HOFA: CHANNEL-WISE HYBRID ATTENTION VIA OUTLIER FACTORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

KV cache memory remains a primary bottleneck during long-context autoregressive decoding. While prior hybrid architectures mitigate this by alternating exact and linear attention across macro structural axes (depth, width, or sequence), they impose a coarse tradeoff between decoding efficiency and fine-grained associative recall. We introduce Hybrid Outlier-Factorized Attention (HOFA), a channel-wise intra-head decomposition. Inside every attention head, HOFA statically allocates a small r-dimensional outlier subspace to exact causal Softmax attention while routing the remaining inlier channels to a recurrent Gated Linear Attention (GLA) state. Hardware profiling demonstrates that this intra-head split reduces the decoding KV-cache footprint by 43.7% and yields up to a +75.9% peak decoding throughput speedup over standard MHA. Pretraining across 70M, 125M, and 350M parameter scales establishes that HOFA matches MHA perplexity and achieves competitive zero-shot reasoning accuracy. We show that this factorization naturally concentrates global associative matching into the outlier channels while the linear recurrence absorbs background context, with retrieval limits closely following a proposed theoretical extreme-value capacity model.

https://anonymous.4open.science/r/OutlierNeuronAttention

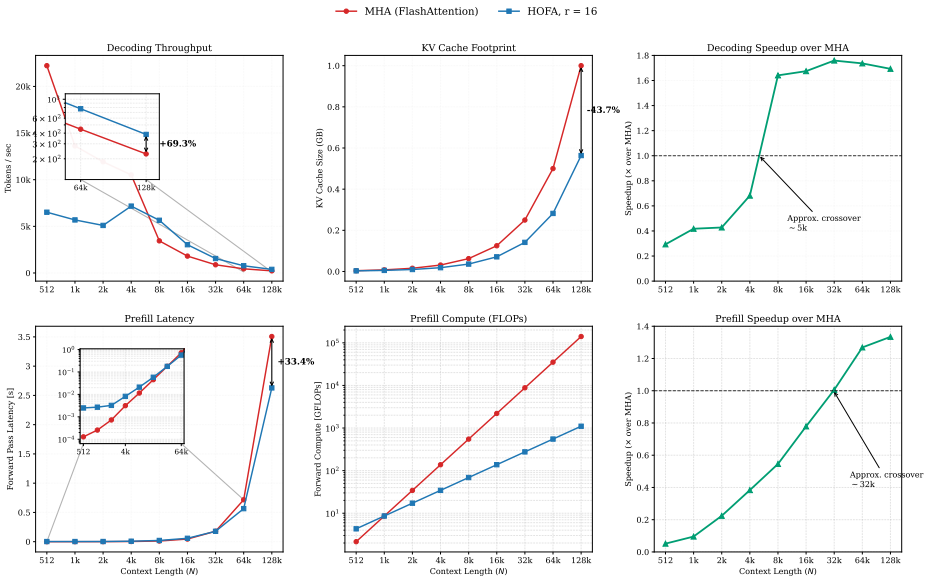


Figure 1: **Hardware Execution & Throughput Benchmarks.** Empirical autoregressive decoding throughput and KV-cache footprint scaling (top row) alongside prefill latency and compute scaling (bottom row) comparing HOFA to FlashAttention-2 MHA.

1 INTRODUCTION & RELATED WORK

Multi-Head Attention (MHA) (Vaswani et al., 2017) forms the computational core of modern Large Language Models (LLMs), but decoding is memory-bandwidth bottlenecked over long contexts. For

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

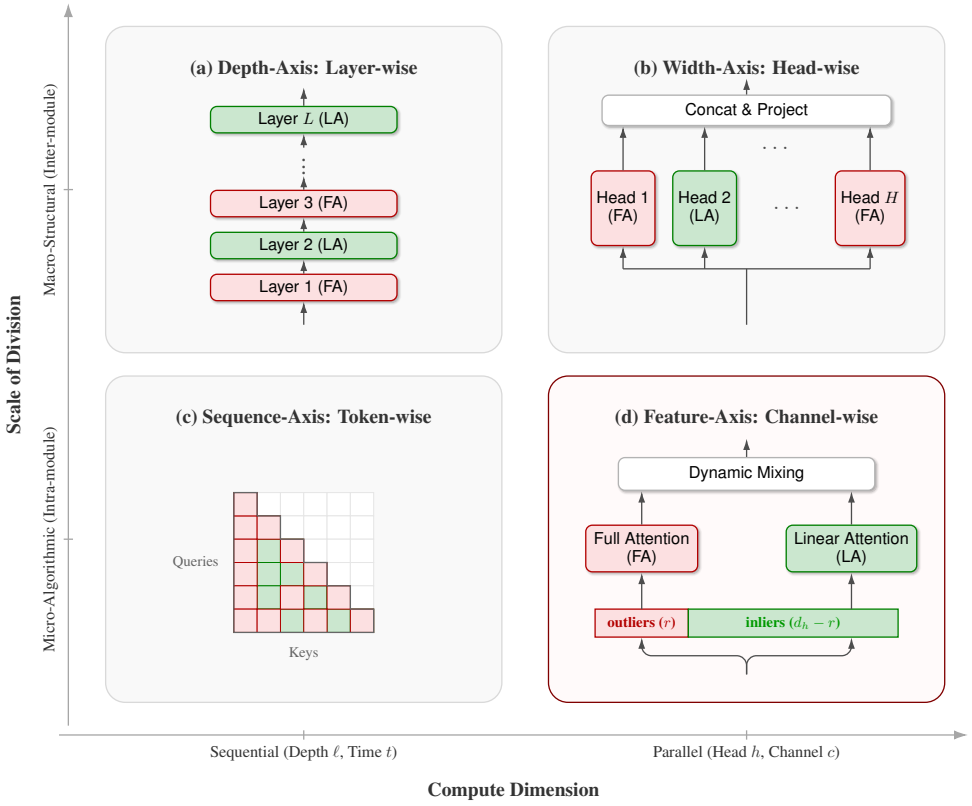


Figure 2: **The Architectural Taxonomy of Hybrid Attention.** Existing hybrid models decompose the attention operation along the (a) layer, (b) head, or (c) token dimension. HOFA introduces the (d) feature dimension axis, processing every token simultaneously through disjoint exact (**Softmax**) and linear (**GLA**) subspaces within the same head.

a sequence length N, standard MHA must cache d_h-dimensional Key and Value vectors for every token across all heads and layers, incurring an $\mathcal{O}(NLHd_h)$ memory footprint per sequence. During generation ($N = 1$), memory bandwidth is dominated by streaming this large-scale KV-cache from High-Bandwidth Memory (HBM) into SRAM, limiting decoding throughput (Dao et al., 2023).

To address this memory bottleneck, recent research has explored hybrid architectures that combine exact $\mathcal{O}(N^2)$ Softmax attention with $\mathcal{O}(1)$ linear-time State-Space Models (SSMs) (Gu & Dao, 2023; Dao & Gu, 2024) or linear attention mechanisms (Peng et al., 2023; Yang et al., 2023; 2025). Existing hybrid approaches decompose attention along three macro structural axes (Figure 2): (i) **depth-axis hybrids** alternating pure Softmax and linear layers sequentially across depth (e.g., StripedHyena (Poli et al., 2023), Jamba (AI21 Labs, 2024)); (ii) **width-axis hybrids** assigning distinct full-attention and linear-attention heads in parallel within each layer (e.g., Hymba (Dong et al., 2024)); and (iii) **sequence-axis hybrids** dynamically routing token positions or sliding-window blocks to exact attention while processing background context linearly (e.g., Longformer (Beltagy et al., 2020), Native Sparse Attention (Yuan et al., 2025)).

While existing hybrids multiplex across depth, width, or sequence dimensions (Figure 2a-c), HOFA introduces the fourth architectural axis: channel-wise intra-head decomposition (Figure 2d). Inside every attention head, HOFA statically splits the hidden dimension d_h into a small r-dimensional outlier channel ($\widetilde{Q}_{\mathcal{O}}, \widetilde{K}_{\mathcal{O}}$) and a remaining $(d_h - r)$-dimensional inlier channel ($\widetilde{Q}_{\mathcal{I}}, \widetilde{K}_{\mathcal{I}}$). The r outlier dimensions receive exact causal Softmax attention with Rotary Position Embeddings (RoPE) (Su et al., 2024), while the inlier dimensions are processed by a data-dependent GLA state.

