CAPACITY-AWARE HOLOGRAPHIC MEMORY FOR LONG-CONTEXT SEQUENCE MIXING

Anonymous authors
Paper under double-blind review

ABSTRACT

Holographic Reduced Representations (HRRs) provide an efficient mechanism for content-addressable sequence interaction, but their associative memory suffers from interference as the number of superposed bindings grows. We show analytically that, under standard random-vector assumptions, retrieval signal-to-noise ratio scales as $d/(N-1)$, identifying the number of bindings assigned to each memory trace as the key capacity bottleneck. Motivated by this analysis, we introduce Pyramidal Holographic Memory (PHM), a capacity-controlled multi-resolution architecture that partitions sequences across memory traces and scales, keeping the binding load per trace bounded independently of sequence length. PHM therefore preserves the computational advantages of FFT-based holographic binding while avoiding the capacity collapse of a globally overloaded trace. Across controlled associative-recall experiments, length extrapolation, long-range sequence benchmarks, speech classification, and language modeling, PHM maintains stable recall at evaluation lengths up to $64\times$ those seen during training and increasingly outperforms optimized scaled dot-product attention at long contexts, while remaining competitive with recurrent state-space models on associative recall. These results show that explicit control of associative memory load can turn holographic binding from a capacity-limited representation into a practical primitive for long-context sequence modeling.

1 INTRODUCTION

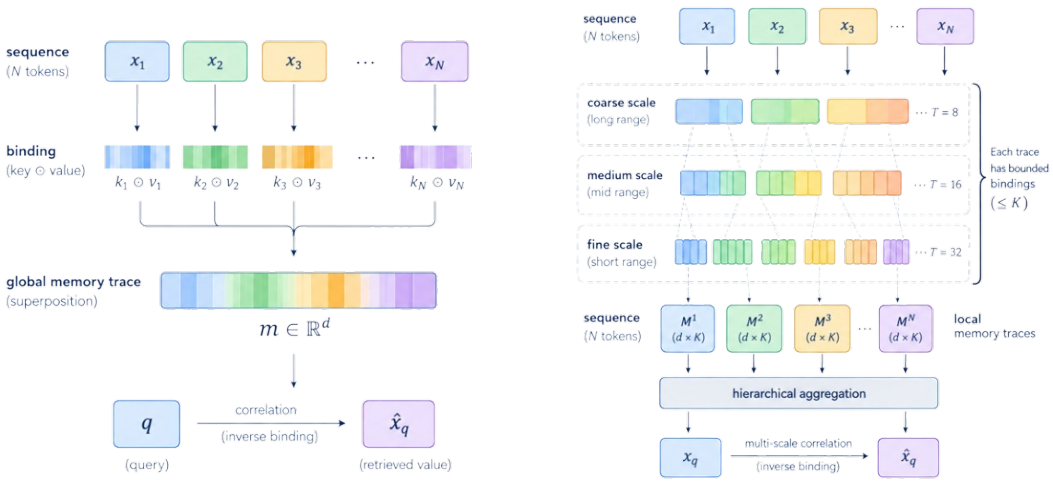
Transformer architectures have become the dominant paradigm for sequence modeling, largely due to the ability of self-attention to directly model interactions between arbitrary pairs of tokens. However, this flexibility comes at a fundamental computational cost: for a sequence of length N, standard self-attention requires $O(N^2)$ pairwise interactions Vaswani et al. (2017). While a number of optimized implementations reduce constant factors and memory overhead, the quadratic dependence on sequence length remains problematic as context windows grow to tens or hundreds of thousands of tokens Dao et al. (2022). This has motivated increasing interest in attention-free sequence mixers with approximately linearly or near-linearly computational cost.

Several alternatives have emerged, including Fourier-based mixers (Lee-Thorp et al., 2022), state-space models such as S4 (Gu et al., 2021), and recent selective state-space architectures such as Mamba (Gu & Dao, 2024). These approaches demonstrate that useful long-range sequence representations can be obtained without explicitly constructing the dense token-to-token interaction matrix of attention. Nevertheless, they rely on fundamentally different mechanisms for propagating information through a sequence. In particular, state-space approaches are well suited to sequential state propagation, whereas Fourier mixing provides efficient global interaction but lacks an explicit content-addressable memory mechanism. This raises a broader question: can an efficient sequence mixer provide direct associative retrieval of information while avoiding the quadratic cost of attention?

Holographic Reduced Representations (HRR) (Plate, 1995) provide a natural basis for such a mechanism. HRR represents information by binding keys and values into high-dimensional vectors and superposing multiple bindings into a shared memory trace. A query can subsequently retrieve an approximately associated value from this trace using an inverse binding operation. Importantly, circular convolution and correlation underlying HRR can be implemented efficiently in the Fourier

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.

domain, giving rise to $O(N \log N)$ sequence operations. This makes holographic representations particularly attractive for long-context sequence modeling, where explicit pairwise attention is expensive.



(a) (b)

Figure 1: Comparison of global HRR memory and the proposed PHM principle. (a) Global HRR superposes all bindings into a single memory trace, causing cross-talk to increase as the sequence length grows. (b) PHM distributes bindings across bounded local memories and multiple temporal scales, preventing the number of bindings per trace from growing with the full sequence length.

