GABA: GEOMETRY-AWARE BILINEAR ATTENTION FOR LINEAR TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear attention makes global context tractable in transformers by replacing dense pairwise attention with a compressed $K^\top V$ state. Recent work has improved what this state encodes through rank augmentation, gating, or multiple local summaries, but the *readout* remains a coordinate-aligned query–key contraction whose interaction geometry is shared across output channels, limiting output-dependent cross-channel interactions and the attainable representation rank. A second, orthogonal gap concerns fusion: the value representation feeding global aggregation and its locally enhanced counterpart evolve independently and meet only at the output, leaving their relative geometry unconstrained. We introduce *Geometry-Aware Bilinear Attention* (GABA), which addresses both within a single associative block. An output channel-conditioned low-rank bilinear refinement allows different output channels to realize distinct query–key geometries and contributes complementary state directions. A contrast-guided anti-symmetric interaction displaces the global and local pathways equally and oppositely about their shared midpoint; although their sum is preserved, subsequent branch-specific transformations induce a non-trivial refinement. The all-linear formulation retains linear sequence complexity while keeping model-level FLOPs essentially unchanged. GABA consistently outperforms strong linear-attention baselines across model scales; the hybrid model achieves 82.03% Top-1 test accuracy on ImageNet-1K, outperforming the second-closest baseline by 0.50, as shown in Fig. 1.

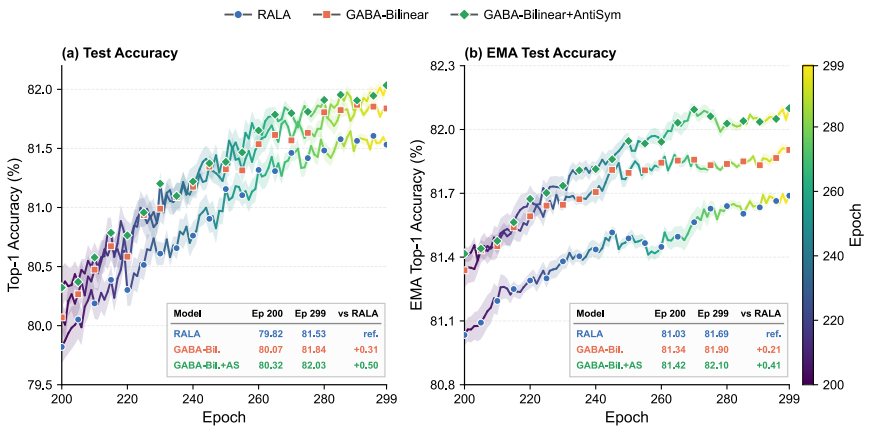

Figure 1: ImageNet-1K top-1 accuracy over the final 100 epochs for RALA, GABA-Bilinear, and GABA-Bilinear+AntiSym. (a) Online weights; (b)EMA weights. Shaded bands denote the local variation envelope; insets give epoch 200/299 values.

1 INTRODUCTION

Linear attention makes global context tractable by replacing the quadratic pairwise computation of Softmax attention with a compact $K^\top V$ state that can be retrieved associatively in time linear in se-

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.

quence length (Katharopoulos et al., 2020; Choromanski et al., 2020). Recent work has substantially narrowed the resulting quality gap by improving *what* linear attention represents through sharper feature mappings, stabilized normalization, polarity- and magnitude-aware similarity, stronger locality, and explicit state-rank augmentation (Zhang et al., 2024; Fan et al., 2025b;a; Meng et al., 2025b; Han et al., 2024a). We ask a complementary question: once information is compressed, *how expressively is it retrieved, and how is it integrated with related representations?* These questions expose two complementary limitations beyond state compression itself.

Retrieval expressiveness. The first concerns retrieval. Standard associative linear attention retrieves the shared state through coordinate-aligned query–key contractions and applies the same interaction geometry across output channels. Learned projections can mix input dimensions, but the retrieval operator itself provides no explicit mechanism for different output channels to exploit distinct cross-channel query–key relations. Moreover, a bilinear transformation shared across all output channels applies a common transformation to the state and cannot increase its rank; therefore, enriching what the state contains does not, by itself, widen the retrieval geometry through which it is read. We therefore introduce an *output-channel-conditioned low-rank bilinear refinement*, allowing different output channels to realize distinct interaction geometries $Q_i M^{(r)} K_j^\top$. This explicitly introduces cross-channel query–key interactions and can contribute complementary state directions while preserving associative computation.

Global–local coordination. A second, orthogonal limitation concerns fusion. The value representation used for global aggregation and its locally enhanced counterpart originate from the same stream but evolve through separate transformations before additive fusion, leaving their relative geometry uncontrolled. We introduce a *contrast-guided anti-symmetric interaction* that applies displacements of $+\Delta$ and $-\Delta$ to the two pathways relative to their original midpoint. Although this exchange preserves their direct sum, it is not functionally neutral: the displaced representations subsequently undergo different global and local transformations, producing a non-trivial branch-dependent refinement. The mechanism therefore coordinates their relative geometry without introducing an independent residual branch or auxiliary separation objective.

Retrieval expressiveness *Is compressed information retrieved through a sufficiently expressive interaction space?*	$\Longrightarrow$	**Bilinear Attention** Enrich cross-channel Q–K interactions and attainable representation rank.
Global–local coordination *Are the global value and local residual pathways explicitly coordinated?*	$\Longrightarrow$	**Anti-Symmetric Interaction** Control their relative representation geometry through a coupled anti-symmetric exchange.

Together, these mechanisms form *Geometry-Aware Bilinear Attention* (GABA), jointly enriching retrieval interactions and global–local representation geometry within a unified attention framework.

Our contributions are: ❶ **Retrieval-Space Insight.** We identify an expressiveness limitation beyond state compression: because conventional associative retrieval applies a channel-shared interaction geometry, the attainable representation rank remains bounded regardless of how rich the compressed state is. ❷ **Geometry-Aware Refinement.** We introduce output-channel-conditioned low-rank bilinear retrieval and contrast-guided anti-symmetric global–local coordination; theoretically, the former can contribute up to R complementary state directions whereas a shared bilinear transformation cannot augment state rank. ❸ **Efficient Unified Formulation.** Both mechanisms integrate into a single associative block, with the all-linear formulation preserving linear sequence complexity at essentially unchanged model-level FLOPs. ❹ **Empirical Validation.** GABA consistently improves strong linear-attention baselines across model scales and vision and language tasks, while state- and representation-level analyses show higher effective dimensionality, broader spectral utilization, and reduced redundancy.

Related Work and Motivation. Due to space constraints, detailed related work and additional motivation are provided in Appendix Sections B and C, respectively, and will be incorporated into the main text in the camera-ready version.

2 METHODOLOGY

GABA addresses two complementary limitations of associative linear attention (Fig. 2). First, an output-channel-conditioned bilinear interaction enriches the query–key retrieval geometry and can

contribute complementary state directions. Second, a contrast-guided anti-symmetric exchange explicitly coordinates the global value and local residual representations around their shared midpoint. Together, these mechanisms enrich retrieval interactions and global–local representation geometry while preserving associative computation. Algorithm 1 summarizes the complete forward pipeline. Throughout, D denotes the block width, h the number of heads, and $d = D/h$ the per-head dimension. Unless stated otherwise, the analysis is written for a single head; kernel and positional transformations are absorbed into Q and K, while unchanged normalization, output gating, and projection are omitted for clarity. Implementation correspondence is provided in Appendix D.8.

2.1 Output-Conditioned Bilinear Retrieval and Rank Enrichment

Baseline rank constraint. For one attention head, let $Q, K, V \in \mathbb{R}^{N\times d}$. Standard associative linear attention compresses the token sequence into the shared key–value state $S = K^\top V$ and retrieves it through Q:

$$S = K^\top V \in \mathbb{R}^{d\times d}, \qquad Y_{\text{base}} = QS, \\ \operatorname{rank}(Y_{\text{base}}) \le \min\{\operatorname{rank}(Q), \operatorname{rank}(S)\}. \tag{1}$$

Thus, rank deficiency in the compressed state directly constrains the attainable output rank (Appendix D.1). Beyond this constraint, conventional retrieval applies one interaction geometry across output channels, providing no explicit mechanism for different channels to exploit distinct cross-channel query–key relations.

Output-channel-conditioned bilinear retrieval. We augment the associative pathway with learnable factors $A, B \in \mathbb{R}^{d\times R}$ and $C \in \mathbb{R}^{R\times d}$, where $R \ll d$; the key and value head dimensions coincide in our configurations, and the appendix retains the general d_v:

$$Y_{\text{bil}} = QS + \gamma(QA)\big[(B^\top S)\odot C\big], \\ M^{(r)} = A\,\operatorname{diag}(C_{:,r})\,B^\top. \tag{2}$$

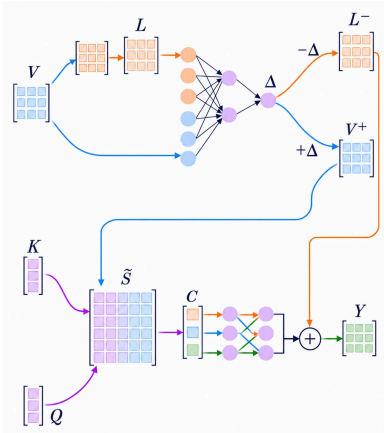

Figure 2: **Overview of GABA. Top:** a local stencil extracts L from the global feature V; a contrast MLP predicts a displacement Δ that updates both branches anti-symmetrically, $V^+ = V + \Delta$ and $L^- = L - \Delta$, preserving their midpoint. **Bottom:** K and V^+ form the associative state $\widetilde{S} = K^\top V^+$, from which Q retrieves under the output-conditioned interaction C; the response is fused with L^- to give Y.

For output channel r, the additional interaction is therefore $Q_i M^{(r)} K_j^\top$. Expanding $M^{(r)}$ yields products $Q_{ic}K_{jk}$ for both $c = k$ and $c \neq k$, while $C_{:,r}$ allows the effective bilinear geometry to vary with the output channel. Hence, A and B define a compact rank-R interaction basis, whereas C conditions that basis on the retrieved output channel. Appendix D.2 gives the pairwise expansion.

This conditioning is essential within the proposed bilinear construction. If the same bilinear transformation is shared across every output channel, the refinement reduces to a common left transformation of S and cannot augment its rank; the formal result is given in Appendix D.4.

Rank enrichment. Equation 2 admits an equivalent refined-state interpretation:

$$Y_{\text{bil}} = Q\widetilde{S}, \qquad \widetilde{S} = S + \Delta S, \quad \Delta S = \gamma A\big[(B^\top S)\odot C\big], \\ \operatorname{rank}(\Delta S) \le R, \qquad \operatorname{rank}(\widetilde{S}) - \operatorname{rank}(S) \le R. \tag{3}$$

Every column of ΔS lies in $\operatorname{col}(A)$, giving the rank-R bound above. The refinement can therefore contribute up to R complementary state directions. Because $M^{(r)}$ varies across output channels, these directions need not collapse to a common transformation of S and may extend beyond its existing column space. Strict augmentation occurs when the update contributes linearly independent directions outside $\operatorname{col}(S)$; if those directions are preserved by Q, the enrichment also appears in the retrieved representation. Appendix D provides sufficient conditions, a constructive strict-rank example, and a tightness result showing that the $+R$ gain is attainable.

2.2 CONTRAST-GUIDED ANTI-SYMMETRIC GLOBAL–LOCAL INTERACTION

Unlike the retrieval analysis above, the anti-symmetric exchange operates on the full-width value representation $V \in \mathbb{R}^{N\times D}$ prior to head splitting; the head-wise scale α_h is subsequently broadcast across the channels of head h.

Global–local decomposition. Let $V \in \mathbb{R}^{N\times D}$ denote the value representation and $L = F(V)$ its locally enhanced counterpart, where $F(\cdot)$ is a 5×5 depthwise convolution. Omitting the unchanged output gate and projection, the baseline response is

$$L = F(V), \qquad Y_{\text{base}} = \mathcal{A}_{Q,K}(V) + L, \tag{4}$$

where $\mathcal{A}_{Q,K}(\cdot)$ denotes associative linear attention for fixed Q and K. Although V and L originate from the same value stream, they subsequently follow distinct global and local pathways and are combined only through additive fusion; their relative geometry is therefore not explicitly controlled.

Contrast-guided displacement. We derive the exchange from the normalized global–local discrepancy so that its direction is scale-insensitive, spatially local, and bounded. For $X \in \{V, L\}$, define

$$\begin{aligned} \rho_i(X) &= \sqrt{\frac{1}{D}\sum_{c=1}^{D} X_{ic}^2 + \epsilon}, & \widehat{X}_{ic} &= \frac{X_{ic}}{\rho_i(X)}, \\ P &= \tanh\Big(F_{\text{dw}}\Big(\text{sg}(\widehat{V} - \widehat{L})\Big)\Big), & \bar{\rho} &= \text{sg}\left[\frac{\rho(V) + \rho(L)}{2}\right], \\ \alpha_h &= \alpha_{\max}\sigma(a_h), & \Delta &= \alpha_h\, P \odot \bar{\rho}. \end{aligned} \tag{5}$$

Here F_{dw} is a learnable bias-free 3×3 depthwise convolution and $\text{sg}(\cdot)$ denotes stop-gradient. RMS normalization removes most per-location magnitude variation from the contrast, F_{dw} allows the displacement to respond to the local discrepancy pattern, and $\tanh(\cdot)$ together with the bounded head-wise scale α_h controls its magnitude. Stop-gradient prevents the displacement-generation pathway from creating an auxiliary gradient route through $\widehat{V}$ and $\widehat{L}$. F_{dw} is zero-initialized, so $P = \tanh(0) = 0$ and hence $\Delta = 0$; the exchange is therefore inactive at initialization and is learned progressively during training.

We then apply the *same* displacement to the two pathways with opposite signs:

$$V^+ = V + \Delta, \qquad L^- = L - \Delta. \tag{6}$$

Unlike two independently learned residual corrections, the exchange is therefore explicitly coupled.

Midpoint preservation and non-trivial refinement. The anti-symmetric construction preserves the shared midpoint while modifying the relative displacement:

$$\frac{V^+ + L^-}{2} = \frac{V + L}{2}, \qquad V^+ - L^- = (V - L) + 2\Delta. \tag{7}$$

Preserving the direct sum does not make the operation functionally neutral. Since $\mathcal{A}_{Q,K}(\cdot)$ is linear in its value argument for fixed Q and K,

$$\begin{aligned} Y_{\text{anti}} &= \mathcal{A}_{Q,K}(V + \Delta) + (L - \Delta) \\ &= Y_{\text{base}} + \big[\mathcal{A}_{Q,K}(\Delta) - \Delta\big]. \end{aligned} \tag{8}$$

Thus, the same displacement is transformed by global attention in one branch while entering directly with opposite sign in the local branch. The resulting term $\mathcal{A}_{Q,K}(\Delta) - \Delta$ constitutes a branch-dependent refinement rather than an independent residual representation. The exchange therefore modifies the *relative* global–local geometry around a fixed shared anchor.

2.3 UNIFIED GABA FORMULATION

GABA-M1. GABA-M1 applies the two proposed mechanisms within every linear-attention block. For one head, the anti-symmetric exchange first gives $\widetilde{V} = V + \Delta$ and $\widetilde{L} = L - \Delta$, from which

$$\begin{aligned} \widetilde{S} &= K^\top \widetilde{V}, \\ Y_{\text{M1}} &= Q\widetilde{S} + \gamma_{qv}\big(\overline{Q} \odot \widetilde{V}\big) + \gamma(QA)\big[(B^\top \widetilde{S}) \odot C\big] + \widetilde{L}. \end{aligned} \tag{9}$$

Algorithm 1 GABA block (one forward pass)

Require: $Q, K, V \in \mathbb{R}^{N\times D}$, $D = hd$, grid $N = HW$
Require: F (5×5 depthwise), F_{dw} (3×3 depthwise, bias-free)
Require: $A_j, B_j \in \mathbb{R}^{d\times R}$, $C_j \in \mathbb{R}^{R\times d}$ for $j = 1, \ldots, h$
Require: head-wise scale $\alpha \in \mathbb{R}^h$ (broadcast over channels), scalars γ, γ_{qv}
Full-width anti-symmetric exchange
$L \leftarrow F(V)$
$\rho_V \leftarrow \mathrm{rms}(V), \quad \rho_L \leftarrow \mathrm{rms}(L)$
$P \leftarrow \tanh(F_{\mathrm{dw}}(\mathrm{sg}[V/\rho_V - L/\rho_L]))$
$\bar{\rho} \leftarrow \mathrm{sg}[(\rho_V + \rho_L)/2]$
$\Delta \leftarrow \alpha \odot P \odot \bar{\rho}$
$\widetilde{V} \leftarrow V + \Delta, \quad \widetilde{L} \leftarrow L - \Delta$
Split $Q, K, \widetilde{V}, \widetilde{L}$ into h heads
Per-head output-conditioned bilinear retrieval
for each head $j = 1, \ldots, h$ **in parallel do**
  $\widetilde{S}_j \leftarrow K_j^\top \widetilde{V}_j$
  $\overline{Q}_j \leftarrow Q_j/(\mathrm{mean}_c\, Q_j + \epsilon)$
  $Y_j \leftarrow Q_j \widetilde{S}_j + \gamma_{qv}(\overline{Q}_j \odot \widetilde{V}_j) + \gamma(Q_j A_j)\big[(B_j^\top \widetilde{S}_j) \odot C_j\big] + \widetilde{L}_j$
end for
return $\mathrm{Concat}(Y_1, \ldots, Y_h)$

Here $\overline{Q}$ denotes the normalized query used by the token-wise Q–V refinement. This Q–V term is retained from the underlying baseline and directly couples each query with its position-aligned value, complementing the globally shared associative state. Its $\mathcal{O}(Nd)$ cost is linear in the number of tokens.

For fixed Q and K, all three value-dependent retrieval terms in Eq. 9 are linear in $\widetilde{V}$. Consequently, the non-cancellation property in Eq. 8 also holds for the complete M1 retrieval operator: the anti-symmetric displacement is transformed differently by the associative, bilinear, token-wise, and local pathways rather than disappearing under fusion.

GABA-M2. GABA-M2 additionally introduces a gated token-specific correction:

$$\mathcal{G}(X) = \sigma(W_g X) \odot UX, \qquad Y_{\mathrm{M2}} = Y_{\mathrm{M1}} + \mathcal{G}(X), \tag{10}$$

where $W_g, U \in \mathbb{R}^{D\times D}$ are learnable channel projections and $\sigma(\cdot)$ denotes the sigmoid function. Since $\mathcal{G}(X)$ operates independently over tokens, it preserves linear scaling with sequence length. Its constant factor is nonetheless substantial: $\mathcal{G}(X)$ contributes approximately $2ND^2$ multiply–accumulate operations against the $2ND^2/h$ of the associative interaction, and $2D^2$ parameters against the $4D^2$ of the shared projections. GABA-M2 therefore trades a channel-wise increase in computation and parameters for an additional token-specific refinement path, and we report both variants throughout so that this trade-off remains visible. Detailed FLOP and parameter accounting for M1 and M2 is provided in Appendix E.

All-linear and hybrid backbones. The M1/M2 distinction is independent of backbone composition. We evaluate two compositions. In the *all-linear* setting, denoted by the subscript L, every attention block uses the proposed linear formulation. In the *hybrid* setting, selected blocks instead use standard Softmax attention, following the placement strategy of RAVLT (Fan et al., 2025b) and MAViT (Fan et al., 2025a). When M2 is used, the gated correction is applied to Softmax blocks as well. All-linear GABA therefore retains linear complexity in sequence length, whereas hybrid variants inherit the quadratic token dependence of their Softmax blocks. Claims of linear sequence scaling throughout the paper refer to the all-linear variants.

3 Experiments

We evaluate GABA on image classification, object detection, instance segmentation, semantic segmentation, and autoregressive language modeling, together with controlled all-linear evaluations that isolate the proposed linear-attention mechanism. Unless otherwise noted, baseline results are

Table 1: **ImageNet-1K comparison with hybrid backbones.** Entries marked * and GABA report Top-1 for EMA weights; the remaining entries are taken from the cited works.

