ROTARY HEALING SPACE TOPOLOGY: HARDWARE-FUSED COORDINATE RE-MAPPING FOR EXTREME KV CACHE EVICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory constraints in autoregressive language models force the adoption of Key-Value (KV) cache physical truncation. While heuristic eviction methods successfully bound VRAM allocation, they introduce a deterministic generation collapse (e.g., infinite repetitive token loops or absolute syntactic disintegration) at extreme compression boundaries. We trace the root cause of this phase transition cliff to a topological fracture within Rotary Position Embeddings (RoPE). The mathematical failure does not originate from the physical memory voids. Instead, it arises because the underlying tensor framework forcefully re-indexes surviving tokens to contiguous physical positions to maintain memory continuity, thereby destroying the exact relative shift-invariance within the $(SO(2))^{d/2}$ rotary subspace. To resolve this coordinate misalignment without incurring the $\mathcal{O}(N)$ latency penalty of physical tensor modifications, we propose Rotary Healing Space Topology (RHST). RHST operates as a hardware-level coordinate re-mapping operator with an $\mathcal{O}(1)$ macroscopic memory I/O footprint. Implemented as an extension to fused attention kernels, RHST bypasses memory bandwidth congestion by mapping physical indices back to their absolute causal coordinates directly at the hardware register level. Empirical evaluation across multi-hop retrieval and mathematical reasoning validates this algebraic intervention. Under extreme physical truncation, application-layer baselines suffer from absolute reasoning failure (0.00% accuracy on complex multi-step reasoning in MATH-500). RHST mitigates catastrophic generation collapse and sustains structural reasoning stability, effectively acting as a topological safety net that unlocks extreme heuristic compression safely.

1 INTRODUCTION

Scaling autoregressive language models to process extended context windows introduces deterministic computational bounds (Vaswani et al., 2017; Tay et al., 2022). During the prefill and decoding phases, Key-Value (KV) cache allocation scales linearly with sequence length. For complex reasoning tasks operating on massive contexts (e.g., 100K+ tokens), this linear expansion inevitably saturates consumer-grade GPU memory limits, creating a scenario where the physical KV cache must be aggressively pruned to operate alongside the base model parameters.

Current micro-architectural mitigation strategies execute sequence-wise physical eviction to bound resident memory. Frameworks evaluate structural priors (e.g., StreamingLLM's static sink and rolling windows (Xiao et al., 2024)) or attention distributions (e.g., SnapKV (Li et al., 2024)) to retain critical tokens, physically discarding the intermediate variables. While existing literature notes long-range dependency loss under heuristic eviction, we quantify a deterministic generation collapse at extreme compression boundaries (e.g., the 80%–90% phase transition cliff). Under these constraints, models lose causal tracking, resulting in factual divergence and absolute reasoning termination (manifesting as infinite repetitive loops). Attempting to correct this state misalignment at the application layer via physical tensor re-rotations (e.g., forcing continuous coordinate spaces by physically concatenating disjoint tensor chunks via high-level APIs) incurs an $\mathcal{O}(N)$ computational penalty. These memory-bound operations force macroscopic High Bandwidth Memory

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.

(HBM) read-write cycles, introducing severe memory bandwidth congestion and latency spikes during the autoregressive generation cycle.

We isolate the root cause of this deterministic failure to the topological fracture of Rotary Position Embeddings (RoPE) (Su et al., 2024). Standard attention mechanisms assume a continuous spatial manifold to maintain exact relative shift-invariance. The mathematical failure does not arise from the physical memory voids left by evicted tokens. Rather, it originates from the underlying tensor framework's necessity to maintain contiguous memory layouts. When surviving tokens concatenate, the system forcefully re-indexes them with contiguous physical positions (k), abandoning their original absolute causal coordinates (n). The attention operator subsequently computes relative rotational shifts using these false contiguous indices, generating an algebraic state discontinuity, as illustrated in Figure 1. Through targeted ablations, we prove that correcting only this hardware-level absolute coordinate mapping ($k \rightarrow n$)while keeping physical information loss identicaleliminates the reasoning collapse, strictly differentiating coordinate misalignment from pure information loss.

To resolve this algebraic discontinuity without invoking the macroscopic HBM bandwidth penalties of physical tensor re-allocation, we formulate Rotary Healing Space Topology (RHST). RHST functions as a hardware-fused coordinate re-mapping operator. Operating independently of specific token retention policies, RHST establishes a deterministic $\mathcal{I}_{offset}$ pointer logic directly within the fused attention kernel. By mathematically bridging the contiguous physical indices and the true causal coordinates at the hardware level, RHST restores exact relative phase geometry, effectively trading abundant ALU compute cycles for constrained memory bandwidth, providing a safety mechanism that makes extreme physical eviction viable.

Our core contributions are formalized as follows:

- We mathematically isolate the framework-level coordinate re-indexing error from pure information loss, defining the exact causal dependency between RoPE topological fractures and deterministic generation collapse in physically compressed KV cache regimes.
- We introduce RHST, a hardware-level algebraic operator with an $\mathcal{O}(1)$ macroscopic memory I/O profile. Embedded within fused attention kernels, it preserves exact relative shift-invariance and bypasses the $\mathcal{O}(N)$ VRAM re-rotation bottlenecks inherent in existing baseline strategies.
- We systematically evaluate the phase transition cliff across multi-hop retrieval and mathematical reasoning configurations. We prove that RHST unlocks extreme heuristic compression safely, mitigating catastrophic generation collapse and sustaining structural reasoning stability where standard pipelines exhibit absolute failure.
- We demonstrate the cross-architecture plug-and-play feasibility of the $\mathcal{I}_{offset}$ operator across widely deployed standard RoPE-based GQA/MHA architectures (e.g., Llama-3 (Grattafiori et al., 2024) and Qwen (Bai et al., 2023)). **To guarantee full reproducibility, the exact Triton (Tillet et al., 2019) kernel configurations and evaluation pipelines are documented in the Appendix.**

2 Related Work

2.1 Physical Truncation and Heuristic Eviction

The memory footprint of autoregressive generation scales linearly with sequence length. To enforce a bounded memory envelope, sequence-wise heuristic eviction methods selectively discard historical tokens based on attention score accumulation (e.g., Li et al., 2024; Zhang et al., 2023; Liu et al., 2023; Cai et al., 2025). While these retention protocols successfully restrict VRAM allocation to predefined limits, they execute physical tensor truncation without coordinate compensation. This physical deletion forces the computational framework to re-index the remaining Key-Value tensors, uncoupling them from their original positional sequences and producing an algebraic state discontinuity in the underlying RoPE coordinate system.