This architectural factorization is motivated by systematic feature outliers in trained Transformers (Dettmers et al., 2022; Pierro & Abreu, 2024; An et al., 2025). Quantization studies (Dettmers et al., 2022; Pierro & Abreu, 2024) demonstrate that extreme activation magnitudes cluster within a few systematic feature dimensions, which route specific semantic sub-tasks (An et al., 2025).

Optimization concentrates discriminative factual matching into a small subset of outlier channels ($\widetilde{Q}_{\mathcal{O}}, \widetilde{K}_{\mathcal{O}}$), while offloading background context to a recurrent GLA inlier state. Here, "outlier" refers to the high-magnitude feature subspace learned by the model; the allocation is architectural rather than determined by a runtime outlier detector. By establishing a dedicated spatial channel (r) inside every attention head, HOFA allows these feature outliers to operate with exact causal Softmax attention while avoiding full KV-cache memory overheads.

We formalize our contributions across three core axes. First, we model HOFA's context capacity limits via the Welch bound, consistent with empirical induction sequence degradation. Second, hardware profiling demonstrates a 43.7% KV-cache footprint reduction and a +75.9% peak decoding throughput speedup over MHA. We motivate this architectural choice by analyzing attention mass distribution across production pretrained models. Finally, pretraining across 70M, 125M, and 350M scales demonstrates perplexity parity with MHA and strong zero-shot reasoning performance under extended $2\times$ Chinchilla-optimal budgets (Hoffmann et al., 2022).

2 Methodology & Theoretical Framework

2.1 Architectural Formulation

Problem Setting. Standard Multi-Head Attention (MHA) computes per-head Softmax attention $Y = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_h}} + \mathbf{M}\right) V$, incurring an $\mathcal{O}(N)$ KV-cache memory footprint per token that limits decoding throughput. Conversely, pure linear attention mechanisms compress historical context into an $\mathcal{O}(1)$ state S_t, but sacrifice exact token retrieval over long noise gaps. HOFA seeks to bridge the gap between linear complexity and exact recall by structurally dividing per-head capacity into isolated functional pathways.

Figure 3 illustrates the HOFA token mixer architecture. Inside every attention head, representations are factorized into isolated exact and linear processing streams. Following linear projection, Query and Key representations are statically partitioned into an r-dimensional outlier subspace ($\widetilde{Q}_{\mathcal{O}}, \widetilde{K}_{\mathcal{O}}$) and a $(d_h - r)$-dimensional inlier subspace ($\widetilde{Q}_{\mathcal{I}}, \widetilde{K}_{\mathcal{I}}$). The outlier stream executes exact causal Softmax attention with Rotary Position Embeddings (RoPE) to isolate high-magnitude associative matching, while the inlier stream drives a Gated Linear Attention (GLA) recurrence stabilized by RMSNorm (Zhang & Sennrich, 2019) and scaled by a learnable channel-wise scale vector $\mathbf{s}_{\mathcal{I}}$ to absorb diffuse background context. A learned, context-aware gate α_h dynamically blends both pathway outputs before final projection $\mathbf{W}_O$.

Let $\mathbf{X} \in \mathbb{R}^{B\times N\times d_{\text{model}}}$ denote the input sequence. For each attention head $h \in \{1, \ldots, H\}$, query, key, and value representations are projected via linear weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{model}}\times d_h}$:

$$\widetilde{Q} = \mathbf{X}\mathbf{W}_Q, \quad \widetilde{K} = \mathbf{X}\mathbf{W}_K, \quad V = \mathbf{X}\mathbf{W}_V \quad \in \mathbb{R}^{B\times N\times d_h} \tag{1}$$

Inside each head, we partition the channels into the first r outlier dimensions and the remaining $j = d_h - r$ inlier dimensions:

$$\widetilde{Q} = \left[\widetilde{Q}_{\mathcal{O}} \parallel \widetilde{Q}_{\mathcal{I}}\right], \quad \widetilde{K} = \left[\widetilde{K}_{\mathcal{O}} \parallel \widetilde{K}_{\mathcal{I}}\right] \tag{2}$$

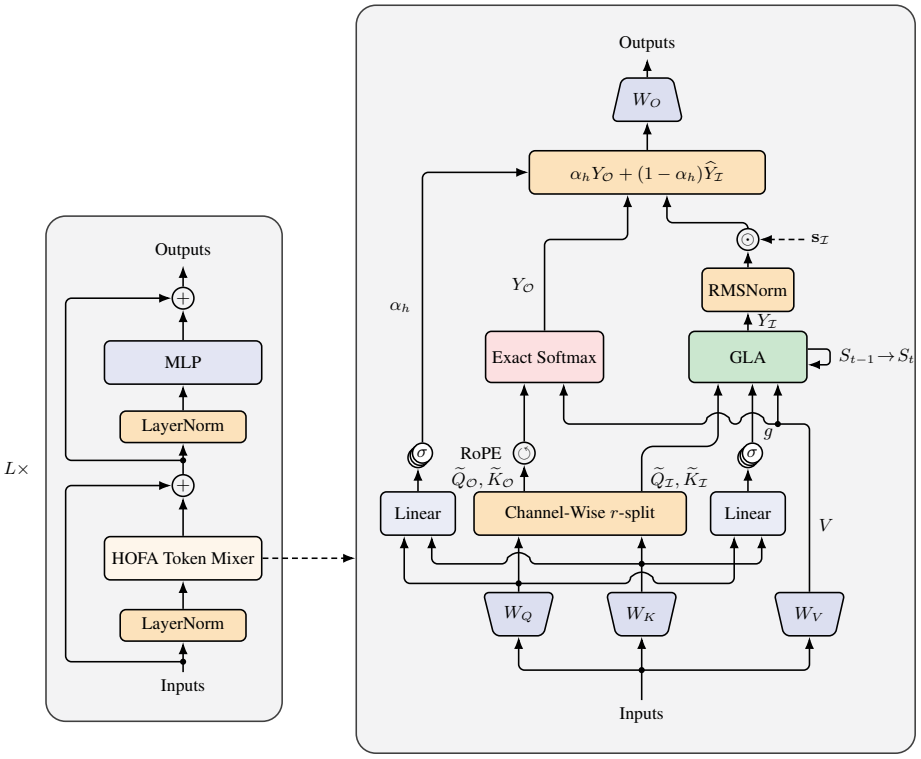
For the outlier pathway, we apply RoPE exclusively to the r-dimensional subspace: $\widehat{Q}_{\mathcal{O}} = \mathrm{RoPE}(\widetilde{Q}_{\mathcal{O}})$, $\widehat{K}_{\mathcal{O}} = \mathrm{RoPE}(\widetilde{K}_{\mathcal{O}})$. The outlier output $Y_{\mathcal{O}}$ is computed using exact causal Softmax attention with variance normalization factor $S = \sqrt{d_h/r}$:

$$Y_{\mathcal{O}} = \mathrm{softmax}\left(\frac{S}{\sqrt{d_h}}\widehat{Q}_{\mathcal{O}}\widehat{K}_{\mathcal{O}}^T + \mathbf{M}\right) V = \mathrm{softmax}\left(\frac{1}{\sqrt{r}}\widehat{Q}_{\mathcal{O}}\widehat{K}_{\mathcal{O}}^T + \mathbf{M}\right) V \tag{3}$$

where $\mathbf{M}$ is the standard causal mask ($\mathbf{M}_{i,j} = 0$ for $i \geq j$, $-\infty$ otherwise).

For the inlier pathway, $\widetilde{Q}_{\mathcal{I}}$ and $\widetilde{K}_{\mathcal{I}}$ drive a GLA recurrence kernel with data-dependent decay $g_t = \sigma(\mathbf{W}_g[\widetilde{Q}_t, \widetilde{K}_t] + \mathbf{b}_g)$:

$$S_t = g_t \odot S_{t-1} + \widetilde{K}_{\mathcal{I},t}^T V_t, \quad Y_{\mathcal{I},t} = \widetilde{Q}_{\mathcal{I},t} S_t \tag{4}$$



(a) Hybrid HOFA Transformer **(b) HOFA Token Mixer (intra-head)**

Figure 3: **Architecture of HOFA. (a)** Pre-LN Transformer layer containing the HOFA token mixer. **(b)** Inside the per-head mixer, projected queries and keys are statically split into r outlier channels (processed by exact causal Softmax attention) and remaining inlier channels (processed by Gated Linear Attention). Both pathways process the full value matrix V, with inlier outputs stabilized via RMSNorm and dynamically fused using a context-aware mixing gate α_h.

