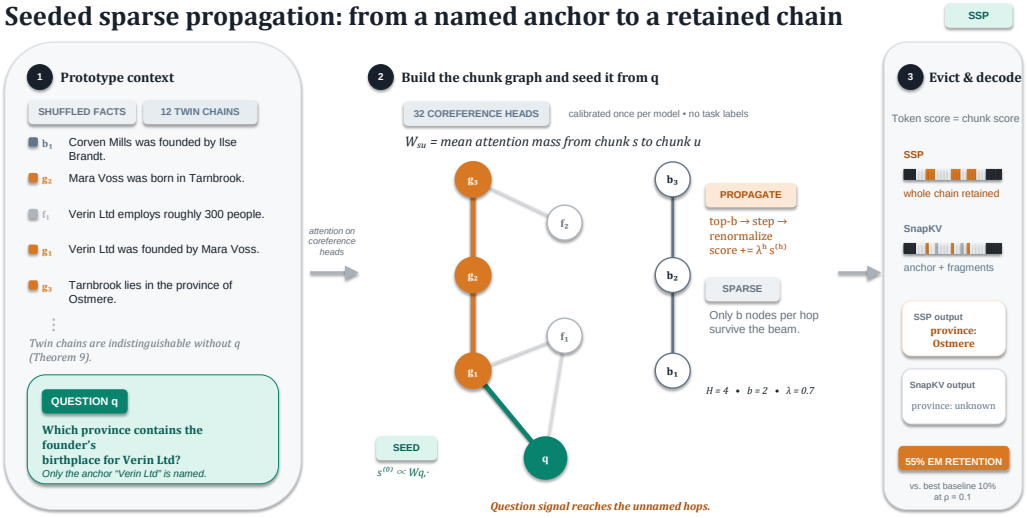

Figure 2: Seeded sparse propagation on a BRIDGEBENCH-REAL instance. The context holds structurally identical chains (gold $g_1g_2g_3$, twin $b_1b_2b_3$, fillers f) and a question q naming only the first hop, so no unseeded score separates g from b (Theorem 9). Coreference heads, calibrated once per model without labels, give a chunk graph W along shared entities; a walk seeded at q keeps a beam of b chunks per hop and discounts by λ^h, so mass reaches g_2, g_3 and not the twins. Chunk scores become token scores, the cache is physically pruned to ρL tokens plus the pinned instruction, question and local window, and the model decodes from what remains. Right: EM retention at ρ=0.1 (Table 1).

6 EXPERIMENTS

Setup. Models: Qwen2.5-3B/7B/14B-Instruct (Qwen Team, 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), bf16, greedy decoding; Qwen2.5-32B-Instruct on a reduced protocol in Appendix J. BRIDGEBENCH-REAL: 30 instances per hop count $k \in \{2, 3, 4\}$ (90 per model), budgets $\rho \in \{0.1, 0.2, 0.3, 0.5\}$ of the fact list. Policies: H2O, SnapKV (last 32 queries, pool 8), PyramidKV (Cai et al., 2024) (β=20), ChunkKV (Liu et al., 2025) (fixed 10-token chunks) and *ChunkKV-sent* (our sentence chunks, all heads), StreamingLLM, Random, three Stream-Influence variants (original, normalised and λ=0, with K=8, γ=0.998, λ=0.5 and sinks excluded as in Wu et al. (2026)) and SSP (H=4, b=2, λ=0.7, 32 heads) with its ablations *SSP-dense* (no beam), *SSP-1hop* (H=0: the question's attention on the coreference heads, chunk-wise) and *SSP-allheads* (all heads). SSP's hyper-parameters were fixed on nine prototype instances with disjoint seeds (baseline details: Appendix C). On BRIDGEBENCH-SYN (Appendix F) centrality recovers ≥ 0.97 of the important set on the chain-structured families where H2O reaches 0.42–0.68; H2O wins only on the heavy-hitter-friendly definition–use family, as Corollary 2 predicts.

6.1 BRIDGEBENCH-REAL: WHAT SURVIVES PHYSICAL EVICTION

Unseeded scores collapse together. At $\rho \leq 0.3$ H2O, StreamingLLM and all three Stream-Influence variants retain only 2–17% of full-cache accuracy (Table 1). The Stream-Influence variants are indistinguishable from random eviction (paired EM difference at ρ=0.1: -0.4 ± 1.6 points, 95% bootstrap CI over 450 prompts). This is Theorem 9 in action: the twelve chains are isomorphic, so no 1-local score can prefer the gold bridges, and every unseeded score retains 8–10% of bridge tokens at ρ=0.1, the random rate (Table 4); H2O's is lowest because it spends the budget on early facts (Lemma 6). Original and normalised Stream-Influence rank identically in practice: with sinks excluded the hub term is nearly constant, although the original score still grows as 1.25^t with the sink in the graph (Appendix E).

One hop finds the anchor, not the chain. SnapKV, ChunkKV and SSP-1hop are the 1-local scores that *use* the question; they retain the anchor fact (evidence recall 34–55% vs. 11% for random) and

Table 1: BRIDGEBENCH-REAL (five models, 90 instances each). **Left:** EM *retention* (%; EM under eviction / full-cache EM) at document budgets ρ, mean over $k \in \{2, 3, 4\}$. **Right:** retention and chain integrity by hop count at ρ=0.2 (>100: eviction removed interfering distractors).

Policy \ ρ	10%	20%	30%	50%
Random	2	8	12	29
StreamingLLM	5	10	17	46
H2O	7	10	16	40
SnapKV	10	23	34	49
H2O (p.h.)	8	21	39	61
ChunkKV	5	18	30	50
ChunkKV-sent	8	25	39	61
SI (orig.)	2	6	11	32
SI-λ0	2	8	11	33
SSP-1hop	34	45	58	74
SSP-allheads	15	28	43	66
SSP-dense	45	55	63	88
SSP (ours)	**55**	**81**	**94**	**98**
Full EM (abs.)	77	77	77	77

	2-hop		3-hop		4-hop	
Policy	Ret.	Chain	Ret.	Chain	Ret.	Chain
Random	10	8	8	4	5	4
StreamingLLM	11	17	14	18	5	17
H2O	20	10	6	2	3	3
SnapKV	34	60	23	50	9	29
H2O (p.h.)	37	32	19	24	5	16
ChunkKV	30	52	15	41	5	28
ChunkKV-sent	37	60	23	51	12	32
SI (orig.)	10	5	5	6	4	6
SI-λ0	14	5	7	3	3	5
SSP-1hop	89	86	19	60	20	39
SSP-allheads	55	68	21	55	3	30
SSP-dense	106	93	25	67	24	53
SSP (ours)	110	93	90	84	33	69

are the only baselines clearly above random. But the question names only the anchor, so SnapKV's chain integrity falls from 60% at k=2 to 29% at k=4 and its retention at ρ=0.1 is 10%. Chunking alone does not help: ChunkKV retains 5% and ChunkKV-sent (SSP's sentence chunks scored on all heads) 8% (-1 ± 3 vs. SnapKV); the same chunks scored on the coreference heads (SSP-1hop) reach 34% ($+20 \pm 4$): the heads, not the chunks, carry the question's signal. PyramidKV's schedule starves upper layers and lands at the random rate (1%; Appendix G).

Seeded sparse propagation retains the chain. SSP retains 55% of full-cache EM at ρ=0.1 and 81% at ρ=0.2 (mean over five models), against 10%/25% for the best baseline (SnapKV, resp. ChunkKV-sent) and 34%/45% for SSP-1hop; the paired gain over SnapKV is $+34 \pm 5$ EM points at ρ=0.1 and $+43 \pm 5$ at ρ=0.2. At ρ=0.3 and 0.5 SSP is at 94% and 98% of the full cache while no baseline exceeds 39% and 61%. Its gold-chunk recall is 78% at ρ=0.1, where about five facts must cover up to four gold facts. The remaining errors concentrate in k=4 at the smallest budgets (Table 1, right), where the budget barely covers the chain; chain integrity (69% vs. 29% for SnapKV) shows that SSP fails late, with the first fields right and a missing final fact yielding "`province: unknown`", whereas baselines fail at the first hop.

Ablations isolate the three ingredients. Removing coreference heads (SSP-allheads) drops retention at ρ=0.1 from 55% to 15%: the chunk graph loses its coreference edges. Removing the beam (SSP-dense) drops it to 45% (-7 ± 4 EM points, paired): mass leaks into fillers that mention chain entities, and the loss grows with hop count as Proposition 10 predicts: at ρ=0.2 SSP-dense's gold-chunk recall falls from 99% (k=2) to 72% and 56% (k=3, 4) while SSP stays at 100/90/74%. Removing propagation (SSP-1hop) drops retention to 34%. The ordering SSP > SSP-dense > SSP-1hop > SSP-allheads holds for every model (Appendix G).