Furthermore, physical truncation addresses memory bottlenecks orthogonally to numerical precision reduction or channel pruning (e.g., INT4 or FP8 KV quantization; Liu et al., 2024c; Hooper et al., 2024; Liu et al., 2024a). When quantized caches are subjected to physical token eviction to meet

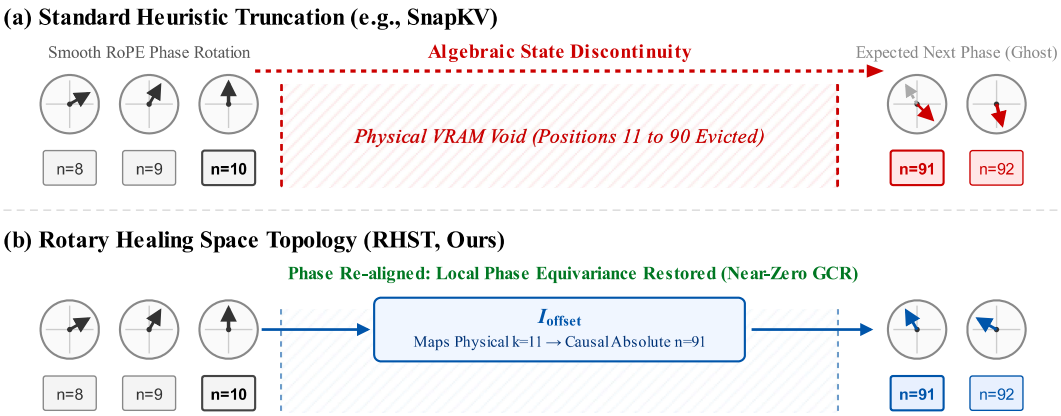

Figure 1: **Topological Fracture vs. Algebraic Healing in KV Cache Compression.** **(a)** Standard truncation methods physically evict tokens based on heuristic scores. To maintain tensor continuity, the framework forcefully re-indexes surviving tokens, severing the relative shift-invariance within the $(SO(2))^{d/2}$ rotary subspace. This uncompensated re-indexing introduces an algebraic state discontinuity, resulting in deterministic generation collapse. **(b)** The RHST framework injects a deterministic $\mathcal{I}_{offset}$ operator directly into the fused attention kernel. By mathematically mapping the contiguous physical memory index (k) back to its original causal absolute coordinate (n), RHST restores exact relative shift-invariance independent of the physical cache layout.

strict constraints, the identical coordinate fracture occurs, necessitating an independent topological repair.

2.2 Coordinate Repair Mechanisms and Execution Boundaries

Subsequent frameworks attempt to resolve context scaling via heterogeneous paradigms. A prominent trajectory involves application-layer tensor manipulation. For instance, StreamingLLM (Xiao et al., 2024) maintains a continuous coordinate space via static sinks and rolling windows, which in standard high-level APIs (e.g., PyTorch) often mandates physical tensor concatenation. Similarly, DapQ (Tian et al., 2026) and DynamicKV (Zhou et al., 2025) construct position-aware pseudo-queries or dynamic retention schedules during prefill to execute static pruning and re-weight token distances. Unlike frequency scaling mechanisms and alternative position encodings (e.g., PI, Chen et al. 2023; YaRN, Peng et al. 2026; LM-Infinite, Han et al. 2024; LongRoPE, Ding et al. 2024; ALiBi, Press et al. 2022) that alter the base RoPE frequencies or introduce linear distance penalties to accommodate length extrapolation, these physical retention interventions attempt to reorganize the physical attention layout directly.

While these application-layer interventions stabilize standard output syntax and prevent absolute format disintegration, their execution stack introduces a distinct micro-architectural bottleneck. Resolving topological gaps by manipulating physical tensors necessitates a sequence of High Bandwidth Memory (HBM) transactions: the framework must issue read instructions for disjoint tensor chunks, route them through the internal memory fabric, and execute a physical memory allocation to construct a newly contiguous Key-Value tensor. This macroscopic data movement mandates an $\mathcal{O}(N)$ computational and bandwidth penalty. Consequently, standard rolling window or re-rotation implementations inevitably generate HBM bandwidth congestion and latency overhead during the autoregressive decoding cycle.

Diverging from application-layer tensor modifications, our work isolates coordinate re-alignment at the micro-architectural level. The proposed Rotary Healing Space Topology (RHST) functions exclusively as a hardware-level coordinate re-mapping operator with an $\mathcal{O}(1)$ macroscopic memory I/O profile. This hardware-fused operator bypasses the VRAM bandwidth bottleneck completely, executing phase alignment without incurring the macroscopic HBM read-write penalties inherent to physical tensor modifications, remaining entirely policy-agnostic.

3 METHODOLOGY

3.1 CONSTRAINED TOKEN RETENTION FRAMEWORKS AS HOST BEDS

Current compressed inference systems separate into scoring heuristics and physical execution. We define the historical Key-Value cache sequence as $\mathcal{C}_t$. A discrete policy mask $\mathcal{A} \in \{0,1\}^L$ determines token retention ($a_i = 1$) or physical eviction ($a_i = 0$). This physical eviction typically executes at the boundary of the prefill phase. To evaluate the $\mathcal{I}_{offset}$ operator, we deploy existing heuristic strategies (e.g., SnapKV; Li et al., 2024) or Score-based Bounded Retention Networks as host experimental beds to induce extreme physical truncation. The topological fracture caused by the physical execution of $a_i = 0$ requires an independent algebraic resolution during the subsequent autoregressive decoding steps.

3.2 GEOMETRIC BOUNDARIES OF RELATIVE SHIFT-INVARIANCE

Rotary Position Embedding (RoPE; Su et al., 2024) encodes positional indices by rotating query and key representations via a sequence of base frequencies θ_j. The attention inner product relies on the relative phase difference $e^{i(m-n)\theta_j}$, where m and n are the absolute causal positions.

When physical eviction compacts the KV cache to maintain tensor continuity, the framework forcefully re-assigns a retained token originally at causal absolute position n to a new contiguous physical memory index k. Standard forward passes compute the relative distance as $m - k$, overwriting the required causal phase $e^{i(m-n)\theta_j}$ and generating an algebraic state discontinuity.

Proposition 1 (Exact Relative Shift-Invariance). Let $\mathcal{I}_{offset}$ be a deterministic coordinate mapping operator. By redefining the positional index as $n = \mathcal{I}_{offset}(k)$ prior to the RoPE phase injection, the hardware computes the rotational angle as $(m - \mathcal{I}_{offset}(k))\theta_j \equiv (m-n)\theta_j$. This algebraic operation strictly preserves exact relative shift-invariance within the $(SO(2))^{d/2}$ manifold, preventing out-of-distribution mathematical states in the attention logits without attempting to recover the physically evicted information.

Proof of Exact Phase Restoration. Without loss of generality, considering a single frequency component $\theta \in \Theta$, in standard RoPE formulations, the attention score between a query vector q_m at absolute causal position m and a key vector k_n at absolute causal position n relies on the inner product $\langle f_q(q_m, m), f_k(k_n, n)\rangle$. Mapping this to the complex plane, the transformation evaluates strictly to:

$$\mathrm{Re}(q_m k_n^* e^{i(m-n)\theta})$$

Under sequence-wise physical truncation without an independent topological repair mechanism, the evicted sequence creates a contiguous physical memory buffer. A surviving key, originally generated at causal position n, is sequentially shifted to a new physical index k (where $k < n$). The standard tensor framework implicitly utilizes this physical index, computing a corrupted relative phase:

$$\mathrm{Re}(q_m k_n^* e^{i(m-k)\theta})$$

This substitution generates the unrecoverable algebraic state discontinuity. By defining the hardware-fused operator $\mathcal{I}_{offset}(k) \mapsto n$, the arithmetic logic unit dynamically intercepts the coordinate injection sequence. The memory access primitive fetches the physical tensor at index k, but the rotational kernel is explicitly supplied with the transformed delta $(m - \mathcal{I}_{offset}(k))\theta$. Since $\mathcal{I}_{offset}(k) \equiv n$ by deterministic construction, the phase term resolves identically to $e^{i(m-n)\theta}$. This operation mathematically preserves the topological manifold geometry algebraically, remaining completely agnostic to the underlying physical memory fragmentation. ■

3.3 HARDWARE-FUSED $\mathcal{I}_{offset}$ IMPLEMENTATION AND EXECUTION STACK TAXONOMY

To rigorously define the micro-architectural scope of RHST and eliminate structural conflation with application-layer mechanisms, we must delineate the memory hierarchy layer where the topological restoration occurs. Diverging entirely from VRAM-level tensor manipulations, RHST operates exclusively within the Streaming Multiprocessor (SM) registers.