However, HRR introduces a less visible but fundamental scaling problem. A memory trace has finite representational capacity: as more key-value bindings are superposed into the same vector, the retrieved signal becomes increasingly contaminated by cross-talk from unrelated bindings (see Fig 1). Classical HRR analysis characterizes this behavior through a signal-to-noise ratio that scales approximately as

$$\mathrm{SNR} \propto \frac{d}{N-1}, \tag{1}$$

where d denotes the representation dimension and N is the number of superposed bindings. Thus, although HRR operations remain computationally efficient as sequence length increases, a global HRR memory does not necessarily remain information-efficient. Increasing the number of tokens while keeping the memory dimension fixed eventually causes retrieval quality to deteriorate. Existing holographic sequence mixers therefore face a fundamental capacity bottleneck Alam et al. (2023b): the computational complexity scales favorably, but the amount of information assigned to a single memory trace does not.

This observation suggests a different approach to efficient sequence mixing. Rather than attempting to increase the dimensionality of a global holographic memory indefinitely, we ask whether the memory itself can be structured so that each trace receives only a bounded number of bindings. The resulting design principle is simple: control the number of bindings per memory trace rather than allowing the trace capacity to scale with the entire sequence length. Under this formulation, sequence length can increase by introducing additional memory regions instead of continuously increasing the load placed on an existing trace. Moreover, multiple memory granularities can be used simultaneously to capture interactions occurring at different ranges.

Based on this principle, we introduce the **Pyramidal holographic memory (PHM)**, an attention-free sequence mixing architecture designed around capacity-aware holographic memory. PHM replaces a globally superposed HRR trace with a hierarchical memory consisting of bounded local chunks and multiple temporal scales. Fine-grained traces preserve short-range associative information, while coarser traces provide progressively larger receptive fields. Consequently, the number of bindings contributing to an individual trace remains bounded even as the overall sequence becomes substantially longer. This converts the capacity constraint of HRR from a limitation into an explicit architectural design criterion.

We evaluate PHM across long-range classification, controlled associative recall, speech classification, and language modeling tasks. Experimental results show that PHM is particularly effective when the task benefits from efficient global associative retrieval, while experiments on scan-heavy sequential reasoning reveal limitations relative to state-space approaches. We therefore view PHM not as a universal sequence-modeling solution, but as an alternative memory primitive whose central advantage arises from explicitly matching holographic memory granularity to its theoretical capacity. We make following contributions (1) We identify the finite-capacity behavior of global HRR superposition as a fundamental scaling bottleneck for holographic sequence mixers and connect the degradation of retrieval quality to the classical HRR signal-to-noise relationship. (2) We introduce a capacity-aware $O(N \log N)$ sequence mixer that organizes holographic memory into bounded local regions and multiple scales, preventing the number of bindings per memory trace from growing with the total sequence length. (3) We combine holographic associative retrieval with depthwise local convolution and multiplicative gating, allowing PHM to jointly model local context and content-addressable long-range information within a single efficient mixing block.

2 Related Work

Efficient attention alternatives. Recent advancements try to reduce or eliminate the quadratic cost of self-attention. Sparse patterns such as BigBird Zaheer et al. (2020) and Longformer Beltagy et al. (2020) restrict attention to local windows and selected global tokens, retaining exact attention within each window but sacrificing full-sequence coverage. Low-rank approaches project the key-value space: Linformer Wang et al. (2020) uses random projections, while Performer Choromanski et al. (2021) and Linear Attention Katharopoulos et al. (2020) approximate the softmax kernel with positive random features or ϕ-mapped dot products. FNet Lee-Thorp et al. (2022) takes a more radical approach, eliminating learned mixing entirely by applying the 2D discrete Fourier transform over sequence and feature dimensions. FlashAttention Dao et al. (2022) reduces the I/O cost of exact attention, retaining $\mathcal{O}(N^2)$ complexity but drastically lowering wall-clock time.

State-space models. S4 Gu et al. (2021) and its successors cast the sequence as a linear recurrence parameterized by a continuous-time state-space model, achieving $\mathcal{O}(N \log N)$ training via the FFT and $\mathcal{O}(1)$ per-step inference. Mamba Gu & Dao (2024) extends this with input-dependent gating. Hyena Poli et al. (2023) implicitly parameterizes a long convolution filter in the frequency domain. These models excel on LRA and audio tasks but operate via fundamentally different mechanisms: recurrence and learned convolution rather than one-shot associative memory. A parallel line of linear-recurrent and gated-linear-attention models—RWKV Peng et al. (2023), RetNet Sun et al. (2023), and gated linear attention (GLA) Yang et al. (2024)—similarly achieves sub-quadratic cost through decaying recurrent state or chunked linear attention; like SSMs they compress context into a fixed-size state rather than a content-addressable associative memory, and the Zoology analysis Arora et al. (2024) shows this class trades off recall fidelity.

Vector Symbolic Architectures. VSA Gayler (2003); Kanerva (2009) have historically been studied in cognitive science and hyperdimensional computing. The key algebraic insight is that circular convolution acts as a binding operation in the frequency domain: $\mathcal{F}(x \circledast y) = \mathcal{F}(x) \odot \mathcal{F}(y)$, allowing arbitrary key-value pairs to be superposed and individually retrieved from a single fixed-size vector. HRRFormer Alam et al. (2023a) integrated HRR binding into a Transformer backbone, achieving competitive LRA results at large head dimensions ($d_h = 128$) but collapsing to near-random accuracy at smaller budgets. Crucially, HRRFormer uses HRR only to compute a per-token *similarity score*—it binds each key-value pair, retrieves with the query, and then applies softmax attention over the resulting fidelity scores—so it remains an attention model with an HRR-flavored kernel. Furthermore, the capacity issue is also left unaddressed by HRRFormer: at $d_h = 32$, the retrieval SNR is $d_h/N = 32/2000 \approx 0.016$ on ListOps—well below the noise floor. A complementary line reduces the cost and instability of the binding operator itself: the Hadamard-derived linear binding (HLB) Alam et al. (2024) replaces FFT circular convolution with an $\mathcal{O}(d)$ element-wise product that is numerically stabler and improves VSA retrieval fidelity, though it has not been applied to sequence mixing.