Scale and per-head baselines. SSP's eviction quality is nearly model-independent: gold-chunk recall at ρ=0.1 is 74–82% across the five models, tracking the best coreference head's MRR at calibration (0.77–0.82); absolute EM under SSP grows with scale within the Qwen family (34/40/46% at ρ=0.1), while unseeded scores do not improve with scale at all (0–4% retention for every model): the obstruction is structural. Qwen2.5-32B confirms this (SSP 53%/69% at ρ=0.1/0.2, best baseline 12%/29%; Appendix J). SSP's scoring costs 12–21 ms per eviction decision, 14% of an SDPA prefill at L=4.6k (Appendix M). Per-(layer, KV-head) keep sets at the same total memory ("p.h."; Appendix G) help H2O at $\rho \geq 0.2$ and leave SnapKV unchanged or worse; neither escapes the obstruction ($\leq 8\%$ retention at ρ=0.1), and SSP stays $+35 \pm 5$ EM points above the best per-head baseline at ρ=0.1.

Sensitivity and context length. Table 16 (Appendix K) sweeps the SSP hyper-parameters on fresh instances (gold-chunk recall, no decoding). Every setting with a beam ($b \leq 3$), $H \geq 2$ and ≤ 64 heads lies within a few points of the default; the beam beats dense propagation at every H, all heads collapse to the SnapKV level, and fixed 16-token spans cost $\sim$25 points. From 12 to 48 chains ($L \approx$1.3k–4.6k) SSP stays at 80–85% recall and SnapKV at 32–35%.

Table 2: Natural-text transfer (300 questions per model, five models, coreference heads from BRIDGEBENCH-REAL *unchanged*). HotpotQA bridge questions: F1 retention (%) and recall of the two gold sentences. MuSiQue-Ans ($k \in \{2, 3, 4\}$): F1 retention and recall of intermediate (bridge) and final answer tokens. Full-cache F1: 59 / 35. Full tables in Appendix G.

	HotpotQA (2-hop bridge)			MuSiQue (2/3/4-hop)			
	F1 ret.		Supp.	F1 ret.		Bridge	Answer
Policy	10%	20%	rec.@10%	10%	20%	rec.@10%	rec.@10%
Random	52	60	0	49	53	11	9
StreamingLLM	51	57	10	52	61	13	11
H2O	58	68	0	51	59	18	10
SnapKV	89	98	12	91	96	60	47
ChunkKV	85	96	18	89	97	55	44
ChunkKV-sent	82	93	47	79	94	45	42
SI-NORM	54	62	0	51	60	13	8
SSP-1hop	**96**	100	**74**	**94**	97	**68**	**61**
SSP (H=2)	92	**100**	67	90	**99**	64	56
SSP (ours)	89	98	64	89	98	62	55

6.2 TRANSFER TO NATURAL TEXT: HOTPOTQA AND MUSIQUE

BRIDGEBENCH-REAL is templated by design, so we run the identical protocol on HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022): Wikipedia paragraphs, sentence chunks, instruction, titles and question pinned, and each model's coreference heads from its BRIDGEBENCH-REAL calibration prompts (Table 2). *(i) The obstruction transfers.* H2O, StreamingLLM and Stream-Influence keep 0–18% of the gold sentences at ρ=0.1 and lie within 0–5 F1 points of random on both datasets (paired, 1500 questions each). *(ii) Seeded chunk selection transfers.* SSP-1hop keeps 74% of the HotpotQA supporting sentences where SnapKV keeps 12%, ChunkKV 18% and ChunkKV-sent 47%, and 68%/61% of the MuSiQue bridge/answer tokens against 60%/47% (SnapKV); it is the best policy at ρ=0.1 on both datasets ($+3.6 \pm 1.7$ F1 points over SnapKV on HotpotQA, $+5.4 \pm 1.5$ over the best baseline on MuSiQue): the coreference heads matter on natural text too, and the gap widens at ρ=0.02 ($+7.4 \pm 2.0$ points; Appendix H). PyramidKV is 3–7 F1 points below SnapKV. *(iii) Propagation is not needed here.* H=2 and H=4 cost 2.2 ± 1.2 and 3.8 ± 1.4 F1 points at ρ=0.1 relative to H=0 on HotpotQA and are within noise at ρ=0.2 (Tables 17, 18). This is the converse of Theorem 9: these questions name an entity of *every* hop, so the seed alone reaches the chain and extra hops spend budget on neighbours; propagation pays when the question under-specifies later hops, as in BRIDGEBENCH-REAL (+21 retention points). All three findings recur on four LongBench QA subsets at 4.5k–7.7k tokens (Appendix H). Conversely, on RULER variable tracking, whose question names only the first link of the chain, propagation pays again: at ρ=0.02 SSP retains 78% of full-cache accuracy against 61% for SSP-1hop and 49% for SnapKV (Appendix I).

7 CONCLUSION

The folklore about attention-graph centrality is half right. Stream-Influence is a causal HITS whose sparsification removes the causal-mass bias of accumulated attention and recovers every bridge of a single chain; but the published recursion diverges under attention sinks, and no unseeded score, centrality included, can tell *which* of several identical chains the question is about. Physical eviction on five models shows both, and the symmetry argument that explains the failure prescribes the remedy: propagate the question sparsely over coreference edges. Seeded sparse propagation is the only policy in our study that retains multi-hop evidence at a 10% budget, its seeded selection transfers to natural text, and propagation pays exactly where the question leaves later hops unnamed.
Limitations. BRIDGEBENCH-REAL and RULER-VT, the two settings where propagation helps, are synthetic; on natural QA ($\leq$7.7k tokens) the question names every hop. 10^5-token contexts and streaming eviction are untested. Theorem 9 assumes exact twins; near-symmetric distractors only bound an unseeded score's gain.

REPRODUCIBILITY STATEMENT

All statements in Section 4 are proved in Appendix A. BRIDGEBENCH is fully procedural (Appendix B: templates, entity pools, prompt format, seeds); models are public checkpoints (Ap-

pendix C); the eviction protocol, every baseline, the head calibration and the scripts producing every table and figure are in the anonymous supplementary code.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing and to draft sections of the paper, and as a coding assistant for the experimental scripts, the procedural generator of BRIDGEBENCH and the drawing script of Figure 2. We have not used generative AI tools for research ideation, for proving mathematical claims, or for retrieval and discovery of related work; the synthetic data in BRIDGEBENCH is produced by deterministic templates, not by a generative model. We have reviewed all AI-assisted work: all proofs were written and checked by the authors, LLM-assisted code was verified against unit tests and hand-computed cases, and all drafted text was edited and fact-checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Zefan Cai, Wen Xiao, Hanshi Sun, Cheng Luo, Yikai Zhang, Ke Wan, Yucheng Li, Yeyang Zhou, Li-Wen Chang, Jiuxiang Gu, Zhen Dong, Anima Anandkumar, Abedelkadir Asi, and Junjie Hu. R-KV: Redundancy-aware KV cache compression for reasoning models. *arXiv preprint arXiv:2505.24133*, 2025.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. *arXiv preprint arXiv:2407.11550*, 2024.

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*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Hyun-rae Jo and Dongkun Shin. A2SF: Accumulative attention scoring with forgetting factor for token pruning in transformer decoder. *arXiv preprint arXiv:2407.20485*, 2024.

Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5): 604–632, 1999.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, 2024.

Xuelin Li, Xiangqi Jin, and Linfeng Zhang. GraphKV: Breaking the static selection paradigm with graph-based KV cache eviction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

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*, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. In *Transactions of the Association for Computational Linguistics*, 2024.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, and Xiaowen Chu. ChunkKV: Semantic-preserving KV cache compression for efficient long-context LLM inference. *arXiv preprint arXiv:2502.00299*, 2025.

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*, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Danning Ke, Shikuan Hong, Yiwu Yao, and Gongyi Wang. RazorAttention: Efficient KV cache compression through retrieval heads. In *International Conference on Learning Representations*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, 2025.

Yusen Wu, Yinjun Huang, Jia Yee Tan, Hao Li, Rongfeng Guo, Liang Fan, Jiachen Luo, Guangyuan Dong, and Sike Xiang. CentralKV: Compressing long-context LLMs via attention-graph centrality. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL `https://openreview.net/forum?id=qUhv4vBrTR`.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

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*, 2023.