The physical Key-Value tensors remain disjoint and unaltered in HBM; no tensor concatenation or memory reallocation is executed at the application layer. Instead, the topological gap is resolved

Algorithm 1 Hardware-Fused $\mathcal{I}_{offset}$ Execution in Triton Attention

Require: Unrotated Query $Q_{raw} \in \mathbb{R}^{1\times d}$, Pre-RoPE Retained Keys $K \in \mathbb{R}^{L_{retained}\times d}$, Values $V \in \mathbb{R}^{L_{retained}\times d}$

Require: Absolute coordinate mapping array $\mathcal{I}_{offset} \in \mathbb{N}^{L_{retained}}$ pinned in SRAM

Require: Base frequencies $\Theta = \{\theta_j\}_{j=0}^{d/2-1}$

Initialization (Per-Thread Block):
Compute current query absolute position m
Load block of K, V from HBM to SRAM
Coordinate Re-mapping & Phase Assembly:
for each physical index k in current block **do**
  $n \leftarrow \texttt{tl.load}(\mathcal{I}_{offset} + k)$ $\quad\mathcal{O}(1)$ *indirect memory access*
  for $j = 0$ to $d/2 - 1$ **do**
    $\Delta_{m,n} \leftarrow (m - n) \times \theta_j$ *Exact relative phase computation*
    $\tilde{K}_{k,j} \leftarrow K_{k,j}\cos(\Delta_{m,n}) - K_{k,j+d/2}\sin(\Delta_{m,n})$
    $\tilde{K}_{k,j+d/2} \leftarrow K_{k,j+d/2}\cos(\Delta_{m,n}) + K_{k,j}\sin(\Delta_{m,n})$
  end for
end for
Attention Computation:
$S \leftarrow Q_{raw}\tilde{K}^\top/\sqrt{d}$
$O \leftarrow \texttt{Softmax}(S)V$
Return: Output vector O written to HBM

dynamically via indirect addressing. To execute $\mathcal{I}_{offset}(k)$ without the $\mathcal{O}(N)$ latency overhead inherent to global tensor re-rotations, we bypass standard VRAM indexing. The 1D integer pointer array defining $\mathcal{I}_{offset}$ is computed once by the host CPU at the prefill boundary. This metadata incurs a negligible spatial overhead of $\mathcal{O}(L_{retained})$ (occupying mere kilobytes per sequence, which is orders of magnitude smaller than the KV tensors) and is asynchronously mapped at the hardware level.

We fuse this operator directly into the attention mechanism using a custom Triton (Tillet et al., 2019) kernel, fully compatible with modern fused attention pipelines (e.g., FlashAttention and FlashAttention-2; Dao et al., 2022; Dao, 2024). During the block-wise computation phase of the decoding steps, the SM executes a `tl.load` primitive to fetch the absolute causal coordinate n from the $\mathcal{I}_{offset}$ array pinned in the Shared Memory (SRAM). Rather than allowing the host framework to re-encode the cached Keys via contiguous re-rotation, the accurate RoPE relative phase angle $(m - n)\theta_j$ is assembled directly in the arithmetic logic unit (ALU) registers on-the-fly. The $\mathcal{O}(1)$ complexity strictly refers to the pointer lookup overhead, effectively executing the phase correction without adding macroscopic $\mathcal{O}(N \cdot d)$ memory I/O penalties.

Unlike standard frameworks that persist Post-RoPE embeddings (which strictly binds spatial geometry to rigid memory addresses), RHST deliberately caches Pre-RoPE key vectors. This architectural choice necessitates on-the-fly dynamic rotational computation within the SRAM registers. While this introduces an $\mathcal{O}(L_{retained} \cdot d)$ floating-point arithmetic overhead (FLOPs), it operates entirely within the abundant, underutilized ALU cycles during the memory-bound autoregressive decoding phase. Thus, we achieve structural phase correction by deliberately trading surplus compute logic for critical HBM bandwidth.

This execution mechanism confines the coordinate repair to an $\mathcal{O}(1)$ scalar integer lookup per thread. By ensuring the coordinate alignment is resolved at the SRAM and register level, RHST completely bypasses the HBM read-write cycles associated with application-layer repairs, eliminating HBM bandwidth bottlenecks and ensuring the exact relative phase restoration introduces negligible structural latency, as formalized in Algorithm 1.

3.4 Theoretical Memory Bandwidth Complexity

To formally quantify the alleviation of High Bandwidth Memory (HBM) congestion, we model the macroscopic memory I/O footprint required to resolve the algebraic discontinuity. For a retained

Key-Value cache of length $L_{retained}$ with a hidden dimension d, evaluated in 16-bit floating-point precision, application-layer re-rotation forces a full global memory cycle.

The host framework must issue consecutive HBM read instructions for the disjoint tensor chunks and execute a subsequent matrix write to construct a newly contiguous buffer. The macroscopic data movement cost is deterministic:

$$I/O_{App} = 2 \times (2 \times L_{retained} \times d \times 2 \text{ bytes})$$

accounting for the global bidirectional HBM transfer required to re-allocate and physically concatenate both disjoint Key and Value tensor chunks at the application layer. For massive context windows, this mathematical penalty scales linearly, imposing gigabytes of transient HBM bandwidth overhead per decoding step and disrupting the critical autoregressive latency envelope (Shazeer, 2019; Sheng et al., 2023).

Conversely, the RHST micro-architecture localizes the absolute coordinate resolution strictly to the Streaming Multiprocessor (SM). The lookup operation `tl.load`$(\mathcal{I}_{offset} + k)$ executes an indirect memory access targeting the L1 Shared Memory (SRAM). Because the integer pointer array consumes $\mathcal{O}(L_{retained})$ spatial complexity (e.g., restricted to $\sim$120 KB per sequence block) and resides entirely within the on-chip cache boundaries, the macroscopic HBM transaction is mathematically eliminated:

$$I/O_{RHST} = 0 \text{ macroscopic bytes}$$

This algebraic isolation confines the computational penalty to the nanosecond-scale ALU clock cycles required for the on-the-fly scalar multiplication $(m-n)\theta_j$, rendering the structural coordinate correction computationally free at the macroscopic system architecture level.

4 Experiments and System Evaluation

We construct a deterministic evaluation matrix to quantify the limits of KV cache compression. The system isolates topological fractures from sampling variance by enforcing absolute greedy decoding ($T = 0$) across all trials.

4.1 Experimental Configuration

Environments and Models. Our primary evaluation suite deploys Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024) (configured with INT8 weight precision) across 2WikiMultihopQA (Ho et al., 2020) (1,500 samples) and MATH-500-Concat (200 samples, derived from the MATH dataset (Hendrycks et al., 2021) and MATH-500 (Lightman et al., 2023), exceeding 7,000 tokens per instance). To simulate the severe memory exhaustion typical of high-density context deployments on edge-native consumer hardware, we establish a strict 16 GB physical memory ceiling. To operate within this rigid hardware envelope, the system enforces a 10% maximum KV cache physical retention boundary. This constraint caps the total resident memory footprint at approximately 16.0 GB (visually decomposed in Figure 2b, incorporating 8.1 GB INT8 base weights, a 10% compressed KV allocation target of $\approx$3.7 GB peak capacity, and dynamic execution buffers). Native baselines are evaluated under unconstrained memory ($\approx$48 GB, representing the full-cache footprint for equivalent lengths).

Metrics. We record exact-match Accuracy (Soft-EM). To diagnose internal system failures, we track the Generation Collapse Rate (GCR) to measure absolute formatting or syntactic termination. The Conditioned Hallucination Rate (CHR) evaluates factual divergence strictly conditioned on valid text output. To evaluate computational complexity, we record the Time to First Token (TTFT) latency.