The inlier output is normalized via Root Mean Square Normalization (RMSNorm) and scaled by a learnable channel-wise inlier scale vector $\mathbf{s}_{\mathcal{I}} \in \mathbb{R}^{d_h}$ (initialized to 1):

$$\widehat{Y}_{\mathcal{I}} = \mathbf{s}_{\mathcal{I}} \odot \mathrm{RMSNorm}(Y_{\mathcal{I}}) \tag{5}$$

Finally, both streams are dynamically blended via scalar gating $\alpha_h = \sigma(\mathbf{W}_\alpha[\widetilde{Q}, \widetilde{K}])$:

$$Y = \alpha_h \odot Y_{\mathcal{O}} + (1-\alpha_h) \odot \widehat{Y}_{\mathcal{I}}, \quad \text{Output} = Y\mathbf{W}_O \tag{6}$$

2.2 Computational Complexity & Limiting Regimes

For sequence length N, layer count L, head count H, and hidden dimension d_h, standard causal MHA evaluates $QK^\top$ and AV over $\frac{1}{2}N^2$ causal token pairs, requiring $\mathrm{FLOPs}_{\mathrm{MHA,attn}}(N) = 2LHd_hN^2$ FLOPs. In HOFA, exact causal Softmax evaluates $Q_{\mathcal{O}}K_{\mathcal{O}}^\top$ on the r-dimensional outlier subspace, and multiplies attention weights by the full V, yielding an attention-core cost of $LH(r+d_h)N^2$. Accounting for chunkwise inlier GLA recurrence ($\approx 6LH(d_h-r)d_hN$), total forward prefill FLOPs is:

$$\mathrm{FLOPs}_{\mathrm{HOFAPrefill}}(N) = 2PN + LH(r+d_h)N^2 + 6LH(d_h-r)d_hN \tag{7}$$

As $N \to \infty$, the quadratic term dominates, yielding the asymptotic attention-core prefill speedup ceiling relative to causal MHA:

$$\lim_{N\to\infty} \mathrm{Speedup}_{\mathrm{Prefill,attn}}(N) = \frac{2LHd_hN^2}{LH(r+d_h)N^2} = \frac{2d_h}{r+d_h} \tag{8}$$

For autoregressive decoding ($N = 1$), memory bandwidth becomes the primary bottleneck. HOFA mitigates this by storing only r-dimensional keys ($K_{\mathcal{O}}$) alongside full Values (V) and a compact $j \times d_h$ linear state (S), resulting in the footprint ratio:

$$\lim_{N\to\infty} \frac{\text{KV}_{\text{HOFA}}}{\text{KV}_{\text{MHA}}} = \frac{r + d_h}{2d_h} \tag{9}$$

and asymptotic decoding speedup limit:

$$\lim_{N\to\infty} \text{Speedup}_{\text{Decode}}(N) = \frac{2d_h}{r + d_h} \tag{10}$$

For $d_h = 128$ and $r = 16$, Equation 9 predicts a footprint ratio of $\frac{16+128}{256} = 0.5625$ (-43.7% memory reduction) and Equation 10 yields a maximum theoretical decoding throughput speedup ceiling of $\frac{256}{144} \approx 1.777\times$ (+77.7%).

Limiting Architectural Regimes. The HOFA formulation exhibits smooth mathematical continuity between two canonical limiting regimes. In **Pure Linear Attention** ($r = 0$), the outlier subspace vanishes, reducing output to the normalized inlier stream $Y = \mathbf{s}_{\mathcal{I}} \odot \text{RMSNorm}(Y_{\mathcal{I}})$; the quadratic vanishes, collapsing to exact linear attention compute of $6LHd_h^2N$ FLOPs (or $(2P + 6LHd_h^2)N$ FLOPs for prefill). In **Dense Softmax Attention** ($r = d_h$), the inlier subspace vanishes, the linear term cancels to zero, and the quadratic term resolves to $LH(d_h + d_h)N^2 = 2LHd_hN^2$, recovering standard causal Multi-Head Attention FLOPs.

2.3 Capacity-Inspired Retrieval Model

We model HOFA's retrieval limits by framing the exact pathway as a continuous associative memory constrained by the aggregate interference floor of the Welch bound (Welch, 1974). As derived in Appendix E, geometric cross-talk bounds the naive SNR ceiling by $\text{SNR}_{\text{naive}} \leq \frac{r}{cN-r}$, which isotropic scaling cannot overcome. Cross-entropy optimization drives an asymmetric gain factor γ within the r outlier channels against $n = N/d_h - 1$ competing distractors, yielding effective $\text{SNR}_{\text{eff}} \approx \gamma \frac{r}{cN-r}$ and logit mean $\mu = \sqrt{\text{SNR}_{\text{eff}}}$. Framing target retrieval as an M-ary hypothesis testing problem governed by Gumbel extreme value statistics (Fisher & Tippett, 1928) yields retrieval probability:

$$P(\text{correct}) \approx \Phi\left(\frac{\mu - b_n}{\sigma}\right) \tag{11}$$

where $b_n \approx \sqrt{2\ln(N/d_h - 1)}$ is the leading-order Gumbel threshold and σ is an empirical calibration parameter. Extended proofs and theoretical derivations are detailed in Appendix E.

3 Experimental Results

3.1 Hardware Execution & Efficiency

Experimental Settings. Appendix A (Table 2) details model configurations and full pretraining hyperparameters. All models are trained using PyTorch bfloat16 mixed-precision and AdamW (WD $= 0.1$) with Cosine Annealing. Guided directly by our empirical Feature-Wise Activation Norm Decomposition analysis across model depth (Section 3.2), we adopt a heterogeneous routing schedule. Concentrating outlier capacity ($r = 64, 32$) at boundary layers captures uncompressed lexical embeddings and next-token logit projections, while the middle $\sim$ 50–70% semantic core tapers to $r = 16$. Full layer-by-layer r-schedules for 70M ($L = 8$), 125M ($L = 12$), and 350M ($L = 24$) models are detailed in Appendix A (Table 3).

Computational Efficiency & Hardware Profiling. We benchmark HOFA against FlashAttention-2 MHA up to $N = 128$k on an NVIDIA RTX 4060 GPU using PyTorch mixed-precision (bfloat16) under dimensions $d_{\text{model}} = 2048, H = 16, d_h = 128, r = 16$. To isolate the core attention mechanism, we evaluate a single attention layer block with identity projections. During autoregressive decoding ($N = 1$), HOFA achieves a 43.7% KV-cache reduction and sustains a +69.3% throughput speedup over MHA at 128k context length, reaching a +75.9% peak speedup at 32k tokens (Figure 1, top), coming close to the $1.77\times$ theoretical bound (Equation 10). During prefill ($N > 1$), HOFA yields a +33.4% latency speedup at 128k tokens (Figure 1, bottom).

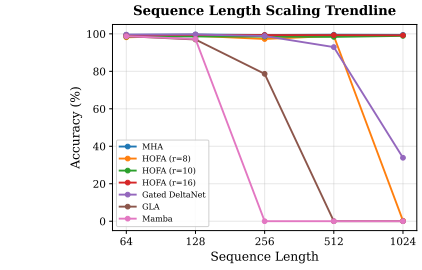
3.2 Mechanistic Validation & Capacity Limits

The Associative Recall Hierarchy. To evaluate sequence-scaling limits on associative recall, we benchmark 4-layer models ($L = 4, d_{\text{model}} = 128, H = 4, d_h = 32$) comparing Dense MHA, HOFA ($r = 8, 10, 16$), Gated DeltaNet, GLA, and Mamba-2 on synthetic Induction Head retrieval (Olsson et al., 2022) across context windows $N \in [640, 1200]$. As shown in Figure 4a, the evaluated architectures fall into four empirical performance regimes, which we refer to as an Associative Recall Hierarchy: Class I (Exact Softmax: MHA, HOFA $r \geq 10$) maintains full retrieval accuracy up to $N = 1024$; Class II (Dynamic Matrix Gated: Gated DeltaNet) delays memory decay via matrix updates, achieving 38% accuracy at $N = 1024$; Class III (Scalar Gated: GLA) transitions to 0% for $N \gtrsim 512$; and Class IV (Pure SSM: Mamba-2) reaches state capacity limits for $N \gtrsim 256$.