A PROOFS

A.1 PROPOSITION 1 (CAUSAL HITS)

Unroll equation 2: $S_j(t) = \sum_{t' \le t} \gamma^{t-t'} \bar{A}_{t',j} \hat{S}(t')$, which is the authority equation of equation 5 with $h_{t'} = \hat{S}(t')$; equation 3 is the hub equation. Uniqueness and the one-pass claim: order the unknowns $h_1, \{a_j(1)\}_j, h_2, \{a_j(2)\}_j, \ldots$. h_t depends only on $\{a_v(t-1)\}$, and $a_j(t)$ only on $a_j(t-1)$ and h_t, so the system is lower-triangular in this order and each unknown is determined by the previous ones; no iteration is needed. Per step the update touches the K entries of $\mathcal{N}(t)$ for the max and the K scatter-adds, i.e. $O(K)$ work, using the invariant $\tilde{S}_j = S_j/\gamma^t$ so that the decay is implicit (Wu et al., 2026). □

A.2 COROLLARY 2

Divide and multiply the unrolled sum by $\bar{H}^{\gamma}_j(t) = \sum_{t'} \gamma^{t-t'} \bar{A}_{t',j}$; $\alpha_j(t)$ is a convex combination of $\{h_{t'}\}$ and each $h_{t'} \in [1, 1 + \lambda M(t'-1)]$. □

A.3 PROPOSITION 3 (DIVERGENCE)

For $t > t_0$, $S_s(t) = \gamma S_s(t-1) + \bar{A}_{t,s} h_t \ge \gamma S_s(t-1) + a(1 + \lambda S_s(t-1)) = r S_s(t-1) + a$ since $s \in \mathcal{N}(t)$ contributes to the max in h_t. Induction gives $S_s(t) \ge a \sum_{i < t - t_0} r^i$. For the second claim, once $S_s(t-1) \ge S_v(t-1)$ for all v (which holds after finitely many steps because every other token receives weight at most $1-a$ per step and $r > 1$), $h_t = 1 + \lambda S_s(t-1) = \Theta(r^t)$, and substituting into the unrolled sum gives the second claim. □

A.4 PROPOSITION 4

With equation 6, $0 \le \max_v a_v(t-1)/M(t-1) \le 1$, so $h_t \in [1, 1+\lambda]$. Then $S_j(t) \le (1+\lambda) \sum_{t' \le t} \gamma^{t-t'} \bar{A}_{t',j} \le (1+\lambda) \sum_{i \ge 0} \gamma^i = (1+\lambda)/(1-\gamma)$ using $\bar{A}_{t',j} \le 1$. Tracking M: the stored invariant is $\tilde{S}_j = S_j/\gamma^t$; the ratio $\tilde{S}_v / \max_j \tilde{S}_j$ equals S_v/M, and $\max_j \tilde{S}_j$ is a running maximum because updates are additive (renormalisation rescales all entries by the same factor). □

A.5 LEMMA 6

For a non-bridge j, $H_j(L) = \sum_{t=j+1}^{L} A_{t,j}$ where $A_{t,j} = 1/(t-1)$ for non-bridge rows and $(1-c)/(t-1)$ for the $n+1$ chain rows. Hence $H_j(L) = \sum_{t=j+1}^{L} \frac{1}{t-1} - c \sum_{\text{chain rows } t > j} \frac{1}{t-1} = \ln(L/j) + O(1/j) + O(n/L)$, the last term because each chain row contributes at most $c/(t-1) \le c/(L/5)$ (bridges lie in $[L/5, 4L/5]$ in our generator). For a bridge b_k, the successor row adds exactly c. □

A.6 THEOREM 7

(b) By Lemma 6, a background token j out-ranks bridge b_k iff $\ln(L/j) > \ln(L/b_k) + c + o(1)$ iff $j < b_k e^{-c}(1 + o(1))$. The number of such tokens is $\lfloor b_k e^{-c} \rfloor (1 + o(1))$; H2O keeps b_k iff this is below the free budget $B - w = \rho L - o(L)$, giving the condition $b_k \le e^c \rho L (1 + o(1))$. With w_{sys} sinks pinned the count starts at w_{sys}, absorbed in the $o(1)$. *(a)* In a background row every predecessor has weight $1/(t-1)$; the top-K set is a tie-break among them, so each background token is a target of at most K rows in expectation under random tie-breaking, receiving at most $K \cdot \max_t 1/(t-1) \cdot h_{\max}$ of sparsified mass, i.e. at most $K(1+\lambda)/(L/5-1)$ after decay under normalisation (bridges lie beyond $L/5$), and $O(K h_{\max}/L)$ otherwise. A bridge b_k is the unique maximum of its successor's row (weight $c > 1/(b_{k+1} - 1)$), so it is in the top-K set for any $K \ge 1$ and receives at least $c\gamma^L \ge c\varepsilon$. For L large enough that $c\varepsilon > K(1+\lambda)/(L/5-1)$, every bridge out-ranks every background token; with n bridges and w pinned tokens any budget $B \ge n + w$ retains them all. We used $K \ge 2$ in the statement only so that the final query's row, which has *two* distinguished targets in the multi-hop-fork family, is covered by the same argument. □

A.7 Refinements of Theorem 7

Proposition 11 (Robustness of sparsification)**.** *If the background is not uniform but any distribution whose per-row maximum over non-chain tokens is at most $\beta/(t-1)$, part (a) of Theorem 7 holds with K replaced by βK. Conversely, if a distractor d receives weight $\geq c$ from a bridge, both scores retain d; sparsification cannot separate bridges from* structurally equivalent *distractors, only from mass.*

Proof. Replace $1/(t-1)$ by the bound $\beta/(t-1)$ on non-chain weights in the argument for (a); the background mass received by any non-bridge token is at most $\beta K(1+\lambda)/(L/5-1)$. For the converse, if a bridge row places weight $\geq c$ on a distractor d, then d is in the top-K set of that row and receives at least $c\gamma^L$ under Stream-Influence, and at least c under H2O; both scores rank d with the bridges. □

Proposition 12 (One-hop influence of λ)**.** *Let u be a token read only by b_1 (the first hop) with weight c', and v a token read only by a background query with the same weight. Then under normalised Stream-Influence, $S_u/S_v = h_{b_1}/h_{bg} \geq 1$, with equality iff $\lambda = 0$. Hence $\lambda > 0$ is necessary for Stream-Influence to distinguish tokens whose importance is only visible through the importance of their readers.*

Proof. u has a single in-edge from b_1 with weight c', so $S_u = \gamma^{L-b_1} c' h_{b_1}$; likewise $S_v = \gamma^{L-t_v} c' h_{t_v}$ for a background query t_v. Ignoring the decay difference (or taking $t_v = b_1$), $S_u/S_v = h_{b_1}/h_{t_v}$. Under normalisation $h_{b_1} = 1 + \lambda S_{b_0}(b_1 - 1)/M \geq 1$ where b_0 is the sink, and $h_{t_v} = 1 + \lambda \max_j S_j/M$ over the uniform-row top-K targets of t_v, which are background tokens with $S_j = O(K(1+\lambda)/L)$, so $h_{t_v} \to 1$. Equality iff $\lambda = 0$. □

A.8 Theorem 9

Let S be automorphism-invariant: $S_{\pi(v)}(G) = S_v(\pi^{-1}G)$ for every automorphism π of $G \setminus \{q\}$ extended by $\pi(q) = q$. Fix $v \in C_1$ not adjacent to q and an automorphism π_i with $\pi_i(C_1) = C_i$. The graph $\pi_i G$ differs from G only in the out-edges of q, which now land in C_i. Since S is 1-local in q, S_v depends on q's edges only through the edge (q, v), which is absent in both G and $\pi_i G$; hence $S_v(\pi_i G) = S_v(G)$. By invariance, $S_{\pi_i(v)}(G) = S_v(\pi_i^{-1}G) = S_v(G)$, the last step because $\pi_i^{-1}G$ again differs from G only at q's out-edges, none of which touch v. The retention claim follows: the m twins of v have equal scores, so a rank-based policy with chain-independent tie-breaking retains each with equal probability, which cannot exceed the fraction of the twin class that fits in the budget. □

Which scores are 1-local. H2O: $H_v = \sum_t A_{t,v}$ involves q only through $A_{q,v}$. SnapKV: max over the last w rows, likewise. StreamingLLM: position only. Stream-Influence (any variant): q is the final node; by equation 4 it contributes $\bar{A}_{q,v} h_q$ to v and nothing to other tokens; h_q depends on q's targets' authorities, which do not depend on q. Seeded propagation with H hops is $(H+1)$-local and escapes the theorem for $H \geq 1$.

A.9 Proposition 10

Under top-1 coreference, at hop h the beam-1 walk holds all its mass on the h-th gold chunk g_h; the step $P^\top s$ moves mass to the neighbours of g_h in proportion to $\bar{W}_{g_h,\cdot}$, and the next pruning keeps the largest, which is g_{h+1} by assumption (g_{h-1} is excluded as "its own hop" by the visited mask in the implementation, or by the assumption's exclusion clause). Thus the walk visits $g_1, \ldots, g_h$ and the accumulated score is positive exactly on these chunks (plus the seed's direct targets). Under independent per-link success with probability p_b for beam b, all h links succeed with probability $\geq p_b^h$. For dense propagation, if a fraction ε of the mass at each hop leaves the chain, the mass on g_h is at most $(1-\varepsilon)^h$ of the seed mass while the anchor keeps weight $\Theta(1)$ from the seed, so the h-th chunk's score relative to the anchor decays geometrically in h; the beam walk renormalises after pruning and does not decay. □

B BRIDGEBENCH DETAILS

Templates. Chain facts: *"{company} was founded by {founder}."*, *"{founder} was born in the town of {town}."*, *"The town of {town} lies in the province of {province}."*, *"The province of {province} is administered from the city of {capital}."* Fillers (eight templates) mention one or two entities of a randomly chosen chain without linking them to another entity, e.g. *"{founder} published a memoir in {year}."* Questions: k=2 *"In which town was the founder of {company} born?"*; k=3 *"In which province is the birthplace of the founder of {company} located?"*; k=4 *"From which city is the province that contains the birthplace of the founder of {company} administered?"*

Entities. Companies: 28 root words × 12 suffixes ("Halcyon Robotics"); founders: 24 first × 24 last names; towns: 32 roots × 12 suffixes ("Brightwick"); provinces: 8 prefixes × 28 roots × 6 suffixes ("Old Dalbymark"); capitals: 28 roots × 8 suffixes. Every root word is used at most once per document so that chains are not confusable by a shared substring.