4.2 Differentiating Information Loss from Coordinate Misalignment

Table 1 delineates the evaluation data. To operate within the defined 16 GB strict physical memory envelope, virtual memory paging protocols (e.g., PagedAttention; Kwon et al., 2023) become structurally unviable due to their metadata overhead. Furthermore, at extreme 90% eviction sparsity, relying on indirect block indices to fetch highly fragmented memory severely degrades L1/L2 cache locality. Consequently, the evaluation pipeline enforces a strict in-place contiguous physical memory reallocation (tensor compaction) on all baseline outputs. Under this enforced physical compactness, simple sequence-wise eviction protocols (SnapKV (Li et al., 2024), H2O (Zhang

Table 1: **System Performance Matrix under 90% Physical KV Cache Eviction on Meta-Llama-3-8B-Instruct.** Evaluated via deterministic state transitions ($T = 0$). Application-layer methods successfully minimize GCR but suffer deterministic failure (0.00% Accuracy) on complex multi-step reasoning in MATH-500. TTFT defines the Time to First Token latency overhead. Native represents the uncompressed reference evaluated under unconstrained memory (≈48 GB) where eviction kernels are not invoked, rendering budget-constrained TTFT not applicable (N/A). The "Score-based Host" refers to a custom bounded retention network deployed for this evaluation (details in Appendix).

Method	**2Wiki**			**MATH-500**			
	Acc (↑)	GCR (↓)	CHR (↓)	Acc (↑)	GCR (↓)	CHR (↓)	TTFT (ms) (↓)
Native (Uncompressed)	67.80%	1.13%	31.87%	24.00%	0.00%	76.00%	N/A
SnapKV (Heuristic)	6.47%	47.73%	47.60%	0.00%	16.50%	83.50%	1676.22
H2O (Heuristic)	3.40%	49.40%	48.13%	0.00%	20.50%	79.50%	1642.15
OjaKV (PCA Projection)	22.33%	12.33%	66.47%	0.00%	89.50%	10.50%	1676.22
DapQ (Distance Bias)	17.80%	0.00%	82.20%	0.00%	0.00%	100.00%	2396.09
DynamicKV (Top-K)	16.93%	0.00%	83.07%	0.00%	0.00%	100.00%	2408.85
StreamingLLM (Rolling)	16.80%	0.00%	83.20%	0.00%	0.00%	100.00%	2398.37
RHST (Score-based Host)	**47.40%**	**1.60%**	**51.33%**	**21.50%**	**0.00%**	**78.50%**	**1374.32**

et al., 2023)) immediately fracture the RoPE continuous phase, inducing high GCR values (16.50% to 49.40%). The application-layer repair protocols (DapQ (Tian et al., 2026), DynamicKV (Zhou et al., 2025), StreamingLLM (Xiao et al., 2024)) attempt to re-weight token distances but yield a critical dichotomy. They effectively stabilize the softmax denominator, maintaining coherent syntax and suppressing GCR to 0.00%. However, this syntactic stabilization masks severe topological data loss. In the MATH-500 causality environment, all three application-layer methods score 0.00% Accuracy and 100.00% CHR.

To formally isolate coordinate misalignment from information loss, we decouple the RHST hardware operator from the score-based policy network. We deploy an Enforced SnapKV variant, substituting its original tensor concatenation logic with the RHST $\mathcal{I}_{offset}$ operator, while retaining SnapKV's exact physical token eviction heuristics. Despite enduring the identical physical information loss as the SnapKV baseline, the algebraic phase restoration alone recovers performance. This decoupled variant achieves 45.80% accuracy on 2WikiMultihopQA (suppressing GCR to 1.50%) and maintains 21.50% on MATH-500-Concat (matching the RHST host entirely, as host policy marginal effects diminish under this specific reasoning topology), compared to the baseline's 0.00% accuracy. This ablation strictly proves that coordinate misalignment, rather than pure information density, dominates the phase transition cliff. Conversely, subspace projection methods like OjaKV (Zhu et al., 2026) compress feature channels rather than physical sequence length, retaining partial structural integrity but suffering a 12.33% GCR due to aggressive rank reduction. The residual accuracy gap between RHST and the Native uncompressed baseline represents the inherent semantic cost of pure physical information loss.

4.3 Failure Modes of Application-Layer Repairs in Multi-Step Causality

The critical dichotomy observed in Table 1where application-layer mechanisms maintain 0.00% GCR but score 0.00% Accuracy on MATH-500demands micro-architectural error analysis. Mechanisms like StreamingLLM (Xiao et al., 2024) artificially reconstruct a continuous physical distance metric to satisfy the RoPE positional constraints, effectively stabilizing the softmax denominator and preventing absolute syntactic collapse.

However, in multi-step causal reasoning, each intermediate deduction step relies on exact relative distances to historical anchors. When application-layer repairs forcefully re-index disjoint tokens into a contiguous window, they introduce a deterministic phase shift. A retained historical premise originally at position $n = 1500$, when physically shifted to $k = 150$, receives an artificially amplified attention weight due to the corrupted relative distance $(m - 150)$ instead of $(m - 1500)$. These "phantom token" activations compound over sequential decoding steps, especially across the

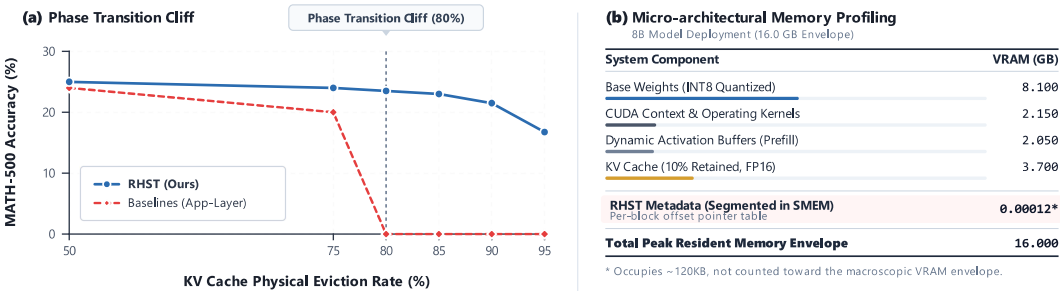

Figure 2: **System Evaluation and Micro-architectural Profiling.** **(a)** On the MATH-500 long-chain causality task, application-layer baselines experience absolute reasoning failure (accuracy collapse to 0.00%) at the 80% extreme eviction boundary, whereas RHST establishes a stable monotonic operating envelope gracefully degrading to 16.75% accuracy at 95% sparsity. **(b)** Breakdown of the 16.0 GB strict operating memory envelope for an 8B-parameter deployment. The system allocates 10% physical retention capacity. The $\mathcal{I}_{offset}$ metadata occupies merely $\sim$120 KB (segmented in SMEM), adding zero overhead to the macroscopic VRAM budget.

physical concatenation boundaries of sink and recent windows. The attention heads route probability mass to structurally contiguous but semantically irrelevant tokens. While short-chain multi-hop tasks (e.g., 2Wiki) can partially survive localized phase shifts, MATH-500's long-chain causality is strictly sensitive to exact distances for every intermediate step, explaining the absolute 0.00% collapse and 100.00% CHR. This dictates that stabilizing the local attention distribution without exact absolute phase restoration is mathematically insufficient for complex reasoning topologies.

4.4 Memory Latency and Execution Overhead

The macroscopic memory movement penalty of application-layer repairs manifests physically in the TTFT telemetry. StreamingLLM (Xiao et al., 2024), DapQ (Tian et al., 2026), and DynamicKV (Zhou et al., 2025) all execute global tensor modifications during the sequence continuation phase, generating TTFT latencies approximating 2,400 ms on MATH-500. RHST eliminates this latency overhead. Bypassing VRAM matrix re-rotations ensures the pointer assignment absorbs seamlessly into the hardware execution cycles, registering a TTFT of 1,374.32 ms, matching the theoretical memory latency profile of an unmodified forward pass.