Tiny (∼2.5G)				Small (∼5G)				Base (∼10G)				Large (∼15G)			
Model	P(M)	F(G)	Top-1	Model	P(M)	F(G)	Top-1	Model	P(M)	F(G)	Top-1	Model	P(M)	F(G)	Top-1
RMT-T	14	2.5	82.4	RMT-S	27	4.5	84.1	RMT-B	54	9.7	85.0	RMT-L	95	18.2	85.5
FL-PVT-T	12	2.0	77.8	FL-CSwin-T	21	4.3	83.1	FL-CSwin-S	35	6.9	83.8	FL-CSwin-B	75	15.0	84.5
FL-PVTv2-B1	13	2.2	79.5	VMamba-T	30	4.9	82.6	VMamba-S	50	8.7	83.6	InLine-CSwin-B	73	14.9	84.5
MSVMamba-M	12	1.5	79.8	MSVMamba-T	32	5.1	83.0	MSVMamba-S	50	8.8	84.1	VMamba-B	89	15.4	83.9
RAVLT-T	14.9*	2.6*	82.31*	RAVLT-S	25.6*	4.9*	84.16*	MILA-S	43	7.3	84.4	MSVMamba-B	91	16.3	84.4
MAViT-T	16.5*	2.7*	82.42*	MAViT-S	27.2*	4.9*	84.33*	RAVLT-B	48.3*	10.6*	85.26*	RAVLT-L	95.3*	17.3*	85.49*
MHLA-VLT-T	16.3*	2.4*	82.55*	MHLA-VLT-S	27.2*	4.6*	84.57*	MAViT-B	49.8*	10.7*	85.40*	MAViT-L	97.8*	17.3*	85.64*
GABA-M1-T	14.9	2.6	<u>82.80</u>	**GABA-M1-S**	25.6	4.9	<u>84.83</u>	**GABA-M1-B**	48.2	10.6	<u>85.72</u>	**GABA-M1-L**	95.2	17.2	<u>86.03</u>
GABA-M2-T	16.8	2.9	**82.95**	**GABA-M2-S**	29.2	5.5	**84.99**	**GABA-M2-B**	55.4	11.8	**85.83**	**GABA-M2-L**	110	19.7	**86.13**

Table 2: **ImageNet-1K comparison with all-linear backbones.** Every attention block uses linear attention, and Top-1 is reported for EMA weights. MHLA_L entries are omitted at Base and Large, for which no corresponding configuration is publicly available.

Tiny (∼2.5G)				Small (∼5G)				Base (∼10G)				Large (∼15G)			
Model	P(M)	F(G)	Top-1	Model	P(M)	F(G)	Top-1	Model	P(M)	F(G)	Top-1	Model	P(M)	F(G)	Top-1
RAVLT_L	14.9*	2.6*	81.69*	RAVLT_L	25.6*	4.8*	83.36*	RAVLT_L	48.3*	10.4*	84.49*	RAVLT_L	95.3*	16.9*	84.72*
MAViT_L	16.5*	2.6*	81.76*	MAViT_L	27.2*	4.8*	83.44*	MAViT_L	49.8*	10.5*	84.60*	MAViT_L	97.8*	16.9*	84.81*
MHLA_L	16.4*	2.6*	81.85*	MHLA_L	26.8*	5.2*	83.50*	MHLA_L	–	–	–	MHLA_L	–	–	–
GABA-M1$_L$	14.9	2.6	<u>82.10</u>	**GABA-M1**$_L$	25.7	4.8	<u>83.76</u>	**GABA-M1**$_L$	48.4	10.4	<u>84.88</u>	**GABA-M1**$_L$	95.5	16.9	<u>85.09</u>
GABA-M2$_L$	16.8	2.9	**82.21**	**GABA-M2**$_L$	29.3	5.4	**83.87**	**GABA-M2**$_L$	55.6	11.8	**85.00**	**GABA-M2**$_L$	110	19.3	**85.22**

taken from the corresponding cited works; * denotes results reproduced under the same training protocol as GABA. Throughout, P(M) and F(G) denote parameters in millions and FLOPs in billions, and the best and second-best results in each table are shown in **bold** and <u>underlined</u>, respectively. Component-wise ablations isolating each mechanism, together with a sweep over the bilinear rank R, are reported in Appendix F.

3.1 Image Classification

Settings. Following DeiT (Touvron et al., 2021), RAVLT (Fan et al., 2025b), and MAViT (Fan et al., 2025a), we train GABA and all reproduced baselines (marked by *) from scratch on ImageNet-1K for 300 epochs at 224×224 resolution. We use AdamW with a peak learning rate of 10^{-3}, weight decay 0.05, a global batch size of 1024, cosine learning-rate decay, and the standard DeiT augmentation and regularization strategy. For controlled comparison, MHLA is additionally re-evaluated under the same training configuration.

Results: hybrid backbone. Table 1 compares GABA with representative Transformer, state-space, and linear-attention backbones, including RMT (Fan et al., 2024), FLatten (Han et al., 2023), VMamba (Liu et al., 2024), MSVMamba (Shi et al., 2024), MILA-S from MLLA (Han et al., 2024b), RAVLT (Fan et al., 2025b), MAViT (Fan et al., 2025a), InLine (Han et al., 2024a), and MHLA-VLT-T/S from MHLA (Zhang et al., 2026). Across Tiny, Small, Base, and Large scales, GABA-M1 obtains $82.80/84.83/85.72/86.03\%$ Top-1 accuracy, improving over the strongest competing baseline at each scale by $0.25/0.26/0.32/0.39$ points while remaining essentially matched to RAVLT in parameters and FLOPs. GABA-M2 further reaches $82.95/84.99/85.83/86.13\%$, corresponding to gains of $0.40/0.42/0.43/0.49$ points over the strongest prior result at each scale, at the additional channel-wise cost of its gated token-specific correction. The gains persist across all four capacity regimes, indicating that the proposed refinements are not restricted to a particular model scale.

Results: all-linear backbone. Table 2 provides a stricter evaluation of the proposed mechanism by removing Softmax attention entirely: every attention block uses linear attention. GABA-M1_L achieves $82.10/83.76/84.88/85.09\%$ Top-1 from Tiny to Large while matching RAVLT_L to within 0.2 M parameters and 0.1 G FLOPs, improving it by $0.41/0.40/0.39/0.37$ points. Relative to the strongest competing linear baseline available at each scale, GABA-M1_L improves accuracy by $0.25/0.26/0.28/0.28$ points, while GABA-M2_L increases these margins to $0.36/0.37/0.40/0.41$ points. Because no Softmax-attention block is retained in this setting, these results show that the gains persist within purely linear attention rather than relying on complementary quadratic attention in later stages.

Visual Results: Linear and hybrid backbone. Figure 3 compares query-conditioned interactions on four ImageNet-1K validation images. Among these models, RALA_L spreads interaction mag-

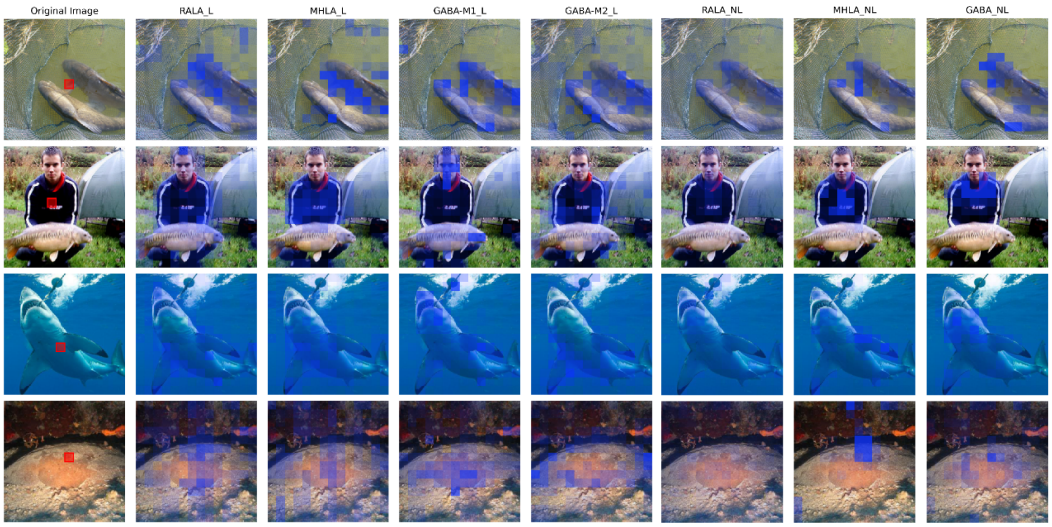

Figure 3: **Qualitative spatial interactions on ImageNet-1K.** The query token is marked by a red box in the first column; query indices are 105, 89, 118, and 91 from top to bottom. The remaining columns show the linear (L) and hybrid (NL) variants of RALA, MHLA, and GABA. All maps come from head 0 of the final block of the third stage, on a 14×14 token grid ($N = 196$). Absolute query-key interaction magnitudes are normalized across key tokens and overlaid under a common opacity mapping; for GABA-M2_L they are first aggregated across output channels.

Table 3: **COCO object detection and 3 segmentation under the $3\times$+MS schedule.** All GABA entries use the GABA-M2 backbone.

Mask R-CNN $3\times$+MS

Backbone	P(M)	F(G)	Box AP AP^b	Box AP AP^b_{50}	Box AP AP^b_{75}	Mask AP AP^m	Mask AP AP^m_{50}	Mask AP AP^m_{75}
Small								
RMT-S	46	262	50.7	71.9	55.6	44.9	69.1	48.4
MLLA-T	44	255	48.8	71.0	53.6	43.8	68.0	46.8
RAVLT-S	44	262	51.2*	72.2*	56.1*	45.2*	69.5*	48.8*
MAViT-S	44	262	51.2*	72.3*	56.0*	45.2*	69.5*	48.8*
GABA-S	48	276	**51.7**	**72.8**	**56.4**	**45.7**	**70.0**	**49.2**
Base								
RMT-B	73	373	52.2	72.9	57.0	46.1	70.4	49.9
MLLA-S	63	319	50.5	71.8	55.2	44.9	69.1	48.2
RAVLT-B	67	372	52.4*	73.3*	57.3*	46.1*	70.4*	49.8*
MAViT-B	67	372	52.9*	73.9*	58.1*	46.7*	71.3*	50.7*
GABA-B	74	402	**53.4**	**74.1**	**58.5**	**47.3**	**71.6**	**51.1**
Large								
Swin-B	107	496	48.6	70.0	53.4	43.3	67.1	46.7
CSWin-B	97	526	50.8	72.1	55.8	44.9	69.1	48.3
RAVLT-L	114	501	53.3*	74.2*	58.5*	47.0*	71.4*	50.8*
MAViT-L	114	501	53.3*	74.1*	58.3*	46.9*	71.3*	51.0*
GABA-L	129	551	**53.7**	**74.8**	**58.7**	**47.5**	**71.8**	**51.3**

Cascade Mask R-CNN $3\times$+MS

Backbone	P(M)	F(G)	Box AP AP^b	Box AP AP^b_{50}	Box AP AP^b_{75}	Mask AP AP^m	Mask AP AP^m_{50}	Mask AP AP^m_{75}
Small								
RMT-S	83	741	53.2	72.0	57.8	46.1	69.8	49.8
FL-Swin-T	87	747	50.8	69.6	55.1	44.1	67.0	48.1
RAVLT-S	82	741	53.6*	72.3*	58.0*	46.3*	70.1*	50.2*
MAViT-S	82	741	53.6*	72.2*	57.9*	46.4*	70.1*	50.4*
GABA-S	86	754	**53.9**	**72.8**	**58.4**	**46.6**	**70.7**	**50.8**
Base								
RMT-B	111	852	54.5	72.8	59.0	47.2	70.5	51.4
FL-Swin-S	108	841	52.2	71.2	56.8	45.4	68.3	49.4
InLine-Swin-S	109	835	52.4	71.0	56.9	45.4	68.8	49.6
RAVLT-B	105	851	55.0*	73.6*	59.7*	47.4*	71.2*	51.7*
MAViT-B	105	851	55.2*	73.8*	60.0*	47.7*	71.5*	52.1*
GABA-B	112	880	**55.8**	**74.2**	**60.3**	**48.1**	**71.7**	**52.6**
Large								
ConvNeXt-B	145	964	52.7	71.3	57.2	45.6	68.9	49.5
CSWin-B	135	1004	53.9	72.6	58.5	46.4	70.0	50.4
RAVLT-L	152	979	55.3*	73.9*	60.1*	47.7*	71.6*	51.9*
MAViT-L	152	979	55.7*	74.4*	60.5*	48.1*	72.2*	52.5*
GABA-L	166	1030	**56.1**	**74.7**	**61.0**	**48.7**	**72.6**	**52.7**

nitude over several regions and MHLA_L responds more locally; GABA-M1_L retains pronounced responses at selected locations under a comparable parameter budget, while GABA-M2_L shows aggregate support spanning the queried object and its surrounding context. GABA_NL follows the same pattern in the hybrid setting, combining a dominant response with additional regions of non-negligible magnitude. These maps illustrate spatial retrieval support; they do not establish that the highlighted regions are causally responsible for the prediction or that broader support is intrinsically preferable. Representation- and interaction-level visualizations for all variants are provided in Appendix G.

3.2 Object Detection and Instance Segmentation

Settings. Following RAVLT and MAViT, we use MMDetection (Chen et al., 2019) and evaluate GABA as the backbone of Mask R-CNN (He et al., 2017) and Cascade Mask R-CNN (Cai & Vasconcelos, 2018) on COCO. We follow the same $3\times$+MS protocol used by the reproduced RAVLT

and MAViT baselines: models are trained for 36 epochs with multi-scale augmentation, randomly resizing the shorter image side within $[480, 800]$ pixels while preserving aspect ratio. The GABA-S/B/L results in this experiment use the GABA-M2 backbone.

Results. Table 3 shows consistent improvements across scales and detection frameworks. Under Mask R-CNN, GABA-S/B/L improve over the stronger of RAVLT and MAViT by $0.5/0.5/0.4$ box AP and $0.5/0.6/0.5$ mask AP, respectively. Under Cascade Mask R-CNN, the corresponding gains are $0.3/0.6/0.4$ box AP and $0.2/0.4/0.6$ mask AP. The improvements also persist at both AP_{50} and the stricter AP_{75} criteria, indicating gains in detection quality, localization, and instance segmentation. These comparisons are not strictly budget-matched because GABA-M2 includes the additional gated correction: under Mask R-CNN this corresponds to approximately 9–13% more parameters and 5–10% more FLOPs than RAVLT across Small–Large scales (e.g., 74 M/402 G versus 67 M/372 G at Base). The consistent gains across both detection frameworks nevertheless show that the improvements observed in classification transfer to dense prediction. Appendix F separates the contributions of the two mechanisms.

3.3 Semantic Segmentation

Table 4: **Semantic segmentation on ADE20K.** Settings: Semantic FPN (80K) and UperNet (160K).

Tiny	Sem. FPN			UperNet		
Model	P	F	mIoU	P	F	mIoU
VAN-B1	18	140	42.9	–	–	–
PVTv2-B1	18	136	42.5	–	–	–
RMT-T	17	136	46.4	–	–	–
MSVMamba-M	–	–	–	42	875	45.1
MAViT-T	18	136	47.6	44	893	48.4
RAVLT-T	18	136	47.9	44	893	49.3
GABA-M1-T	18	138	48.2	43	894	49.6
GABA-M2-T	19	142	**48.3**	45	899	**49.7**

Small	Sem. FPN			UperNet		
Model	P	F	mIoU	P	F	mIoU
MogaNet-S	29	189	47.7	55	946	49.2
RMT-S	30	180	49.4	56	937	49.8
VMamba-T	–	–	–	62	949	47.9
FL-Swin-T	–	–	–	60	946	44.8
MAViT-S	28	180	50.7	55	937	51.0
RAVLT-S	28	180	49.5	55	937	50.7
GABA-M1-S	28	184	50.9	55	941	51.2
GABA-M2-S	32	194	**51.0**	58	951	**51.3**

Base	Sem. FPN			UperNet		
Model	P	F	mIoU	P	F	mIoU
MogaNet-B	–	–	–	74	1050	50.1
RMT-B	57	294	50.4	83	1051	52.0
VMamba-S	–	–	–	82	1028	50.6
FL-Swin-S	–	–	–	82	1038	48.1
MAViT-B	51	292	51.5	77	1050	52.8
RAVLT-B	51	292	51.9	77	1050	52.5
GABA-M1-B	51	304	52.2	78	1062	53.0
GABA-M2-B	58	323	**52.3**	85	1080	**53.1**

Large	Sem. FPN			UperNet		
Model	P	F	mIoU	P	F	mIoU
CSWin-B	81	464	49.9	109	1222	51.1
RMT-L	98	482	51.4	125	1241	52.8
VMamba-B	–	–	–	122	1170	51.0
MLLA-B	–	–	–	128	1183	51.9
MAViT-L	98	424	52.8	125	1182	53.6
RAVLT-L	98	424	52.6	125	1182	53.2
GABA-M1-L	98	438	53.0	126	1196	53.8
GABA-M2-L	113	476	**53.2**	140	1234	**53.9**

Settings. Following RAVLT (Fan et al., 2025b) and MAViT (Fan et al., 2025a), we evaluate transfer to ADE20K semantic segmentation using Semantic FPN and UperNet. ImageNet-1K pretrained GABA-M1 and GABA-M2 models are used as backbones without modifying the segmentation heads. For Semantic FPN, we follow the PVT training configuration and train for 80K iterations; for UperNet, we follow the standard Swin Transformer protocol and train for 160K iterations. All experiments are implemented with MMSegmentation, and performance is measured using mIoU.

Results. Table 4 shows consistent transfer under both segmentation frameworks. At closely matched parameter counts, GABA-M1 outperforms RAVLT and MAViT across all scales under both Semantic FPN and UperNet. GABA-M2 further achieves the best mIoU in all eight scale–decoder settings, improving the stronger of RAVLT and MAViT by 0.3–0.4 mIoU in every case. Specifically, GABA-M2 reaches $48.3/51.0/52.3/53.2$ mIoU with Semantic FPN and $49.7/51.3/53.1/53.9$ with UperNet from Tiny to Large. The consistency across two segmentation heads and four backbone scales indicates that the improved GABA representations transfer from image-level recognition to fine-grained spatial prediction.

3.4 Natural Language Processing

Settings. To test whether the proposed retrieval refinements generalize beyond visual recognition, we additionally evaluate GABA in autoregressive language modeling following the experimental protocol of MHLA. We train the GABA language models from scratch on 10B tokens from FineWeb-Edu (Penedo et al., 2024), using a context length of 2048 and an effective batch size of approximately 0.25M tokens. Optimization uses AdamW with a peak learning rate of 3×10^{-4}, weight decay 0.01, gradient clipping 1.0, and cosine learning-rate decay for 40,960 optimization steps with 2,048 warm-up steps. Following MHLA, we evaluate language modeling perplexity, MMLU, common-sense reasoning, and LongBench (Bai et al., 2024). The local operator $F(\cdot)$ and the displacement convolution F_{dw} are spatial operators defined over a 2-D token grid, and the anti-symmetric exchange is correspondingly motivated by the global–local decomposition specific to vision backbones. The language models therefore instantiate only the output-conditioned bilinear retrieval of GABA-M1 on top of the MHLA backbone, leaving the global–local coordination mechanism to the visual setting. This experiment consequently isolates the transfer of the retrieval refinement itself, which is the component our theoretical analysis in Sec. 2.1 concerns.

Table 5: **NLP evaluation.** Comparison on language modeling, common-sense reasoning, MMLU, and LongBench.