Prompt. Chat template of the model; user turn = instruction (answer format `founder: <name> town: <name> ...`), a one-shot example with disjoint entities, the numbered fact list, the question.

Seeds. Evaluation seeds $1000 + 100k + i$; calibration seeds $9000 + 100k + i$; prototype seeds 5000–8000.

Synthetic families. Generators follow the descriptions in §5; L=1000, chain weight 0.5 (fork: 0.4), 5 seeds.

HotpotQA. Distractor-setting validation split (Yang et al., 2018), questions of type *bridge*, shuffled with seed 0 and the first 300 with prompts ≤ 4096 tokens taken. The user turn is an instruction ("Answer the question using the passages. Reply with only the answer (a short phrase, or yes/no)"), the ten paragraphs as "`[i] Title`" followed by one sentence per line, and "`Question: ...`". Chunks are the dataset's sentences; the instruction, titles and question are pinned, and ρ applies to the sentence tokens. Decoding is greedy for up to 24 tokens; EM and F1 use the SQuAD normalisation (lower-case, strip punctuation and articles). Supporting-fact recall is the fraction of the gold supporting sentences retained entirely. Coreference heads are the 32 heads selected on the three BRIDGEBENCH-REAL calibration instances of the same model; the walk uses the BRIDGEBENCH-REAL defaults (b=2, λ=0.7) with H=4 and with H=2.

MuSiQue. MuSiQue-Ans validation split (Trivedi et al., 2022), answerable questions only, the first 100 per hop count $k \in \{2, 3, 4\}$ after shuffling with seed 0 whose prompts fit in 4096 tokens. Each example provides 20 paragraphs (k gold) and a question decomposition with one intermediate answer per hop; the prompt, chunking, pinning and decoding follow the HotpotQA protocol. Bridge recall is the fraction of tokens of the intermediate answers (hops $1..k-1$), located in their supporting paragraph, that survive eviction; answer recall is the same for the final answer.

LongBench. The four English QA subsets HotpotQA, 2WikiMQA, MuSiQue and Qasper (Bai et al., 2024), all 200 questions each, with the official prompts and generation lengths (32 tokens; 64 for Qasper) and the official metric (SQuAD F1, maximum over reference answers). Contexts are truncated in the middle to 7,500 tokens as in the official evaluation script, giving prompts of 4.5k–7.7k tokens. Chunks are regex sentences of the context (split at sentence-final punctuation and line breaks; 220–330 per prompt); the instruction and question are pinned, ρ applies to the context tokens. Attention is harvested with forward hooks and reduced layer by layer (mean matrix; chunk graph of the calibrated heads), so no $N_{\text{layers}}\, H\, L^2$ tensor is materialised. Since LongBench has no evidence annotation we report F1 retention and, for the questions whose reference answer occurs verbatim in the context, the fraction of the answer's tokens that survive eviction.

C EXPERIMENTAL SETUP

Public checkpoints from HuggingFace; transformers eager attention for prefill (to expose attention probabilities), SDPA for decoding. One NVIDIA GH200 (96 GB) or, for the 3B model, a 64-core CPU node (the protocol is deterministic and device-independent up to bf16/fp32 rounding). Prefill of a 1.3k-token prompt plus the 45 pruned decodes per instance takes $\sim$75 s for 3B on CPU and $\sim$30–90 s on the GPU depending on model size.

Sensitivity protocol. Table 16 uses fresh BRIDGEBENCH-REAL instances (seeds $3000+1000m+100k+i$, $i<10$, for $m \in \{12, 24, 48\}$ chains and $k \in \{2, 3, 4\}$; fillers equal chains, so $L \approx 1.3$k, 2.4k and 4.6k tokens). The attention of one prefill is harvested once and every configuration is scored by gold-chunk recall at the given budget without decoding; "all heads" uses the mean chunk graph over every head. Fixed 16-token chunks re-segment the fact region into consecutive 16-token spans regardless of sentence boundaries.

D COREFERENCE HEADS

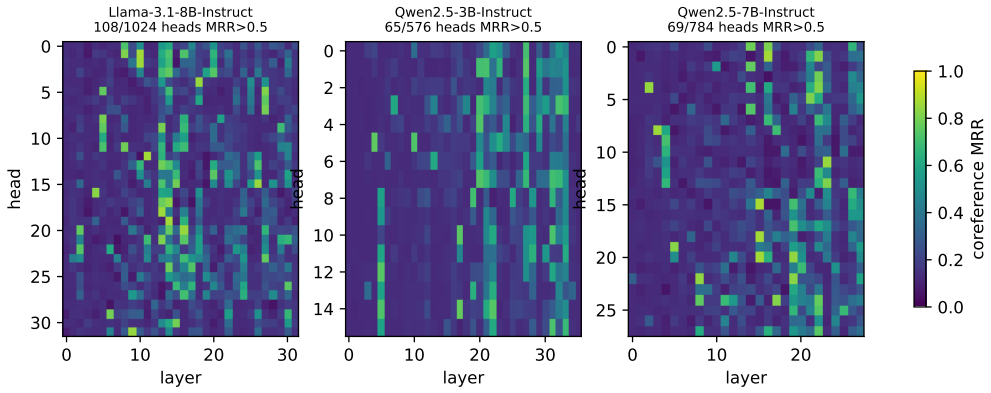

Figure 3: Per-head coreference MRR on three calibration instances for Qwen2.5-3B/7B and Llama-3.1-8B. A minority of heads (66/576, 69/784 and 108/1024 with MRR>0.5) carry the coreference signal; the mean over all heads has MRR 0.37.

Head selection uses *all* entity-sharing fact pairs (gold and distractor chains) on the calibration prompts, so it does not use the question or the gold labels. Selected heads for Qwen2.5-3B concentrate in layers 4–5, 17–22 and 27–32.

E DIVERGENCE MEASUREMENTS

With sinks excluded the remaining dominant token is position 42, the newline closing the instruction; with a single mid-layer head instead of the head-averaged matrix the score stays bounded (≤ 20), because a single head's top-K set is less stable across queries. The normalised and $\lambda{=}0$ variants produce the same ranking on BRIDGEBENCH-REAL because, once sinks are excluded, $\max_{v \in \mathcal{N}(t)} S_v/M$ is close to 1 for nearly every query (the secondary sink is in almost every top-K set), so $h_t \approx 1 + \lambda$ is a constant factor.

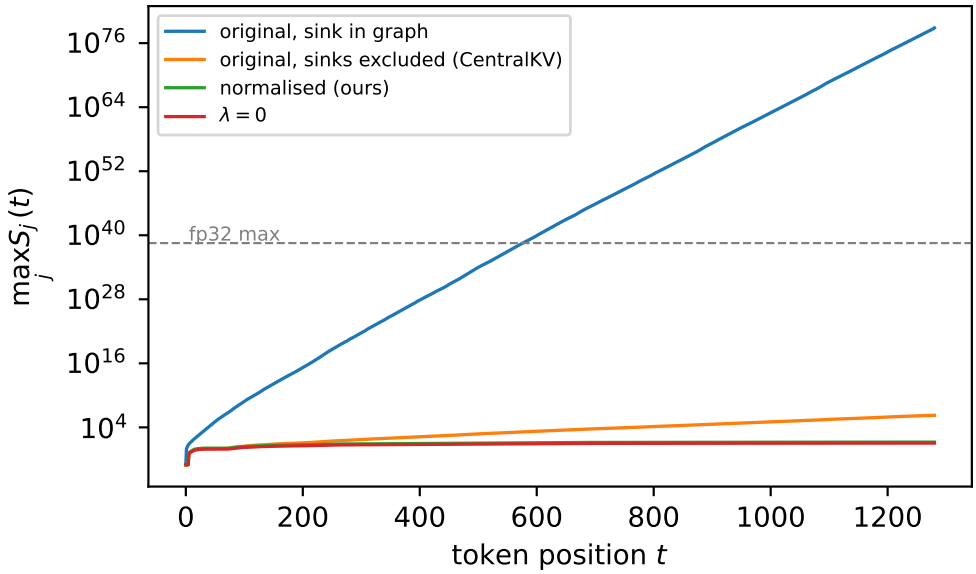

Figure 4: $\max_j S_j(t)$ of the original recursion versus t on a BRIDGEBENCH-REAL prompt (Qwen2.5-3B, mean attention over heads and layers): sink included ($r \approx 1.25$), sinks excluded (secondary sink), and the normalised recursion. Dashed: fp32 maximum.