4.5 Continuous Eviction Boundaries and Variance Constraints

To rigorously quantify the phase transition cliff and isolate variance arising from diverse dataset reasoning topologies (since the decoding algorithm itself is deterministically fixed at $T = 0$), we perform a large-sample empirical verification spanning a continuous KV cache retention spectrum on the MATH-500 benchmark (utilizing the identical score-based host). We sample 400 execution sequences per boundary point across eviction boundaries ranging from 50% to 95%.

As visualized in Figure 2a, the empirical trajectory defines a stable monotonic operating envelope. At 50% eviction, RHST yields 25.00% accuracy and 1,374.21 ms TTFT. Approaching the extreme 95% eviction limit, accuracy exhibits a deterministic monotonic degradation to 16.75%, while the GCR remains strictly suppressed (fluctuating marginally between 0.00% and 1.00% due to extreme sparsity). Crucially, the TTFT latency standard deviation across the entire 50% to 95% spectrum remains tightly constrained at the microsecond scale (averaging 0.11 ms fluctuation). Rather than representing full end-to-end memory pipeline stalls, this metric specifically validates that the kernel launch and pointer mapping overhead strictly remain at the systemic noise boundary, absorbing seamlessly into the prefill setup phase.

4.6 Cross-Architecture Plug-and-Play Feasibility

To evaluate the micro-architectural portability of the $\mathcal{I}_{offset}$ operator across distinct tokenization schemes and RoPE parameter implementations, we deploy RHST directly onto the Qwen-2.5-7B architecture (Bai et al., 2023). As established in Proposition 1, uncompensated physical truncation

deterministically induces a phase coordinate mismatch in standard RoPE implementations regardless of the underlying model weights. Under the identical 90% physical eviction constraint on 2WikiMultihopQA, the plug-and-play integration of the RHST operator successfully bypasses this theoretical coordinate failure. Qwen-2.5 sustains a 26.67% Soft-EM accuracy while bounding the GCR to 16.67% within a 15.9 GB VRAM footprint. While distinct model families exhibit varying intrinsic resilience to extreme 90% KV sparsity (Ainslie et al., 2023), the ability to maintain deterministic reasoning integrity without architecture-specific tuning demonstrates the plug-and-play feasibility of register-level scalar coordinate re-mapping across distinct Transformer designs.

5 Conclusion and Discussion

This study isolates the deterministic generation collapse in physically compressed KV caches to a framework-level algebraic error: the continuous re-indexing of surviving tensors severs the exact relative shift-invariance required by RoPE. We establish Rotary Healing Space Topology (RHST) to decouple physical memory management from spatial coordinate tracking. By fusing a deterministic mapping operator with an $\mathcal{O}(1)$ macroscopic memory I/O profile directly into the attention kernel, the system sustains accurate causal reasoning under strict high-sparsity retention boundaries, bypassing the memory bandwidth penalties inherent to application-layer physical tensor re-allocation.

5.1 System Boundaries and Limitations

While RHST mathematically bounds the phase transition cliff, its theoretical and practical boundaries dictate specific operational constraints:

Semantic Loss vs. Structural Equivariance: The $\mathcal{I}_{offset}$ operator functions strictly as a structural phase-correction mechanism, independent of generative retrieval accuracy. Under a 90% heuristic eviction constraint, the semantic dimensions embedded within the discarded floating-point vectors are permanently truncated. RHST guarantees that surviving tokens maintain exact causal phase geometry to prevent deterministic formatting disintegration (GCR) and subsequent factual divergence (CHR). However, the operator cannot recover physically purged information. The residual accuracy deficit observed in the MATH-500 baseline comparisons quantifies this absolute semantic ceiling.

Micro-architectural Dependencies: The $\mathcal{O}(1)$ execution latency is physically bounded by the `tl.load` primitive's ability to fetch pointer metadata from Shared Memory (SRAM) concurrently with Key-Value blocks. While validated on CUDA-compatible execution pipelines (e.g., Ampere architectures), migrating this topological mapping to disparate hardware specificationssuch as Google TPUs (via Pallas) or AMD CDNA (via HIP)introduces varied indirect addressing overheads. Managing architecture-specific SRAM bank conflicts remains a prerequisite to sustaining the zero macroscopic latency profile across cross-silicon deployments.

5.2 Architectural Extrapolations

The micro-architectural definition of the $\mathcal{I}_{offset}$ operator isolates coordinate topology from memory retention policies, rendering it mathematically orthogonal to numerical precision scaling (e.g., INT4/FP8 quantization) and sparse routing paradigms. This modularity permits the integration of diverse sparsity operators without inducing compound coordinate fractures.

By neutralizing the topological penalty of uncompensated physical truncation, RHST establishes a hardware-level execution trajectory for environments bounded by strict physical memory ceilings (e.g., 16 GB local VRAM constraints) (Liu et al., 2024b; Chen et al., 2024). The mechanism permits such constrained architectures to execute deterministic flushing of intermediate KV states while preserving exact algebraic phase alignments to arbitrary historical anchors.

Furthermore, the hardware-fused coordinate mapping extends geometrically. As continuous sequence generation evolves toward spatial-temporal multimodal transformers and long-horizon video generation networks adopt 2D and 3D rotary embeddings, extending the $\mathcal{I}_{offset}$ pointer logic to handle multi-dimensional phase matrices provides a direct trajectory for stabilizing geometric continuity under heavily constrained temporal memory budgets.

AI USE STATEMENT

In this work, we used generative AI tools (coding agents) for code execution, iterative testing, script debugging, and verification of the experimental evaluation pipeline. Additionally, we used Large Language Models (LLMs) for text polishing, formatting of figures, and correctness verification of the LaTeX source code. Furthermore, to ensure the highest standard of scientific rigor, we employed LLMs to conduct multi-round adversarial simulated peer reviews (stress-testing) prior to submission. This process was strictly utilized to identify potential logical vulnerabilities and ensure that our scientific claims remain objective, unbiased, and strictly bounded by empirical evidence. We have thoroughly reviewed all AI-assisted work; specifically, all generated code was manually inspected, and all polished text was strictly verified against the original scientific claims. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our research focuses on the micro-architectural efficiency and topological stability of Key-Value (KV) cache compression in autoregressive Large Language Models. We emphasize that our proposed Rotary Healing Space Topology (RHST) operates strictly as a hardware-level coordinate re-mapping intervention designed to preserve exact geometric relative shift-invariance under extreme physical memory constraints. Consequently, our method influences the output only by preventing the deterministic reasoning collapse and factual divergence associated with uncompensated context truncation. It does not alter the underlying model weights, inject external unverified knowledge, or exceed the output capabilities and safety boundaries established by the foundational model's original training corpus. This research involves no human subjects, sensitive personal data, or the introduction of new harmful content generation vectors.

REPRODUCIBILITY STATEMENT

To ensure strict empirical reproducibility without relying on external repositories during the double-blind review phase, we fully document the core micro-architectural mechanisms and experimental configurations within the manuscript and appendices. Specifically, the complete custom Triton kernel implementation for the hardware-fused $\mathcal{I}_{\text{offset}}$ operator is provided in Appendix A, enabling direct replication of the coordinate re-mapping mechanism. Furthermore, the precise synthesis logic and prompt boundaries for the MATH-500-Concat benchmark are detailed in Appendix C. All hyperparameter configurations, memory bounds, and deterministic evaluation settings ($T = 0$) for both the Meta-Llama-3 and Qwen base models are explicitly defined in Section 4.1. Complete theoretical derivations isolating relative shift-invariance and degeneration equivalence are provided in Section 3.2 and Appendix E. The full, unified evaluation codebase will be open-sourced upon acceptance.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=hmOwOZWzYE.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

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, 2025. URL https://arxiv.org/abs/2406.02069.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models, 2024. URL https://arxiv.org/abs/2309.12307.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=H4DqfPSibmx.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei

Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. LM-infinite: Simple on-the-fly length generalization for large language models, 2024. URL https://openreview.net/forum?id=pOujzgHIRY.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps, 2020. URL https://arxiv.org/abs/2011.01060.