Associative memory and gated sequence models. Associative memory provides a complementary view of attention, with modern Hopfield networks Ramsauer et al. (2021) showing that a single update of a continuous Hopfield energy is equivalent to the softmax attention operation and offering

exponential storage capacity in the pattern dimension. In parallel, multiplicative gating has become an effective mechanism in sequence models, with SwiGLU Shazeer (2020) and related gated linear units now widely adopted in large language models. Learnable LayerScale Touvron et al. (2021) has further been proposed to stabilize deep architectures by initializing each block close to an identity mapping.

3 METHODS

3.1 OVERVIEW

We introduce Pyramid Holographic Memory (PHM), a linear-complexity sequence mixer that converts holographic reduced representations (HRR) into a apacity-aware, multi-resolution content-addressable memory. Our design is motivated by a fundamental mismatch between HRR and long-context sequence modeling: HRR stores all key-value bindings in a single fixed-dimensional trace, causing retrieval interference to grow with the number of stored bindings. PHM addresses this limitation not by increasing the memory dimension, but by artitioning interference across a hierarchy of bounded-capacity memories.

This principle leads to the following architecture. First, lightweight local mixing provides short-range token interactions. Second, keys are position-bound in the Fourier domain to form position-specific holographic states. Third, these states are aggregated into memories at multiple temporal scales, where each local memory has a controlled maximum number of bindings. Queries then retrieve information from these memories through the exact HRR dual of the binding operation. For autoregressive tasks, the same capacity principle is implemented causally through an energy-gated prefix memory that suppresses low-information positions. Thus, PHM replaces both self-attention and the conventional feed-forward sublayer with a structured memory operation while maintaining $\mathcal{O}(ND)$ activation memory.

3.2 FROM HRR CAPACITY TO CAPACITY-AWARE MEMORY

Holographic Binding and Retrieval. HRR Plate (1995) represents N key-value bindings in a single d-dimensional memory trace:

$$M = \sum_{i=1}^{N} k_i \circledast v_i, \tag{2}$$

where $\circledast$ denotes circular convolution. Given a query key q, the associated value is approximately retrieved through circular correlation:

$$\hat{v}_q \approx M \circledast \overline{k_q}, \tag{3}$$

where $\overline{k_q}$ denotes the approximate inverse of k_q.

The convolution theorem transforms both operations into element-wise products in the Fourier domain giving $\mathcal{O}(d \log d)$ complexity:

$$x \circledast y = \mathcal{F}^{-1}\left(\mathcal{F}(x) \odot \mathcal{F}(y)\right), \tag{4}$$

The Interference Bottleneck. The scalability limitation follows directly from superposition. Consider i.i.d. $x_i, p_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$ and the normalized trace

$$M = \frac{1}{N} \sum_{i=1}^{N} x_i \circledast p_i. \tag{5}$$

Retrieval using p_j decomposes into the desired signal and cross-talk:

$$R_j = \frac{1}{N} x_j + \frac{1}{N} \sum_{i \neq j} \epsilon_i, \tag{6}$$

where the cross-talk terms have zero mean and per-component variance $1/d$. Consequently,

$$\mathrm{SNR} \approx \frac{d}{N-1}. \tag{7}$$

A complete proof is provided in Appendix A.1.

Equation 7 reveals the central design problem: memory interference scales with the number of bindings, not with the semantic complexity of the sequence. For example, at $d = 32$ and $N = 2000$, the theoretical SNR of a single global trace is only ≈ 0.016. Increasing the sequence length therefore progressively converts the associative memory into an interference-dominated representation.

Capacity-aware design principle. Rather than increasing d to accommodate more bindings, PHM bounds the number of bindings contributing to each retrieval memory. For a memory containing at most C bindings, the same analysis gives

$$\mathrm{SNR}_C \geq \frac{d}{C}. \tag{8}$$

This observation motivates the central idea beging PHM: partition the sequence into overlapping memory scales whose individual capacities remain controlled, while combining them to recover long-range context. Importantly, this is not conventional multi-scale feature pooling; the scale hierarchy is introduced specifically to control holographic interference.

3.3 Pyramid Holographic Memory

A PHM block maps $X \in \mathbb{R}^{B\times N\times D}$ to an output representation of the same shape. Rather than computing pairwise attention scores, PHM constructs distributed associative memories by binding position-aware feature states with normalized binding states and retrieving contextual information through holographic correlation.

Local Projection and Mixing. We first transform the input using a fused projection followed by depthwise local mixing:

$$[F, B, R] = \mathrm{Split}_3\left(\mathrm{DWConv}_3\left(XW_{\mathrm{mix}}\right)\right), \qquad W_{\mathrm{mix}} \in \mathbb{R}^{D\times 3D}. \tag{9}$$

Here, F, B, and R denote the *feature*, *binding*, and *retrieval* states, respectively. These states are specific to the associative-memory formulation and are distinct from the query, key, and value components of self-attention. The depthwise convolution uses kernel size 3, padding 1, and $3D$ groups, introducing nearest-neighbor context independently within each channel while adding only $9D$ convolutional parameters. The three states are reshaped into H heads with head dimension $d_h = D/H$. The feature state is scaled as

$$F_h \leftarrow \frac{F_h}{\sqrt{d_h}}, \tag{10}$$

while the binding and retrieval states are L2-normalized:

$$B_h \leftarrow \frac{B_h}{\|B_h\|_2}, \qquad R_h \leftarrow \frac{R_h}{\|R_h\|_2}. \tag{11}$$

This normalization removes magnitude-dependent variation from the associative operations, making binding and retrieval primarily dependent on representation direction.