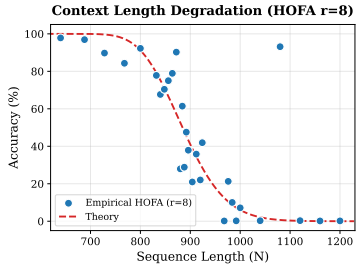
Empirical Capacity Bounds. Figure 4b evaluates the $r = 8$ model via held-out inference with a fixed lookback gap, using average per-token accuracy, the direct empirical estimate of the single retrieval event probability, modeled by Equation 11. Accuracy shows a broad downward trend consistent with the theoretical extreme value capacity model, though individual points, including a localized peak near $N = 1080$, exhibit high variance consistent with the sampling sensitivity of synthetic associative recall tasks noted in Section 4.1.

Feature Norm Disparity. Probing parameter norms across trained induction task checkpoints provides empirical support for the asymmetric amplification mechanism (γ). Standard MHA scatters weight magnitude across all dimensions to capture RoPE frequencies, whereas HOFA ($r = 8, 16$) exhibits a sharp "norm cliff" precisely at index $r-1$, where we observe that optimization concentrates parameter magnitude into the r outlier channels (reaching ≈ 1.85 for $r = 8$) while leaving inlier channels at baseline levels (≈ 0.4, Figure 4c).

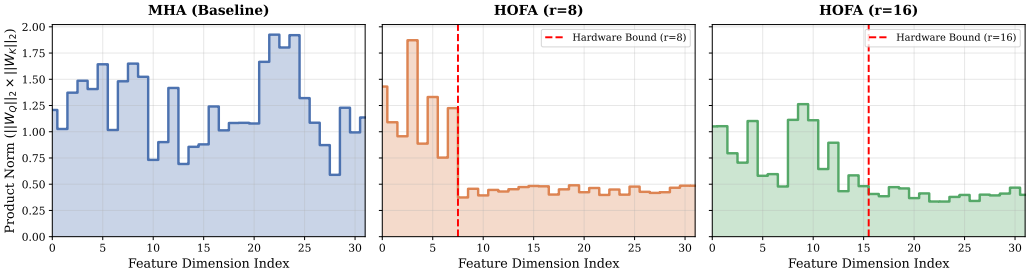
Feature-Wise Activation Norm Concentration. To motivate our choice of outlier capacity, we analyze the feature-wise activation norm decomposition (Vig & Belinkov, 2019; Clark et al., 2019) across eight production Transformer models scaling from 124M to 8B parameters (Appendix F,



(a) Sequence Length Memory Cliff



(b) Empirical vs. Theoretical CDF Model ($r = 8$)



(c) Feature Norm Disparity Across Hidden Dimensions

Figure 4: **Mechanistic Validation & Capacity Dashboard.** (a) Associative recall accuracy vs. sequence length N across 4-layer models ($d_h = 32$). (b) Empirical degradation fitted against the theoretical capacity model. (c) Parameter magnitude (norm) across hidden dimensions in trained weights, showing sharp outlier concentration at index $r - 1$ (dashed red lines mark static hardware boundaries $r = 8, 16$).

Figure 9). Normalizing cumulative feature activation norm by model scale shows that the cumulative feature norm concentrates within approximately the first 16 dimensions ($r \approx 16$). This observation motivates HOFA's static $r = 16$ channel allocation. The decomposition holds across two distinct architectural tracks: (i) pure MHA models (GPT-2 124M, Pythia 410M/1.4B/2.8B) (Xiao et al., 2023; Kovaleva et al., 2019); and (ii) Grouped-Query Attention (GQA) (Ainslie et al., 2023) models (Llama-3.2 1B/3B, Llama-3.1 8B, Qwen2.5 7B). This provides an empirical basis for our symmetrical capacity allocation schedule, descending from $r = 64$ at boundary layers down to $r = 16$ across the middle $\sim$ 50–70% semantic core (Appendix A).

3.3 End-to-End Language Modeling

Pretraining Dynamics. To ensure robust representation learning across the dual pathways, we pretrain 70M, 125M, and 350M models on FineWeb-Edu (Penedo et al., 2024) beyond compute-optimality at 40 tokens per parameter (2× Chinchilla (Hoffmann et al., 2022)). This corresponds to 2.8B, 5B, and 14B tokens at context length $N = 1024$. As shown in Figure 5, 350M HOFA closely tracks standard MHA throughout pretraining, matching final validation perplexity at $\sim 17\%$ fewer total pretraining FLOPs. Because these pretraining runs are conducted at $N = 1024$, where feed-forward layers dominate total compute, this efficiency advantage widens at longer context lengths as attention compute scales toward the asymptotic d_h/r prefill limit (Section 2.2, Equation 8).

Dynamic Mixing Gate Analysis. To characterize the behavior of the dynamic mixing gate α_h, we extract layer-wise distributions from the 350M HOFA model across 1,000 FineWeb-Edu validation sequences (discarding the first 5 attention-sink tokens). Contrary to binary routing expectations, token-level gate values span the full $[0, 1]$ range dynamically across sequence positions, while layer-wise mean gate values across middle and deep layers cluster tightly in a soft-blending regime ($\mu_\alpha \approx 0.38$–0.55, averaging ~ 0.45). Layer 1 exhibits a unique, low-variance concentration at $\mu_\alpha \approx 0.24$, heavily prioritizing the recurrent GLA pathway to establish initial local context (see Appendix C, Figure 7).

Effective Attention Distance: Preservation of Global Reach. To test whether confining exact Softmax to r dimensions limits global attention reach, we measure layer-wise effective attention distance across 1,000 FineWeb-Edu validation sequences. Appendix D (Figure 8) shows 350M HOFA's outlier pathway tracking standard MHA ($d_h = 128$) across all 24 layers (averaging 257.1 vs. 255.9 tokens). Conversely, the GLA inlier pathway operates at a sharp local average of 3.9 tokens. This indicates that HOFA preserves global attention reach, concentrating long-range associative matching in the outlier subspace while the recurrent linear state primarily handles immediate local context.

Zero-Shot Evaluation. To evaluate downstream reasoning capabilities, we benchmark HOFA against standard MHA and established open-source baselines across six common-sense reasoning tasks using `lm-evaluation-harness` (Gao et al., 2021). As detailed in Table 1, 125M HOFA outperforms matched 125M MHA across 4 of 6 benchmarks while remaining competitive with

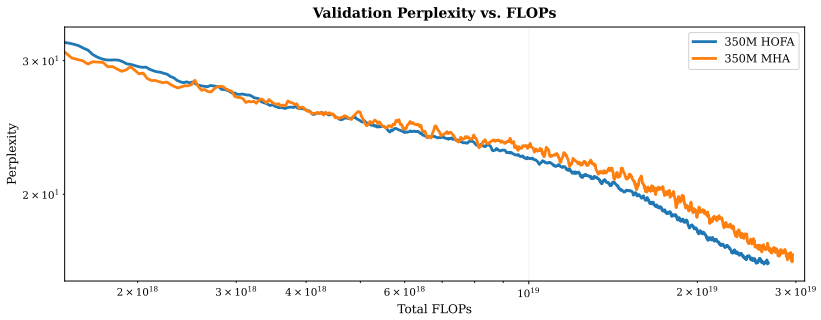


Figure 5: **350M Pretraining Dynamics.** Validation perplexity vs. pretraining FLOPs on FineWeb-Edu. 350M HOFA closely tracks standard MHA throughout training, matching final validation perplexity at $\sim 17\%$ fewer total FLOPs.

Table 1: **Zero-Shot Common-Sense Reasoning Performance (%).** Evaluated across six standard tasks using `lm-evaluation-harness`. All models, including external baselines, were re-evaluated locally under matched evaluation settings. **Bold** denotes overall best in scale class; underline denotes best in 1:1 controlled MHA vs. HOFA pair.