(a) Language modeling and common-sense reasoning

Model	MMLU↑	CSR↑	Wino.↑	PIQA↑	ARC-c↑	OBQA↑	ARC-e↑	BoolQ↑	Wiki.↓	LMB.↓
GLA (340M)	22.9	46.0	50.0	62.9	25.5	31.0	45.8	60.8	41.47	86.98
Transformer++ (340M)	22.9	46.8	49.6	64.4	25.7	32.8	48.1	60.5	**34.57**	60.46
Mamba (390M)	23.5	46.4	50.5	64.1	24.9	32.4	48.3	58.2	38.32	62.43
Mamba2 (340M)	23.0	47.0	49.8	64.6	25.5	32.0	**49.2**	61.2	35.40	**58.51**
GDN (360M)	23.0	46.9	51.3	64.5	25.4	31.4	47.3	**62.0**	35.01	60.16
MHLA (340M)	23.7	47.1	51.3	64.4	25.9	33.4	46.5	61.3	38.31	71.64
GABA (340M)	24.0	47.4	51.6	64.7	26.2	33.7	46.8	61.6	38.19	71.42
GABA (390M)	**24.1**	**47.5**	**51.7**	**64.8**	**26.3**	**33.8**	46.9	61.7	38.14	71.32

(b) LongBench

	Multi-Doc QA			Single-Doc QA		Few-shot		Synthetic		Summarization			Code		AVG
Model	2WM	HQA	Mus	QQA	NQA	SSM	TQA	PEN	PZH	QMS	GvR	MNs	RBP	LCC	
Mamba (360M)	3.37	2.36	1.60	4.57	2.28	5.16	5.49	1.10	0.10	12.23	18.36	14.96	**13.63**	12.33	6.97
GLA (325M)	3.23	2.31	1.67	4.53	2.13	3.94	0.70	**1.98**	0.27	11.42	17.72	15.34	13.59	12.55	6.53
GDN (346M)	2.86	2.24	1.54	**4.73**	**2.48**	**6.85**	**7.61**	0.53	0.41	12.46	17.91	**15.98**	10.42	9.98	6.86
Transformer++ (325M)	**4.97**	2.13	**2.22**	4.45	2.35	6.24	7.47	0.76	1.18	11.75	16.81	15.11	11.56	9.92	6.92
Mamba2 (330M)	3.56	2.38	1.69	4.70	2.20	4.97	7.03	0.72	**1.51**	12.57	17.65	14.00	10.15	9.49	6.62
MHLA (325M)	3.58	2.97	1.87	4.68	2.38	6.41	6.44	1.69	1.49	12.58	18.59	15.01	13.37	12.72	7.41
GABA (340M)	3.59	**2.98**	1.88	4.69	2.39	6.43	6.46	1.70	1.49	12.62	18.65	15.06	13.41	12.76	7.44
GABA (390M)	3.60	**2.98**	1.88	4.70	2.39	6.44	6.47	1.70	1.50	**12.64**	**18.67**	15.08	13.43	**12.78**	**7.45**

Results. Table 5 shows that the bilinear retrieval refinement transfers to the autoregressive setting. At comparable model scale, GABA-340M improves over MHLA-340M on MMLU ($23.7 \rightarrow 24.0$), CSR ($47.1 \rightarrow 47.4$), and every reported component common-sense reasoning benchmark, while also reducing WikiText and LAMBADA perplexity from $38.31/71.64$ to $38.19/71.42$. On LongBench, GABA-340M raises the average score from 7.41 to 7.44 and improves or matches MHLA across the reported tasks. Scaling GABA to 390M further yields 24.1 MMLU, 47.5 CSR, and a 7.45 LongBench average. GABA is not uniformly best among all language architectures: Transformer++, Mamba2, and GDN retain lower language-modeling perplexity at comparable scales, and individual baselines remain stronger on several tasks. We therefore interpret these results as evidence that the retrieval refinement generalizes beyond vision, rather than as a claim of state-of-the-art language modeling.

4 Conclusion

We introduced **Geometry-Aware Bilinear Attention (GABA)**, which addresses two complementary limitations of associative linear attention: the restricted retrieval geometry induced by the compressed $K^\top V$ state, and the absence of explicit coordination between the global and local value pathways. GABA enriches retrieval through an output-channel-conditioned low-rank bilinear interaction and coordinates the two pathways through a contrast-guided anti-symmetric exchange about their shared anchor. Both mechanisms preserve associative computation: the all-linear formulation retains linear scaling in sequence length, at an overhead linear in the rank R and at most 0.33% of model FLOPs at $R{=}4$. Improvements across image classification, object detection, instance and semantic segmentation, and language modeling, together with the state-level spectral and redundancy analyses in Appendices F and G, indicate that the benefit arises from better-utilized representation directions rather than added capacity, a reading reinforced by the rank ablation, where accuracy peaks at a small rank and does not improve as the refinement subspace is enlarged. Several limitations remain. Our rank results are attainability bounds rather than guarantees that every trained layer realizes them. The hybrid variants inherit the quadratic cost of their Softmax blocks, and the gated correction in GABA-M2 is substantially more expensive than the attention it augments. Our evaluation is also confined to the token counts typical of 224×224 recognition; validating the efficiency argument at the long-sequence scales that motivate linear attention remains open. More broadly, GABA suggests that closing the expressiveness gap need not require recovering dense pairwise attention; a richer retrieval geometry over a compact associative state may suffice.

AI USE STATEMENT

We used LLM-based tools in a reviewing capacity for tasks requiring disclosure: they provided feedback on the manuscript, including the presentation and interpretation of reported results and the consistency of reported values with our result files, and checked the written proofs for errors; all such feedback was verified by the authors before use. We did not use generative AI tools to develop the theoretical framework, formulate mathematical claims or construct their proofs, propose hypotheses, design experiments, implement the methods, or generate or process data; translation and qualitative analysis are not applicable to this work. Additionally, we used these tools to edit the paper for readability, revise figure captions, identify candidate related work (each verified against the original publication), and review portions of the analysis code. We reviewed all AI-assisted work and take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

The GABA block is fully specified in Sec. 2 (Eqs. 2–3 and 5–6, Algorithm 1); the initializations are stated in Sec. 2.2 and Appendix D.8, which also maps the formulation to the implemented operations. Training settings for all tasks are given in Sec. 3, and reproduced baselines (*) use the same protocol. Complete proofs and their scope conditions appear in Appendix D, and the operation-counting convention and parameter and FLOP accounting in Appendix E. Appendix H defines every diagnostic, and Appendices F and G state the evaluation protocol for each measurement. All datasets (ImageNet-1K, COCO, ADE20K, FineWeb-Edu) are public and used with standard splits. Code, configurations, and analysis scripts will be released; an anonymized version is included in the supplementary material.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pp. 864–873. PMLR, 2020.

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 17256–17267. IEEE, 2023.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 6154–6162. Ieee, 2018.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12114–12124. IEEE, 2022.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pp. 2793–2803. PMLR, 2021.

Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. Rmt: Retentive networks meet vision transformers. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5641–5651. IEEE, 2024.

Qihang Fan, Huaibo Huang, Yuang Ai, and Ran He. Rectifying magnitude neglect in linear attention. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21505–21514. IEEE, 2025a.

Qihang Fan, Huaibo Huang, and Ran He. Breaking the low-rank dilemma of linear attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25271–25280. IEEE, 2025b.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5938–5948. IEEE, 2023.

Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. *Advances in Neural Information Processing Systems*, 37:79221–79245, 2024a.

Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify mamba in vision: A linear attention perspective. *Advances in neural information processing systems*, 37:127181–127203, 2024b.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

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. PMLR, 2020.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3): 455–500, 2009.

Alexander C Li, Alexei A Efros, and Deepak Pathak. Understanding collapse in non-contrastive siamese representation learning. In *European conference on computer vision*, pp. 490–505. Springer, 2022.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

Weikang Meng, Yadan Luo, Liangyu Huo, Yingjian Li, Yaowei Wang, Xin Li, and Zheng Zhang. Norm $\times$ direction: Restoring the missing query norm in vision linear attention. *arXiv preprint arXiv:2506.21137*, 2025a.

Weikang Meng, Yadan Luo, Xin Li, Dongmei Jiang, and Zheng Zhang. Polaformer: Polarity-aware linear attention for vision transformers. In *International Conference on Learning Representations*, volume 2025, pp. 50020–50037, 2025b.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10968–10977. IEEE, 2020.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*, 2022.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Amartya Sanyal, Philip HS Torr, and Puneet K Dokania. Stable rank normalization for improved generalization in neural networks and gans. *arXiv preprint arXiv:1906.04659*, 2019.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pp. 9355–9366. PMLR, 2021.

Yuheng Shi, Minjing Dong, and Chang Xu. Multi-scale vmamba: Hierarchy in hierarchy visual state space model. *Advances in Neural Information Processing Systems*, 37:25687–25708, 2024.

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, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. *arXiv preprint arXiv:2202.06258*, 2022.

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, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pp. 29687–29707, 2025.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Kewei Zhang, Ye Huang, Yufan Deng, Jincheng Yu, Junsong Chen, Huan Ling, Enze Xie, and Daquan Zhou. Mhla: Restoring expressivity of linear attention via token-level multi-head. In *International Conference on Learning Representations*, volume 2026, pp. 12225–12245, 2026.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *International Conference on Learning Representations*, volume 2024, pp. 53551–53580, 2024.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

A Appendix Overview

This appendix contains the supporting material referenced from the main text.

- **Sec. B (Related Work).** Positions GABA within the linear-attention literature along two less-explored axes: the geometry through which the compressed state is read, and the coordination of representations derived from a shared value stream.
- **Sec. C (Motivations).** Develops the two structural limitations that motivate the proposed mechanisms, each stated as a baseline operation, the property it lacks, and the correction we introduce.
- **Sec. D (Theoretical Analysis).** Establishes the baseline rank constraint, the non-augmentation result for channel-shared transformations, the rank-R gain bound with a sufficient condition and a tightness construction, and the correspondence to the implemented operations; all results concern exact algebraic rank rather than the tolerance-based numerical rank reported empirically.
- **Sec. E (Computational Complexity).** Accounts for the computational and parameter cost of each mechanism, separating the attention mechanism within a block (M1 versus M2) from the backbone composition (all-linear versus hybrid).
- **Sec. F (Ablation Study).** Reports the component ablation, the sweep over the bilinear rank R, and a within-model comparison of the base and bilinearly refined associative quantities under fixed learned parameters.
- **Sec. G (Visual Analysis).** Compares GABA with the corresponding RALA and MHLA baselines at the representation and query–key interaction levels across separately trained models; unlike the within-model comparison in Sec. F, these measurements characterize complete architectures and do not attribute differences to an individual operation.
- **Sec. H (Metrics).** Defines every representation and attention diagnostic used in the two preceding analyses, together with the interpretation each measurement does and does not support.

Find the model-related code here[1]. The full code and checkpoints will be shared after the final publication.

B Related Work

Research on linear attention has advanced through feature mappings, normalization, state updates, rank restoration, and local modeling. We review these developments and position our contribution along two less-explored dimensions: the geometry through which information is *retrieved* from the compressed state and the coordination of global and local representations derived from a shared value stream.

Linear attention and the softmax expressiveness gap. Linear attention replaces quadratic pairwise attention with a decomposable similarity, enabling associative computation through a compact state $S = K^\top V$, where Q and K denote feature-mapped queries and keys (Katharopoulos et al., 2020). This state can also be interpreted as a finite-dimensional fast-weight memory (Schlag et al., 2021). Subsequent work improves expressiveness through feature mappings, attention distributions, normalization, and state dynamics. Performer approximates softmax using positive random features (Choromanski et al., 2020), while cosFormer and Hedgehog improve attention concentration through structured or learned feature maps (Qin et al., 2022; Zhang et al., 2024). Flowformer restores token competition through conservation constraints (Wu et al., 2022), and PolaFormer models same- and opposite-polarity query–key interactions to sharpen attention distributions (Meng et al., 2025b). MALA identifies the loss of query-magnitude information under normalized linear attention (Fan et al., 2025a), while NaLaFormer incorporates query-norm and direction-aware similarity modeling (Meng et al., 2025a). A parallel research direction improves associative memory dynamics: GLA introduces data-dependent state gating (Yang et al., 2023), RetNet employs decayed state accumulation (Sun et al., 2023), Mamba-2 establishes connections between structured state-space

[1] https://anonymous.4open.science/r/GABA-E0B6

updates and attention (Dao & Gu, 2024), and DeltaNet and Gated DeltaNet introduce delta-rule updates that retrieve existing memory content before applying a corrective write (Yang et al., 2024; 2025). These approaches improve feature representations, attention normalization, and memory dynamics through distinct mechanisms. A separate limitation arises in the channel-shared interaction geometry of conventional associative retrieval. Omitting the normalizer for clarity, its output for channel c is

$$o_c = q^\top S e_c = \sum_j (q^\top k_j) v_{j,c}, \tag{11}$$

where e_c selects the c-th output channel. Although the columns of S contain different value information, the underlying query–key contraction $q^\top k_j$ is shared across output channels within each attention head. Our work generalizes this channel-shared retrieval geometry through output-channel-conditioned bilinear interactions while retaining associative computation.

Rank restoration and bilinear interaction. Representational diversity is another important factor governing linear attention. FLatten Transformer introduces rank restoration alongside focused linear attention (Han et al., 2023); RALA analyzes rank deficiency in both the $K^\top V$ state and output representations and augments their representational capacity (Fan et al., 2025b); InLine addresses non-injectivity and insufficient local modeling (Han et al., 2024a); and MHLA mitigates global-context collapse through token-level multi-head computation (Zhang et al., 2026). Related expressiveness limitations have been studied for standard multi-head attention (Bhojanapalli et al., 2020), although its head-dimension bottleneck is distinct from the algebraic rank constraint of associative linear-attention outputs.

These methods improve representational capacity through distinct state, output, and attention-organization mechanisms. Our refinement addresses a different aspect of associative retrieval: the query–key interaction metric can vary across output channels rather than being shared within each attention head. We introduce this channel conditioning directly into the state readout:

$$\widetilde{o}_c = q^\top M_c S e_c, \qquad M_c = I + \gamma A \,\mathrm{diag}(C_{:,c}) B^\top, \tag{12}$$

where $A, B \in \mathbb{R}^{d\times R}$, $C \in \mathbb{R}^{R\times d_v}$, and $R \ll d$. The identity component recovers conventional retrieval, while the channel-dependent rank-R correction enables distinct query–key interaction geometries. This correction also permits additional state directions, with the attainable rank increase bounded by R and further constrained by the available dimensions and alignment of the learned factors (Sec. 2.1).

Higher-order interactions have previously been explored through bilinear pooling and attention. Bilinear Attention Networks model bilinear interactions for multimodal reasoning (Kim et al., 2018), while X-Linear Attention captures second-order spatial and channel-wise interactions (Pan et al., 2020). Both construct bilinear interactions over an explicit token-pair attention map. In contrast, our factorization refines the compact associative state directly, preserving linear complexity in sequence length for fixed feature dimensions and refinement rank.

Global–local modeling and value-side enhancement. Local inductive biases complement global attention in visual representation learning. LePE injects locally transformed value features alongside the attention response (Dong et al., 2022); EfficientViT combines multi-scale local modeling with linear attention (Cai et al., 2023); and FLatten and InLine further emphasize locality in efficient attention (Han et al., 2023; 2024a). MLLA demonstrates that gating and macro block design surrounding the attention operator substantially influence representational quality (Han et al., 2024b). These approaches enhance local modeling through different transformations and fusion mechanisms. Our work instead focuses on explicitly coordinating the global value stream and its local enhancement through a shared, sum-preserving exchange.

Specifically, our contrast-guided anti-symmetric interaction applies a shared displacement Δ, predicted from the contrast between the two representations:

$$\widetilde{V} = V + \Delta, \qquad \widetilde{L} = L - \Delta, \qquad \widetilde{V} + \widetilde{L} = V + L. \tag{13}$$

The exchange preserves the combined branch representation and its midpoint while applying equal-and-opposite updates to the individual branches. Unlike independently generated residual enhancements, these updates are explicitly coupled through the same displacement. This invariant applies to the branch representations themselves, not to the final attention output.

C MOTIVATIONS

Beyond improving the representations encoded in the compressed state, we examine two structural properties of associative linear attention: the interaction geometry through which the state is retrieved and the coordination between global and local representations derived from a shared value stream. Both refinements recover their respective baseline operations when their correction terms vanish, allowing their contributions to be examined through controlled ablations.

Output-channel-conditioned bilinear retrieval. Associative linear attention compresses global interactions into $S = K^\top V$ and retrieves the state through $Y = QS$. Writing the conventional unnormalized readout entrywise, the output for channel c is

$$Y_{ic} = \sum_j \left(\sum_a Q_{ia} K_{ja} \right) V_{jc}, \tag{14}$$

where a indexes the projected query–key dimension. Although the columns of S contain different value information, the query–key contraction is shared across output channels within each attention head. Learned projections can mix input dimensions, but the associative readout itself does not explicitly condition this interaction on the output channel. Moreover, the compressed state constrains the attainable output rank through $\mathrm{rank}(QS) \le \min\{\mathrm{rank}(Q), \mathrm{rank}(S)\}$. State-side rank deficiency can therefore restrict the diversity of retrieved representations, as observed in vision linear attention (Fan et al., 2025b).

These observations motivate an output-channel-conditioned bilinear refinement that augments the shared inner product with a channel-dependent interaction:

$$\widetilde{Y}_{ic} = \sum_j \left(q_i^\top M_c k_j \right) V_{jc}, \qquad M_c = I + \gamma A \,\mathrm{diag}(C_{:,c}) B^\top, \tag{15}$$

where $q_i^\top$ and $k_j^\top$ denote the i-th and j-th rows of Q and K, respectively; $A, B \in \mathbb{R}^{d \times R}$, $C \in \mathbb{R}^{R \times d_v}$, $R \ll d$, and γ is a learnable scalar initialized to zero. The identity component recovers conventional retrieval when $\gamma = 0$, while the rank-R correction permits output channels to assign different weights to interactions between projected query and key dimensions, including cross-dimensional terms $Q_{ia}K_{jb}$ for $a \neq b$.

Importantly, the refinement remains associative:

$$\widetilde{S} = S + \gamma A\big[(B^\top S) \odot C\big], \qquad \widetilde{Y} = Q\widetilde{S}, \tag{16}$$

where $\odot$ denotes the Hadamard product. Refining the precomputed state requires $O(dd_vR)$ additional operations and does not construct a dense token-pair attention matrix; the complete mechanism therefore retains linear complexity in sequence length for fixed feature dimensions and refinement rank.

Unlike a channel-shared left transformation MS, whose rank cannot exceed $\mathrm{rank}(S)$, our channel-conditioned refinement is not generally expressible as a single shared linear transformation of S. By applying different coefficients across output channels, the correction can increase the state rank by up to R, subject to the available dimensions and the alignment of the learned factors. This can relax the state-side rank constraint without removing the bound imposed by Q. We analyze the corresponding rank properties in Sec. 2.1.

Contrast-guided anti-symmetric interaction. In our baseline attention block, the value stream V contributes to global aggregation through $\mathcal{A}_{Q,K}(V)$ and produces a local enhancement $L = F(V)$. Although the branches are trained jointly, their late additive fusion, $\mathcal{A}_{Q,K}(V)+L$, does not explicitly coordinate their representations before aggregation. Unconstrained residual updates could modify both the combined branch representation and the relative allocation of information between the two pathways. We instead impose a sum-preserving exchange that holds their pre-aggregation aggregate fixed while allowing their individual representations to adapt. This provides a structured way to modify the relative branch representations without independently altering their combined sum.

Specifically, we introduce a contrast-guided displacement Δ and apply equal-and-opposite updates:

$$\widetilde{V} = V + \Delta, \qquad \widetilde{L} = L - \Delta, \qquad \widetilde{V} + \widetilde{L} = V + L. \tag{17}$$

The displacement is predicted from the contrast between the value and local representations, allowing the update to depend on their current differences. Contrast guidance determines the input-dependent displacement, whereas the anti-symmetric form constrains how it is distributed between the two branches. Consequently, the exchange preserves their combined representation and common midpoint before aggregation while allowing the branches to undergo coordinated, relative adaptation.

Preserving the pre-aggregation sum does not, however, require the fused output to remain unchanged. For fixed Q and K, the associative aggregation operator $\mathcal{A}_{Q,K}$ is linear in its value argument, yielding

$$\begin{aligned}&\mathcal{A}_{Q,K}(\widetilde{V}) + \widetilde{L} \\ &\quad = \mathcal{A}_{Q,K}(V) + L + [\mathcal{A}_{Q,K}(\Delta) - \Delta]\,.\end{aligned} \tag{18}$$

The correction measures the difference between the globally aggregated and unaggregated displacement and vanishes precisely when $\mathcal{A}_{Q,K}(\Delta) = \Delta$. Thus, the shared displacement can induce a nonzero refinement of the fused output because the two pathways process it differently, despite preserving their pre-aggregation sum. This identity characterizes the isolated value-side interaction before the remaining operations of the full attention block.

D Additional Theoretical Analysis of Bilinear Attention

This section extends the analysis of the output-channel-conditioned bilinear refinement introduced in Sec. 2.1. We address four main questions: (i) what class of query–key interactions is represented by the proposed factorization, (ii) why output-channel conditioning is structurally important, (iii) under what conditions the refinement can strictly increase representation rank, and (iv) how the theoretical formulation corresponds to the implementation. We additionally establish the baseline rank constraint and analyze the computational and parameter complexity. All derivations concern a single attention head and extend independently to the multi-head setting.