Position-Aware Feature Binding. PHM incorporates positional information directly into the holographic representation. For each position i and Fourier frequency f, we construct a unit-modulus complex phase:

$$P_{i,f} = \exp\left(j\frac{i}{10000^{2f/d_h}}\right). \tag{12}$$

For the sinusoidal configuration, the resulting phase sequence is stored as a buffer and dynamically extended when a sequence longer than the initially allocated length is encountered. The feature state is transformed into the Fourier domain and modulated by its corresponding positional phase:

$$\widetilde{F}_{i,h} = \mathcal{F}(F_{i,h}) \odot P_i. \tag{13}$$

The position-aware feature state is then bound with the normalized binding state:

$$H_{i,h} = \widetilde{F}_{i,h} \odot \mathcal{F}(B_{i,h}). \tag{14}$$

Equivalently,

$$H_{i,h} = \mathcal{F}(F_{i,h}) \odot P_i \odot \mathcal{F}(B_{i,h}). \tag{15}$$

Here, $H_{i,h}$ is the resulting *bound holographic state*. The binding operation combines feature content, positional identity, and the binding state into a distributed representation. Consequently, identical feature and binding states occurring at different positions produce distinct bound states due to the position-dependent phase. This positional operation is analogous in spirit to phase-based positional encodings such as RoPE, but the two mechanisms are not algebraically equivalent. RoPE applies position-dependent rotations before dot-product attention, whereas PHM incorporates the positional phase directly into the Fourier-domain holographic binding operation.

Multi-Resolution Capacity-Controlled Memory. The bound holographic states are aggregated into a hierarchy of associative memories. Let

$$\mathcal{S} = \{C_1, \ldots, C_k\} \tag{16}$$

denote a set of temporal chunk sizes. For each scale $C \in \mathcal{S}$, the sequence is partitioned into non-overlapping chunks containing at most C bound states. For position i, let $\overline{H}_{i,h}^{(C)}$ denote the mean bound state of the chunk containing position i. We additionally retain a global memory trace:

$$\overline{H}_h = \frac{1}{N} \sum_{i=1}^{N} H_{i,h}. \tag{17}$$

For the pyramid configuration, the memory visible to position i is the average of the global trace and all applicable local memories:

$$M_{i,h} = \frac{\overline{H}_h + \sum_{\substack{C \in \mathcal{S} \\ C < N}} \overline{H}_{i,h}^{(C)}}{1 + |\{C \in \mathcal{S} : C < N\}|}. \tag{18}$$

The key property of this construction is that every local memory has an explicit capacity determined by its chunk size. A memory at scale C contains at most C superposed holographic bindings. Under the standard HRR interference model, the corresponding retrieval signal-to-noise ratio scales as

$$\mathrm{SNR}_C \gtrsim \frac{d_h}{C}. \tag{19}$$

Thus, each local memory has a bounded interference level that depends on its capacity rather than the total sequence length.

For example, with $d_h = 32$ and $\mathcal{S} = \{32, 128, 512\}$, the finest memory contains at most 32 superposed bindings, whereas a global trace over $N = 2000$ positions contains up to 2000. The pyramid consequently provides multiple effective receptive fields: fine scales preserve lower-interference local associations, intermediate scales provide progressively broader contextual information, and the global trace provides a coarse sequence-level summary. Long-range interaction is therefore obtained through the composition of bounded-capacity associative memories rather than through a single increasingly overloaded memory trace.

Content-Addressed Holographic Retrieval. For every position and head, contextual information is retrieved using the corresponding normalized retrieval state. Specifically,

$$Z_{i,h} = \mathcal{F}^{-1}\left(\mathcal{F}(M_{i,h}) \odot \overline{\mathcal{F}(R_{i,h})}\right). \tag{20}$$

Here, $Z_{i,h}$ denotes the retrieved holographic representation.

This operation is the algebraic dual of the preceding binding operation. During storage, information is encoded through Fourier-domain multiplication, corresponding to circular convolution in the original domain. During retrieval, the conjugate Fourier representation of the retrieval state performs correlation against the memory. Consequently, retrieval is content-addressed rather than based on explicit pairwise attention scores. The retrieved representations from all heads are concatenated and normalized:

$$Z = \mathrm{LayerNorm}\left(\mathrm{Concat}\left(Z_1, \ldots, Z_H\right)\right). \tag{21}$$

Input-Adaptive Memory Integration. Not every input position requires the same amount of retrieved contextual information. PHM therefore computes an input-dependent integration gate directly from the original representation:

$$G = \mathrm{SiLU}\left(XW_g\right). \tag{22}$$

The retrieved representation is then modulated element-wise:

$$U = G \odot Z. \tag{23}$$

Finally, a learned output projection produces the PHM output:

$$Y = W_o U. \tag{24}$$

The resulting architecture separates three principal associative operations. The *binding state* determines how feature information is encoded into distributed holographic representations; the *memory hierarchy* controls how many bound states are superposed at each temporal scale; and the *retrieval state* determines which information is recovered from the resulting memory through correlation. The input-adaptive gate then controls how strongly the retrieved information contributes to the final representation.

4 RESULTS

We evaluate PHM from complementary perspectives: general sequence modeling, long-range benchmark performance, associative-memory capacity, speech classification and length extrapolation. We additionally test causal language modeling and measure computational scaling. Unless otherwise stated, reported results are means over three random seeds. Our results show that PHM works competitive with associative recall tasks.