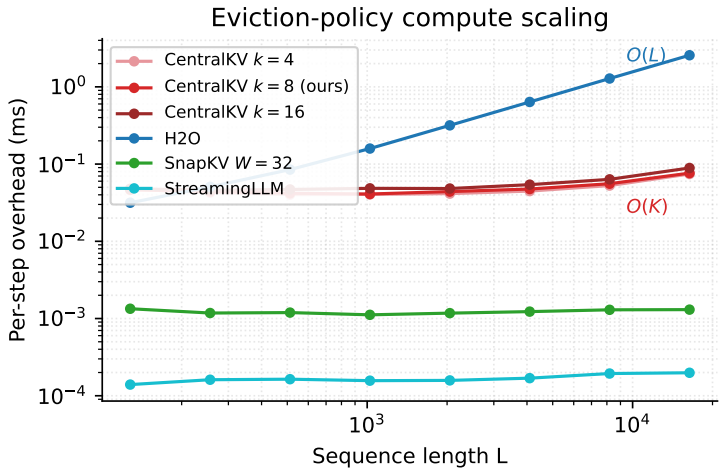

Figure 5: Per-step scoring cost versus context length (CPU, single stream). Stream-Influence is flat in L; H2O's dense accumulation grows linearly.

F SYNTHETIC GRAPHS: THE SCORE IN ISOLATION

Table 3: BRIDGEBENCH-SYN: recall of the ground-truth important set at a 20% budget (L=1000, mean of 5 seeds). Centrality is Stream-Influence with K=8.

Family	Centrality	H2O	SnapKV	StreamingLLM	Random
Broad coverage	0.91	0.25	**0.91**	0.24	0.35
Def–use chain	0.94	**0.97**	0.60	0.40	0.30
Linear bridge chain	**1.00**	0.42	0.42	0.27	0.37
Multi-hop fork	**1.00**	0.50	0.53	0.34	0.47
Needle-in-a-haystack	**1.00**	0.68	**1.00**	0.55	0.58

Table 3 reports recall of the ground-truth important set on BRIDGEBENCH-SYN at a 20% budget. Centrality is at or near 1.0 on the three families with chain structure (linear chain, multi-hop fork,

needle), where H2O ranges from 0.42 to 0.68 and SnapKV from 0.42 to 1.00 (SnapKV finds the needle, a one-hop pattern, but not the chain). On the definition–use family, the heavy-hitter-friendly case, H2O is best (0.97 vs. 0.94): when the important tokens *are* the heavy hitters, sparsification loses a little mass. On broad coverage the two scores that ignore position (centrality and SnapKV) tie at 0.91 while H2O's positional bias drags it to 0.25, below random.

G PER-MODEL RESULTS AND TOKEN-LEVEL RECALL

Table 4: BRIDGEBENCH-REAL at ρ=0.1: token-level bridge recall, evidence recall, whole-fact gold-chunk recall and chain integrity (%, mean over models). Random eviction retains a fraction ρ of bridges in expectation.

Policy	Bridge	Evidence	Gold chunk	Chain integrity
Random	10	11	0	2
StreamingLLM	9	8	7	9
H2O	8	11	0	2
SnapKV	32	34	25	30
H2O (per-head)	2	4	0	8
SnapKV (per-head)	5	5	3	18
PyramidKV	5	5	3	8
ChunkKV	27	28	15	24
ChunkKV (sentence chunks)	35	34	35	35
SI (original)	9	13	1	2
SI-NORM	9	11	0	2
SI-λ0	9	11	0	1
SSP-1hop (coref. heads, H=0)	56	55	57	54
SSP-allheads	33	31	32	32
SSP-dense	70	66	67	65
SSP (ours)	85	77	78	75

Table 5: BRIDGEBENCH-REAL per-model EM retention (%) at $\rho = 0.1$ and 0.2; full-cache EM in the header row.

	Qwen2.5-3B		Qwen2.5-7B		Mistral-7B		Llama-3.1-8B		Qwen2.5-14B	
Full EM	62		68		67		96		94	
Policy \ ρ	10%	20%	10%	20%	10%	20%	10%	20%	10%	20%
Random	2	5	3	7	3	12	0	5	2	9
StreamingLLM	7	9	3	10	5	13	3	7	5	11
H2O	5	7	8	15	8	8	8	13	6	8
SnapKV	9	16	7	18	8	25	14	31	11	25
H2O (per-head)	5	9	7	13	15	30	5	23	11	31
SnapKV (per-head)	0	4	2	7	8	22	7	20	6	25
PyramidKV	2	2	0	0	2	10	2	13	0	5
ChunkKV	2	11	5	20	7	20	6	17	5	20
ChunkKV (sentence chunks)	4	14	8	25	5	30	13	29	9	26
SI (original)	4	5	3	7	0	5	0	6	1	8
SI-NORM	4	4	3	5	3	15	0	6	1	7
SI-λ0	4	5	3	7	3	15	0	6	1	8
SSP-1hop (coref. heads, H=0)	20	29	28	44	48	58	36	44	36	51
SSP-allheads	7	16	11	26	25	42	15	34	14	25
SSP-dense	38	46	43	51	58	73	40	49	45	55
SSP (ours)	55	71	59	87	70	103	43	71	48	71

Table 6: HotpotQA distractor setting, bridge questions (300 per model, mean over five models): F1 retention at every budget and supporting-sentence recall. Coreference heads are transferred from BRIDGEBENCH-REAL without re-calibration.

Policy	F1 retention (%) at budget ρ				Supp.-fact recall (%)	
	10%	20%	30%	50%	10%	20%
Random	52	60	70	80	0	0
StreamingLLM	51	57	65	78	10	19
H2O	58	68	73	82	0	0
SnapKV	89	98	100	99	12	29
PyramidKV	77	88	93	97	3	10
ChunkKV	85	96	97	100	18	38
ChunkKV (sentence chunks)	82	93	98	100	47	73
SI-NORM	54	62	67	80	0	0
SSP-1hop (coref. heads, H=0)	**96**	100	**101**	101	**74**	**89**
SSP-dense	90	98	99	101	65	84
SSP (H=2)	92	**100**	100	**102**	67	88
SSP (ours)	89	98	100	102	64	85
Full-cache F1 / EM (abs.)	59 / 45	59 / 45	59 / 45	59 / 45	100	100

Table 7: HotpotQA (distractor, bridge questions): per-model F1 retention (%) at $\rho = 0.1$ and 0.2; full-cache F1 in the header row.

	Qwen2.5-3B		Qwen2.5-7B		Mistral-7B		Llama-3.1-8B		Qwen2.5-14B	
Full F1	51		60		51		70		65	
Policy \ ρ	10%	20%	10%	20%	10%	20%	10%	20%	10%	20%
Random	33	42	46	56	66	71	69	72	47	58
StreamingLLM	33	46	41	49	65	66	70	70	49	55
H2O	40	54	52	61	68	80	71	77	60	66
SnapKV	72	93	93	98	87	101	96	98	96	98
PyramidKV	41	66	78	88	85	94	94	99	86	94
ChunkKV	65	88	84	97	90	98	93	98	92	97
ChunkKV (sentence chunks)	70	88	80	92	86	95	88	96	85	95
SI-NORM	34	44	46	57	71	72	71	78	49	60
SSP-1hop (coref. heads, H=0)	90	93	96	103	99	102	96	99	97	100
SSP-dense	83	91	86	98	95	101	93	98	93	100
SSP (H=2)	82	97	88	100	97	103	95	99	96	99
SSP (ours)	78	94	87	99	95	103	93	98	92	98

Table 8: MuSiQue (100 questions per hop count, five models): F1 retention at every budget and token-level recall of the intermediate (bridge) and final answers at ρ=0.1, 0.2.

Policy	F1 retention (%) at ρ				Bridge recall		Answer recall	
	10%	20%	30%	50%	10%	20%	10%	20%
Random	49	53	61	74	11	21	9	20
StreamingLLM	52	61	66	79	13	23	11	21
H2O	51	59	66	74	18	32	10	20
SnapKV	91	96	98	100	60	76	47	65
PyramidKV	80	88	91	99	9	29	5	21
ChunkKV	89	97	99	100	55	74	44	63
ChunkKV-sent	79	94	97	100	45	68	42	62
SI-NORM	51	60	68	80	13	25	8	18
SSP-1hop	**94**	97	**100**	100	**68**	**82**	**61**	76
SSP-dense	90	97	98	100	62	77	55	70
SSP (H=2)	90	**99**	98	100	64	79	56	**77**
SSP (ours)	89	98	98	**101**	62	78	55	75
Full F1 (abs.)	35	35	35	35	100	100	100	100

Table 9: MuSiQue by hop count at ρ=0.2: F1 retention (%) and bridge-token recall (%), mean over five models.