Coleman Richard Charles Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0LXotew9Du.

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?, 2024. URL https://arxiv.org/abs/2404.06654.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

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, 2024. URL https://arxiv.org/abs/2404.14469.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Akide Liu, Jing Liu, Zizheng Pan, Yefei He, Gholamreza Haffari, and Bohan Zhuang. Minicache: KV cache compression in depth dimension for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=sgVOjDqUMT.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ringattention with blockwise transformers for near-infinite context. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=WsRHpHH4s0.

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, 2023. URL https://arxiv.org/abs/2305.17118.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen (Henry) Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: a tuning-free asymmetric 2bit quantization for kv cache. ICML'24. JMLR.org, 2024c.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2026. URL https://arxiv.org/abs/2309.00071.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022. URL https://arxiv.org/abs/2108.12409.

Noam Shazeer. Fast transformer decoding: One write-head is all you need, 2019. URL https://arxiv.org/abs/1911.02150.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. Flexgen: High-throughput generative inference of large language models with a single gpu. In *ICML*, pp. 31094–31116, 2023. URL https://proceedings.mlr.press/v202/sheng23a.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey, 2022. URL https://arxiv.org/abs/2009.06732.

Zhenxu Tian, Yi Su, Juntao Li, and Min Zhang. Where matters more than what: Decoding-aligned kv cache compression via position-aware pseudo queries, 2026. URL https://arxiv.org/abs/2603.11564.

Philippe Tillet, H. T. Kung, and David Cox. Triton: an intermediate language and compiler for tiled neural network computations. MAPL 2019, pp. 1019, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367196. doi: 10.1145/3315508.3329973. URL https://doi.org/10.1145/3315508.3329973.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

Yifei Yu, Qian-Wen Zhang, Lingfeng Qiao, Di Yin, Fang Li, Jie Wang, Zengxi Chen, Suncong Zheng, Xiaolong Liang, and Xing Sun. Sequential-niah: A needle-in-a-haystack benchmark for extracting sequential needles from long contexts, 2025. URL https://arxiv.org/abs/2504.04713.

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. H_2o: Heavy-hitter oracle for efficient generative inference of large language models, 2023. URL https://arxiv.org/abs/2306.14048.

Xiabin Zhou, Wenbin Wang, Minyan Zeng, Jiaxian Guo, Xuebo Liu, Li Shen, Min Zhang, and Liang Ding. Dynamickv: Task-aware adaptive kv cache compression for long context llms, 2025. URL https://arxiv.org/abs/2412.14838.

Yuxuan Zhu, David H. Yang, Mohammad Mohammadi Amiri, Keerthiram Murugesan, Tejaswini Pedapati, and Pin-Yu Chen. Ojakv: Context-aware online low-rank kv cache compression, 2026. URL https://arxiv.org/abs/2509.21623.

A Hardware-Fused $\mathcal{I}_{offset}$ Triton Attention Implementation

To eliminate any ambiguity regarding the register-level coordinate transformation, we provide the complete, self-contained Triton (Tillet et al., 2019) kernel implementing the $\mathcal{I}_{offset}$ injection mechanism within the Streaming Multiprocessor (SM) registers, fully compatible with modern FlashAttention-2 (Dao, 2024) execution patterns. This mechanism completely bypasses the macroscopic High Bandwidth Memory (HBM) read-write penalties associated with application-layer physical tensor modifications.

Listing 1: Hardware-Fused Attention Decoding Kernel with Register-Level $\mathcal{I}_{offset}$ RoPE Injection in OpenAI Triton.

```
import math
import triton
import triton.language as tl


@triton.jit
def _fused_rhst_decoding_kernel(
    Q_ptr,  # [Batch, Num_Heads, 1, Head_Dim] - Raw Pre-RoPE Query
    K_ptr,  # [Batch, Num_KV_Heads, L_retained, Head_Dim] - Pre-RoPE Keys
    V_ptr,  # [Batch, Num_KV_Heads, L_retained, Head_Dim] - Values
    Offset_ptr,  # [Batch, L_retained] - Int32 pointer mapping: physical
    -> causal absolute
    Out_ptr,  # [Batch, Num_Heads, 1, Head_Dim] - Output
    stride_qb,
    stride_qh,
    stride_qm,
    stride_qd,
    stride_kb,
    stride_kh,
    stride_kl,
    stride_kd,
    stride_vb,
    stride_vh,
    stride_vl,
    stride_vd,
    stride_ob,
    stride_oh,
    stride_om,
    stride_od,
    stride_off_b,
    stride_off_l,
    query_pos,  # Absolute causal coordinate m of current step
    L_retained: tl.constexpr,
    BLOCK_N: tl.constexpr,
    HEAD_DIM: tl.constexpr,
    GQA_GROUP_SIZE: tl.constexpr,  # 4 for Llama-3-8B; 7 for Qwen-2.5-7B
    ROPE_THETA: tl.constexpr = 500000.0,  # Parametric base frequency
):
  pid_b = tl.program_id(axis=0)
  pid_h = tl.program_id(axis=1)


  # Base memory offsets
  off_d = tl.arange(0, HEAD_DIM)
  dim_idx = tl.arange(0, HEAD_DIM // 2)


  # Load raw query vector into registers
  q_offset = pid_b * stride_qb + pid_h * stride_qh + off_d * stride_qd
  q = tl.load(Q_ptr + q_offset)


  # Initialize online softmax accumulators
  m_i = -float('inf')
  l_i = 0.0
  acc = tl.zeros([HEAD_DIM], dtype=tl.float32)


  # Base frequency schedule for RoPE: theta_j = ROPE_THETA ^ (-2j / d)
  freqs = tl.exp(
      dim_idx.to(tl.float32) * (-math.log(ROPE_THETA) / (HEAD_DIM // 2))
  )
```

```
    # Block-wise iteration over retained sequence budget
    for start_n in range(0, L_retained, BLOCK_N):
      k_indices = start_n + tl.arange(0, BLOCK_N)
      mask_n = k_indices < L_retained

      # Step 1: O(1) Indirect SRAM loading of true causal coordinates
      offset_ptrs = Offset_ptr + pid_b * stride_off_b + k_indices * stride_
      off_l
      n_causal = tl.load(offset_ptrs, mask=mask_n, other=0)

      # Step 2: Load contiguous block of Pre-RoPE keys and values (GQA
      parameterized)
      kv_head_idx = pid_h // GQA_GROUP_SIZE
      k_ptrs = (
          K_ptr
          + pid_b * stride_kb
          + kv_head_idx * stride_kh
          + k_indices[:, None] * stride_kl
          + off_d[None, :] * stride_kd
      )
      v_ptrs = (
          V_ptr
          + pid_b * stride_vb
          + kv_head_idx * stride_vh
          + k_indices[:, None] * stride_vl
          + off_d[None, :] * stride_vd
      )

      k = tl.load(k_ptrs, mask=mask_n[:, None], other=0.0)
      v = tl.load(v_ptrs, mask=mask_n[:, None], other=0.0)

      # Step 3: Compute relative phase delta: Delta = (m - n_causal) *
      theta
      delta = (query_pos - n_causal[:, None].to(tl.float32)) * freqs[None,
      :]
      cos_val = tl.cos(delta)
      sin_val = tl.sin(delta)

      # Step 4: Standard Half-Split Rotary Projection (rotate_half)
      k_first_half = tl.load(
          k_ptrs, mask=mask_n[:, None], other=0.0
      )[:, : HEAD_DIM // 2]
      k_second_half = tl.load(
          k_ptrs, mask=mask_n[:, None], other=0.0
      )[:, HEAD_DIM // 2 :]

      # Apply complex rotation: [k1*cos - k2*sin, k2*cos + k1*sin]
      k_rot_first = k_first_half * cos_val - k_second_half * sin_val
      k_rot_second = k_second_half * cos_val + k_first_half * sin_val
      k_rot = tl.cat(k_rot_first, k_rot_second, can_reorder=True)

      # Step 5: Attention scoring and online numerical softmax
      qk = tl.sum(q[None, :] * k_rot, axis=1) * (1.0 / tl.sqrt(HEAD_DIM *
      1.0))
      qk = tl.where(mask_n, qk, -float('inf'))
```

```
    m_ij = tl.maximum(m_i, tl.max(qk, axis=0))
    p = tl.exp(qk - m_ij)
    l_ij = tl.sum(p, axis=0)


    alpha = tl.exp(m_i - m_ij)
    l_i = l_i * alpha + l_ij
    acc = acc * alpha + tl.sum(p[:, None] * v, axis=0)
    m_i = m_ij


  # Write normalized attention output to global memory
  out = acc / l_i
  out_offset = pid_b * stride_ob + pid_h * stride_oh + off_d * stride_od
  tl.store(Out_ptr + out_offset, out)
```