Associative Recall Under Increasing Memory Load. We test the capacity hypothesis using Multi-Query Associative Recall (MQAR) Arora et al. (2024). Each sequence contains K randomly positioned key-value pairs followed by query markers, requiring the model to retrieve the corresponding values autoregressively. Unlike conventional classification benchmarks, MQAR directly controls the number of simultaneously stored bindings and therefore provides a controlled test of holographic memory interference.

Table 1: Accuracy for MQAR

K	8	16	32
PHM (our)	88.7±0.5	78.5±0.1	62.9±0.7
Mamba Gu & Dao (2024)	89.2±0.3	79.3±0.2	63.6±0.2
Transformer Vaswani et al. (2017)	89.5±0.0	15.1±0.1	11.1±0.0

With $d_h = 32$, the theoretical HRR analysis predicts a characteristic capacity scale of only a few reliable bindings per trace. PHM instead distributes information across multiple causal memory operations and uses energy-based admission to prevent inactive positions from unnecessarily contributing to the prefix memory. The resulting behavior is shown in Table 1. PHM achieves 88.7% accuracy at $K = 8$, 78.5% at $K = 16$, and 62.9% at $K = 32$. Rather than exhibiting an abrupt failure when the number of bindings exceeds the nominal single-trace capacity, performance degrades progressively with increasing memory load. This behavior is particularly informative when compared with the matched baselines. Mamba follows a remarkably similar degradation profile, reaching 89.2%, 79.3%, and 63.6% at the three loads. Thus, graceful degradation under increasing associative load is not unique to PHM; rather, the result establishes that PHM can achieve this behavior using a fundamentally different memory primitive.

The causal Transformer provides a contrasting failure mode. It reaches 89.5% at $K = 8$ but falls to 15.1% and 11.1% at $K = 16$ and $K = 32$, respectively. PHM therefore retains substantial partial-recall capability well beyond the nominal capacity of an individual holographic trace. Importantly, the MQAR results should not be interpreted as proof that the theoretical SNR directly determines task accuracy. Instead, they demonstrate the qualitative prediction of the theory: controlling the effective number of bindings prevents the abrupt interference-driven collapse observed when a single memory becomes overloaded.

Table 3: Performance on the Long Range Arena (LRA) benchmark. Results are test accuracy (%). Higher is better.

Model	ListOps	Text	Retrieval	Image	Path.	Path-X	Avg.
S4 Gu et al. (2021)	59.60	86.82	90.90	88.65	94.20	96.35	86.09
S5 Smith et al. (2022)	62.15	89.31	91.40	88.00	95.33	98.58	87.46
Transformer Vaswani et al. (2017)	36.37	64.27	57.46	42.44	71.40	FAIL	54.39
HRRFormer Alam et al. (2023a)	39.98	65.38	76.15	50.45	72.17	FAIL	60.83
LRU Orvieto et al. (2023)	60.2	89.4	89.9	89.0	95.1	94.2	86.3
RotRNN Biegun et al. (2024)	61.1	89.6	89.9	85.9	93.0	89.2	84.8
PHM (ours)	41.28	81.99	88.90	64.90	62.4	FAIL	69.27

Autoregressive Language Modeling. The preceding experiments establish retrieval and long-context behavior. We next ask whether PHM can serve as a general causal sequence model for next-token prediction. We therefore evaluate byte-level language modeling on enwik8 Mahoney (2011) using the causal energy-gated PHM formulation. As shown in Table 2, PHM reaches 1.544 ± 0.002 test bits-per-character (bpc), substantially outperforming the causal FNet baseline at 3.629 ± 0.001 bpc. A converged SDPA Transformer remains stronger at 1.364 ± 0.002 bpc. We emphasize that this experiment is not intended as a state-of-the-art enwik8 result. Rather, it demonstrates that the holographic retrieval mechanism remains viable when the task requires genuine autoregressive prediction rather than explicit associative recall. The relatively small gap between PHM and the Transformer at this model scale, together with the large improvement over FNet, indicates that PHM is not restricted to synthetic memory benchmarks. The remaining gap to attention also provides a useful limitation of the current method. At moderate sequence lengths, optimized FlashAttention benefits from highly fused kernels, whereas the present PHM implementation performs several FFT and element-wise operations. Thus, the theoretical linear-memory advantage does not imply a constant-factor wall-clock advantage at every sequence length.

Table 2: enwik8 byte-level language modeling (test bpc, mean$\pm$std over 3 seeds; lower is better; $N = 1024$).

Model	Params	Test bpc	Epochs
Transformer Vaswani et al. (2017)	5.13M	$\mathbf{1.364 \pm 0.002}$	60
FNet Lee-Thorp et al. (2022)	3.55M	3.629 ± 0.001	20
PHM(ours)	5.55M	1.544 ± 0.002	60

Long Range Arena. Table 3 evaluates PHM on three Long Range Arena tasks under matched model budgets. PHM achieves the strongest performance among the attention-free models considered on all three tasks, reaching 41.28%, 64.90%, and 81.99% on ListOps, Pixel, and Text, respectively. The gains over our matched FNet Lee-Thorp et al. (2022) are substantial: $+3.25$ pp on ListOps, $+11.49$ pp on Pixel, and $+15.15$ pp on Text. Moreover, PHM exceeds FNet on every individual seed for all three tasks. With only three seeds, we treat this per-seed dominance descriptively rather than as a formal statistical significance claim. PHM also substantially outperforms equal-budget HRRFormer Alam et al. (2023a), particularly on the long-context tasks where a single global holographic trace becomes heavily overloaded. The contrast is consistent with the capacity analysis in Section 3: PHM bounds the number of bindings contributing to each local memory, whereas HRRFormer exposes every retrieval operation to the full sequence-level interference load. However, S4 remains substantially stronger on ListOps and Pixel, with margins of 18.32 and 23.75 pp, respectively, and reaches 86.82% on Text compared with PHM's 81.99%. This provides an important characterization of the PHM regime: its advantage is not universal replacement of sequential state-space models, but efficient *content-addressable* long-range interaction.