Throughout this section, let $Q, K \in \mathbb{R}^{N\times d}$, $V \in \mathbb{R}^{N\times d_v}$, and $S = K^\top V \in \mathbb{R}^{d\times d_v}$, where Q and K denote the effective query and key representations after the feature and positional transformations. Following Sec. 2.1, the associative attention core is

$$Y_{\text{prop}} = QS + \gamma(QA)\left[(B^\top S)\odot C\right], \tag{19}$$

where $A, B \in \mathbb{R}^{d\times R}$ and $C \in \mathbb{R}^{R\times d_v}$. Here, γ denotes the effective bilinear coefficient in the theoretical formulation; its relationship to the learned implementation coefficient is established in Sec. D.8. For convenience, define

$$T = (B^\top S)\odot C, \qquad \Delta_0 = AT, \qquad \widetilde{S}(\gamma) = S + \gamma\Delta_0. \tag{20}$$

The proposed associative output can therefore be expressed as $Y_{\text{prop}} = Q\widetilde{S}(\gamma)$.

D.1 Proof of the Baseline Rank Constraint

The baseline associative attention output is $Y_{\text{base}} = QS$, where $S = K^\top V$ is the compressed key–value state. Its rank satisfies

$$\operatorname{rank}(Y_{\text{base}}) \le \min\{\operatorname{rank}(Q), \operatorname{rank}(S)\}. \tag{21}$$

Proof. For arbitrary conformable matrices X and Z, the standard matrix-product inequality gives $\operatorname{rank}(XZ) \le \min\{\operatorname{rank}(X), \operatorname{rank}(Z)\}$. Applying this inequality to $Y_{\text{base}} = QS$ yields Eq. 21. Moreover, if Q has full column rank, then $\ker(Q) = \{0\}$, so left multiplication by Q is injective and

$$\operatorname{rank}(QS) = \operatorname{rank}(S). \tag{22}$$

More generally, the same equality holds whenever

$$\ker(Q) \cap \operatorname{col}(S) = \{0\}, \tag{23}$$

that is, whenever Q is injective on the state subspace represented by S. □

D.2 Bilinear Interpretation of the Proposed Retrieval

The compact formulation in Eq. 19 preserves associative computation but does not explicitly reveal how the refinement modifies individual query–key interactions. We therefore expand the bilinear contribution for output channel $e \in \{1, \dots, d_v\}$ and low-rank component $m \in \{1, \dots, R\}$.

Proposition D.1 (Output-conditioned pairwise bilinear form)**.** *For output channel e, define*

$$M^{(e)} = A \operatorname{diag}(C_{:,e}) B^\top \in \mathbb{R}^{d \times d}. \tag{24}$$

Then the e-th output channel of Eq. 19 admits the token-pair expansion

$$Y^{\text{prop}}_{i,e} = \sum_{j=1}^{N} \left[Q_i K_j^\top + \gamma Q_i M^{(e)} K_j^\top \right] V_{j,e}. \tag{25}$$

Thus, the baseline uses the same identity interaction matrix for every output channel, whereas the proposed refinement introduces an output-dependent bilinear transformation $M^{(e)}$.

Proof. Let Y^{cross} denote the unscaled bilinear contribution, so that $Y_{\text{prop}} = Y_{\text{base}} + \gamma Y^{\text{cross}}$. For token i and output channel e, this contribution expands as

$$Y^{\text{cross}}_{i,e} = \sum_{m=1}^{R} (QA)_{i,m} (B^\top S)_{m,e} C_{m,e} \tag{26}$$

$$= \sum_{m=1}^{R} \left(\sum_{c=1}^{d} Q_{i,c} A_{c,m} \right) C_{m,e} \left(\sum_{a=1}^{d} B_{a,m} S_{a,e} \right). \tag{27}$$

Using $S_{a,e} = \sum_{j=1}^{N} K_{j,a} V_{j,e}$ gives

$$Y^{\text{cross}}_{i,e} = \sum_{j=1}^{N} \sum_{c=1}^{d} \sum_{a=1}^{d} Q_{i,c} \left[\sum_{m=1}^{R} A_{c,m} C_{m,e} B_{a,m} \right] K_{j,a} V_{j,e} \tag{28}$$

$$= \sum_{j=1}^{N} Q_i A \operatorname{diag}(C_{:,e}) B^\top K_j^\top V_{j,e} \tag{29}$$

$$= \sum_{j=1}^{N} Q_i M^{(e)} K_j^\top V_{j,e}. \tag{30}$$

Multiplying the unscaled contribution by γ and adding the standard associative term $\sum_j Q_i K_j^\top V_{j,e}$ yields Eq. 25. □

Equation 25 makes the change in retrieval geometry explicit. Standard associative attention uses the coordinate-aligned contraction $Q_i K_j^\top = \sum_{c=1}^{d} Q_{i,c} K_{j,c}$, whereas the proposed bilinear interaction is

$$Q_i M^{(e)} K_j^\top = \sum_{c=1}^{d} \sum_{a=1}^{d} M^{(e)}_{c,a} Q_{i,c} K_{j,a}. \tag{31}$$

It therefore permits cross-channel interactions $Q_{i,c} K_{j,a}$ for $c \neq a$ whenever the corresponding coefficients of $M^{(e)}$ are nonzero. The proposed branch extends the coordinate-aligned baseline contraction to an output-conditioned bilinear interaction space.

D.3 A Low-Rank Interaction Tensor View

The three factors A, B, and C define the third-order query–key–output interaction tensor

$$\mathcal{T}_{c,a,e} = \sum_{m=1}^{R} A_{c,m} B_{a,m} C_{m,e}, \tag{32}$$

whose entries satisfy $M^{(e)}_{c,a} = \mathcal{T}_{c,a,e}$. Equivalently,

$$\mathcal{T} = \sum_{m=1}^{R} A_{:,m} \otimes B_{:,m} \otimes C_{m,:}, \tag{33}$$

which is a canonical polyadic decomposition with R components, defining a tensor $\mathcal{T} \in \mathbb{R}^{d\times d\times d_v}$ whose CP rank is at most R (Kolda & Bader, 2009).

An unconstrained family of output-conditioned bilinear matrices $\{M^{(e)}\}_{e=1}^{d_v}$ would require

$$d^2 d_v \tag{34}$$

coefficients per head, whereas the proposed factorization requires only

$$R(2d + d_v) \tag{35}$$

parameters. For $d_v = d$, the interaction parameterization is reduced from $\mathcal{O}(d^3)$ to $\mathcal{O}(dR)$.

Importantly, the low-rank designation refers to the factorized interaction tensor, not to an upper bound of R on the complete attention representation. As shown below, the rank-R correction can introduce directions beyond those represented by the original compressed state, subject to the available dimensions and learned factors.

D.4 Why Output-Channel Conditioning Matters

Each output channel is associated with its own effective interaction matrix $M^{(e)}$. We first consider the case in which these matrices are identical across all output channels. Within the proposed factorization, identical columns of C are a sufficient condition for this case, although distinct columns may also yield the same interaction matrix when the factorization contains redundant components.

Proposition D.2 (Shared bilinear transformations cannot augment state rank)**.** *Suppose* $M^{(1)} = \cdots = M^{(d_v)} = M$*. Then*

$$\widetilde{S} = (I + \gamma M)S \tag{36}$$

and hence

$$\operatorname{rank}(\widetilde{S}) \le \operatorname{rank}(S). \tag{37}$$

If $I + \gamma M$ *is nonsingular, then*

$$\operatorname{rank}(\widetilde{S}) = \operatorname{rank}(S). \tag{38}$$

Proof. For each state column s_e, sharing the same matrix M gives $\widetilde{s}_e = s_e + \gamma M s_e = (I + \gamma M)s_e$. Stacking the refined columns yields Eq. 36. The matrix-product rank inequality gives Eq. 37. If $I + \gamma M$ is invertible, left multiplication preserves rank, yielding Eq. 38. □

Thus, a shared left transformation may reorient or rescale the compressed state but cannot increase its algebraic rank. Output-channel conditioning removes this particular restriction by allowing different state columns to be transformed through different matrices $M^{(e)}$. It does not, however, guarantee rank augmentation for every choice of input and learned factors.

We next examine whether the output-conditioned interaction family can be replaced universally by a single shared query–key projection.

Proposition D.3 (Non-collapsibility into a shared query–key projection)**.** *Assume* $\gamma \neq 0$ *and that two output channels* e_1, e_2 *satisfy* $M^{(e_1)} \neq M^{(e_2)}$*. Then no single matrix* $P \in \mathbb{R}^{d\times d}$ *satisfies*

$$Q_i(I + \gamma M^{(e)})K_j^\top = Q_i P K_j^\top \tag{39}$$

for every $Q_i, K_j \in \mathbb{R}^d$ *and every output channel* e*.*

Proof. Suppose such a matrix P exists. Equality of the bilinear forms for all possible query and key vectors requires $P = I + \gamma M^{(e)}$ for every output channel e. In particular, $I + \gamma M^{(e_1)} = P = I + \gamma M^{(e_2)}$. Since $\gamma \neq 0$, this implies $M^{(e_1)} = M^{(e_2)}$, contradicting the assumption. □

The proposition establishes that genuinely distinct output-conditioned interaction matrices cannot, in general, be represented by one channel-independent query–key transformation for all possible inputs. It does not rule out equivalent representations on restricted query–key subspaces or establish that output-channel conditioning is the only mechanism capable of increasing representation rank.

D.5 RANK-GAIN CAPACITY

We next quantify the maximum additional state and output rank contributed by the rank-R refinement.

Proposition D.4 (Rank-gain bound)**.** *Let*

$$\Delta S = \gamma A \left[(B^\top S) \odot C \right]. \tag{40}$$

Then

$$\operatorname{rank}(\Delta S) \le \min \left\{ \operatorname{rank}(A), \operatorname{rank}\left((B^\top S) \odot C \right) \right\} \le R. \tag{41}$$

Consequently,

$$\operatorname{rank}(\widetilde{S}) - \operatorname{rank}(S) \le R, \tag{42}$$

and

$$\operatorname{rank}(Y_{\text{prop}}) - \operatorname{rank}(Y_{\text{base}}) \le R. \tag{43}$$

Thus, any positive algebraic-rank increase introduced by the branch is at most R.

Proof. Let $T = (B^\top S) \odot C \in \mathbb{R}^{R \times d_v}$. Since $\Delta S = \gamma A T$, the matrix-product rank inequality gives $\operatorname{rank}(\Delta S) \le \min\{\operatorname{rank}(A), \operatorname{rank}(T)\} \le R$. By rank subadditivity,

$$\operatorname{rank}(\widetilde{S}) = \operatorname{rank}(S + \Delta S) \le \operatorname{rank}(S) + \operatorname{rank}(\Delta S), \tag{44}$$

which establishes Eq. 42. Similarly, $Y_{\text{prop}} = Y_{\text{base}} + Q\Delta S$, so

$$\operatorname{rank}(Y_{\text{prop}}) \le \operatorname{rank}(Y_{\text{base}}) + \operatorname{rank}(Q\Delta S) \tag{45}$$

$$\le \operatorname{rank}(Y_{\text{base}}) + R, \tag{46}$$

which yields Eq. 43. □

The bound applies to exact algebraic rank for each state and readout pair; it is distinct from the numerical-rank criterion used in the empirical analysis. The default implementation uses $R = 4$. Sec. F reports mean numerical-rank increases of 1.30 for the state and 2.81 for the readout. These measurements illustrate the rank changes observed under the evaluation criterion, but should not be interpreted as a direct verification of the exact algebraic rank-R bound.

The rank-gain bound does not, by itself, ensure that enrichment of the associative state is preserved after query retrieval. The following corollary specifies a sufficient condition for exact rank preservation.

Corollary D.5 (Transfer of state-rank enrichment to the output)**.** *If*

$$\ker(Q) \cap \operatorname{col}(\widetilde{S}) = \{0\}, \tag{47}$$

then

$$\operatorname{rank}(Y_{\text{prop}}) = \operatorname{rank}(\widetilde{S}). \tag{48}$$

Thus, Q need not have full column rank; injectivity on the refined state subspace is sufficient.

Proof. The stated condition makes the restriction of Q to $\operatorname{col}(\widetilde{S})$ injective. Left multiplication by Q therefore preserves the dimension of this subspace, giving $\operatorname{rank}(Q\widetilde{S}) = \operatorname{rank}(\widetilde{S})$. □

Corollary D.5 gives a sufficient condition for the exact rank of the refined state to be preserved after query retrieval. Empirically, Sec. F reports larger relative increases in the readout's effective and stable ranks than in those of the associative state. These measurements show that the learned correction also changes the retrieved representation, but the averaged numerical-rank values do not establish whether the corollary's injectivity condition holds.

D.6 A Sufficient Condition for Strict Rank Augmentation

The rank-gain bound characterizes the maximum additional rank permitted by the refinement, but does not guarantee a positive increase for every input or parameter configuration. We therefore derive a sufficient geometric condition for strict augmentation of the compressed state.

Let $r = \mathrm{rank}(S)$ and consider the singular-value decomposition

$$S = U_r \Sigma_r V_r^\top, \tag{49}$$

where $\Sigma_r \in \mathbb{R}^{r\times r}$ is nonsingular. Complete U_r and V_r to orthogonal bases $U = [U_r\ U_\perp]$ and $V = [V_r\ V_\perp]$. Express the unscaled refinement $\Delta_0 = A[(B^\top S) \odot C]$ in these bases as

$$U^\top \Delta_0 V = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}, \qquad E_{22} = U_\perp^\top \Delta_0 V_\perp. \tag{50}$$

Proposition D.6 (Transverse refinement implies strict rank gain)**.** *If*

$$\mathrm{rank}\left(U_\perp^\top \Delta_0 V_\perp\right) = t > 0, \tag{51}$$

then, for all sufficiently small nonzero γ,

$$\mathrm{rank}(S + \gamma\Delta_0) \geq r + t > r. \tag{52}$$

Proof. In the singular-vector coordinates of S,

$$U^\top (S + \gamma\Delta_0) V = \begin{bmatrix} \Sigma_r + \gamma E_{11} & \gamma E_{12} \\ \gamma E_{21} & \gamma E_{22} \end{bmatrix}. \tag{53}$$

Because Σ_r is nonsingular, $\Sigma_r + \gamma E_{11}$ remains nonsingular for sufficiently small $|\gamma|$. Block Gaussian elimination and the Schur complement therefore give

$$\mathrm{rank}(S + \gamma\Delta_0) = r + \mathrm{rank}\Big(\gamma E_{22} \tag{54}$$

$$- \gamma^2 E_{21}(\Sigma_r + \gamma E_{11})^{-1} E_{12}\Big). \tag{55}$$

For $\gamma \neq 0$, factoring out γ leaves

$$E_{22} - \gamma E_{21}(\Sigma_r + \gamma E_{11})^{-1} E_{12}. \tag{56}$$

As $\gamma \to 0$, this matrix converges to E_{22}. Since E_{22} has rank t, its t positive singular values remain positive under sufficiently small perturbations. Hence the Schur complement has rank at least t for all sufficiently small nonzero γ, establishing Eq. 52. □

The condition in Eq. 51 has a geometric interpretation. The columns of $U_\perp$ span directions outside the original column space of S, while $V_\perp$ spans its orthogonal complement in the row-space coordinates. Accordingly, $U_\perp^\top \Delta_0 V_\perp$ captures the part of the correction that maps output-channel combinations outside the row space of S into feature directions transverse to its column space. Positive rank of this component is sufficient for strict algebraic-rank augmentation when the effective coefficient is sufficiently small and nonzero.

This condition is sufficient rather than necessary: strict rank augmentation may also arise through higher-order coupling between E_{12} and E_{21} when $E_{22} = 0$. Moreover, the small-γ condition is local; the trained coefficient γ_{impl} is learned and is not constrained by this proposition to remain in that neighborhood. The result therefore establishes a sufficient geometric mechanism for strict rank gain, not a guarantee for every trained coefficient.

D.7 The Rank-R Gain Bound Is Tight

We now establish that the upper bound of Proposition D.4 is algebraically attainable. In contrast to the local sufficient condition above, the following construction realizes the full gain for every nonzero effective coefficient γ within the constructed example.

Proposition D.7 (Tightness of the rank-R augmentation bound)**.** *For every $R \geq 1$ and every $\gamma \neq 0$, there exist matrices Q, K, V, A, B, C for which*

$$\operatorname{rank}(Y_{\text{prop}}) - \operatorname{rank}(Y_{\text{base}}) = R. \tag{57}$$

Proof. Choose

$$N = d = d_v = R + 1, \qquad Q = K = I_{R+1}, \qquad \gamma \neq 0. \tag{58}$$

Let $e_1, \ldots, e_{R+1}$ be the standard basis vectors of $\mathbb{R}^{R+1}$ and define

$$S = V = e_1 \mathbf{1}_{R+1}^\top. \tag{59}$$

Since $K = I_{R+1}$, the state satisfies $K^\top V = S$ and has rank one. Choose

$$A = [e_2 \quad e_3 \quad \cdots \quad e_{R+1}], \tag{60}$$

$$B = e_1 \mathbf{1}_R^\top, \tag{61}$$

and

$$C = [\mathbf{0}_{R\times 1} \quad I_R]. \tag{62}$$

Because every column of S equals e_1, we have $B^\top S = \mathbf{1}_R \mathbf{1}_{R+1}^\top$ and therefore $(B^\top S) \odot C = C$. Hence,

$$\Delta_0 = AC = [0 \quad e_2 \quad e_3 \quad \cdots \quad e_{R+1}]. \tag{63}$$

The refined state becomes

$$\widetilde{S} = [e_1 \quad e_1 + \gamma e_2 \quad e_1 + \gamma e_3 \quad \cdots \quad e_1 + \gamma e_{R+1}]. \tag{64}$$

For every $\gamma \neq 0$, subtracting the first column from the remaining columns yields the linearly independent set $\{e_1, \gamma e_2, \ldots, \gamma e_{R+1}\}$. Therefore $\operatorname{rank}(\widetilde{S}) = R + 1$. Since $Q = I_{R+1}$, the baseline output has rank one and the refined output has rank $R + 1$, establishing Eq. 57. □

This construction treats Q and K as free effective representations and therefore establishes attainability of the algebraic rank bound, rather than guaranteeing that $Q = K = I_{R+1}$ is realizable under every particular feature mapping and positional transformation. It demonstrates that the rank-R branch can supply R additional independent directions when the state and factors permit them. Unlike Proposition D.6, the construction achieves the full bound for every nonzero γ in the specified example; neither result guarantees strict augmentation for every input or trained layer.

D.8 Correspondence to the Implemented Attention

The theoretical formulation isolates the associative bilinear mechanism. The implementation additionally applies state normalization, fixed rank-dependent scaling, a learned branch coefficient, and other attention-block operations. We now identify the corresponding effective coefficient and the scope of the rank analysis.

For a single attention head, the implemented associative state is

$$S_{\text{impl}} = \frac{1}{N} K^\top V, \tag{65}$$

where Q and K are the effective RoPE-transformed query and key representations, and V denotes the value representation entering the global associative pathway. The bilinear branch computes

$$Q_R = QA, \qquad S_R = B^\top S_{\text{impl}}, \qquad \widehat{S}_R = S_R \odot C, \tag{66}$$

followed by

$$Y_{\text{cross}} = \frac{1}{\sqrt{R}} Q_R \widehat{S}_R. \tag{67}$$

Let $D_z = \operatorname{diag}(z_1, \ldots, z_N)$ denote the token-wise normalization used by the linear-attention pathway. Its implemented associative output is

$$Y_{\text{core}} = D_z \left[Q S_{\text{impl}} + \frac{\gamma_{\text{impl}}}{\sqrt{R}} (QA) \left[(B^\top S_{\text{impl}}) \odot C \right] \right]. \tag{68}$$

Thus, the coefficient in the abstract theory corresponds to

$$\boxed{\gamma = \frac{\gamma_{\mathrm{impl}}}{\sqrt{R}}}, \tag{69}$$

where γ_{impl} is the learned implementation coefficient. The theoretical coefficient γ already includes the fixed factor $1/\sqrt{R}$ and must not be multiplied by this factor again when using Eq. 19.

The nonzero scalar factors $1/N$ and $1/\sqrt{R}$ change representation magnitudes but do not alter the corresponding exact algebraic-rank arguments. The same token-wise normalizer is applied to the standard and bilinear associative branches. If every $z_i \neq 0$, then D_z is nonsingular and

$$\mathrm{rank}(D_z Y) = \mathrm{rank}(Y). \tag{70}$$

The positive denominator mapping and numerical stabilizer yield finite, nonzero normalizers under normal execution. This invariance concerns exact algebraic rank; row-wise normalization can still change singular-value magnitudes, effective rank, stable rank, and numerical rank under a floating-point tolerance.

The learned coefficient γ_{impl} is initialized to zero, so the bilinear branch is inactive at initialization and the baseline associative pathway is recovered. During training, the coefficient and the low-rank factors are learned jointly, allowing the output-channel-conditioned correction to become active.