Model	Tokens	ARC-E	ARC-C	PIQA	Wino	OBQA	Hella
125M MHA	5B	**46.2**	24.8	61.3	49.7	29.4	31.2
125M HOFA	5B	44.9	**26.5**	61.2	51.7	**30.2**	31.8
GPT-2 (124M)	10B	39.5	22.7	62.5	51.6	27.2	31.1
Pythia-160M	300B	39.4	23.6	61.2	52.1	26.8	30.1
Mamba-130M	300B	41.9	24.3	**63.2**	**52.5**	28.6	**35.2**
350M MHA	14B	50.4	26.3	66.5	51.1	32.2	41.0
350M HOFA	14B	**53.8**	**28.8**	66.8	51.5	**34.2**	42.1
OPT-350M	180B	40.1	23.7	64.4	52.4	28.4	36.8
Pythia-410M	300B	46.0	24.4	67.1	53.5	29.4	40.6
Mamba-370M	300B	48.2	27.9	**68.4**	**55.6**	31.2	**46.5**

academic baselines such as GPT-2 (Radford et al., 2019), Pythia-160M (Biderman et al., 2023), and Mamba-130M (Gu & Dao, 2023). At the 350M parameter scale, HOFA outperforms matched 350M MHA across all six benchmarks, while scoring higher on ARC-E, ARC-C, and OBQA than higher-token baselines like OPT-350M (Zhang et al., 2022), Pythia-410M (Biderman et al., 2023), and Mamba-370M (Gu & Dao, 2023).

3.4 Ablation Suite & Macro-Axis Structural Controls

Pretraining a 70M parameter ablation suite ($2\times$ Chinchilla-optimal budget) under strict parameter and FLOP parity shows that intra-head HOFA outperforms both Depth-Axis (Jamba-style) and Width-Axis (Hymba-style) hybrids on downstream reasoning, with context-aware gating contributing positively to pathway fusion. Full ablation curves and configurations are detailed in Appendix G.

4 Conclusion, Limitations & Future Work

4.1 Limitations & Future Work

While HOFA establishes an effective channel-wise hybridization paradigm, several limitations should be noted. First, native pretraining is evaluated up to the 350M parameter scale on a single-node setup; scaling end-to-end training to 1B+ parameter models remains for future work. Second, our constant-$r \approx 16$ schedule choice is motivated by post-hoc attention mass analysis of pretrained dense models (Section 3.2), and direct validation of optimal r-schedules during multi-billion parameter pretraining is left for future investigation. Third, pretraining runs and downstream zero-shot reasoning evaluations reflect single-seed executions due to compute constraints. Finally, synthetic associative recall evaluations (such as the standard induction head task) can be prone to sampling variance and local n-gram shortcut reliance. To prevent benchmark hacking artifacts, we position our multi-token block copying benchmarks (Appendix B) strictly as a mechanistic diagnostic probe rather than a standalone performance metric, grounding our main claims in end-to-end pretraining dynamics and zero-shot reasoning.

Looking ahead, extending channel hybridization to composite compression schemes (Appendix H) presents promising directions for extreme long-context scaling. Furthermore, our finding that $r \approx 16$ outlier channels preserve associative recall suggests an alternative design trade-off for HOFA models: reallocating capacity from head width to model depth. Deep, narrow-head HOFA architectures represent an interesting setup for memory-bandwidth-bound inference, particularly on CPU and edge hardware (Dao et al., 2023; Jiang et al., 2025). Systematically mapping this depth–width trade-off in channel-factorized architectures remains an area for future investigation.

4.2 Conclusion

We introduce HOFA, an intra-head hybrid attention layer that factorizes the hidden dimension into exact and recurrent processing streams. A dedicated r-dimensional Softmax subspace isolates and retrieves high-magnitude semantic features, while a Gated Linear Attention state processes the remaining channels to absorb background context. This decouples global associative recall from temporal state decay, recovers both pure linear and dense attention as limiting architectural cases, and enables the exact-attention pathway to preserve long-range associative retrieval while the linear pathway handles local and background context. Under matched pretraining up to 350M parameters, HOFA matches the dense MHA baseline across language modeling, commonsense reasoning, and synthetic retrieval, while reducing the decoding KV-cache footprint and improving peak throughput. Mechanistic analyses show that both pathways contribute to the hybrid state, with optimization naturally concentrating the long-range associative capacity into the Softmax outlier channels.

AI use statement

In this work, we used generative AI tools for drafting experimental code, creating and modifying scientific figures, drafting parts of the paper, formatting references, and proposing keywords. We have not used generative AI tools for helping develop theoretical models, proving mathematical claims, proposing hypotheses, designing research methodology or experiments, or interpreting results. Furthermore, generating synthetic data sets, assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. We rigorously tested all software implementations, and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

To ensure reproducibility, we document all hyperparameters in Appendix A and extended theoretical proofs in Appendix E. Complete anonymized source code for model architectures, training routines, and custom Triton kernels is provided in the supplementary material and available at `https://anonymous.4open.science/r/OutlierNeuronAttention`.

References

AI21 Labs. Jamba: A hybrid transformer-mamba architecture. *arXiv preprint arXiv:2403.19887*, 2024.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yongqi An et al. Systematic outliers in large language models. *International Conference on Learning Representations (ICLR)*, 2025.

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

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning (ICML)*, 2023.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, 2019.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *International Conference on Machine Learning (ICML)*, 2024.

Tri Dao, Daniel Haziza, Francisco Massa, and Grigory Sizov. Flash-decoding for long-context inference. PyTorch Blog, 2023. URL https://pytorch.org/blog/flash-decoding/.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35, 2022.

Xin Dong, Yonggan Ying, Yao Lu, Yuxiong Chen, Peisong Dong, Yingyan Liang, Jan Kautz, Zhuang Liu, and Pavlo Molchanov. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*, 2024.

Ronald A Fisher and Leonard HC Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190, 1928.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Goldings, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *NeurIPS Datasets and Benchmarks*, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Chaoyi Jiang, Lei Gao, Hossein Entezari Zarch, and Murali Annavaram. Kvpr: Efficient llm inference with i/o-aware kv cache partial recomputation. In *Findings of the Association for Computational Linguistics (ACL)*, 2025.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *Proceedings of EMNLP-IJCNLP*, 2019.

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

Guilherme Penedo, Hynek Kydlíček, Antonio Loureiro, Thomas Wolf, and Leandro von Werra. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Alessandro Pierro and Steven Abreu. Mamba-ptq: Outlier channels in recurrent large language models. *arXiv preprint arXiv:2407.12397*, 2024.

Michael Poli, Thomas Wang, Stefano Massaroli, Jeffrey Quesnelle, Ryan Carlow, Eric Nguyen, and Armin W. Thomas. Stripedhyena: Moving beyond transformers with hybrid signal processing models. *Together AI Technical Report*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

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

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 63–76, 2019.

Lloyd Welch. Lower bounds on the maximum cross-correlation of signals. *IEEE Transactions on Information Theory*, 20(3):397–399, 1974.

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

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *International Conference on Learning Representations (ICLR)*, 2025.

Jingyang Yuan et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Full Hyperparameter & Model Configuration Tables

Table 2 lists full pretraining hyperparameters, and Table 3 details the exact layer-by-layer heterogeneous r-schedules across model scales.

Table 2: **Full Pretraining Hyperparameters.**

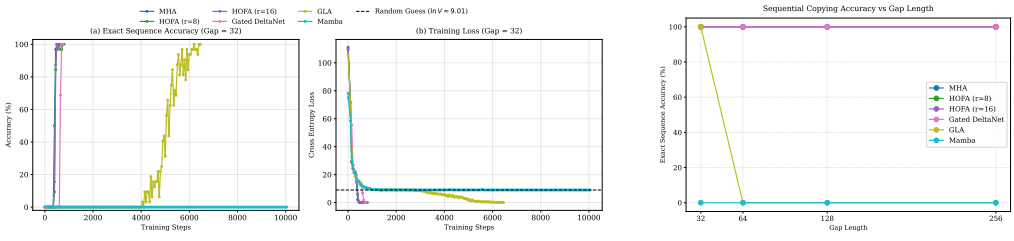
Parameter	**70M**	**125M**	**350M**
Layers (L)	8	12	24
Hidden Dim (d_{model})	512	768	1024
Heads (H)	8	6	8
Training Tokens	2.8B	5.0B	14.0B
Batch Size (sequences)	64	64	128
Sequence Length (N)	1024	1024	1024
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.1	0.1	0.1
Adam β_1, β_2	0.9, 0.95	0.9, 0.95	0.9, 0.95
Warmup Steps	2000	2000	2000
Precision	bfloat16	bfloat16	bfloat16

Table 3: **Heterogeneous Outlier Routing Schedules (r-schedules) Across Model Scales.**

Model Scale	**Layers (L)**	**Layer-by-Layer Outlier Dimension Schedule (r)**
70M	8	$[64, 32, 16, 16, 16, 16, 32, 64]$
125M	12	$[64, 32, 16, 16, 16, 16, 16, 16, 16, 16, 32, 64]$
350M	24	$[64, 32, 32, \underbrace{16, \ldots, 16}_{\times 18}, 32, 32, 64]$

B Extended Multi-Token Block Copying Benchmarks

To isolate gap driven temporal state decay from pattern capacity limits, we evaluate 4-layer models ($L = 4, d_{\text{model}} = 128, H = 4, d_h = 32$, vocab $= 8192$) on multi-token block copying across noise gaps $G \in \{32, 64, 128, 256\}$ at pattern length $P = 2$. As shown in Figure 6a, Class I (MHA, HOFA) and Class II (Gated DeltaNet) converge within ~ 500 steps at $G = 32$, whereas Class III (GLA) requires ~ 5000 steps. Across longer gaps $G \geq 64$ (Figure 6b), GLA collapses to 0% accuracy from recurrent state collisions, while HOFA reliably maintains exact recall up to $G = 256$.