Speech Commands. LRA is partly synthetic, so we additionally evaluate PHM on the natural long-sequence classification task of Speech Commands. The dataset consists of 1-second waveforms of 35 spoken English words sampled at 16 kHz. The task is to classify the word from its sampled waveform. PHM achieves 91.94% accuracy (see Table 4), compared with 96.8% for the best reported

Table 4: Results on Speech Commands classification

Model	Params	Acc (%)
S4 Gu et al. (2021)	307k	96.1
S4D Gu et al. (2022)	306k	95.8
Liquid S4 Hasani et al. (2022)	224k	96.8
S5 Smith et al. (2022)	280k	96.8
RotRNN Biegun et al. (2024)	284k	95.2
PHM	**214k**	**91.94**

result. The gap of 4.86 percentage points indicates that PHM still trails the strongest specialized approaches, but its performance demonstrates that holographic mixing can support effective sequence classification without quadratic self-attention.

Length Extrapolation Beyond Training Context. MQAR varies the number of bindings at fixed sequence length. We next isolate the complementary question: *does retrieval quality remain stable as the total sequence length increases while the number of relevant bindings remains fixed?* We use Length-Invariant Associative Recall (LIAR), training all models at $N = 512$ and evaluating without retraining at lengths up to 32,768, a $64\times$ extrapolation. Each sequence contains $K = 8$ relevant key-value pairs, while additional positions are padding.

Table 5: LIAR sweep A: recall accuracy (%) vs. evaluation length.

Model	$N = 512$	1K	2K	4K	8K	16K	32K
Flash Attention Dao et al. (2022)	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Mamba Gu & Dao (2024)	91.6	91.6	91.6	91.6	91.7	91.7	91.6
GRU Chung et al. (2014)	12.5	12.7	12.2	12.3	12.4	12.6	12.4
PHM (ours)	90.5	90.4	90.1	90.5	90.3	90.4	90.5

The results in Table 5 provide direct evidence for the length-independence predicted by the PHM memory construction. PHM remains essentially constant across the entire range, achieving 90.5% at $N = 512$ and 90.5% at $N = 32{,}768$, with less than 0.4 pp variation across the evaluated lengths. Mamba exhibits similarly stable behavior, reaching approximately 91.6% throughout the sweep. Flash attention achieves exact recall at all tested lengths, but retains quadratic attention computation. GRU remains at chance, illustrating the difficulty of one-shot associative retrieval using a fixed-size recurrent state. The key distinction is therefore computational rather than absolute recall accuracy: PHM preserves associative retrieval across a $64\times$ increase in sequence length while avoiding the quadratic interaction matrix. This behavior follows naturally from Eq. 8: the interference of the finest memory depends on its chunk capacity rather than the total number of tokens.

Long-Context Computational Scaling. We finally evaluate the computational behavior that motivates PHM as a long-context mixer and compare with Scaled Dot-Product Attention (SDPA) Vaswani et al. (2017). Table 6 reports forward-plus-backward latency on an RTX 3090 for sequence lengths from 1K to 131K (at batch=4, D = 256). At short contexts, PHM incurs a constant-factor overhead from FFTs and projection operations, making optimized FlashAttention faster. As sequence length increases, however, the quadratic scaling of attention Vaswani et al. (2017) becomes dominant and PHM becomes increasingly advantageous. At 131K tokens, PHM provides a $7.57\times$ latency advantage over SDPA under the reported configuration.

Table 6: Forward+backward latency and speedup.

N	PHM (ms)	SDPA (ms)	Speedup
1,024	3.4	2.1	0.62×
2,048	6.1	4.4	0.72×
4,096	11.4	9.9	0.87×
8,192	22.4	25.7	**1.15**×
16,384	43.8	77.4	**1.77**×
32,768	88.8	261.8	**2.95**×
65,536	225.1	979.5	**4.35**×
131,072	518.6	3925.4	**7.57**×

The empirical scaling is consistent with the theoretical complexity: over the long-context regime, PHM grows approximately as $N \log N$, whereas attention exhibits the expected quadratic growth. PHM therefore trades a modest short-context kernel overhead for substantially better scaling at long sequence lengths. PHM does require more intermediate memory than the highly optimized SDPA implementation at the tested lengths because the current implementation materializes FFT buffers. This is a practical implementation limitation rather than a change in asymptotic complexity. At 131K tokens the model remains executable on the evaluated GPU, while larger configurations become dominated by intermediate-buffer memory requirements. We therefore interpret the current measurements as evidence of favorable asymptotic scaling rather than as an optimized systems implementation.

5 Conclusion

We presented PHM, a pyramidal holographic sequence mixer that explores circular-convolution-based associative memory as an alternative to quadratic self-attention. Across long-range sequence tasks, PHM achieves competitive performance while offering favorable scaling at long sequence lengths. Our experiments also demonstrate robust length extrapolation and competitive efficiency compared with attention-based and attention-free baselines. Overall, the results suggest that hierarchical holographic memory is a promising direction for efficient long-context sequence modeling.

Reproducibility Statement.

We provide the mathematical formulation, theoretical assumptions, architectural details, and training configurations necessary to reproduce the results reported in the main paper and the supplementary material. The appendix specifies the implementation details for PHM, including memory construction, retrieval, positional encoding, and causal processing, as well as the dataset-specific model configurations. Source code and configuration files are provided as supplementary material where permitted.