The complete GABA block also applies the anti-symmetric value-side refinement $\widetilde{V} = V + \Delta$ before associative aggregation. The preceding algebraic results remain valid when V is replaced by $\widetilde{V}$, because they hold for an arbitrary value matrix. In the implemented state,

$$S_{\mathrm{impl}} = \frac{1}{N} K^\top \widetilde{V}. \tag{71}$$

The full block additionally includes the existing Q–V refinement, local enhancement, residual correction, multiplicative output gating, and final projection. These operations are excluded from the isolated rank analysis because additive residual terms and element-wise output gating can themselves change the final representation rank. The propositions therefore characterize the associative bilinear mechanism before these downstream operations.

D.9 Computational Complexity and Lightweight Integration

The bilinear refinement reuses the compressed associative state $S = K^\top V$ and does not materialize an $N \times N$ token-pair attention matrix. For one head, its additional operations are QA, $B^\top S$, $(B^\top S) \odot C$, and $(QA)[(B^\top S) \odot C]$, with respective computational costs

$$\mathcal{O}(NdR), \qquad \mathcal{O}(Rdd_v), \qquad \mathcal{O}(Rd_v), \qquad \mathcal{O}(NRd_v). \tag{72}$$

The total additional computational complexity per head is therefore

$$\mathcal{O}\left(NdR + Rdd_v + NRd_v\right). \tag{73}$$

For $d_v = d$, this simplifies to

$$\mathcal{O}\left(2NdR + Rd^2\right) = \mathcal{O}\left(NdR + Rd^2\right), \tag{74}$$

which remains linear in sequence length N for fixed feature dimensions and interaction rank.

The bilinear factors introduce

$$dR + dR + Rd_v = R(2d + d_v) \tag{75}$$

parameters per head, reducing to

$$3dR \tag{76}$$

when $d_v = d$. For h heads with total embedding dimension $D = hd$, this gives $3DR$ factor parameters, in addition to the learned coefficient(s) controlling the bilinear branch. The default configuration uses $R = 4$, corresponding to $3dR = 768$ bilinear-factor parameters per head for $d = d_v = 64$.

The principal additional intermediate representations have dimensions $N \times R$ and $R \times d_v$ per head. Excluding outputs already present in the baseline computation, their activation-storage cost is

$$\mathcal{O}(NR + Rd_v). \tag{77}$$

Thus, the bilinear branch preserves associative computation and linear scaling with respect to token count without introducing a quadratic attention-map requirement. Its zero-initialized learned coefficient also permits integration without activating the correction at initialization.

D.10 SUMMARY

The analysis establishes three structural properties of the proposed bilinear refinement. First, its R-component tensor factorization permits output-conditioned, cross-channel query–key interactions while retaining associative computation. Second, a channel-shared left transformation cannot increase the compressed-state rank, whereas output-channel conditioning removes this particular restriction. Third, the proposed correction can augment the exact algebraic rank of both the associative state and global readout by at most R, and the bound is attainable in the constructed example for every nonzero effective coefficient γ.

The transverse condition in Proposition D.6 is a local sufficient condition for strict rank gain at sufficiently small nonzero γ; it does not guarantee augmentation at the learned coefficient of every trained layer. The implementation correspondence and complexity analysis show that this representational capacity is introduced through a lightweight, rank-R associative branch whose additional computation remains linear in sequence length. The within-model numerical measurements of Sec. F illustrate the spectral changes realized by the trained refinement, while remaining distinct from the exact algebraic-rank guarantees established here.

E COMPUTATIONAL COMPLEXITY ANALYSIS

We analyze the computational and memory complexity of the proposed attention mechanism and compare it with representative linear-attention methods, including RALA from RAVLT (Fan et al., 2025b), MALA from MAViT (Fan et al., 2025a), and MHLA (Zhang et al., 2026). The comparison is conducted on the all-linear backbones, so that the attention mechanism itself is isolated from the choice of hybrid composition.

Scope: two independent architectural axes. Our models vary along two axes that must not be conflated. The first is the *attention mechanism* within a block: GABA-M1 denotes the proposed linear-attention block, and GABA-M2 additionally applies the gated token-specific correction $\mathcal{G}(X) = \sigma(W_g X) \odot UX$. The second is the *block composition* of the backbone: the subscript L denotes an all-linear backbone in which every block uses linear attention, whereas NL denotes the hybrid backbone in which selected blocks use Softmax attention. These axes are independent: GABA-M2$_L$ is an all-linear architecture that includes the gated correction and therefore retains linear scaling in N, whereas any NL variant inherits the quadratic token dependence of its Softmax blocks regardless of whether $\mathcal{G}(X)$ is used. The analysis below treats the two axes separately.

Notation and accounting conventions. To avoid overloading the bilinear output-conditioning factor C used in Sec. D, we denote the embedding width in this section by D. Let N denote the number of tokens, h the number of channel heads, and $d = D/h$ the dimensionality of each head. For GABA, R denotes the rank of the output-conditioned bilinear refinement; we use the fixed value $R = 4$ throughout our experiments, with $R \ll d$. For MHLA, M denotes the number of token pieces. All expressions are stated per head in terms of d and then summed over h heads.

Arithmetic costs are reported as multiply–accumulate operations (MACs), with GMAC denoting 10^9 such operations; this is the convention used for the operation counts reported throughout the paper. All percentage overheads are computed using the same convention in the numerator and the denominator.

We focus on the attention interaction and report the shared Q/K/V and output-projection costs separately from the mechanism-specific operations. Feed-forward and other backbone operations are excluded from this interaction-level comparison; any architectural differences outside the attention interaction are reflected in the reported full-model parameter and operation counts. Method-specific operations—GABA's bilinear branch, the Q–V refinement, the anti-symmetric exchange, and the gated correction of GABA-M2—are accounted for explicitly. Unless stated otherwise, "forward memory" refers to asymptotic inference-time activation and state storage; parameter storage and autograd buffers are discussed separately.

Our backbones are hierarchical, so N, D, h, and d vary across the four stages. At 224×224 input the stage token counts are $N \in \{3136, 784, 196, 49\}$, and the per-head dimension is $d = 64$ at every stage of the Tiny and Small models and $d \in \{96, 96, 64, 64\}$ for Base and Large. We therefore report stage-wise quantities where the variation is material rather than quoting a single representative value.

Table 6: **Theoretical complexity of attention mechanisms.** N denotes the number of tokens, D the embedding width, h the number of heads, $d = D/h$, R the low-rank dimension of GABA, and M the number of token pieces in MHLA. Shared Q/K/V and output projections and the batch dimension are omitted; "extra params" counts method-specific weights per block. "Compact state" denotes the storage required by the associative attention state, whereas "forward memory" additionally includes token representations and method-specific intermediate activations. The Softmax row is a reference point for the quadratic token dependence that all linear mechanisms avoid. GABA-M2 differs from GABA-M1 only by the gated correction, which is token-wise and therefore preserves linear scaling. The MHLA entries follow the piece-wise formulation described in the text; its method-specific parameter count depends on the mixing configuration and is not tabulated here.

Method	Attention Time	Compact State	Forward Memory	Extra Params	Scaling in N
Softmax	$\mathcal{O}(N^2D)$	—	$\mathcal{O}(ND+hN^2)$	0	$\mathcal{O}(N^2)$
RALA$_L$	$\mathcal{O}(Nhd^2)$	$\mathcal{O}(hd^2)$	$\mathcal{O}(ND+hd^2)$	$\mathcal{O}(D)$	$\mathcal{O}(N)$
MALA$_L$	$\mathcal{O}(Nhd^2)$	$\mathcal{O}(hd^2)$	$\mathcal{O}(ND+hd^2)$	$\mathcal{O}(D)$	$\mathcal{O}(N)$
MHLA$_L$	$\mathcal{O}(hNd^2+hM^2d^2)$	$\mathcal{O}(hMd^2)$	$\mathcal{O}(ND+hMd^2)$	—	$\mathcal{O}(N)^\dagger$
GABA-M1$_L$	$\mathcal{O}(Nhd^2+NhdR+hRd^2)$	$\mathcal{O}(hd^2+hRd)$	$\mathcal{O}(ND+hd^2+NhR+hRd)$	$3DR+9D+h+2$	$\mathcal{O}(N)$
GABA-M2$_L$	$\mathcal{O}(Nhd^2+NhdR+hRd^2+ND^2)$	$\mathcal{O}(hd^2+hRd)$	$\mathcal{O}(ND+hd^2+NhR+hRd)$	$3DR+9D+h+2+2D^2$	$\mathcal{O}(N)$

†MHLA treats M as a fixed hyperparameter, so hM^2d^2 is an additive constant with respect to N and linear scaling holds for any fixed M. The condition $M^2 \le N$ ensures that this constant does not dominate the hNd^2 term at the sequence lengths considered.

Comparison with existing linear attention. RALA and MALA exploit the associative ordering $Q(K^\top V)$ and therefore avoid explicitly constructing the $N \times N$ query–key interaction matrix. For each head, forming the compressed associative state

$$S = K^\top V \in \mathbb{R}^{d\times d} \tag{78}$$

and subsequently retrieving it through QS each require $\mathcal{O}(Nd^2)$ computation. Summed across h heads, their dominant attention cost is

$$\mathcal{O}(Nhd^2) = \mathcal{O}\left(\frac{ND^2}{h}\right), \tag{79}$$

with compact-state storage $\mathcal{O}(hd^2)$. The additional importance- or magnitude-dependent operations in RALA and MALA are token-wise and contribute lower-order $\mathcal{O}(ND)$ computation, leaving the leading sequence dependence unchanged.

It is useful to state the comparison with Softmax attention explicitly. Dense attention requires $\mathcal{O}(N^2hd) = \mathcal{O}(N^2D)$ computation against the $\mathcal{O}(Nhd^2) = \mathcal{O}(NdD)$ of associative linear attention. Under this interaction-only MAC accounting, the costs coincide at $N = d$, and associative linear attention has a lower arithmetic count when $N > d$; this crossover does not, by itself, determine realized runtime or complete-block efficiency. At 224×224 input, the interaction-level arithmetic ratio N/d is approximately 49, 12, and 3 across the first three stages of Tiny and Small, and 33, 8, and 3 for Base and Large. In contrast, the final stage operates at $N = 49 < d = 64$, where dense attention has the lower interaction-level arithmetic count. The arithmetic advantage of associative attention therefore arises primarily in the earlier, high-token stages and increases with input resolution.

The memory distinction is asymptotic: conventional materialized Softmax attention requires $\mathcal{O}(hN^2)$ attention-map storage, whereas associative linear attention uses an $\mathcal{O}(hd^2)$ compact state whose size is independent of N. At the final stage, where $N < d$, the materialized Softmax map can be smaller than the associative state; memory-efficient Softmax kernels can further avoid storing the full attention matrix. These comparisons concern the attention-specific storage only, since both block types additionally retain token activations.

MHLA increases retrieval flexibility by partitioning the sequence into M pieces and maintaining multiple piece-wise associative summaries. Under the same per-head notation and the piecewise formulation described above, its attention computation is

$$T_{\mathrm{MHLA}} = \mathcal{O}\big(hNd^2 + hM^2d^2\big), \tag{80}$$

while the corresponding compact-state storage grows from $\mathcal{O}(hd^2)$ to $\mathcal{O}(hMd^2)$. MHLA therefore seeks to improve token-level retrieval by expanding the number of associative summaries, at an M-fold increase in compact-state storage.

Complexity of GABA-M1. GABA-M1 follows a complementary strategy: rather than maintaining multiple full $d \times d$ associative states, it retains the original shared state and introduces a low-rank output-conditioned refinement. As derived in Sec. D, the base associative branch computes

$$S = K^\top V, \qquad Y_{\text{base}} = QS, \tag{81}$$

with complexity $\mathcal{O}(Nhd^2)$. The bilinear refinement operates as

$$Q_R = QA, \qquad S_R = B^\top S, \qquad Y_{\text{bil}} = Q_R(S_R \odot C), \tag{82}$$

where $A, B \in \mathbb{R}^{d\times R}$ and $C \in \mathbb{R}^{R\times d}$ for each head. The implementation computes $B^\top S$ from the already formed associative state, contributing $\mathcal{O}(hRd^2)$ computation. The algebraically equivalent token-side ordering $(KB)^\top V$ requires two contractions totalling $\mathcal{O}(NhdR)$, so the state-side ordering is cheaper whenever $N > d/2$; this holds at every stage of our configurations, including the final stage where $N = 49$ against $d/2 = 32$.

The rank-R query projection and the reduced-state retrieval each require $\mathcal{O}(NhdR)$ computation, whereas the state reduction requires $\mathcal{O}(hRd^2)$. Ignoring constant factors, the additional bilinear cost is

$$T_{\text{bil}} = \mathcal{O}\big(NhdR + hRd^2\big) = \mathcal{O}\left(NDR + \frac{RD^2}{h}\right). \tag{83}$$

The Q–V refinement consists of a channel-mean normalization of the query followed by an element-wise product with the refined value representation, and the anti-symmetric global–local exchange uses element-wise operations together with a bias-free 3×3 depthwise convolution contributing $9ND$ multiply–accumulate operations. Both are $\mathcal{O}(ND)$ in arithmetic cost and do not alter the leading complexity. We note that these operations have low arithmetic intensity and are expected to be bandwidth- rather than compute-bound, so their realized cost need not be proportional to their arithmetic share. The resulting attention complexity of GABA-M1 is

$$T_{\text{GABA-M1}} = \mathcal{O}\big(Nhd^2 + NhdR + hRd^2\big) = \mathcal{O}\left(\frac{ND^2}{h} + NDR + \frac{RD^2}{h}\right). \tag{84}$$

For fixed D, h, and R, GABA-M1 therefore remains linear in the number of tokens and never materializes an $N \times N$ attention matrix.

Rank-controlled complexity–capacity trade-off. A central property of the proposed formulation is that its additional computation is explicitly controlled by the low-rank dimension R. Consider the two dominant operations of conventional associative linear attention, $K^\top V$ and $Q(K^\top V)$, whose combined arithmetic cost is approximately $2Nhd^2$. The bilinear refinement requires approximately $2NhdR + hRd^2$, so its overhead relative to the associative attention core is

$$\frac{T_{\text{bil}}}{T_{\text{base}}} \approx \frac{R}{d} + \frac{R}{2N}. \tag{85}$$

The first term arises from the two query–rank interactions and is independent of sequence length; the second arises from the one-off state reduction $B^\top S$ and is negligible relative to the first whenever $2N \gg d$.

Table 7 instantiates Eq. 85 for our configurations with $R = 4$. The overhead is 4.2%–7.3% at the first three stages, where the query–rank term dominates as expected. At the final stage the condition $2N \gg d$ no longer holds—$N = 49$ against $d = 64$—and the state-reduction term contributes 4.1 of the 10.3 percentage points. Weighting each stage's overhead by its total baseline associative-attention cost, including the stage depth, token count, and head configuration, gives model-level overheads of 6.9%, 6.9%, 5.2%, and 5.5% for Tiny, Small, Base, and Large, respectively; the final stage carries only 49 tokens and therefore contributes little to these aggregates despite its larger local overhead.

Relative to the complete attention block the share is far smaller, because the shared projections, local residual pathway, and feed-forward network are unchanged: the bilinear branch accounts for

Table 7: **Stage-wise bilinear overhead** from Eq. 85 with $R = 4$ at 224×224 input. The rightmost column is the overhead on the associative attention core for the complete model, obtained by weighting each stage by its total baseline associative-attention cost.

	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Model**
N	3136	784	196	49	—
Tiny / Small (d=64)	6.31%	6.51%	7.27%	10.33%	6.9%
Base / Large (d=96, 96, 64, 64)	4.23%	4.42%	7.27%	10.33%	5.2–5.5%

0.83% of block operations at the widest-overhead stage and 0.17% at the narrowest. At the model level it contributes 0.0084, 0.0153, 0.0278, and 0.0376 GMACs at Tiny, Small, Base, and Large, i.e. 0.22–0.33% of total model operations. This is why GABA-M1$_L$ and RAVLT$_L$ report identical FLOPs in Tab. 2: the difference is real but falls below the reported precision.

The same parameter R also bounds the rank-enrichment capacity of the refinement. As established in Sec. D, $\mathrm{rank}(\Delta S) \le R$ and consequently $\mathrm{rank}(S + \Delta S) - \mathrm{rank}(S) \le R$; this is an upper bound rather than a guarantee that every input or trained layer realizes the full increase. The accompanying sufficient-condition result identifies a geometric condition for strict rank augmentation, while the tightness construction establishes that the full R-dimensional gain is algebraically attainable. For the trained models, the within-model analysis in Sec. F compares the base and refined associative quantities under fixed learned parameters and reports increases in the effective and stable ranks of both the associative state and the global readout. That comparison isolates the immediate effect of the correction on a single forward pass; it is not a retraining ablation, and it therefore does not establish how the representations would differ had the model been trained without the branch. The role of R is thus twofold: it linearly controls the computational and parameter overhead while bounding the additional rank capacity. The sensitivity of accuracy to R is reported separately in Sec. F.

Memory complexity. The baseline associative state of GABA-M1 is $S \in \mathbb{R}^{h\times d\times d}$, requiring $\mathcal{O}(hd^2) = \mathcal{O}(D^2/h)$ storage. The bilinear branch additionally introduces

$$Q_R \in \mathbb{R}^{h\times N\times R}, \qquad S_R \in \mathbb{R}^{h\times R\times d}, \tag{86}$$

requiring $\mathcal{O}(NhR)$ and $\mathcal{O}(hRd)$ storage. Including the ordinary token representations, the forward activation and state complexity is

$$\mathcal{O}\big(ND + hd^2 + NhR + hRd\big)\,. \tag{87}$$

Because $R \ll d$, the rank-dependent intermediates remain substantially smaller than the ordinary $N \times D$ token representations: Q_R occupies a fraction $hR/D = R/d$ of a single token representation, i.e. between 4.2% and 6.3% in our configurations. The asymptotic order with respect to sequence length is therefore $\mathcal{O}(ND + D^2/h)$, matching conventional associative linear attention.

The state-storage comparison also highlights a structural distinction between GABA and MHLA. RALA and MALA retain one $d \times d$ associative state per head, while MHLA maintains M piecewise states. GABA-M1 instead preserves a single full associative state and augments it with only a rank-reduced $R \times d$ state together with the $N \times R$ query representation. GABA therefore enriches retrieval through a compact low-rank auxiliary representation rather than by replicating the complete associative state.

Training memory. Beyond the baseline activations, GABA-M1 introduces the rank-reduced query and state representations and the intermediates of the anti-symmetric exchange. Stop-gradient on the normalized contrast and shared magnitude prevents gradient propagation through their generating operations, although downstream operations may still retain their values for backward computation. The additional activation storage scales as $\mathcal{O}(NhR + hRd + ND)$ per layer, excluding implementation-dependent temporary buffers, and introduces no quadratic term in N.

Parameter and model-level overhead. For equal query/key/value head dimensionality, the three bilinear factors introduce

$$h(dR + dR + Rd) = 3hdR = 3DR \tag{88}$$

learnable parameters per block. The anti-symmetric interaction additionally uses a bias-free 3×3 depthwise convolution containing $9D$ weights, one bounded scale per attention head, and two scalar branch coefficients, giving $3DR + 9D + h + 2$ additional parameters per block.

The relative size of this overhead depends on the stage, and we report it rather than quoting only its most favourable value. Compared with the $4D^2$ parameters of the shared Q/K/V and output projections, the addition ranges from 8.2% at the narrowest stage ($D = 64$) down to 0.8% at the widest ($D = 640$); expressed against the complete block including the feed-forward network, the range is 2.5% to 0.3%. The overhead is largest where D is smallest, because the $9D$ depthwise term scales linearly while the projections scale quadratically. Aggregated over all blocks, GABA-M1$_L$ adds 0.062M, 0.110M, 0.194M, and 0.323M parameters at Tiny, Small, Base, and Large, i.e. 0.3–0.4% of total model size, which is why GABA-M1$_L$ and RAVLT$_L$ differ by at most 0.2M in Tab. 2. We report these exact deltas so that the matched budgets are not mistaken for an exactly zero-parameter modification.

We note for reproducibility that the bilinear branch is implemented with `einsum` contractions, which some operation-counting utilities do not instrument. The reported operation counts were obtained using `fvcore.nn.FlopCountAnalysis` with an explicit handler for `aten::einsum`; the bilinear-branch complexity was additionally calculated from the closed-form operation counts in Eq. 84.