(a) Convergence ($G = 32$). (b) Accuracy vs. noise gap G.

Figure 6: **Multi-Token Block Copying Benchmarks** ($P = 2$)**. (a)** Class I (MHA, HOFA) and Class II (Gated DeltaNet) converge in ~ 500 steps at $G = 32$, while Class III (GLA) requires ~ 5000 steps and Class IV (Mamba-2) flatlines. **(b)** Exact recall accuracy vs. noise gap G; by $G \geq 64$, GLA collapses to 0%, while HOFA preserves retrieval up to $G = 256$.

C Dynamic Mixing Gate Distribution

To characterize dynamic mixing gate behavior, we extract layer-wise distributions from the 350M HOFA model ($L = 24$) across 1,000 validation sequences. As shown in Figure 7, token-level gate values span the full $[0, 1]$ range, demonstrating dynamic, context-dependent blending rather than hard binary routing. Across middle and deep layers, mean gate values cluster in a soft-blending regime. Layer 1 displays an anomalous low-variance distribution, reflecting a strong initial reliance on the recurrent GLA state for local context aggregation.

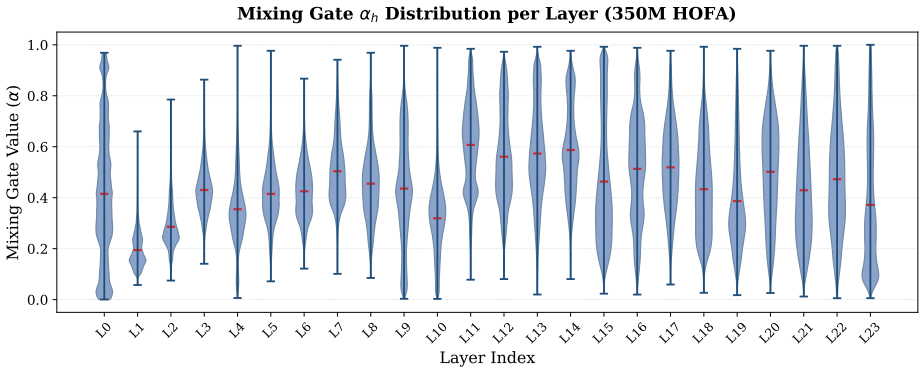


Figure 7: **Mixing Gate α_h Distribution per Layer (350M HOFA).** Layer-wise empirical distribution of the dynamic mixing gate across 24 layers. Token-level gate values span $[0, 1]$ dynamically, while layer-wise means across middle and deep layers cluster in a soft-blending regime ($\mu_\alpha \approx 0.38$–0.55, horizontal ticks). Layer 1 exhibits anomalous, low-variance reliance on the inlier GLA pathway ($\mu_\alpha \approx 0.24$).

D Effective Attention Distance Analysis

To evaluate whether HOFA maintains long range context dependency comparable to full Softmax attention, we compute the layer wise effective attention distance d_l across 1,000 validation sequences from FineWeb-Edu following standard attention distance methodologies. For standard MHA and HOFA's exact outlier pathway, d_l is defined as:

$$d_l^{\text{outlier}} = \frac{1}{N-5} \sum_{i=6}^{N} \sum_{j \le i} p_{ij}^{(l)} (i-j), \quad \text{where } p_{ij}^{(l)} = \text{softmax}_{j \le i} \left(\frac{1}{\sqrt{r}} \widetilde{Q}_{\mathcal{O},i}^{(l)} (\widetilde{K}_{\mathcal{O},j}^{(l)})^T + M_{ij} \right) \quad (12)$$

discarding the first 5 attention-sink tokens (Xiao et al., 2023). Because the unrolled linear attention state yields unbounded and potentially negative values, we introduce a magnitude-normalized pseudo-probability $P_{ij}^{(l)}$ to enable direct distance comparisons with Softmax distributions. This is derived from the inlier feature content $A_{ij}^{(l)}$ and the cumulative gate decay matrix $D_{ij}^{(l)}$:

$$A_{ij}^{(l)} = \widetilde{Q}_{\mathcal{I},i}^{(l)} (\widetilde{K}_{\mathcal{I},j}^{(l)})^T, \qquad D_{ij}^{(l)} = \prod_{s=j+1}^{i-1} \sigma(-g_s^{(l)}) \quad (j < i), \quad (13)$$

$$P_{ij}^{(l)} = \frac{|D_{ij}^{(l)} A_{ij}^{(l)}|}{\sum_{k<i} |D_{ik}^{(l)} A_{ik}^{(l)}|}, \qquad d_l^{\text{inlier}} = \frac{1}{N-5} \sum_{i=6}^{N} \sum_{j<i} P_{ij}^{(l)} (i-j). \quad (14)$$

Figure 8 shows the 350M HOFA outlier pathway tracking standard MHA across all 24 depth layers (averaging 257.1 vs. 255.9 tokens). Conversely, the GLA pathway operates at a sharp local average of 3.9 tokens. This provides empirical evidence that long-range associative matching is concentrated in the outlier subspace, while the recurrent linear state primarily captures immediate local context.

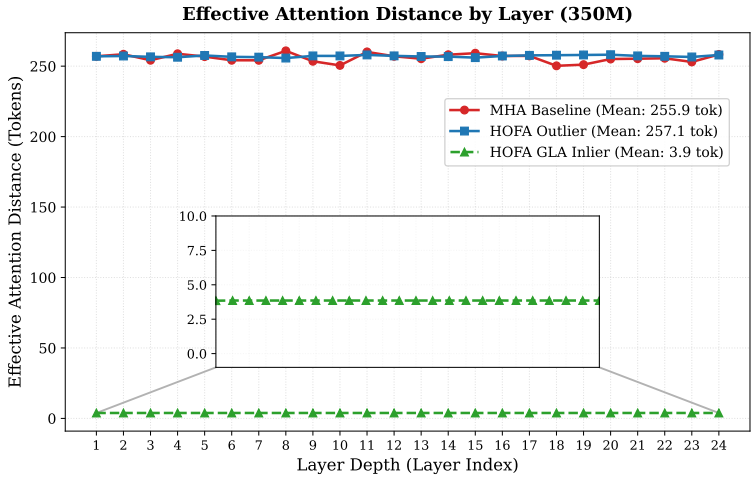


Figure 8: **Layer-Wise Effective Attention Distance (350M Model).** Comparison of standard MHA ($d_h = 128$), HOFA Outlier pathway, and HOFA GLA Inlier pathway across all 24 Transformer layers. The outlier pathway preserves MHA's full global attention reach (257.1 vs. 255.9 tokens), while the GLA state absorbs local contextual background (3.9 tokens).