AI use statement

We used generative AI tools for writing assistance, including drafting, editing, and polishing this paper's text. The authors reviewed and revised all AI-assisted text and take full responsibility for the final content, including its accuracy, claims, and presentation.

References

Mohammad Mahmudul Alam, Edward Raff, Stella Biderman, Tim Oates, and James Holt. Recasting self-attention with holographic reduced representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 490–507. PMLR, 2023a.

Mohammad Mahmudul Alam, Edward Raff, Stella Biderman, Tim Oates, and James Holt. Recasting self-attention with holographic reduced representations. In *International Conference on Machine Learning*, pp. 490–507. PMLR, 2023b.

Mohammad Mahmudul Alam, Alexander Oberle, Edward Raff, Stella Biderman, Tim Oates, and James Holt. A walsh hadamard derived linear vector symbolic architecture. In *Advances in Neural Information Processing Systems*, 2024.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Re. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2024.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Kai Biegun, Rares Dolga, Jake Cunningham, and David Barber. Rotrnn: Modelling long sequences with rotations. *arXiv preprint arXiv:2407.07239*, 2024.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Ross W Gayler. Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience. In *ICCS/ASCS International Conference on Cognitive Science*, pp. 133–138, 2003.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in neural information processing systems*, 35:35971–35983, 2022.

Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models. *arXiv preprint arXiv:2209.12951*, 2022.

Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159, 2009.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165, 2020.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. FNet: Mixing tokens with Fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4296–4313, 2022.

Matt Mahoney. Large text compression benchmark. `http://mattmahoney.net/dc/text.html`, 2011. enwik8: first 10^8 bytes of English Wikipedia.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International conference on machine learning*, pp. 26670–26698. PMLR, 2023.

Bo Peng, Eric Alcaide, Quentin Anthony, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, 2023.

Tony A. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3): 623–641, 1995.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078, 2023.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations (ICLR)*, 2021.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

A Appendix

A.1 HRR Capacity Theorem: Full Proof

Setup. Let

$$\{x_i\}_{i=1}^N, \qquad \{p_i\}_{i=1}^N$$

be independent draws from

$$\mathcal{N}\left(0, \frac{1}{d} I_d\right).$$

We use the standard HRR circular-convolution convention and define the mean-pooled memory trace as

$$M = \frac{1}{N} \sum_{i=1}^{N} x_i \circledast p_i.$$

Retrieval of item j using the conjugate key is

$$R_j = M \circledast \tilde{p}_j,$$

where $\tilde{p}_j$ denotes the conjugate/reversal of p_j.

Step 1: Decomposition into signal and cross-talk. By linearity of circular convolution,

$$R_j = \frac{1}{N} \sum_{i=1}^{N} (x_i \circledast p_i) \circledast \tilde{p}_j.$$

Separating the $i = j$ term gives

$$R_j = \frac{1}{N} (x_j \circledast p_j) \circledast \tilde{p}_j + \frac{1}{N} \sum_{i \neq j} (x_i \circledast p_i) \circledast \tilde{p}_j.$$

Under conjugate unbinding, the first term approximately reproduces x_j. Hence,

$$R_j = \frac{1}{N} x_j + \eta_j,$$

where

$$\eta_j = \frac{1}{N} \sum_{i \neq j} \epsilon_{ij}, \qquad \epsilon_{ij} = (x_i \circledast p_i) \circledast \tilde{p}_j$$

is the aggregate cross-talk noise.

Step 2: Signal energy. Since

$$x_j \sim \mathcal{N}\left(0, \frac{1}{d} I_d\right),$$

we have

$$\mathbb{E}\|x_j\|_2^2 = d \cdot \frac{1}{d} = 1.$$

Therefore, the expected squared norm of the retrieved signal is

$$P_{\text{sig}} = \mathbb{E}\left\| \frac{1}{N} x_j \right\|_2^2 = \frac{1}{N^2}.$$

Step 3: Cross-talk energy. For $i \neq j$, the vectors x_i, p_i, p_j are independent. Under the standard HRR normalization, conjugate unbinding produces a cross-talk vector satisfying

$$\mathbb{E}[\epsilon_{ij}] = 0, \qquad \mathbb{E}\|\epsilon_{ij}\|_2^2 \approx \frac{1}{d}$$

for large d. The cross-talk terms are therefore approximately uncorrelated, giving

$$\mathbb{E}\|\eta_j\|_2^2 \approx \frac{1}{N^2}\sum_{i\neq j}\mathbb{E}\|\epsilon_{ij}\|_2^2 = \frac{N-1}{dN^2}.$$

Thus,

$$P_{\text{noise}} \approx \frac{N-1}{dN^2}.$$

Step 4: Signal-to-noise ratio. Defining SNR as the ratio of total expected signal energy to total expected cross-talk energy,

$$\text{SNR} = \frac{P_{\text{sig}}}{P_{\text{noise}}},$$

we obtain

$$\text{SNR} \approx \frac{1/N^2}{(N-1)/(dN^2)} = \boxed{\frac{d}{N-1}}.$$

Thus, the characteristic capacity of mean-pooled HRR memory grows linearly with the representation dimension:

$$N = O(d).$$

Step 5: Heuristic reliability bound. For sufficiently large N, the aggregate cross-talk can be approximated by a Gaussian random vector by the central limit theorem. If a particular retrieval procedure requires an effective SNR of at least $\log(1/\delta)$ to achieve an error probability on the order of δ, then

$$\frac{d}{N-1} \gtrsim \log\frac{1}{\delta},$$

which gives the heuristic condition

$$\boxed{N \lesssim 1 + \frac{d}{\log(1/\delta)}}.$$

This final expression is a Gaussian-tail-based reliability heuristic rather than a direct consequence of the SNR identity. The rigorous result established above is the $d/(N-1)$ SNR scaling and the resulting $O(d)$ capacity scaling.