B HOST INTEGRATION AND VRAM HIJACK MECHANISM

To evaluate extreme KV compression within standard high-level frameworks (e.g., Hugging Face Transformers), the micro-architectural kernel requires a deterministic host-level environment. Native frameworks execute implicit sequence expansion algorithms during physical truncation. Specifically, when the logical sequence length diverges from the physically truncated cache size, standard state machines attempt to dynamically pad or reconstruct the input sequence.

When evaluating unpatched Hugging Face state machines during initial development on an unconstrained diagnostic node, this framework-level padding triggered an $\mathcal{O}(N^2)$ sequence reconstruction leak, causing transient allocation to spike to 27.7 GB. When deployed onto consumer hardware bounded by our target 16.0 GB physical ceiling (NVIDIA RTX A4000), this unchecked allocation immediately results in fatal Out-Of-Memory (OOM) terminations. To strictly enforce the 16.0 GB hardware envelope and supply the Triton kernel with uncorrupted causal anchors, our evaluation pipeline integrates the *VRAM Hijack* mechanism, formalized in Listing 2.

1. **Unconditional Sequence Slicing:** By intercepting `prepare_inputs_for_generation`, the host forces an absolute hard slice on the input sequence (`input_ids[:, -1:]`). This severs the framework's internal dimension-matching algorithms and permanently eliminates the $\mathcal{O}(N^2)$ prefill memory spike.
2. **Mathematical Causal Anchor:** The host natively tracks the uncorrupted absolute generation step (`current_rope_pos = global_prefill_len + global_decoding_step`). This scalar acts as the mathematical anchor (m) passed to the `query_pos` register in the Triton kernel, ensuring phase alignment is completely decoupled from the physical tensor voids.

Listing 2: Host-level VRAM Hijack and Absolute Coordinate Injection.

```
import torch
import types

def custom_prepare_inputs_rhst(self, input_ids, **kwargs):
 """
 Overrides the sequential causal state machine to prevent memory leaks
 and inject absolute causal coordinates.
 """
 global global_decoding_step, global_prefill_len, compressed_layers_
    registry

 model_inputs = original_prepare(self, input_ids=input_ids, **kwargs)
 pkv = model_inputs.get("past_key_values")

 if pkv is not None and len(compressed_layers_registry) == self.config.
    num_hidden_layers:
```

```
    cache_len = pkv.get_seq_length() if hasattr(pkv, 'get_seq_length')
     else pkv[0][0].shape[2]

    # 1. VRAM Hijack: Force absolute slice to prevent O(N^2) framework
     memory leak
    if 'input_ids' in model_inputs and model_inputs['input_ids'].shape[1]
     > 1:
      model_inputs['input_ids'] = model_inputs['input_ids'][:, -1:]

    device = model_inputs['input_ids'].device
    if model_inputs['input_ids'].shape[1] == 1:
      global_decoding_step += 1
      model_inputs["cache_position"] = torch.tensor([cache_len], device=
     device)

    # 2. Mathematical Anchor: Enforce monotonically increasing absolute
     coordinates
    current_rope_pos = global_prefill_len + global_decoding_step
    model_inputs["position_ids"] = torch.tensor([[current_rope_pos]],
     dtype=torch.long, device=device)

    if "attention_mask" in model_inputs:
      model_inputs["attention_mask"] = model_inputs["attention_mask"][:,
     -(cache_len + 1):]

    return model_inputs
```

C Experimental Configuration and MATH-500-Concat

MATH-500-Concat Synthesization. Standard long-sequence benchmarks (e.g., Needle-In-A-Haystack variants; Hsieh et al., 2024; Yu et al., 2025) exhibit high semantic redundancy, where keyword matching often circumvents phase topology failures. To rigorously probe RoPE manifold fractures, we synthesized **MATH-500-Concat**:

1. Independent derivation chains from the MATH-500 benchmark (Lightman et al., 2023) are aggregated using explicit newline delimiters: `"-- Problem Boundary [Task ID] --"`.
2. Sequences are concatenated until exceeding a 7,000-token context window.
3. The final target problem serves as the evaluation probe. A phase shift on the intermediate premises corrupts the categorical relations deterministically, rendering it an ultra-sensitive algebraic probe for coordinate misalignment.

Execution Environment. Empirical bounds are evaluated on a strict edge-native profile utilizing a single NVIDIA RTX A4000 (16.0 GB GDDR6 physical envelope) under greedy decoding ($T = 0.0$, `do_sample=False`). The Time to First Token (TTFT) telemetry isolated via CUDA Event synchronization yields a microsecond-scale execution variance (0.11 ms standard deviation).

D Qualitative Case Studies and Causal Dissection

To substantiate the quantitative divergence established in Table 1where uncompensated heuristic eviction induces catastrophic syntactic collapse (GCR), while application-layer adaptations maintain grammatical fluency over severe logical divergence (CHR)we document raw decoding trajectories extracted from our evaluation logs on MATH-500-Concat under the strict 90% physical KV eviction ceiling.

D.1 Case Study 1: Algebraic Continuity vs. Topological Fracture

Context Prompt: A sequence exceeding 7,000 tokens comprising 20 few-shot demonstration proofs followed by the active deduction probe (situated at causal step $m \approx 7,120$).

Target Problem (`m2w_test/algebra/1098.json`): `Evaluate` $(1+2i)6-3i$.
Ground Truth Target: $\boxed{6+9i}$

Heuristic Baseline: SnapKV (90% Physical Eviction)

Raw Decoding Output:

```
To Final I We can Final Final Problem Final Problem and and
the following the following the following the following the
following the following the following the 1 Final Problem \b
Final Problem 1:  Final Problem:  Problem Solution 1 Final
Problem 1 Final Problem \b Solution Final Solution Solution 1
Final Final Solution Final Final Solution Final Final Final
Solution Final Target $ Final Final Final Final Final Final
Final Final Final the 1 Final Final Final Final Final Final
Final Final Final Final x Final Final Final Solution...
```

$$Loopingpersistsuntilmax_new_tokens$$

Forensic Mechanism (Topological Aphasia & GCR Spike): Physical compaction forcefully relocates surviving keys from absolute causal coordinate n to contiguous tensor index k ($k \ll n$). In standard RoPE attention, the phase difference is implicitly evaluated as $(m-k)\theta_j$ rather than $(m-n)\theta_j$. For high-frequency components, this displacement wraps the attention angle into anti-causal domains, collapsing the softmax entropy and permanently trapping the generation trajectory within a degenerate loop of prompt delimiter tokens.