Cost of the gated token-specific correction (GABA-M2). GABA-M2 adds to each block the gated correction

$$\mathcal{G}(X) = \sigma(W_g X) \odot UX, \qquad W_g, U \in \mathbb{R}^{D\times D}, \tag{89}$$

which contributes approximately $2ND^2$ multiply–accumulate operations and $2D^2$ parameters. Since both projections are token-wise, the correction preserves linear scaling in N: GABA-M2$_L$ remains an $\mathcal{O}(N)$ architecture.

Its constant factor, however, is substantial, and we state this rather than describe the correction as a minor addition. Relative to the associative attention interaction, whose cost is $2ND^2/h$, the gated correction costs a factor of exactly

$$\frac{T_{\text{gate}}}{T_{\text{base}}} = h \tag{90}$$

within a block, ranging from $h = 1$ at the first stage to $h = 10$ at the last stage of the Large model. Because the early stages carry most of the tokens, the aggregate ratio over the complete model is 2.5–$3.6\times$: the correction adds 0.31, 0.64, 1.43, and 2.45 GMACs at Tiny, Small, Base, and Large, against 0.12, 0.22, 0.54, and 0.68 GMACs for the entire associative attention core. In parameters it adds 1.98M, 3.73M, 7.41M, and 15.12M, i.e. $2D^2$ per block against the $4D^2$ of the shared projections.

GABA-M1 is therefore the computationally economical formulation, and GABA-M2 should be read as trading a substantial channel-wise increase in computation and parameters for an additional token-specific refinement path. Tab. 2 reports both so that this trade-off is visible at every scale.

Cost of hybrid block composition. The NL backbones replace selected linear blocks with conventional Softmax-attention blocks, which explicitly compute pairwise token interactions at $\mathcal{O}(N^2hd) = \mathcal{O}(N^2D)$ per block. This composition, rather than the gated correction, is what introduces quadratic token dependence. Let $\mathcal{L}_{\text{lin}}$ and $\mathcal{L}_{\text{sa}}$ denote the sets of GABA linear and Softmax blocks, with N_ℓ, D_ℓ, h_ℓ, $d_\ell = D_\ell/h_\ell$ the dimensions of block ℓ. Ignoring the shared projections,

$$\begin{aligned} T_{\text{hybrid}} = &\sum_{\ell\in\mathcal{L}_{\text{lin}}} \mathcal{O}\big(N_\ell h_\ell d_\ell^2 + N_\ell h_\ell d_\ell R + h_\ell R d_\ell^2\big) + \sum_{\ell\in\mathcal{L}_{\text{sa}}} \mathcal{O}\big(N_\ell^2 D_\ell\big) \\ &+ \mathbb{1}[\text{M2}] \sum_{\ell} \mathcal{O}\big(N_\ell D_\ell^2\big)\,. \end{aligned} \tag{91}$$

The practical magnitude of the quadratic term depends on where the Softmax blocks are placed. Confining Softmax attention to the final stage, where the token count at 224×224 has been reduced to $N_\ell = 49$, keeps the $N_\ell^2 D_\ell$ term at 0.12–0.17% of total model operations; extending it to the third stage as well raises this to 3.1–4.2%. Because the term grows quadratically with input resolution, the corresponding shares at 1024×1024 become 2.5–3.5% and 40–48% respectively. The placement of the Softmax blocks is therefore the dominant factor in whether the hybrid variants remain practical at high resolution.

The memory behaviour follows the same distinction: a GABA linear block retains linear activation and state scaling, whereas a materialized Softmax attention map requires $\mathcal{O}(h_\ell N_\ell^2)$ additional storage. Memory-efficient Softmax kernels avoid explicitly materializing the $N_\ell \times N_\ell$ matrix during

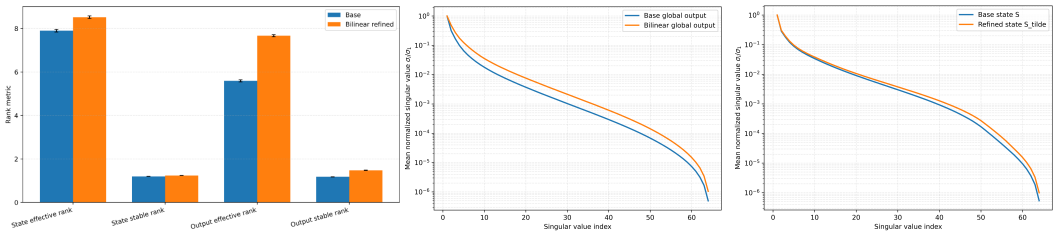

Figure 4: **Dimensional and spectral analysis of the bilinear refinement in GABA-M2$_L$.** All measurements are taken from the final block of the third stage, operating on $N = 196$ tokens. The analysis uses the same anti-symmetrically refined value representation $\widetilde{V} = V + \Delta$ for both paths and directly compares the base and bilinearly refined associative responses; the Q–V refinement, local residual addition, gated token-specific correction, output gate, and final projection are excluded. **Left:** Effective and stable rank of the base associative state S and global output Y_{base} against their bilinearly refined counterparts $\widetilde{S}$ and Y_{bil}. Effective rank is $r_{\text{eff}} = \exp(-\sum_i p_i \log p_i)$, where $p_i = \sigma_i / \sum_j \sigma_j$. Metrics are computed independently per attention head and image and averaged over a class-balanced set of 1,000 ImageNet-1K validation images; error bars denote 95% confidence intervals across validation images. Numerical rank is reported in Table 10 rather than in the figure. **Middle:** Normalized singular-value spectra of the base and refined global responses, $Y_{\text{base}} = (Q_{\text{rope}} S) \odot z$ and $Y_{\text{bil}} = (Q_{\text{rope}} \widetilde{S}) \odot z$, where z denotes the shared linear-attention normalizer. **Right:** Corresponding spectra of $S = K_{\text{rope}}^\top \widetilde{V}/N$ and $\widetilde{S} = S + \Delta S$ before query retrieval, where $\Delta S = \gamma A\big[(B^\top S) \odot C\big]$. For the middle and right panels, singular values are normalized by the leading component, $\bar{\sigma}_i = \sigma_i / \sigma_1$, averaged across heads and the same validation set, and shown on a logarithmic scale.

execution, but this implementation optimization does not change the underlying $\mathcal{O}(N_\ell^2 D_\ell)$ arithmetic complexity.

Hybrid variants should therefore be interpreted as an accuracy–efficiency trade-off rather than as linear-complexity architectures; the all-linear GABA-M1$_L$ and GABA-M2$_L$ are the variants for which the linear-scaling claims of this section hold without qualification.

Summary. The comparison reveals three different strategies for increasing the expressiveness of associative linear attention. RALA and MALA retain a single global $K^\top V$ state and modify its weighting or normalization; MHLA increases token-level retrieval flexibility through multiple piece-wise associative states; GABA-M1 instead preserves a single full $K^\top V$ state and enriches its retrieval through a compact rank-R output-conditioned interaction.

With $R = 4$, GABA-M1 introduces an output-conditioned bilinear refinement with an analytical overhead of 5.2–6.9% relative to the associative attention core and 0.22–0.33% of total model operations across the evaluated backbones. The branch preserves linear scaling in token count and never materializes an $N \times N$ attention matrix. These figures describe the bilinear branch alone; the gated token-specific correction of GABA-M2 is a separate and substantially larger cost, adding 2.5–3.6× the operations of the associative attention core and $2D^2$ parameters per block, as quantified above. These cost estimates complement the representation and downstream-task results reported in our experiments.

F Ablation Study

Component contributions. Table 8 isolates the two proposed mechanisms on the all-linear Tiny backbone under an identical training schedule. Both improve the baseline independently: the anti-symmetric exchange contributes $+0.22/+0.16$ points (online/EMA), while the bilinear branch contributes $+0.31/+0.21$. Their combination reaches $+0.50/+0.41$, approximately matching the summed individual gains ($+0.53/+0.37$); the combined variant is marginally sub-additive under online evaluation and marginally super-additive under EMA. This approximate additivity is consistent with the mechanisms addressing complementary limitations: the bilinear branch enriches retrieval from the compressed state, whereas the anti-symmetric exchange coordinates the global and local value pathways before state formation.

Table 8: **Component contributions.** ImageNet-1K Top-1 at the final epoch using the all-linear Tiny backbone. Δ is measured relative to the baseline. All entries are single training runs under an identical schedule. RALA denotes the RAVLT$_L$-Tiny baseline of Table 2, whose EMA values match those reported there.

	Online		EMA	
Variant	Top-1	Δ	Top-1	Δ
RALA	81.53	—	81.69	—
+AntiSym	81.75	+0.22	81.85	+0.16
+Bilinear	81.84	+0.31	81.90	+0.21
+Both	**82.03**	+0.50	**82.10**	+0.41

Because each entry corresponds to a single training run, individual differences of this magnitude are comparable to typical run-to-run variation on ImageNet-1K, and we therefore do not treat any single pairwise gap as statistically established. Two forms of internal consistency nevertheless support the reported ordering. First, the same ranking holds under both online and EMA evaluation, which respond differently to late-training fluctuation. Second, the ordering persists throughout the final 100 training epochs (Fig. 1) rather than appearing only at the final checkpoint. The component-level conclusions should be read as consistent directional evidence under a fixed schedule rather than as seed-averaged measurements.

Effect of the bilinear rank. Table 9 varies R while holding all other components fixed. Computational cost grows approximately linearly with R: each additional rank contributes about 8.8 K parameters and 0.0021 G FLOPs. Accuracy, however, does not improve monotonically with R. The highest accuracy occurs at R=4, while R=2 and R=8 are nearly indistinguishable (81.90 and 81.89 online), and R=16 is $0.23/0.24$ points below the peak despite a fourfold larger refinement rank. Online and EMA evaluation exhibit the same ordering across all settings. Since $\mathrm{rank}(\Delta S) \le R$, the choice of R controls both the capacity and the structural constraint of the refinement. As in Table 8, these are single-run measurements and the differences between adjacent settings are small; we therefore draw only the weaker conclusion supported by the sweep, namely that enlarging the refinement subspace within the tested range does not improve accuracy and that a compact correction suffices. We use R=4 throughout as the best-performing tested setting, adding only 0.24% parameters and 0.33% FLOPs relative to the R=0 reference.

Table 9: **Effect of the bilinear rank R.** All-linear Tiny backbone. R=0 removes the bilinear branch while retaining all other components (14.865 M, 2.546 G); percentages are relative to this reference. All entries are single training runs.

R	Added params	Added FLOPs	Top-1 (%)	EMA (%)
2	0.018 M (+0.12%)	0.004 G (+0.16%)	81.90	81.97
4	0.035 M (+0.24%)	0.008 G (+0.33%)	**82.03**	**82.10**
8	0.071 M (+0.48%)	0.017 G (+0.66%)	81.89	81.94
16	0.141 M (+0.95%)	0.034 G (+1.32%)	81.80	81.86

Isolating the channel-conditioned correction. The component ablation above compares separately trained models, which differ in their learned parameters as well as in the presence of the bilinear branch. To isolate the correction itself, we additionally compare the base and refined associative quantities *within the same trained GABA-M2$_L$ model*, using identical inputs and learned parameters (protocol in Sec. H.10). Since $\widetilde{S}$ and S differ only by $\Delta S = \gamma A[(B^\top S) \odot C]$, this comparison measures the immediate effect of the channel-conditioned correction on the state and global readout under fixed learned representations; it is a controlled forward-pass comparison rather than an ablation in which the model is retrained without the branch, and it therefore does not establish how the representations would differ under retraining.

Table 10: **Base versus bilinear-refined quantities within GABA-M2_L.** Metrics are averaged across heads within each image and then over 1000 class-balanced validation images; values are means with 95% confidence intervals across validation images. S and $\widetilde{S}$ denote the associative state before and after the channel-conditioned correction, while Y_{base} and $\widetilde{Y}$ denote their corresponding global readouts. Δ gives the relative change calculated from the displayed means. The reported intervals quantify uncertainty in each mean across validation images and are not confidence intervals for the paired base-to-refined difference.

Quantity	Metric	Base	Refined	Δ
State $S \to \widetilde{S}$	Numerical rank	58.99 ± 0.05	60.29 ± 0.04	$+2.2\%$
	Effective rank	7.90 ± 0.06	8.52 ± 0.06	$+7.8\%$
	Stable rank	1.196 ± 0.004	1.238 ± 0.005	$+3.5\%$
Readout $Y_{\text{base}} \to \widetilde{Y}$	Numerical rank	54.29 ± 0.13	57.10 ± 0.10	$+5.2\%$
	Effective rank	5.59 ± 0.05	7.67 ± 0.05	$+37.2\%$
	Stable rank	1.175 ± 0.004	1.475 ± 0.010	$+25.5\%$

Effective and stable rank of the refined state and output. Numerical rank becomes uninformative near the algebraic ceiling, since a formally full-rank representation may still concentrate most of its spectral mass in a few dominant directions. We therefore measure entropy-based effective rank and energy-based stable rank before and after bilinear refinement (Table 10; Fig. 4, left). The base state is nonetheless rank-deficient relative to its 64-dimensional head capacity, and the refinement raises its mean numerical rank by 1.30, with the corresponding readout rising by 2.81; both increases lie within the bound $\operatorname{rank}(\Delta S) \le R$ at $R = 4$, although a tolerance-based count does not by itself establish an increase in exact algebraic rank. At the associative state, effective rank increases from 7.90 to 8.52 ($+7.8\%$) and stable rank from 1.196 to 1.238 ($+3.5\%$), indicating a modest but consistent spectral broadening. The effect is substantially stronger after query retrieval: effective rank increases from 5.59 to 7.67 ($+37.2\%$) and stable rank from 1.175 to 1.475 ($+25.5\%$). The 95% confidence intervals of the base and refined means do not overlap for any of these six comparisons. Thus, although the refinement broadens the state spectrum only modestly, its contribution remains expressed and becomes considerably more pronounced in effective-dimensionality terms after multiplication by Q. The absolute stable ranks remain low (≈ 1.2–1.5), reflecting a strong leading spectral mode; the refinement reduces, rather than eliminates, this concentration.

Where the spectral enrichment appears. The scalar rank measures summarize overall dimensionality but not where the change occurs in the spectrum. At the state level (Fig. 4, right), S and $\widetilde{S}$ are nearly indistinguishable over the leading ~ 10 singular values and separate progressively through the middle and tail, where $\widetilde{S}$ retains larger normalized components. The refinement therefore increases the relative strength of directions that are weakly represented in the original state, broadening the spectrum away from its dominant modes. This behavior is consistent with $\Delta S = \gamma A\big[(B^\top S) \odot C\big]$, where the output-channel-dependent factor C allows the refinement to contribute directions that cannot in general be reduced to a single shared transformation of S (Sec. 2.1).

Enrichment survives query retrieval. State-level enrichment need not automatically appear at the output because Q may suppress directions introduced by ΔS. Figure 4 (middle) shows that the enrichment survives retrieval: relative to $Y_{\text{base}} = (Q_{\text{rope}} S) \odot z$, the refined response $Y_{\text{bil}} = (Q_{\text{rope}} \widetilde{S}) \odot z$ exhibits slower decay across nearly the entire non-leading spectrum. The separation is also more pronounced at the output, consistent with the larger effective-rank increase observed in the left panel. Because each spectrum is normalized by its own leading singular value, this difference reflects broader *relative* spectral utilization rather than a uniform increase in activation magnitude. Together with the rank measurements, these results show that the state-level enrichment remains expressed after query retrieval rather than being absorbed at readout.

G Visual Analysis

Evaluation protocol. We compare the proposed GABA variants with RALA and MHLA under matched architectural settings: GABA-M1_L and GABA-M2_L against RALA_L and MHLA_L in the linear-only setting, and GABA_NL against RALA_NL and MHLA_NL in the

Table 11: **Representation diagnostics for the linear models.** Metrics are averaged over 1000 class-balanced ImageNet-1K validation images (one per class, identical indices across models); values are means with 95% confidence intervals across validation images. Class separability ($\mathcal{S}_{\text{cls}}$) is evaluated separately over the full validation set (50,000 images). r_{eff} and r_{s} denote effective and stable rank, respectively; `OffDiag` denotes mean absolute off-diagonal channel correlation, and `TokenSim` denotes mean absolute off-diagonal token cosine similarity. All evaluated linear-model representations attain numerical rank 128 on each of the 1000 validation images. The reported intervals quantify uncertainty in each model's mean across validation images and are not confidence intervals for paired between-model differences. Best reported value per column is shown in bold.

Model	$r_{\text{eff}} \uparrow$	$r_{\text{s}} \uparrow$	OffDiag↓	TokenSim↓	$\mathcal{S}_{\text{cls}} \uparrow$
RALA_L	77.31 ±0.33	3.351 ±0.061	0.1637 ±0.0014	0.1869 ±0.0021	0.3075
MHLA_L	79.97 ±0.33	3.442 ±0.073	0.1477 ±0.0011	0.1786 ±0.0018	0.2761
GABA-M1_L	82.72 ±0.29	5.262 ±0.079	0.1437 ±0.0011	0.1895 ±0.0020	0.3503
GABA-M2_L	**84.04** ±0.30	**6.578** ±0.083	**0.1259** ±0.0009	**0.1726** ±0.0017	**0.3510**

hybrid linear–nonlinear setting. All analyses use the final attention block of the third stage (`layers.2.blocks.5`) operating on $N = 196$ tokens. Query–key interaction analyses use head 0; representation analyses concatenate two consecutive heads starting from head 0, each of dimension $d_h = 64$, to form $Y \in \mathbb{R}^{196\times128}$ immediately before the final attention output projection.

Three evaluation scales are used. Scalar representation statistics for the linear models are computed independently for each image and averaged over 1000 class-balanced ImageNet-1K validation images (one per class), using identical image indices across models; we report the mean with a 95% confidence interval across validation images. These representation statistics reproduce across an independent class-balanced sample of the same size drawn from the same trained checkpoints, with all reported model orderings preserved and all corresponding confidence intervals overlapping between the two samples; this indicates stability with respect to the validation-image sample rather than across training runs or checkpoint selection. Class separability is evaluated separately over the full validation set (50 images per class, 50,000 images). Heatmaps, spatial overlays, single-query profiles, query–key interaction statistics, and the reported hybrid-model representation values are single-image illustrations; per-panel annotations refer to the displayed image. Metric definitions follow Sec. H.

Two scope conditions apply throughout. First, interaction visualizations characterize the query–key component of each model-specific attention operator rather than the complete block output, which may additionally include local enhancement, gating, and residual pathways. For GABA-M2_L, absolute output-channel-conditioned interaction magnitudes are averaged across channels before key-wise normalization; its aggregate profile therefore does not establish the selectivity of individual output channels. Second, cross-model measurements compare separately trained architectures and do not independently isolate the contribution of an individual operation. The within-model analysis in Sec. F instead isolates the immediate effect of the bilinear correction under fixed learned parameters.