	2-hop		3-hop		4-hop	
Policy	Ret.	Bridge	Ret.	Bridge	Ret.	Bridge
Random	57	23	59	20	45	20
StreamingLLM	62	27	67	21	56	21
H2O	63	29	65	34	51	35
SnapKV	97	90	92	71	97	69
PyramidKV	89	34	88	22	85	30
ChunkKV	94	88	96	66	101	69
ChunkKV-sent	91	80	97	60	96	64
SI-NORM	63	28	66	24	52	24
SSP-1hop	97	94	95	76	98	76
SSP-dense	95	91	100	69	97	71
SSP (H=2)	99	92	97	70	101	76
SSP (ours)	100	92	97	69	95	73

H LongBench QA subsets

Table 10: LongBench (four English QA subsets, 200 questions each, five models, official prompts and metric; protocol in Appendix B). F1 retention (%) at $\rho = 0.1$ and 0.2 of the context tokens, and the fraction of the reference answer's tokens that survive eviction at ρ=0.1 (questions whose answer occurs verbatim in the context: 73/90/66/42%). Contexts are 4.5k–7.7k tokens with 220–330 sentence chunks; coreference heads are transferred from BridgeBench-Real unchanged.

	HotpotQA			2WikiMQA			MuSiQue			Qasper		
	F1 ret.		Ans.	F1 ret.		Ans.	F1 ret.		Ans.	F1 ret.		Ans.
Policy	10%	20%	@10%	10%	20%	@10%	10%	20%	@10%	10%	20%	@10%
Random	76	83	11	78	85	10	55	71	10	62	70	12
StreamingLLM	75	79	2	77	81	3	52	57	2	57	65	4
H2O	88	93	49	82	84	31	77	85	34	70	79	18
SnapKV	**101**	**100**	**74**	99	99	71	**96**	99	57	93	**97**	61
ChunkKV	99	99	72	99	99	63	93	98	55	88	94	55
ChunkKV-sent	96	99	51	96	100	45	93	**99**	49	89	93	54
SI-norm	76	82	4	77	86	8	57	63	3	58	67	6
SSP-1hop	97	99	68	98	98	**71**	95	97	**60**	**95**	96	**72**
SSP (H=2)	97	98	61	**100**	**100**	66	92	96	56	94	95	68
SSP (ours)	97	99	58	98	100	64	91	96	53	93	96	67
Full F1 (abs.)	46		–	44		–	24		–	41		–

Table 11: LongBench: per-model F1 retention (%) at ρ=0.1 (Q = Qwen2.5, M = Mistral, L = Llama-3.1).

	HotpotQA					2WikiMQA					MuSiQue					Qasper				
Policy	Q3B	Q7B	M7B	L8B	Q14B	Q3B	Q7B	M7B	L8B	Q14B	Q3B	Q7B	M7B	L8B	Q14B	Q3B	Q7B	M7B	L8B	Q14B
Random	63	69	83	86	77	69	74	84	86	76	40	47	61	68	59	53	61	69	66	61
StreamingLLM	57	66	83	86	81	67	74	79	87	77	37	47	64	60	54	52	55	64	61	55
H2O	77	82	94	95	92	74	81	84	89	83	62	84	84	79	75	66	70	74	74	68
SnapKV	104	98	97	104	103	95	96	98	104	101	93	95	96	99	97	85	93	95	97	95
ChunkKV	101	96	98	103	99	94	103	94	102	101	89	99	95	89	95	79	89	85	94	91
ChunkKV-sent	90	91	96	104	99	96	97	93	97	95	103	92	87	94	92	78	89	93	95	88
SI-norm	61	70	84	85	78	70	71	81	87	78	41	55	63	70	56	51	57	65	63	56
SSP-1hop	92	97	96	103	98	95	98	97	101	100	98	102	89	93	94	83	93	97	101	100
SSP (H=2)	95	96	96	101	98	100	101	99	99	102	93	93	92	87	94	76	94	98	98	102
SSP (ours)	93	97	95	102	98	98	100	95	96	98	90	95	90	86	93	77	95	97	97	99
Full F1 (abs.)	41	49	41	49	52	37	42	37	47	57	18	23	21	29	31	37	43	35	46	45

Table 10 extends the natural-text transfer to LongBench contexts three to six times longer than the HotpotQA and MuSiQue prompts of §6. The picture is the same at this scale. *Unseeded scores stay at the random rate*: Stream-Influence-norm and StreamingLLM are within noise of, or below, random eviction on every subset (paired F1 difference to random at ρ=0.1: -0.4 ± 0.9 for Stream-Influence-norm over 4000 questions) and keep 2–8% of the answer tokens; H2O is $+4.0 \pm 0.9$ F1 points above random but 10–25 retention points below every seeded score. *Seeded scores are tied*: SnapKV, ChunkKV and SSP-1hop all retain 91–101% of full-cache F1 at ρ=0.1, and SSP-1hop is within noise of SnapKV (-0.3 ± 0.5 F1 points, 4000 questions) while keeping the most answer tokens on 2WikiMQA, MuSiQue and Qasper (71/60/72% vs. 71/57/61% for SnapKV). Propagation is again unnecessary (H=4 vs. H=0: -0.6 ± 0.5 F1 points at ρ=0.1, $+0.1 \pm 0.4$ at 0.2): LongBench questions name the entities of every hop, and Qasper questions are single-hop. The transferred coreference heads thus do no harm on single-document scientific text (Qasper: SSP-1hop 95% vs. SnapKV 93%, $+0.9 \pm 1.0$) and the chunk-level walk scales to 330 chunks at no measurable cost (Appendix M). The advantage of propagation is specific to the regime Theorem 9 describes and BridgeBench-Real instantiates, namely questions that name only the first hop of a chain hidden among structurally identical distractors, which none of the public benchmarks we know of isolates.

Low eviction budgets. At ρ=0.1 every seeded score is near the full-cache F1 on LongBench, so Table 12 lowers the budget to 2% and 5% of the context tokens (90–150 tokens at 2% on LongBench, kept in addition to the pinned instruction and question). The seeded scores now separate. SSP-1hop

Table 12: Low budgets $\rho \in \{0.02, 0.05\}$ on the same questions and models as Tables 6 and 10: F1 retention (%), mean over five models (HotpotQA-dist. is the distractor setting of §6; StreamingLLM and Stream-Influence-norm were not run there).

	HotpotQA-dist.		HotpotQA		2WikiMQA		MuSiQue		Qasper	
Policy \ ρ	2%	5%	2%	5%	2%	5%	2%	5%	2%	5%
Random	44	43	68	69	77	76	53	53	51	55
StreamingLLM	–	–	69	72	75	74	54	52	49	52
H2O	48	52	79	84	77	81	68	73	58	63
SnapKV	51	70	92	**99**	**94**	**97**	82	**91**	75	86
ChunkKV	54	71	92	98	84	94	80	88	68	81
ChunkKV-sent	53	66	86	93	82	93	74	85	68	81
SI-NORM	–	–	69	69	74	76	54	53	50	54
SSP-1hop	**64**	**84**	**92**	98	91	97	**84**	88	85	**93**
SSP (H=2)	61	79	86	95	86	92	70	83	**85**	88
SSP (ours)	60	78	84	93	85	91	67	78	83	88
Full F1 (abs.)	59		46		44		24		41	

is the best or tied-best policy on HotpotQA-distractor, MuSiQue and Qasper and retains $+2.5 \pm 0.8$ F1 points more than SnapKV pooled over 5500 paired questions at either budget; the gain is concentrated where the answer sentence is not lexically close to the question ($+7.4 \pm 2.0$ / $+7.9 \pm 2.0$ on HotpotQA-distractor, $+3.9 \pm 1.4$ / $+2.9 \pm 1.3$ on Qasper) and is within noise on LongBench HotpotQA, 2WikiMQA and MuSiQue. Against ChunkKV-sent, which scores the same sentence chunks with all heads, SSP-1hop gains $+4.6 \pm 0.7$ points on every dataset, so the coreference heads carry the signal at low budgets too. Propagation costs 2–3 points here (H=2: -1.9 ± 0.6, H=4: -2.6 ± 0.6 vs. H=0 at ρ=0.02), for the reason given above: with about a hundred tokens of budget the neighbours reached by the walk displace the answer sentence itself.

I PUBLIC BENCHMARKS WITH CHAINED FACTS: RULER-VT AND BABILONG

Two public benchmarks come closest to the regime of Theorem 9, a chain of facts whose later links the question does not name: RULER variable tracking (Hsieh et al., 2024) and BABILong (Kuratov et al., 2024). We run the identical eviction protocol on both, with each model's coreference heads transferred from BRIDGEBENCH-REAL unchanged.

RULER variable tracking. We use RULER's official generator (Hsieh et al., 2024): one chain of k+1 assignments (`VAR X = 12345`, `VAR Y = VAR X`, ...) inserted at random positions among 150 copies of the noise sentence (∼3.8k tokens, 154 sentence chunks), and the question "find all variables that are assigned the value 12345", which names the value of the first link only. We keep RULER's answer prefix as the start of the assistant turn and add the sentence "Answer with the variable names only, separated by commas" to the instruction, since the chat models otherwise write numbered explanations; with this prompt full-cache accuracy is 90–100% on all five models. Accuracy is the fraction of the k+1 gold variable names in the output (64 generated tokens), and "chain recall" the fraction of the k+1 assignment lines whose tokens survive eviction. 60 instances per $k \in \{2, 4\}$ and model.