A.2 HRR Capacity Empirical Verification

We empirically verify the SNR prediction by storing N random bindings in a D-dimensional HRR trace and measuring retrieval accuracy (Table 7, Figure 2). Accuracy collapses rapidly as N approaches D: at $N = D = 16$, retrieval drops to 13.5%. This validates the theoretical capacity limit and directly explains HRRFormers's failure at $d_h = 32$ on long sequences.

Table 7: HRR retrieval accuracy (%) vs. binding count at trace dimension $D = 16$ (the full $D \in \{16, \ldots, 512\}$ sweep is in Figure 2).

D	N	Cosine sim	Acc (%)
16	1	0.769	99.0
16	4	0.446	43.1
16	16	0.234	13.5
16	32	0.163	7.8
16	64	0.118	5.6

A.3 Hyperparameters

Table 8 provides the hyperparameters used in the expreriments.

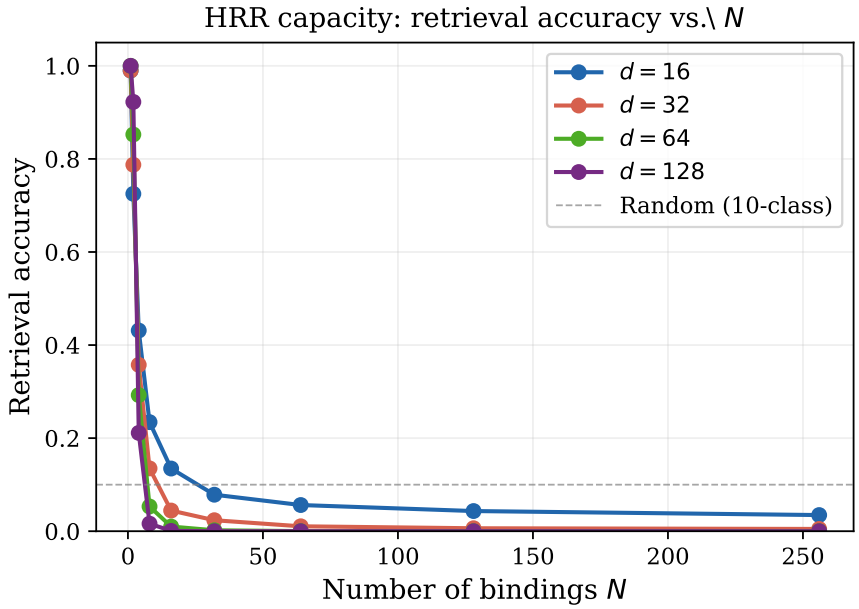


Figure 2: HRR retrieval accuracy vs. number of bindings for varying trace dimensions.

Table 8: Hyperparameters for all experiments.

Task	D	L	H	lr	wd	dropout	epochs
Speech Cmds	320	6	8	2e-3	0.1	0.2	50
LRA (all)	128	4	4	1e-3	0.01	0.1	50
MQAR	256	6	8	1e-3	0.05	0.1	100

A.4 Parameter Counts

Table 9 provides a per-component parameter breakdown for a single PHM block.

Table 9: Parameters per PHM block ($D = 128$).

Component	Formula	Count
W_{mix} (Q/K/V proj.)	$D \times 3D$	49,152
Depthwise conv ($k = 3$)	$3 \times 3D$	1,152
W_{gate}	$D \times D$	16,384
Post-mix LayerNorm	$2D$	256
W_{out}	$D \times D$	16,384
LayerScale γ	D	128
Mixer subtotal	$\sim 5D^2$	83,456
FeedForward ($4\times$ MLP)	$\sim 8D^2$	131,072
Block total	$\sim 13D^2$	214,528

Full model parameter counts: HRR_LRA_Pixel $\approx$ 870K, HRR_LRA_Text $\approx$ 900K, HRR_LRA_ListOps $\approx$ 870K. The Speech Commands model ($D = 320$, $L = 6$) has $\sim$ 5.3M parameters.

A.5 Complexity and Scaling

PHM avoids constructing an $N \times N$ interaction matrix. The projection operations require $\mathcal{O}(ND^2)$ computation, while Fourier-domain binding and retrieval require $\mathcal{O}(ND \log D)$. Hence, for fixed representation dimension,

$$\text{Time}\mathcal{O}(ND^2 + ND \log D), \qquad \text{Memory}\mathcal{O}(ND). \tag{25}$$

In contrast, self-attention requires $\mathcal{O}(N^2 D)$ mixing computation and $\mathcal{O}(N^2 + ND)$ activation memory. The key scaling property of PHM is therefore not merely the absence of quadratic attention. It is the simultaneous separation of *sequence length* from *associative-memory interference*: computational and activation costs remain linear in N, while the finest memory's interference is controlled by its fixed capacity $C_{\min}$ rather than by the total number of tokens.

For the LRA configuration with $D = 128$ and $L = 4$, the PHM mixer contains approximately $5D^2$ parameters, while the accompanying feed-forward component contains approximately $8D^2$, resulting in approximately $13D^2 \approx 213$K parameters per block and ~ 870K parameters for the complete model. Thus, the defining contribution of PHM is a *capacity-aware hierarchical associative memory*: rather than scaling HRR by increasing its dimensionality, PHM controls interference structurally through bounded local memories, combines those memories across temporal resolutions, and selectively admits informative states in the causal setting.