E Extended Mathematical Proofs & Derivations

The Welch Bound and the Scaling Paradox. Packing N key vectors into a restricted r-dimensional routing subspace creates inevitable geometric interference. By Welch's theorem (Welch, 1974), the aggregate off-diagonal squared inner products of N unit vectors satisfy:

$$\sum_{i=1}^{N}\sum_{j\neq i}^{N} |\langle x_i, x_j\rangle|^2 \geq \frac{N^2}{r} - N = \frac{N(N-r)}{r} \tag{15}$$

Whereas classical coding theory often considers pairwise maximum correlation (Welch, 1974), associative retrieval in attention aggregates interference across all $N-1$ competing tokens simultaneously. The expected per-vector interference power seen by a query vector is:

$$I = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}^{N} |\langle x_i, x_j\rangle|^2 \geq \frac{N-r}{r} \tag{16}$$

which scales linearly with sequence length N. Assuming unit signal alignment for the target key ($\langle q, k_0\rangle^2 = 1$), Equation 16 dictates that the naive Signal-to-Noise Ratio is bounded from above by an achievable capacity ceiling:

$$\text{SNR}_{\text{naive}} = \frac{1}{I} \leq \frac{r}{N-r} \tag{17}$$

To model structured representations in multi-head attention, we introduce the semantic subspace partitioning hypothesis: keys are distributed across d_h approximately orthogonal feature directions, concentrating query competition into an active subspace pool of cN tokens with distractor density $c = 1/d_h$. Within this shared semantic subspace, the naive SNR ceiling becomes:

$$\text{SNR}_{\text{naive}} \leq \frac{r}{cN-r} \tag{18}$$

While continuous interference variance scales with $(cN - r)$, discrete target retrieval faces $n = cN - 1 = N/d_h - 1$ competing distractors. This establishes an apparent scaling paradox: if the network attempts to boost signal separation by uniformly scaling projection weights by an isotropic magnitude factor S ($q \leftarrow S\hat{q}, k \leftarrow S\hat{k}$), target signal power scales as S^2, but distractor noise variance scales proportionally as $S^2 I \geq S^2 \frac{cN-r}{r}$. The factor S^2 cancels algebraically in the SNR ratio, leaving the naive SNR ceiling strictly invariant to uniform magnitude scaling.

Asymmetric Gradient Gain. We model HOFA's feature crystallization as an asymmetric optimization process. Cross-entropy loss rewards increasing the true target logit z_0 relative to distractors

z_i, applying an asymmetric, anisotropic gradient pressure. The optimizer selectively amplifies discriminative subspace directions responsible for pattern matching, while ambient distractor directions receive diffuse gradients under the Softmax partition function. We define $\gamma \gg 1$ as this loss-driven asymmetric gain factor, yielding the effective Signal-to-Noise Ratio:

$$\text{SNR}_{\text{eff}} \approx \gamma \frac{r}{cN - r}, \quad \mu = \sqrt{\text{SNR}_{\text{eff}}} \approx \sqrt{\gamma \frac{r}{cN - r}} \tag{19}$$

This anisotropic gain is empirically manifested by the sharp "norm cliff" in outlier feature projections (Section 3.2, Figure 4c), where optimization concentrates high-magnitude associative matching into the r outlier channels.

M-ary Hypothesis Testing & The Gumbel Threshold. For successful retrieval, the target logit $z_0 \sim \mathcal{N}(\mu, 1)$ must exceed the maximum distractor logit $M_n = \max_{i=1}^{n} z_i$ across $n = N/d_h - 1$ competing keys. Under the Fisher-Tippett-Gnedenko theorem (Fisher & Tippett, 1928), the maximum of n Gaussian distractors converges to a Gumbel distribution with leading-order threshold $b_n \approx \sqrt{2 \ln n}$. The retrieval probability is:

$$P(\text{correct}) = P(z_0 > M_n) \approx \Phi\left(\frac{\mu - b_n}{\sigma}\right) \tag{20}$$

Here, σ is an empirical calibration parameter. While standard extreme value theory assumes independent unit-variance distractors, real transformer representations exhibit feature correlations and non-unit variance; σ calibrates the effective transition width without altering the underlying Gumbel scaling trajectory.

F Attention Mass Distribution Across Model Scales

To quantify how attention weight concentrates across latent feature dimensions, we compute the cumulative Probability Mass Decomposition across query heads:

$$\mathcal{C}(r) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{k=1}^{r} p_q^{(k)} \tag{21}$$

where $p_q^{(k)}$ denotes the sorted attention weight allocated to feature dimension k.

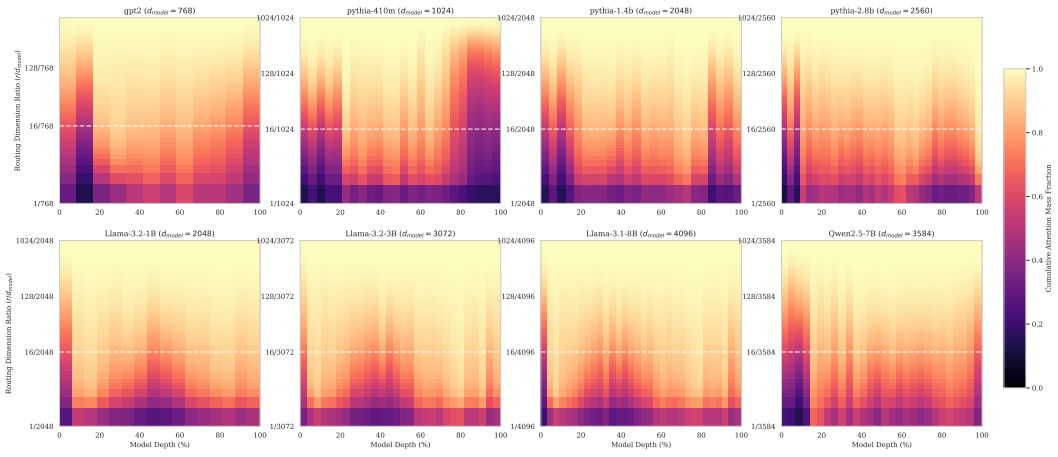


Figure 9: **Feature-Wise Activation Norm Decomposition Across 8 Production Models (2x4 Grid).** Empirical motivation for HOFA's static r-schedule. Top row: Pure MHA track (GPT-2 124M to Pythia-2.8B). Bottom row: GQA spatial compression track (Llama-3.2 1B/3B, Llama-3.1-8B, Qwen2.5-7B). The routing dimension $r \approx 16$ (dashed white line) captures a similar fraction of cumulative attention mass as d_{model} scales up to 8B parameters.

G Extended Ablation Suite & Macro-Axis Controls

To evaluate the intra-head channel-wise hybridization axis against macro-structural alternatives, we pretrain a 70M parameter ablation suite ($L = 8, d_{\text{model}} = 512, H = 8, d_h = 64$) on 2.8B tokens ($2\times$ Chinchilla-optimal budget, 40 tokens/parameter). Zero-shot reasoning benchmarks (Table 4) and validation perplexity curves (Figure 10) reflect strict parameter and FLOP parity across all 70M variants.

Table 4: **70M Parameter Scale Zero-Shot Common-Sense Reasoning Performance.** Evaluated across six standard tasks using `lm-evaluation-harness` ($2\times$ Chinchilla-optimal budget, 2.8B tokens).

Model (70M Scale)	ARC-E	ARC-C	PIQA	Wino	OBQA	Hella
MHA Baseline	**41.2**	**25.4**	57.7	50.6	28.6	28.2
Depth-Axis Hybrid (GLA + MHA)	40.8	22.7	59.1	48.8	28.8	28.0
Width-Axis Hybrid (GLA + MHA)	40.4	23.3	59.0	50.9	30.2	27.7
HOFA (Flat $r = 32$ + Dynamic Gate)	41.1	24.1	57.9	50.8	30.6	27.8
Canonical HOFA (U-Shaped + Dynamic Gate)	40.6	23.7	**60.2**	**51.1**	**30.8**	**28.5**
Static-Blend HOFA ($-$ Dynamic Gate)	39.7	24.6	58.8	**51.1**	30.6	28.2

Capacity Matching (Average $r = 32$). To establish a fair comparison, all macro axis baselines are parameterized to match Canonical HOFA's total exact attention capacity. Canonical HOFA employs a heterogeneous, boundary heavy routing schedule ($[64, 32, 16, 16, 16, 16, 32, 64]$; Table 3), yielding a cumulative budget of 256 exact attention channels across 8 layers (an average of $r = 32$ per layer). Consequently, the Uniform HOFA control fixes $r = 32$ across all layers, while the Depth-Axis and Width-Axis hybrids allocate exactly 50% of their total capacity (256 out of 512 total channels) to exact MHA, ensuring $1:1$ parameter and FLOP equivalence.