G.1 Representation Geometry

Spectral utilization. The displayed pre-projection representations attain numerical rank 128 (Fig. 5); furthermore, all 1000 evaluated representations attain numerical rank 128 for each linear model (Table 11). Thus, numerical rank does not distinguish the evaluated linear-model representations under the stated criterion, whereas effective rank differentiates how singular-value magnitude is distributed across their available directions. Effective rank increases from 77.31 for RALA_L and 79.97 for MHLA_L to 82.72 for GABA-M1_L and 84.04 for GABA-M2_L, corresponding to increases of approximately 7.0% and 8.7% relative to RALA_L. Stable rank shows a larger separation, increasing from 3.351 and 3.442 for the baselines to 5.262 and 6.578 for GABA-M1_L and GABA-M2_L, respectively, corresponding to increases of approximately 57.0% and 96.3% over RALA_L. The normalized singular-value spectra in Fig. 7 (middle) illustrate corresponding differences in the selected image, with GABA-M2_L retaining stronger non-leading components across much of the plotted range. GABA-M1_L exhibits this improvement under a parameter budget com-

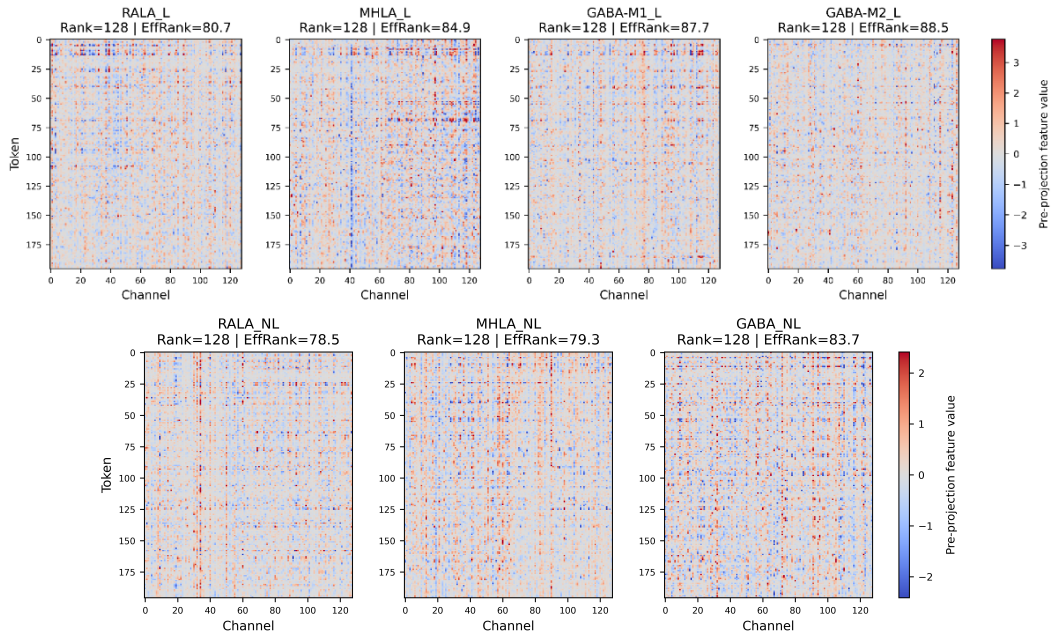

Figure 5: **Pre-projection representations and effective dimensionality of GABA and the corresponding baselines.** The top row compares RALA_L, MHLA_L, GABA-M1_L, and GABA-M2_L; the bottom row compares RALA_NL, MHLA_NL, and GABA_NL on the same ImageNet-1K validation image. Features are extracted immediately before the final attention output projection from the final block of the third stage operating on $N = 196$ tokens, concatenating two consecutive heads starting from head 0, each of dimension $d_h = 64$, to form $Y \in \mathbb{R}^{196\times128}$. Because the pre-projection activations are signed, panels use a diverging colormap on a common symmetric scale within each architectural setting. Numerical rank and entropy-based effective rank are computed from the original unscaled feature matrices. Their annotations refer to the displayed image, while dataset-level statistics for the linear models appear in Table 11. All displayed matrices attain numerical rank 128, while their effective ranks differ.

parable to RALA_L, demonstrating that broader spectral utilization does not require a substantial increase in model size. The further spectral margin for GABA-M2_L is consistent with the direction of the isolated within-model bilinear-refinement effect reported in Sec. F, although the cross-model comparison does not independently establish that the correction accounts for the full difference. In the corresponding single-image hybrid comparison, GABA_NL obtains effective rank 83.7 and stable rank 4.82, compared with 78.5 and 4.57 for RALA_NL and 79.3 and 4.71 for MHLA_NL.

Channel and token redundancy. Channel correlation (Fig. 6) and token similarity (Fig. 8) complement the spectral analysis by characterizing pairwise alignment along the feature and token axes, respectively (Zbontar et al., 2021; Bardes et al., 2021; Dong et al., 2021). Across the 1000 evaluated images, mean absolute off-diagonal channel correlation decreases from 0.1637 for RALA_L and 0.1477 for MHLA_L to 0.1437 for GABA-M1_L and 0.1259 for GABA-M2_L, corresponding to reductions of approximately 12.2% and 23.1% relative to RALA_L (Table 11). Thus, both GABA variants exhibit lower mean pairwise channel correlation than the corresponding linear baselines under the stated extraction protocol.

The token-level results exhibit a different pattern. GABA-M1_L does not reduce token similarity relative to RALA_L (0.1895 versus 0.1869) or MHLA_L (0.1786), whereas GABA-M2_L obtains 0.1726, approximately 7.7% below RALA_L and lower than both evaluated baselines. This distinction is consistent with the additional representational flexibility permitted by output-channel-conditioned bilinear refinement, but the comparison between separately trained models does not isolate its contribution to token similarity. In the selected hybrid example, GABA_NL obtains channel correlation 0.163 against 0.172 for RALA_NL and 0.166 for MHLA_NL, while its token similarity is 0.158 against 0.232 and 0.223, respectively. Together, the spectral, channel-correlation,

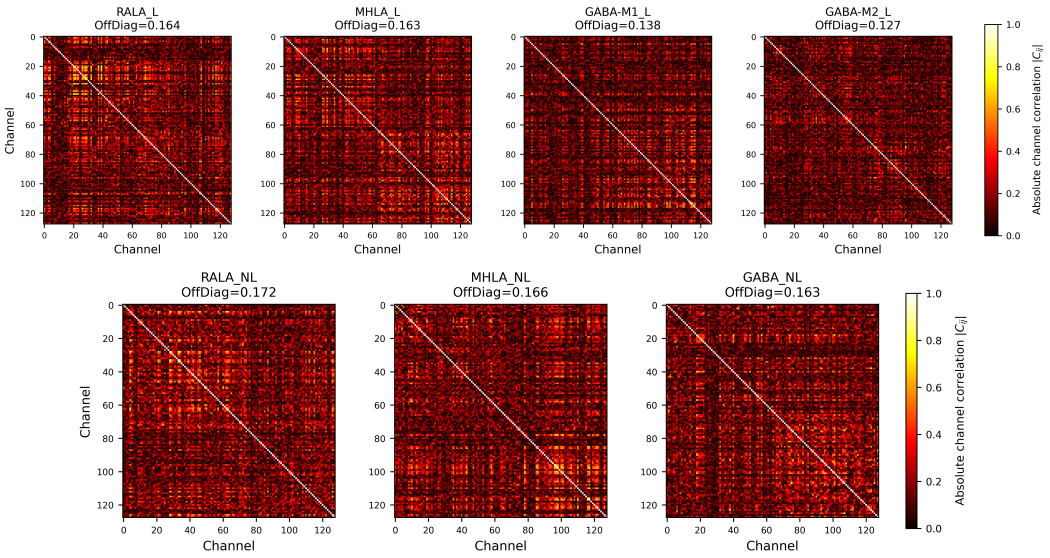

Figure 6: **Channel-correlation structure of GABA and the corresponding RALA and MHLA baselines.** The top row compares the linear-only variants RALA_L, MHLA_L, GABA-M1_L, and GABA-M2_L; the bottom row compares the hybrid variants RALA_NL, MHLA_NL, and GABA_NL. Representations are extracted immediately before the final attention output projection from the final block of the third stage operating on $N = 196$ tokens, using two consecutive heads of dimension $d_h = 64$ to form $Y \in \mathbb{R}^{196\times128}$. Each channel is centered and standardized across tokens before computing the pairwise channel-correlation matrix; panels visualize absolute correlations on the common range $[0, 1]$. `OffDiag` denotes mean absolute off-diagonal correlation, where lower values indicate reduced pairwise linear correlation between feature-channel responses across tokens. Annotated values correspond to the displayed image, while the averaged linear-model statistics appear in Table 11.

and token-similarity diagnostics characterize complementary properties of the evaluated representations; lower pairwise correlation or similarity does not by itself establish statistical independence, semantic complementarity, or improved prediction quality.

Class-discriminative geometry. Figure 7 (left) compares the class-discriminative geometry of GABA with the corresponding RALA and MHLA baselines over the full ImageNet-1K validation set. In the linear-only setting, RALA_L and MHLA_L obtain class-separability scores of 0.3075 and 0.2761, respectively, whereas GABA-M1_L and GABA-M2_L obtain 0.3503 and 0.3510. In the hybrid setting, GABA_NL obtains 0.5813, compared with 0.4187 for RALA_NL and 0.3326 for MHLA_NL. The GABA variants therefore exhibit greater between-class dispersion relative to within-class variation than their corresponding baselines under the stated evaluation protocol. Notably, MHLA_L exhibits higher effective rank than RALA_L but lower class separability, illustrating that broader spectral utilization does not necessarily imply stronger class-discriminative geometry. The GABA results exhibit both properties under the reported measurements, without establishing that every additional singular direction is individually discriminative.

G.2 Query–Key Interaction Structure

Query-scale response. Figure 9 compares the responses of GABA and its corresponding RALA and MHLA baselines to query-scale perturbations. In the displayed linear-only example, MHLA_L exhibits relatively small changes across scales, whereas RALA_L shows a visible redistribution of interaction magnitude. GABA-M1_L progressively concentrates more magnitude on several dominant keys, while GABA-M2_L changes the relative contributions of multiple key regions while retaining broader aggregate support. In the corresponding hybrid comparison, MHLA_NL develops a strongly dominant response, whereas GABA_NL exhibits a dominant interaction alongside secondary responses at lower query scales. For linear attention, the experiment evaluates $\phi(aq)$ rather than $a\phi(q)$ and therefore characterizes the response of the model-specific kernelized retrieval mech-

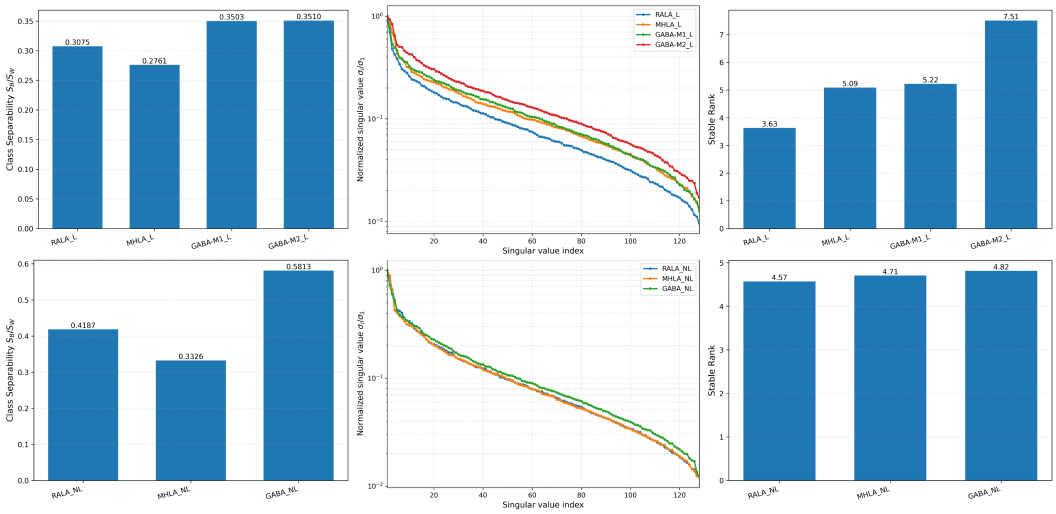

Figure 7: **Class-discriminative and spectral analysis of GABA and the corresponding baselines.** The top row compares RALA_L, MHLA_L, GABA-M1_L, and GABA-M2_L; the bottom row compares RALA_NL, MHLA_NL, and GABA_NL. Representations are extracted immediately before the final attention output projection from the final block of the third stage operating on $N = 196$ tokens, using two consecutive heads of dimension $d_h = 64$ to form $Y \in \mathbb{R}^{196\times128}$. **Left:** Class separability, evaluated on ImageNet-1K using 50 validation images per class (50,000 images) after global token averaging and ℓ_2 normalization, as the ratio of sample-weighted between-class to within-class scatter. **Middle:** Singular values normalized by the leading component, $\bar{\sigma}_i = \sigma_i/\sigma_1$, on a logarithmic scale; slower decay indicates stronger non-leading components relative to the leading singular value rather than larger absolute singular values. **Right:** Stable rank of the extracted representation, characterizing spectral energy relative to the dominant singular component. The spectral plots illustrate the selected representation, whereas class separability is evaluated over the full validation set. The displayed single-image stable ranks differ in magnitude from the dataset-level values in Table 11, although the model ordering is preserved.

anism rather than directly testing the query-magnitude invariance analyzed by MALA (Fan et al., 2025a). For a fixed-key Softmax operator, query scaling also changes the logit scale and can sharpen the interaction distribution independently of the architecture. The observed differences therefore describe query-scale responses in the selected examples rather than establishing that greater sensitivity is intrinsically preferable.

Interaction support and concentration. Output-channel-conditioned retrieval permits different output channels to emphasize different keys, potentially producing broader aggregate interaction support. The fixed-query profiles in Fig. 10 illustrate the corresponding aggregate structure in the selected image: RALA_L distributes interaction magnitude across several keys, MHLA_L exhibits a more concentrated profile, GABA-M1_L retains pronounced key-specific responses, and GABA-M2_L exhibits a broader multi-peak aggregate profile. Because the GABA-M2_L profile is computed after aggregation across output channels, it characterizes aggregate interaction support and does not establish whether individual channels are sharply selective or consistently emphasize different key regions. In the hybrid comparison, Fig. 11 reports normalized row-wise entropies of 0.935 for RALA_NL, 0.704 for MHLA_NL, and 0.739 for GABA_NL in the displayed image, indicating more concentrated interaction magnitude for GABA_NL than RALA_NL while retaining higher entropy than MHLA_NL. These observations describe interaction support in the selected examples and are not dataset-level measurements; they do not identify an optimal concentration regime.

H Metrics Used in Attention-Space Analysis

Task-level accuracy provides the ultimate measure of recognition performance, but it alone does not reveal whether an attention mechanism improves the internal representation as intended by its

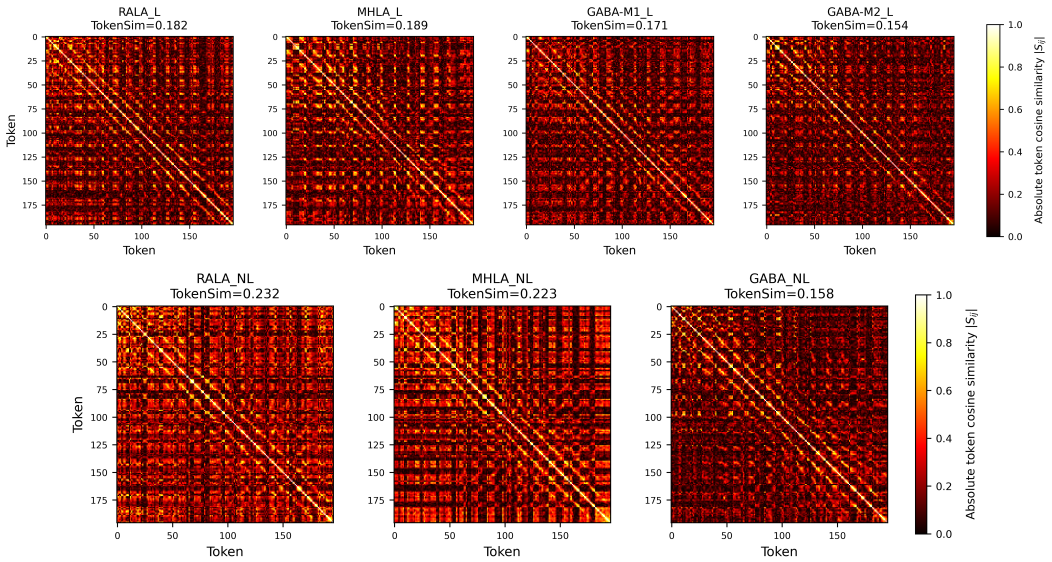

Figure 8: **Token-wise similarity of GABA and the corresponding RALA and MHLA baselines.** The top row compares RALA_L, MHLA_L, GABA-M1_L, and GABA-M2_L; the bottom row compares RALA_NL, MHLA_NL, and GABA_NL on the same ImageNet-1K validation image. Pre-projection features are extracted from the final block of the third stage operating on $N = 196$ tokens, with two consecutive heads of dimension $d_h = 64$ concatenated to form $Y \in \mathbb{R}^{196\times128}$. Each token representation is ℓ_2-normalized before computing pairwise cosine similarity; panels show absolute similarities on the common range $[0, 1]$. `TokenSim` denotes mean absolute off-diagonal token similarity, where lower values indicate weaker pairwise alignment between spatial-token representations. Annotated values correspond to the displayed image. For this metric the displayed image is not representative of the dataset-level ordering: averaged over 1000 validation images, GABA-M1_L exhibits higher token similarity than both RALA_L and MHLA_L, whereas GABA-M2_L remains lowest (Table 11).

design. This distinction matters for linear attention, where computational efficiency comes from associative factorization. However, the resulting representations may suffer from restricted rank, redundant feature directions, token homogenization, or poorly concentrated attention distributions. Prior analyses have linked rank degeneration to limited modeling capacity in linear attention (Fan et al., 2025b), shown an intrinsic tendency of self-attention toward token uniformity and rank collapse (Dong et al., 2021), and shown that the effective dimensionality of learned representations can strongly inform their downstream quality (Li et al., 2022). Complementary work further emphasizes redundancy reduction across representation dimensions (Zbontar et al., 2021; Bardes et al., 2021) and the importance of concentrated, low-entropy interactions for reproducing the behavior of Softmax attention (Zhang et al., 2024). These observations lead us to analyze the learned attention representations from four complementary perspectives: *dimensionality*, *feature redundancy*, *discriminative geometry*, and *attention concentration and sensitivity*. We then provide a mechanism-specific analysis of the bilinear retrieval state and a qualitative visualization protocol.

Let $Y = [y_1, \ldots, y_N]^\top \in \mathbb{R}^{N\times d}$ denote the pre-projection representation produced by the selected attention block. Here, N denotes the number of spatial tokens and d the analyzed channel dimension. We evaluate the representation immediately before the final output projection, avoiding additional channel mixing that could obscure the representation directly generated by the attention mechanism. For comparisons involving multiple attention heads, the corresponding head representations are concatenated along the feature dimension to construct the same $N \times d$ analysis matrix for every model.

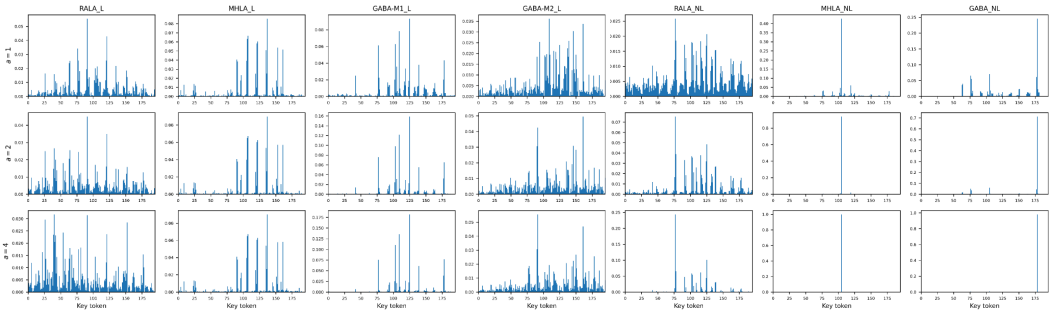

Figure 9: **Query-scale sensitivity of GABA and the corresponding RALA and MHLA baselines.** For the same ImageNet-1K validation image, we scale a selected query by $a \in \{1, 2, 4\}$ and visualize its normalized absolute interaction magnitudes over all $N = 196$ key tokens. The first four columns compare RALA_L, MHLA_L, GABA-M1_L, and GABA-M2_L; the last three compare RALA_NL, MHLA_NL, and GABA_NL. Rows denote increasing query scales. All models use head 0, query token $i = 105$, and the final block of the third stage operating on $N = 196$ tokens. For linear attention, scaling is applied to the pre-kernel query, followed by recomputation of the model-specific feature mapping, positional transformation, query-dependent quantities, and normalization. For Softmax attention, the projected query is scaled before recomputing the query–key logits. The resulting interaction magnitudes are normalized across keys and characterize changes in relative retrieval support rather than raw attention-output magnitude.