Table 13 shows the same ordering as BRIDGEBENCH-REAL once the budget is tight. At ρ=0.02 and k=4, SSP retains 78% of full-cache accuracy against 49% for SnapKV ($+28.7 \pm 2.9$ points, 300 paired instances), 69% for ChunkKV ($+8.5 \pm 3.3$) and 61% for SSP-1hop ($+16.9 \pm 2.8$): propagation, not the seed alone, recovers the later links of the chain, and it keeps 82% of the assignment lines where SnapKV keeps 42%. Unseeded scores (H2O, StreamingLLM, Stream-Influence-norm) are again at or near the random rate, as the theorem predicts for a chain whose links are twins of one another. The effect is confined to the tightest budget: at ρ=0.05 every seeded policy exceeds 85%, and ChunkKV, whose fixed 10-token chunks happen to match the assignment lines, is 4.4 ± 2.1 points above SSP at k=4; at ρ=0.1 all seeded scores saturate. RULER-VT is thus an easier instance of the obstruction than BRIDGEBENCH-REAL: the noise is a single repeated sentence, so any query-

Table 13: RULER variable tracking (one chain of $k{+}1$ assignments among 150 noise lines, 60 instances per k, five models): accuracy retention (%) at $\rho \in \{0.02, 0.05, 0.1\}$ of the context tokens and recall of the assignment lines at $\rho{=}0.02$. At 2% the budget ($\sim$75 tokens) is about the size of the chain itself.

	2 hops				4 hops			
	Acc. retention			Chain	Acc. retention			Chain
Policy	2%	5%	10%	rec.@2%	2%	5%	10%	rec.@2%
Random	0	1	3	0	0	0	1	0
StreamingLLM	2	6	17	1	2	6	12	1
H2O	28	66	83	1	16	51	75	0
SnapKV	87	100	100	78	49	86	98	42
ChunkKV	91	**100**	**100**	76	69	**97**	**99**	43
ChunkKV-sent	78	99	100	77	50	85	99	46
SI-NORM	4	10	22	0	1	5	12	0
SSP-1hop	94	99	99	96	61	89	97	60
SSP ($H{=}2$)	97	99	99	**99**	69	90	96	72
SSP (ours)	**98**	100	99	**99**	**78**	93	97	**82**
Full-cache acc. (abs.)		98		100		98		100

Table 14: BABILong qa2/qa3 at 4k and 8k tokens (100 questions each, five models, official zero-shot prompt): accuracy retention (%) and recall of the bAbI fact sentences that survive eviction, at $\rho \in \{0.05, 0.1\}$. Full-cache accuracy is only 25–44%, so retention is noisy and bounded by the models' own reasoning; fact recall is the informative column.

	qa2 (2 facts), 4k				qa2 (2 facts), 8k				qa3 (3 facts), 4k				qa3 (3 facts), 8k			
	Acc. ret.		Fact rec.		Acc. ret.		Fact rec.		Acc. ret.		Fact rec.		Acc. ret.		Fact rec.	
Policy	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
Random	33	39	0	0	28	39	0	0	76	83	0	0	59	74	0	0
StreamingLLM	46	61	5	10	43	56	5	10	70	115	6	11	78	98	5	11
H2O	46	61	1	3	55	68	1	3	71	73	1	2	71	71	1	3
SnapKV	102	100	40	62	99	99	58	76	98	100	19	36	98	99	32	54
ChunkKV	99	101	33	53	99	101	47	66	93	99	17	31	93	99	26	45
ChunkKV-sent	93	103	40	63	102	103	63	82	95	92	16	30	93	96	29	48
SI-NORM	32	50	0	0	27	62	0	2	83	119	0	0	64	105	0	3
SSP-1hop	100	101	72	87	97	98	**84**	**94**	92	92	**41**	**62**	98	99	**59**	**77**
SSP ($H{=}2$)	103	102	**76**	**88**	102	100	82	92	93	88	40	58	88	101	56	71
SSP (ours)	103	100	76	88	98	100	81	90	91	85	40	57	86	95	55	69
Full-cache acc. (abs.)	44		100		40		100		25		100		26		100	

seeded score locates the assignment lines by their lexical form, and only when the budget is close to the chain size does the choice among them matter.

On BABILong (Table 14) the full-cache models themselves solve only 25–44% of the questions, so accuracy retention is close to 100% for every seeded policy and indistinguishable within noise (SSP-1hop vs. SnapKV: -0.8 ± 1.2 points over 2000 paired questions). The eviction decision is nevertheless visible in what survives: SSP-1hop keeps 72–84% (qa2) and 41–59% (qa3) of the bAbI fact sentences at $\rho{=}0.05$, against 19–58% for SnapKV and 16–63% for ChunkKV-sent ($+26.6 \pm 0.8$ and $+27.1 \pm 0.8$ recall points), while the unseeded scores keep 0–6%. As on LongBench, propagation does not add to this (-1.1 ± 0.9 recall points): the bAbI facts all name the person the question asks about, so the seed already reaches them.

J SCALE: QWEN2.5-32B

We repeat the BRIDGEBENCH-REAL, HotpotQA-distractor and RULER-VT experiments with Qwen2.5-32B-Instruct (64 layers, 40 heads; bf16 on one GH200), calibrating its coreference heads on the same three prompts (best head MRR 0.81, as for 14B). Its full-cache accuracy is much higher than the smaller models' (BRIDGEBENCH-REAL EM 99% vs. 77%, HotpotQA F1 75 vs. 59, VT 100%), so retention is measured against a stronger reference. Table 15 shows the three findings of

Table 15: Qwen2.5-32B-Instruct next to the mean of the five 3B–14B models of the main text, same instances, budgets and protocol. BB: BRIDGEBENCH-REAL EM retention (%) at ρ=0.1, 0.2 and gold-chunk recall at 0.1 (90 instances); HP: HotpotQA-distractor F1 retention and supporting-fact recall at ρ=0.1 (300 questions); VT: RULER variable tracking (k=4, 60 instances) accuracy retention at ρ=0.02, 0.05. Bold: best 32B value per column. SSP-dense was not run on VT.

	BB ret. 10%		BB ret. 20%		BB gold rec. 10%		HP F1 ret. 10%		HP supp. rec. 10%		VT ret. 2%		VT ret. 5%	
Policy	32B	3–14B	32B	3–14B	32B	3–14B	32B	3–14B	32B	3–14B	32B	3–14B	32B	3–14B
Random	2	2	8	8	0	0	48	52	0	0	0	0	0	0
StreamingLLM	2	5	11	10	7	7	43	51	10	10	3	2	7	6
H2O	10	7	13	10	0	0	63	58	0	0	25	16	52	51
SnapKV	12	10	25	23	31	25	93	89	15	12	75	49	98	86
ChunkKV	6	5	24	18	14	15	88	85	20	18	**77**	69	**99**	97
ChunkKV-sent	7	8	29	25	34	35	81	82	47	47	42	50	93	85
SI-NORM	2	2	9	7	0	0	54	54	0	0	2	1	7	5
SSP-1hop	42	34	54	45	70	57	**95**	96	**79**	74	59	61	93	89
SSP-dense	43	45	52	55	72	67	91	90	71	65	–	–	–	–
SSP (ours)	**53**	55	**69**	81	**83**	78	88	89	69	64	68	78	90	93
Full-cache (abs.)	99	77	99	77	100	100	75	59	100	100	100	98	100	98

§6 unchanged at this scale. *The obstruction does not weaken with scale*: H2O, StreamingLLM and Stream-Influence-norm retain 2–13% of BRIDGEBENCH-REAL EM and keep 0–7% of the gold chunks, and lie at the random rate on HotpotQA (H2O is +15 F1-retention points above random, StreamingLLM −5). *Seeded propagation retains the chain*: SSP keeps 53%/69% of EM at ρ=0.1/0.2 against 12%/25% for SnapKV and 7%/29% for ChunkKV-sent ($+40 \pm 11$ and $+43 \pm 11$ EM points over SnapKV, paired over 90 instances), with the same ablation ordering (SSP > SSP-dense ≈ SSP-1hop; $+11 \pm 8$ over SSP-1hop at ρ=0.1). Its gold-chunk recall is 83% (78% for 3B–14B). The remaining errors are again at k=4, where SSP's EM is 7%/20% at ρ=0.1/0.2 against 100% at k=2 and 50%/83% at k=3; because the 32B full cache solves 97% of the k=4 instances, the retention denominator is larger than for the smaller models and the 4-hop row pulls the mean down. *Seeded chunk selection transfers* to HotpotQA: SSP-1hop keeps 79% of the supporting sentences (SnapKV 15%, ChunkKV-sent 47%) and is within noise of SnapKV on F1 ($+1.4 \pm 3.2$) while $+10.5 \pm 3.9$ above ChunkKV-sent; propagation costs 5.0 ± 3.3 F1 points here, as for the smaller models.