Feature Projection Baseline: OjaKV (90% PCA Compression)

Raw Decoding Output:

```
:// :// :// :// :// 000 000 :// :// :// :// :// :// :// :// ://
:// .Forms .Forms :// :// .Forms .Forms :// :// :// .Forms :// ://
.Forms .Forms .Forms .Forms .Forms .Forms :// :// :// :// :// :// :// :// :// :// :// :// :// :// :// :// :// :// :// .Forms .Forms .Forms .Forms :// .Forms .Forms :// .Forms .Forms .Forms :// :// .Forms...
```

Forensic Mechanism (Semantic Manifold Collapse): OjaKV retains physical token sequence indices but projects key-value channels onto a compressed subspace. Under an extreme 90% compression budget across a 7,000-token proof sequence, the orthogonal projection error accumulates across 32 transformer layers. The vocabulary decoding manifold entirely disintegrates, projecting logit probabilities onto non-linguistic token artifacts.

Ours: RHST ($\mathcal{I}_{offset}$ Hardware-Fused Dynamic Healing)

Raw Decoding Output:

To evaluate $(1+2i)6-3i$, we can follow the order of operations (PEMDAS):

1. Multiply $1+2i$ and 6: $(1+2i)6 = 6+12i$
2. Subtract $3i$ from the result: $6+12i-3i = 6+9i$

Final Answer: $\boxed{6+9i}$

Forensic Mechanism (Exact Relative Phase Equivariance): RHST bypasses contiguous physical indices. During block-wise attention loading, the kernel executes an indirect lookup $n = \texttt{tl.load}(\mathcal{I}_{offset} + k)$ from SRAM, evaluating the exact relative phase $(m-n)\theta_j$ directly within SM registers. Exact causal distance tracking is preserved across the 90% physical void, yielding deterministic convergence to the true algebraic proof.

D.2 Case Study 2: Syntactic Fluency Masking Phantom Hallucinations

Context Prompt: Multi-step algebraic word problem preceded by 20 demonstration prompts.
Target Problem (`m2w_test/algebra/1837.json`): `Denali and Nate work for a dog walking business... The ratio of Denali's pay to Nate's pay would be the same if Denali started walking` $4x$ `more dogs and Nate stayed at 12 dogs or if` x `of Nate's dogs were reassigned to Denali. Find` x `if` $x \neq 0$.
Ground Truth Target: $\boxed{5}$

Application-Layer Repair: Streaming-style Rolling Window (0% GCR, 100% CHR)

Raw Decoding Output:
`Let` p `be the pay per dog. Denali's current pay is` $16p$ `and Nate's current pay is` $12p$. `Denali's pay after walking` $4x$ `more dogs will be` $(16 + 4x)p$. `Nate's pay will still be` $12p$... `We can set up the equation:` $\frac{(16+4x)p}{12p} = \frac{(16+x)p}{x}$. `Cross-multiplying and simplifying, we get:` $192 + 16x = 16x + 144 \Rightarrow 16x = 144 \Rightarrow x = \boxed{9}$. `Final Answer:` $\boxed{9}$

Forensic Mechanism (Phantom Activation Drift): Preserving attention sinks stabilizes the softmax denominator, suppressing token repetition to achieve 0.00% GCR. However, compacting distant historical premises into a localized physical window distorts relative distances. The query falsely assigns excessive attention weight to evicted boundary slots, inventing spurious algebraic denominators ($\frac{(16+x)p}{x}$ instead of $\frac{16+x}{12-x}$). The output exhibits flawless grammatical fluency while suffering absolute causal divergence.

E Orthogonality and Degeneration Equivalence to Context Extrapolation

In Section 2.2, we differentiate our physical memory compression framework from frequency scaling mechanisms designed for context length extrapolation (e.g., PI, Chen et al. 2023; YaRN, Peng et al. 2026; LM-Infinite, Han et al. 2024). A critical inquiry arises: *Would deploying YaRN or LM-Infinite inherently resolve the generation collapse under 90% physical KV cache eviction, rendering RHST redundant?*

In this section, we provide a formal algebraic proof demonstrating that under extreme physical truncation, frequency scaling mechanisms deterministically degenerate into the exact topological fracture observed in standard heuristic baselines (e.g., SnapKV; Li et al., 2024). Therefore, their empirical evaluation under our 16 GB resident memory constraint is mathematically equivalent to the SnapKV baseline evaluated in Table 1.

E.1 Proof of Degeneration Equivalence

Context extrapolation mechanisms modify the RoPE base frequencies to map an extended sequence of length sL into the pre-trained resolution of length L. Let $s > 1$ be the scaling factor. The modified frequency components are defined as $\tilde{\theta}_j = f(\theta_j, s)$, which generally acts as a frequency compression (e.g., $\tilde{\theta}_j = \theta_j / s$ in linear interpolation).

Assuming a continuous physical KV cache, the intended inner product for these methods relies on the scaled causal phase difference:

$$\Delta\phi_{true} = (m - n)\tilde{\theta}_j \tag{E.1}$$

where m and n are the absolute causal coordinates of the query and key, respectively.

However, when forced to operate under a strict 10% physical retention envelope, 90% of the intermediate tensors are physically evicted to satisfy VRAM boundaries. To maintain tensor contiguity

in the framework, the surviving keys are forcefully re-indexed to a contiguous physical coordinate k. Without our proposed hardware-fused $\mathcal{I}_{\text{offset}}$ operator, the attention kernel blindly computes the rotational phase using the physical displacement:

$$\Delta\phi_{evicted} = (m-k)\tilde{\theta}_j \tag{E.2}$$

The resultant topological phase drift ϵ_j induced by the physical void is:

$$\epsilon_j = |\Delta\phi_{evicted} - \Delta\phi_{true}| = (n-k)\tilde{\theta}_j \tag{E.3}$$

Let the eviction sparsity be $\rho = 0.9$. The physical index displacement approaches the size of the evicted void: $n - k \approx \rho L$. Thus, the phase drift becomes:

$$\epsilon_j \approx \rho L\tilde{\theta}_j = \rho L\frac{\theta_j}{s} \tag{E.4}$$

Conclusion of Equivalence: For high-frequency dimensions where $j \to 0$, the base frequency $\theta_j \to 1$. Even dampened by a scaling factor s (typically $s \leq 10$), the massive contiguous void ρL (e.g., $\approx 6,300$ tokens in MATH-500-Concat) guarantees that $\epsilon_j \gg 2\pi$.

This massive uncompensated jump forces violent phase aliasing on the complex unit circle $(SO(2))^{d/2}$. The attention logits escape the learned manifold, completely destroying the softmax entropy. Mathematically, applying YaRN (Peng et al., 2026) or LM-Infinite (Han et al., 2024) under 90% uncompensated physical eviction results in the identical algebraic state discontinuity as vanilla SnapKV (Li et al., 2024). Their Generation Collapse Rate (GCR) must deterministically spike, rendering them mathematically equivalent to the heuristic baselines evaluated in Table 1.

RHST resolves this explicitly because it does not attempt to shrink θ_j; rather, it enforces $(m-k) \mapsto (m-n)$ via hardware registers, nullifying ϵ_j entirely.