Controlled Depth-Axis Baseline. The Depth-Axis Hybrid (Jamba-style) alternates 4 pure MHA layers ($d_h = 64$) and 4 pure GLA layers ($d_h = 64$) using a symmetric $[0, 64, 64, 0, 0, 64, 64, 0]$ schedule. We deliberately omit full attention from the boundary layers (Layers 0 and 7) so that the depth baseline does not implicitly benefit from HOFA's boundary heavy capacity thesis (Section 3.2, Figure 9). Although standard depth-hybrids employ much sparser exact attention ratios (e.g., 1:7), we enforce a strict 1:1 ratio (50% capacity) for this baseline. This ensures that downstream performance differences isolate the architectural routing geometry, rather than variations in total exact attention capacity.

Micro-Architecture Control (GLA vs. Mamba). Across all hybrid baselines (Depth-Axis and Width-Axis), the linear pathway is implemented using our exact Gated Linear Attention (GLA) kernel. Holding the linear micro-architecture constant isolates the structural hybridization axis (layer-wise vs. head-wise vs. channel-wise) as the sole independent variable.

Static vs. Dynamic Blending. In addition to macro axis baselines, we evaluate a Static-Blend HOFA variant where the dynamic token-level mixing gate α_h is replaced by a fixed, layer-wise scalar, isolating the contribution of context-aware dynamic gating.

Ablation Results & Analysis. Results from our ablation suite, detailed in Table 4 and Figure 10, demonstrate the advantages of channel-wise hybridization. While the Homogeneous ($r = 32$) HOFA baseline performs competitively, upgrading to the heterogeneous, boundary-heavy U-shaped r-schedule yields higher downstream accuracy. This aligns with our attention mass decomposition (Figure 9), confirming that allocating exact capacity to boundary layers outperforms a flat distribution. Canonical HOFA outperforms both Depth-Axis (Jamba-style) and Width-Axis (Hymba-style) hybrids across 4 of 6 downstream reasoning benchmarks (including PIQA, WinoGrande, OBQA, and HellaSwag), while optimizing faster to reach target perplexities earlier under matched compute (Figure 10). Finally, removing the dynamic mixing gate (Static-Blend HOFA) degrades accuracy, indicating that dynamic per-token routing projection benefits exact-linear pathway fusion over a static scalar.

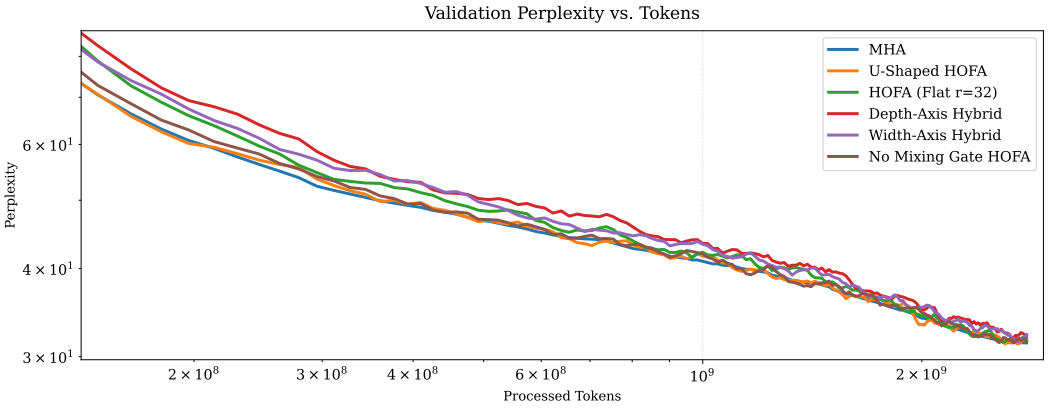


Figure 10: **Validation Perplexity Curves for the 70M Parameter Ablation Suite.** Models are trained on 2.8B tokens (2× Chinchilla-optimal). All hybrid baselines are restricted to an identical 50% exact attention capacity budget. U-Shaped HOFA and the No Mixing Gate variant track the pure MHA baseline with similar optimization trajectories, whereas the Depth-Axis and Width-Axis hybrids exhibit slower initial optimization during the first 10^9 tokens. (A rolling smoothing window is applied to all curves to filter batch-wise variance).

H Orthogonal Architectural Extensions: Composite GQA and MLA Formulations

Multi-Head Latent Attention (MLA) (DeepSeek-AI, 2024) and Grouped-Query Attention (GQA) (Ainslie et al., 2023) are prominent strategies for KV-cache memory reduction, reducing memory footprint via low-rank projection or spatial key sharing. These mechanisms operate along structural dimensions orthogonal to HOFA's intra-head channel decomposition. Below, we propose conceptual composite formulations showing how HOFA can be combined with both mechanisms.

Composite HOFA–GQA. Grouped-Query Attention (GQA) achieves spatial compression by forcing G Query heads to share single Key and Value heads. In contrast, HOFA provides temporal compression by offloading non-outlier channels to a recurrent GLA state. In a composite HOFA–GQA layer, G Query heads share a single r-dimensional Softmax Key head alongside a shared $(d_h - r)$-dimensional GLA inlier state, combining spatial head-sharing with temporal channel compression. A practical consideration is that aggressive spatial head sharing ($G \geq 8$) may constrain key diversity across heads, requiring appropriate selection of r.

Composite HOFA–MLA. Multi-Head Latent Attention (MLA) provides low-rank latent compression by projecting key-value states into compact latent spaces. Specifically, MLA requires a decoupled positional key path because low-rank compression and rotary embeddings do not commute; since HOFA already confines RoPE exclusively to the r-dimensional outlier subspace, this decoupling is inherently satisfied without additional positional projection machinery. Low-rank latent compression can therefore be applied directly to the unpartitioned Value stream V alone by caching a single shared low-rank latent vector $c_t^V \in \mathbb{R}^{d_c}$ per token (where $d_c < Hd_h$), reconstructing per-head value states at read time. This reduces the per-token cache footprint from $H(r + d_h)$ to $Hr + d_c$, showing that intra-head channel factorization is structurally compatible with latent compression. Retaining uncompressed outlier keys ($K_{\mathcal{O}}$) separates retrieval precision from value-stream compression, while decompressing value vectors introduces additional reconstruction operations during decoding.

We emphasize that these composite formulations serve as conceptual blueprints that suggest HOFA may be structurally compatible with orthogonal KV-compression mechanisms. We leave empirical training and hardware benchmarking of these composite configurations to future work.

I IMPLEMENTATION & TRITON KERNEL ENGINEERING

To optimize the exact outlier pathway under FlashAttention execution logic, we implement Causal Loop Peeling (Dao, 2023) in a custom Triton kernel; complete kernel implementations and evaluation suites are provided in the linked anonymized repository. Below is a streamlined structural excerpt of the fused forward kernel (`_fwd_kernel`) illustrating the two-phase loop peeling execution (unmasked block-sparse loop below diagonal and masked diagonal block):

```
@triton.jit
def _fwd_kernel(
   Q, K, V, Out, mix_g_ptr, sm_scale,
   stride_qz, stride_qh, stride_mix_g_n, ....,
   Z, H, N_CTX, r,
   BLOCK_M: tl.constexpr, BLOCK_N: tl.constexpr
):
   start_m = tl.program_id(0)
   off_hz = tl.program_id(1)

   # Unmasked blocks below diagonal
   for start_n in range(0, start_m * BLOCK_M, BLOCK_N):
      k = tl.load(k_ptrs, mask=mask_qk[:, None], other=0.0)
      qk = tl.dot(q, k, allow_tf32=True) * sm_scale
      m_ij = tl.maximum(m_i, tl.max(qk, 1))
      p = tl.exp(qk - m_ij[:, None])
      l_i = l_i * tl.exp(m_i - m_ij) + tl.sum(p, 1)
      acc = acc * tl.exp(m_i - m_ij)[:, None] + tl.dot(p.to(v.dtype), v)
      m_i = m_ij

   # Masked diagonal block
   for start_n in range(start_m * BLOCK_M, (start_m + 1) * BLOCK_M,
       BLOCK_N):
      k = tl.load(k_ptrs, mask=mask_k, other=0.0)
      qk = tl.dot(q, k, allow_tf32=True) * sm_scale
      qk = tl.where(offs_m[:, None] >= offs_n_curr[None, :], qk,
          float("-inf"))
      m_ij = tl.maximum(m_i, tl.max(qk, 1))
      p = tl.exp(qk - m_ij[:, None])
      acc = acc * tl.exp(m_i - m_ij)[:, None] + tl.dot(p.to(v.dtype), v)
```