RULER-VT is the one place where the 32B result differs. At k=4 and ρ=0.02, propagation still helps relative to the seed alone (SSP 68% vs. SSP-1hop 59%, $+8.7 \pm 5.6$), but SnapKV (75%) and ChunkKV (77%) are now above SSP (-7.3 ± 5.3 vs. SnapKV), whereas they were at 49% and 69% for the five smaller models. Two things change. SnapKV's token-level selection keeps 66% of the gold assignment tokens for 32B (46–56% for the smaller models) while keeping the same 42% of whole lines, and the 32B model reads the variable names off these fragments where the smaller models do not. SSP's walk, in turn, keeps 66% of the assignment lines for 32B against 70–99% for the others: the chain lines are twins of one another up to the variable names, and the 32B coreference heads spread more mass among them. At ρ=0.05 every seeded policy is above 90%. On BRIDGEBENCH-REAL, where the distractors are structurally identical chains with distinct entities, the gap does not shrink (Table 15, left).

K SSP SENSITIVITY AND CONTEXT LENGTH

Per-head baselines. The per-(layer, KV-head) variants of H2O and SnapKV were run on the GPU for all models; for Qwen2.5-3B, whose main run was on CPU (fp32), the full-cache EM differs slightly between the two runs (0.59 vs. 0.62), and we normalise the per-head rows by the main run's full-cache EM.

PyramidKV and ChunkKV. PyramidKV (Cai et al., 2024) follows the official budget schedule: with n dynamic slots per layer on average and β=20, layer ℓ of N receives $n_{\max} - \ell\,(n_{\max} - n_{\min})/(N-1)$ slots with $n_{\min} = n/\beta$ and $n_{\max} = 2n - n_{\min}$ (same total memory as a uniform budget); within a layer each KV head keeps its own top set under the SnapKV score. Cache entries are physically removed per layer; decoding attends to caches of different lengths per layer. ChunkKV (Liu et al., 2025) sums the observation-window attention (last 32 queries, mean over heads and

Table 16: Gold-chunk recall (%) of SSP under hyper-parameter changes and context length, on fresh BRIDGEBENCH-REAL instances (10 per hop count $k \in \{2, 3, 4\}$; mean over Qwen2.5-7B, Llama-3.1-8B and Qwen2.5-14B, the 48-chain row over the first two), measured without decoding. One factor is changed at a time from the default.

Configuration	ρ=10%	ρ=20%
Default (32 heads, b=2, H=4, λ=0.7, sentence chunks)	83	89
coreference heads: 4 / 8 / 16 / 64 / all	87 / 86 / 85 / 80 / 40	91 / 91 / 91 / 87 / 57
beam b: 1 / 3 / dense	85 / 81 / 70	88 / 89 / 78
hops H: 0 / 1 / 2 / 6	62 / 73 / 82 / 83	71 / 77 / 88 / 90
discount λ: 0.5 / 0.9 / 1.0	79 / 85 / 86	88 / 90 / 90
fixed 16-token chunks	55	66
SnapKV (token level)	34	41
Context length (m chains): 12 / 24 / 48		
SSP default	80 / 85 / 82	89 / 89 / 86
SnapKV	32 / 35 / 34	43 / 39 / 41

Table 17: HotpotQA: supporting-fact recall (%) of the seeded walk as a function of hops H, discount λ and beam b (mean over models and 300 questions each; no decoding). H=0 is SSP-1hop.

	ρ=10%, beam b =				ρ=20%, beam b =			
H, λ	1	2	3	dense	1	2	3	dense
H=0	74	74	74	74	90	90	90	90
H=1, λ=0.3	73	73	73	74	89	90	90	89
H=1, λ=0.7	67	68	69	72	88	89	89	89
H=1, λ=1.0	63	65	67	70	87	87	88	88
H=2, λ=0.3	73	73	73	73	89	90	90	89
H=2, λ=0.7	67	68	69	69	86	88	88	87
H=2, λ=1.0	62	65	67	66	84	86	87	85
H=3, λ=0.3	72	72	73	73	89	90	90	89
H=3, λ=0.7	64	65	66	67	85	87	87	86
H=3, λ=1.0	57	60	60	61	81	83	84	82
H=4, λ=0.3	72	72	73	73	89	90	90	89
H=4, λ=0.7	63	65	65	66	84	86	87	85
H=4, λ=1.0	56	58	58	57	80	81	83	78

layers) over fixed 10-token chunks (the paper's default) and keeps whole chunks; *ChunkKV-sent* replaces the fixed chunks by the sentence / fact chunks SSP uses, so that it differs from SSP-1hop only in scoring on all heads rather than on the calibrated coreference heads (and in using the trailing window rather than the question chunk as the seed). Both were run on the same instances as the main runs (Table 5 and the natural-text tables).

Table 18: MuSiQue: bridge-token recall (%) of the seeded walk as a function of hops H, discount λ and beam b (mean over models and 300 questions each; no decoding). As on HotpotQA, $H{=}0$ is best and the loss from propagation grows with H and λ: the question already names every hop.

	$\rho{=}10\%$, beam b =				$\rho{=}20\%$, beam b =			
H, λ	1	2	3	dense	1	2	3	dense
$H{=}0$	68	68	68	68	82	82	82	82
$H{=}1$, $\lambda{=}0.3$	67	67	67	68	82	82	82	82
$H{=}1$, $\lambda{=}0.7$	65	65	65	66	80	81	82	81
$H{=}1$, $\lambda{=}1.0$	64	64	64	65	79	80	81	80
$H{=}2$, $\lambda{=}0.3$	67	67	67	67	82	82	82	81
$H{=}2$, $\lambda{=}0.7$	62	64	64	64	78	79	80	80
$H{=}2$, $\lambda{=}1.0$	59	61	62	62	76	78	79	77
$H{=}3$, $\lambda{=}0.3$	67	67	67	67	81	82	82	81
$H{=}3$, $\lambda{=}0.7$	61	63	62	63	77	79	79	78
$H{=}3$, $\lambda{=}1.0$	57	59	59	57	73	76	76	74
$H{=}4$, $\lambda{=}0.3$	67	67	67	67	81	82	82	81
$H{=}4$, $\lambda{=}0.7$	60	62	62	62	77	78	78	77
$H{=}4$, $\lambda{=}1.0$	55	58	58	54	71	75	75	71

L Implementation notes

Stream-Influence stores $\tilde{S}_j = S_j/\gamma^t$ and renormalises every 1024 steps; the normalised variant additionally tracks the running maximum of $\tilde{S}$ over non-pinned positions. Scores are kept in fp64 on the host for the original recursion so that the ranking of divergent scores is exact; in fp32 the original recursion returns `inf` for the sink after $\sim$400 tokens, at which point its ranking is undefined. SSP's chunk graph is accumulated from the same top-K edge stream by adding $A_{t,j}$ to $W_{\mathrm{seg}(t),\mathrm{seg}(j)}$ for the selected heads; the walk is run once per eviction decision.

M Runtime and memory overhead

Table 19: Wall-clock cost of one eviction decision on BridgeBench-Real prompts (one GH200 GPU, bf16, mean over five prompts after warm-up). *Prefill*: one forward pass with SDPA kernels (no attention output) vs. the eager pass with `output_attentions` used by our protocol. *Attn. memory*: peak extra memory of the eager pass (all layers and heads) vs. the 32 calibrated head matrices SSP needs. *Scoring*: time to turn the harvested attention into a ranking (Stream-Influence with $K{=}8$ on the host as in the reference implementation; SSP: chunk graph of the 32 heads on the GPU, beam walk, chunk-to-token broadcast). *Prune*: gathering the kept entries of the KV cache.

		Prefill (ms)		Attn. mem. (GB)		Scoring (ms)					Prune
Model	L	SDPA	+attn	all	32 heads	H2O	SnapKV	ChunkKV	Stream-Influence	SSP	(ms)
Qwen2.5-7B	1275	46	87	2.8	0.1	0	1	1	75	12	1
	2355	80	191	9.5	0.3	0	2	2	143	14	1
	4590	150	450	35.6	1.3	1	4	4	297	21	1
Qwen2.5-14B	1275	92	186	6.6	0.1	0	1	1	76	12	1
	2355	160	416	22.1	0.3	0	2	2	149	13	1

Table 19 reports the cost of one eviction decision. Every attention-based policy shares the harvest: with the eager pass the prefill is 2–3$\times$ slower than SDPA and the full attention tensor ($N_{\text{layers}}\, H\, L^2$ in bf16) dominates memory (35.6 GB at $L{=}4.6$k for Qwen2.5-7B). This is a property of our counterfactual protocol, which needs identical inputs for every policy, not of SSP: SSP reads only the 32 calibrated heads (1.3 GB at $L{=}4.6$k; a fused kernel would emit their chunk-level sums directly, $O(S^2)$ memory), and its scoring (segment-matrix products for 32 heads plus a walk over an $S{\times}S$ matrix with $S \leq 97$ chunks) takes 12–21 ms, i.e. 14% of the SDPA prefill at $L{=}4.6$k and independent of the model size (the walk itself is $\sim$1.5 ms). H2O, SnapKV and ChunkKV cost ≤ 4 ms; Stream-Influence with $K{=}8$ costs 75–300 ms in the host reference implementation. Head calibration is a one-off cost of 6 s (7B) and 11 s (14B) for three prompts. Physical pruning is a gather of 1 ms. SSP adds no decoding cost: after eviction every policy decodes over the same number of cache entries.