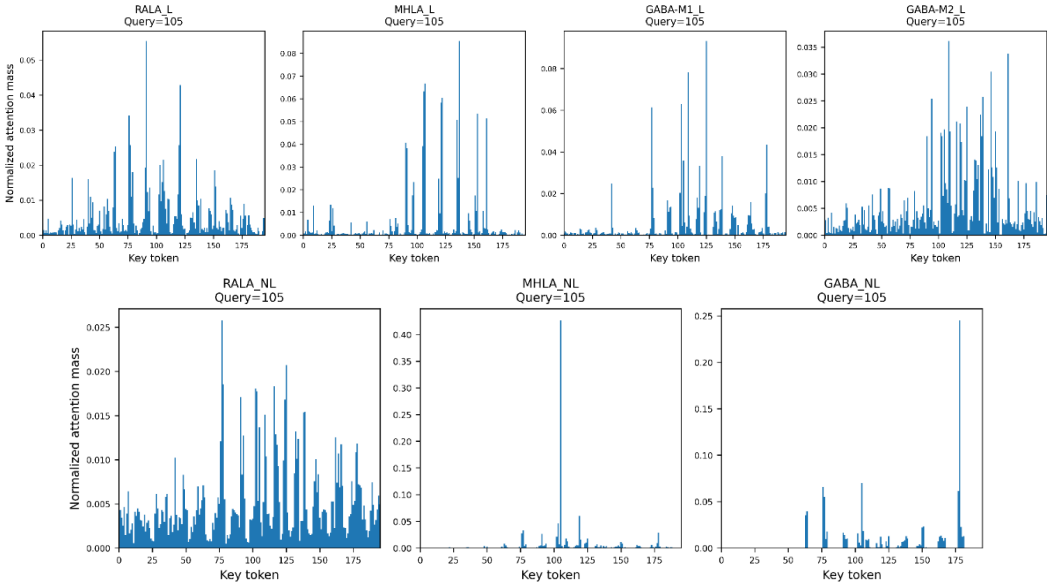

Figure 10: **Normalized query–key interaction magnitudes for a fixed query.** For the same ImageNet-1K validation image, we visualize the relative interaction magnitude between query token $i = 105$ and all $N = 196$ key tokens. The top row compares RALA_L, MHLA_L, GABA-M1_L, and GABA-M2_L; the bottom row compares RALA_NL, MHLA_NL, and GABA_NL. All models use head 0 of the final block of the third stage operating on $N = 196$ tokens. Absolute query–key interaction magnitudes are ℓ_1-normalized across keys so that the bars describe relative retrieval support rather than raw activation magnitude. For GABA-M2_L, absolute output-channel-conditioned interaction magnitudes are averaged across output channels before key-wise normalization; its displayed profile therefore represents aggregate channel-conditioned retrieval.

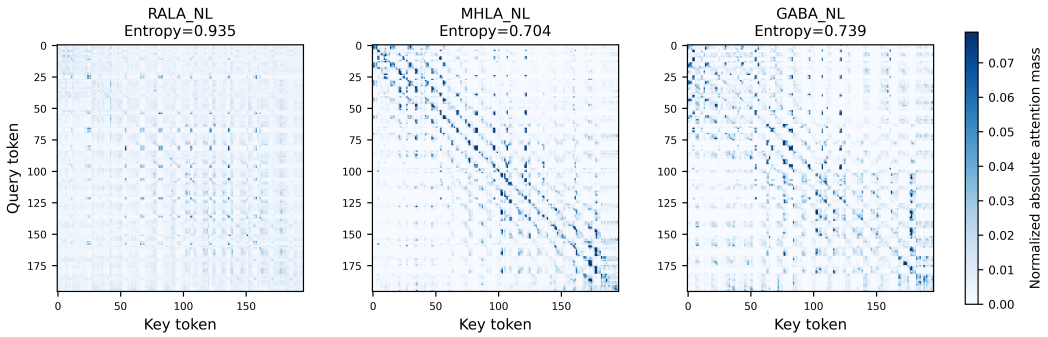

Figure 11: **Query–key interaction structure of GABA_NL and its hybrid baselines.** We compare RALA_NL, MHLA_NL, and GABA_NL on the same ImageNet-1K validation image using head 0 of the final block of the third stage operating on $N = 196$ tokens. Each panel visualizes the absolute query–key interaction matrix under a common color scale. The reported entropy is computed from the corresponding row-normalized nonnegative interaction magnitudes, $H = -\frac{1}{N \log N} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} \log(A_{ij} + \epsilon)$ with $\epsilon = 10^{-12}$, where each row of A is normalized across key tokens. The reported entropy characterizes the concentration of row-normalized interaction magnitudes under the selected model-specific attention operator; lower entropy does not, by itself, imply better predictive performance.

H.1 Matrix Rank

Definition. The algebraic rank measures the number of linearly independent directions contained in a representation. Given the singular value decomposition

$$Y = U \Sigma V^{\top}, \qquad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), \tag{92}$$

with $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$, the algebraic rank $\mathrm{rank}(Y)$ is the number of nonzero singular values of Y. In numerical evaluation we use the tolerance-dependent rank

$$\mathrm{rank}_{\tau}(Y) = |\{i : \sigma_i > \tau\}|, \tag{93}$$

where τ is the tolerance applied by the matrix-rank estimator. The maximum attainable algebraic rank is

$$r_{\max} = \min(N, d). \tag{94}$$

Throughout the paper, reported "numerical rank" refers to rank_{τ}.

Why it matters. Rank provides the most direct measure of how many independent feature directions are available to represent token information. This is especially relevant to linear attention. (Fan et al., 2025b) identify low-rank feature representations as an important source of the performance gap between linear and Softmax attention and explicitly analyze both the compressed KV state and output representations. More generally, (Dong et al., 2021) show that repeated attention operations can drive token representations toward a rank-one configuration, establishing a direct connection between rank degeneration and loss of token diversity. For our analysis, rank therefore provides a first-order indication of whether the attention representation spans a sufficiently rich feature subspace. A higher rank indicates that a larger number of linearly independent directions are retained and is therefore preferable when evaluating *representational capacity*. However, rank alone is a coarse diagnostic: a matrix may be formally full rank while most of its energy remains concentrated in only a few singular directions. Consequently, we report rank jointly with effective rank, stable rank, and the complete singular-value spectrum.

Interpretation. Higher is preferable for representational dimensionality, up to $r_{\max}$. Nevertheless, full numerical rank should not by itself be interpreted as evidence of better recognition performance, since weak or noise-dominated singular directions can also make a matrix formally full rank under a given tolerance.

H.2 Effective Rank

Definition. To characterize the *effective*, rather than merely algebraic, dimensionality of Y, we use the entropy-based effective rank introduced by (Roy & Vetterli, 2007) and subsequently adopted for representation-quality analysis by (Garrido et al., 2023). We first normalize the singular values as

$$p_i = \frac{\sigma_i}{\sum_{j=1}^{r} \sigma_j + \epsilon}, \tag{95}$$

and compute their spectral entropy

$$H_\sigma(Y) = -\sum_{i=1}^{r} p_i \log(p_i + \epsilon). \tag{96}$$

The effective rank is then

$$r_{\text{eff}}(Y) = \exp(H_\sigma(Y)). \tag{97}$$

Unlike discrete rank, r_{eff} accounts for the relative magnitude of all singular directions. If one singular value dominates, r_{eff} approaches one; if the spectral energy is distributed comparably over r directions, r_{eff} approaches r.

Why it matters. This distinction is important in our setting because nominally full-rank representations can still behave as effectively low-dimensional matrices. RankMe demonstrates that effective rank is informative of representation quality and can track downstream performance without labels (Garrido et al., 2023), while analyses of dimensional collapse similarly associate concentrated spectra with impoverished learned representations (Li et al., 2022). Effective rank is therefore a stronger diagnostic than numerical rank for characterizing how representational energy is distributed. An increase in r_{eff} indicates that the singular-value distribution becomes less concentrated in a small number of dominant directions, even when numerical rank remains unchanged. This measures broader spectral utilization rather than establishing that individual directions carry additional discriminative information; the latter question is addressed separately in Sec. H.7.

Interpretation. Higher is preferable for effective representational dimensionality. A larger r_{eff} indicates a more evenly utilized spectrum and less concentration in a small number of dominant directions.

H.3 Stable Rank

Definition. We additionally report the stable rank,

$$r_{\text{stable}}(Y) = \frac{\|Y\|_F^2}{\|Y\|_2^2} = \frac{\sum_{i=1}^{r} \sigma_i^2}{\sigma_1^2}, \tag{98}$$

where $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius and spectral norms, respectively. For every nonzero Y, stable rank satisfies

$$1 \le r_{\text{stable}}(Y) \le \text{rank}(Y), \tag{99}$$

where $\text{rank}(Y)$ denotes exact algebraic rank rather than the tolerance-dependent estimate $\text{rank}_\tau(Y)$; singular values falling below τ are excluded from rank_τ but still contribute to the stable rank. Stable rank is a continuous measure of spectral dimensionality and has previously appeared in deep-learning analyses as a measure of effective matrix complexity (Sanyal et al., 2019).

Why it matters. Whereas effective rank measures the entropy of the normalized singular-value spectrum, stable rank compares the *total spectral energy* with the energy carried by the largest singular direction. It therefore directly reveals whether the representation is dominated by its principal component. For example, two matrices may have identical numerical rank while having substantially different stable ranks if one exhibits a much sharper leading singular value.

In our representation analysis, a larger stable rank indicates that information is distributed across multiple directions rather than being overwhelmingly concentrated along a single dominant mode. This complements effective rank and provides a second continuous measure of dimensional utilization.

Interpretation. For the *feature representations analyzed here*, **higher stable rank indicates a more distributed and less spectrally dominated representation**. This interpretation is specific to representation expressiveness. Stable rank is a general matrix statistic, and previous work has also regularized the stable rank of *weight operators* for different objectives (Sanyal et al., 2019); therefore, "higher is better" should not be treated as a universal statement outside the present representation-level analysis.

H.4 Singular-Value Spectrum

Definition. Rank-based scalar measures compress the spectrum into a single number and may therefore hide substantial differences in spectral geometry. We consequently inspect the complete normalized singular-value spectrum,

$$\widetilde{\sigma}_i = \frac{\sigma_i}{\sigma_1 + \epsilon}, \qquad i = 1, \dots, r. \tag{100}$$

The spectrum is visualized on a logarithmic vertical scale to expose differences among both dominant and low-energy directions.

Why it matters. A rapidly decaying spectrum indicates that most representational energy lies in a small subspace even when the matrix is technically full rank. In contrast, a slower spectral decay indicates that a broader set of directions contributes meaningfully to the representation. Such spectral diagnostics are widely used to identify partial dimensional collapse (Li et al., 2022; Garrido et al., 2023), and are particularly informative for our higher-rank analysis because they reveal *where* in the spectrum a change in effective dimensionality originates.

A useful refinement should therefore not merely perturb the leading directions; it should strengthen previously weak spectral directions sufficiently to produce a broader spectral tail.

Interpretation. There is no single scalar "higher/lower" criterion. **A slower decay and larger normalized tail singular values indicate stronger utilization of multiple representation directions**. Conversely, a steep decay toward zero indicates effective low-dimensionality. As with effective rank, an artificially flat spectrum caused by noise is not intrinsically desirable, so the spectral analysis should be interpreted jointly with discriminative performance.

H.5 Channel Redundancy

Definition. To determine whether different representation dimensions carry linearly redundant information, we measure pairwise channel correlation. Each channel is first standardized over the N tokens,

$$\widehat{Y}_{ij} = \frac{Y_{ij} - \mu_j}{s_j + \epsilon}, \tag{101}$$

where

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} Y_{ij}, \qquad s_j^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_{ij} - \mu_j)^2. \tag{102}$$

The channel-correlation matrix is

$$C_{\mathrm{ch}} = \frac{1}{N}\widehat{Y}^\top \widehat{Y} \in \mathbb{R}^{d \times d}. \tag{103}$$

We summarize redundancy through the mean absolute off-diagonal correlation,

$$\rho_{\mathrm{ch}} = \frac{1}{d(d-1)} \sum_{i \neq j} \left| C_{\mathrm{ch},ij} \right|. \tag{104}$$

Why it matters. High dimensionality alone does not guarantee expressive representations if many channels encode nearly identical directions. Redundancy reduction has consequently become an important principle in representation learning. For example, Barlow Twins explicitly drives off-diagonal cross-correlations toward zero to reduce redundancy among representation components (Zbontar et al., 2021), while VICReg independently penalizes covariance between different representation variables to prevent dimensional collapse (Bardes et al., 2021). Our metric is not their

training objective; rather, it adopts the same underlying diagnostic principle to evaluate redundancy within the attention representation.

This metric is particularly complementary to rank analysis. Higher effective rank together with lower channel correlation provides considerably stronger evidence of enriched representation geometry than either property alone: the former indicates that multiple directions carry spectral energy, while the latter indicates that those directions are less linearly redundant.

Interpretation. Lower is preferable when measuring channel redundancy. A smaller ρ_{ch} indicates weaker average pairwise linear correlation between feature-channel responses across tokens. It does not by itself establish statistical independence or semantic complementarity between channels, and values need not approach zero, since correlated channels may legitimately encode related visual structures.

H.6 Token-to-Token Similarity

Definition. To measure redundancy across spatial tokens rather than feature channels, each token representation is ℓ_2-normalized,

$$\widehat{y}_i = \frac{y_i}{\|y_i\|_2 + \epsilon}, \tag{105}$$

and the token-similarity matrix is computed as

$$T_{ij} = \widehat{y}_i^{\top} \widehat{y}_j. \tag{106}$$

We report the mean absolute off-diagonal similarity

$$\rho_{\mathrm{tok}} = \frac{1}{N(N-1)} \sum_{i \neq j} |T_{ij}| \,. \tag{107}$$

Why it matters. Token diversity is fundamental to dense visual representation learning because different image regions should retain sufficiently distinct representations to encode heterogeneous spatial content. (Dong et al., 2021) theoretically show that self-attention possesses an inductive tendency toward *token uniformity*, with repeated pure-attention operations converging toward a rank-one token representation. Token similarity therefore provides a direct complementary view of the same degeneration captured spectrally by rank.

In the context of linear attention, a high average token similarity suggests that spatial positions are mapped to nearly parallel representation vectors, reducing their capacity to encode distinct structures. Lower token similarity, when accompanied by strong task performance, indicates that the model preserves a richer set of token-dependent responses.

Interpretation. Lower is generally preferable for token diversity, since it indicates less token homogenization. This should again not be interpreted as requiring mutually orthogonal tokens: semantically related image regions can naturally produce similar representations. The metric is intended to detect excessive uniformity rather than to enforce zero similarity.

H.7 Class Separability

Definition. Rank and redundancy measure representational richness without using labels, but they do not establish whether the additional variation is discriminative. We therefore complement them with a Fisher-style between-/within-class separability criterion. For each image, the token representations are first globally averaged,

$$x_n = \frac{1}{N} \sum_{i=1}^{N} y_{n,i}, \tag{108}$$

followed by ℓ_2 normalization. Let μ_c denote the mean embedding of class c, μ the global mean, and n_c the number of evaluated images belonging to class c. We compute the within-class scatter

$$S_W = \frac{1}{M} \sum_{c} \sum_{x_n \in c} \|x_n - \mu_c\|_2^2 \,, \tag{109}$$

and between-class scatter

$$S_B = \frac{1}{M} \sum_c n_c \left\| \mu_c - \mu \right\|_2^2, \tag{110}$$

where $M = \sum_c n_c$. The class-separability score is

$$\mathcal{S}_{\text{cls}} = \frac{S_B}{S_W + \epsilon}. \tag{111}$$

Why it matters. A useful representation should simultaneously exhibit compact structure within semantic classes and sufficient separation between different classes. This geometric principle is closely related to the extensive literature on discriminative feature learning and to observations underlying Neural Collapse, where well-trained classifiers exhibit decreasing within-class variability and highly separated class means (Zhu et al., 2021). Our objective here is not to claim Neural Collapse, but to use the more general geometric principle that discriminative representations should exhibit favorable between-class separation relative to within-class dispersion.

This metric is especially important when interpreting the rank-based diagnostics: broader spectral utilization is useful only if the corresponding variation is organized along discriminative directions rather than being unstructured. Higher class separability therefore indicates that richer representation geometry is consistent with stronger class-discriminative organization under the evaluated criterion.

Interpretation. Higher is preferable for discriminative representation quality. It can arise through larger between-class separation, smaller within-class scatter, or both.

H.8 Normalized Attention Entropy

Common attention-mass normalization. Softmax attention produces nonnegative normalized weights, whereas effective interactions reconstructed from linear attention may be signed and need not form a probability distribution. To enable a common concentration analysis, we convert the effective attention matrix $A \in \mathbb{R}^{N \times N}$ into row-normalized interaction magnitude,

$$P_{ij} = \frac{|A_{ij}|}{\sum_{k=1}^{N} |A_{ik}| + \epsilon}. \tag{112}$$

For nonzero interaction rows this yields $\sum_j P_{ij} \approx 1$, with the small deviation arising from the numerical stabilizer ϵ. For Softmax attention the construction reduces to the ordinary attention distribution; for signed linear attention it measures the distribution of *interaction magnitude*. Consequently, these diagnostics intentionally assess concentration rather than the polarity of individual interactions.

Definition. For each query token, we compute Shannon entropy and normalize it by the entropy of the uniform distribution:

$$\mathcal{H}_i = -\frac{\sum_{j=1}^{N} P_{ij} \log(P_{ij} + \epsilon)}{\log N}. \tag{113}$$

The reported attention entropy is

$$\mathcal{H}_{\text{attn}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{H}_i, \tag{114}$$

which is approximately bounded between 0 and 1 under the stated normalization.

Why it matters. Attention concentration has been explicitly investigated in recent linear-attention work. Hedgehog identifies low-entropy or "spiky" attention weights as a property of Softmax attention that is missing from several linear alternatives and shows that recovering this behavior substantially narrows the performance gap (Zhang et al., 2024). Related linear-attention designs similarly seek stronger concentration while retaining linear complexity (Qin et al., 2022). Entropy therefore provides a direct quantitative measure of how interaction magnitude is distributed across keys: a value close to one corresponds to nearly uniform interaction mass, whereas a smaller value indicates increasingly selective retrieval.

Interpretation. Lower entropy indicates greater concentration of interaction magnitude, whereas higher entropy indicates broader interaction support. Neither extreme is universally preferable: concentrated interactions can emphasize selected keys, while broader support can preserve multiple relevant regions. We therefore interpret entropy together with qualitative spatial structure and recognition performance rather than treating lower entropy alone as evidence of a better attention mechanism.

H.9 Sensitivity to Query Magnitude

Motivation. Attention concentration should not only depend on the angular relation between a query and its keys; the magnitude of the query can also alter the confidence of retrieval. For Softmax attention, scaling a query q_i by a positive factor a scales its logits,

$$P_{i,:}^{(a)} = \mathrm{Softmax}\left(\frac{a\, q_i K^\top}{\sqrt{d}}\right), \tag{115}$$

which generally sharpens the resulting distribution as a increases. (Fan et al., 2025a) identify the neglect of query magnitude as a fundamental limitation of conventional normalized linear attention: because query scaling can cancel between the numerator and denominator, the resulting attention distribution may remain nearly invariant despite large changes in query magnitude.

Evaluation. To test whether an attention mechanism retains this sensitivity, we perturb a selected *pre-kernel* query using

$$q_i^{(a)} = a\, q_i, \qquad a \in \{1, 2, 4\}, \tag{116}$$

and reconstruct the complete effective attention distribution after reapplying the corresponding kernel mapping, positional transformation, and linear-attention normalization. Importantly, all query-dependent quantities are recomputed after perturbation. Thus, when the model also uses the query to modulate its key representation, the altered key weighting is included rather than being held fixed. For the output-channel-conditioned bilinear variant, both the base linear interaction and the bilinear interaction are recomputed before aggregation.

We then visualize

$$P_{i,:}^{(a)} = \frac{\left|A_{i,:}^{(a)}\right|}{\left\|A_{i,:}^{(a)}\right\|_1 + \epsilon} \tag{117}$$

for increasing a.

Why it matters. This experiment probes a fundamentally different property from static attention concentration. Entropy describes *what* the current distribution looks like, whereas query-scale analysis measures whether the attention mechanism can *adapt* its retrieval distribution when the query strength changes. Such adaptive behavior is an important distinction between Softmax and magnitude-insensitive linear formulations (Fan et al., 2025a).

Interpretation. The query-scale experiment characterizes how the relative interaction distribution responds to changes in query magnitude. A nearly unchanged distribution indicates magnitude insensitivity under the tested perturbation, whereas changes in key-wise interaction mass indicate scale-dependent retrieval. Neither greater sensitivity nor stronger concentration is intrinsically preferable; the resulting behavior must be interpreted alongside the model's retrieval structure and recognition performance.

H.10 Direct Rank Analysis of the Bilinear Retrieval State

The preceding metrics characterize the complete pre-projection representation and are therefore influenced by all components within the attention block. To isolate whether the bilinear retrieval mechanism itself enriches the global attention representation, we additionally analyze its state and output *before* subsequent local enhancement, residual correction, output gating, and final projection. This subsection defines the protocol; the corresponding measurements are reported in Sec. F.

For each head, let Q_{rope} and K_{rope} denote the position-aware query and key features and let V denote the value representation entering the global associative pathway. The base associative state

is

$$S = \frac{K_{\text{rope}}^{\top} V}{N}. \tag{118}$$

For the output-channel-conditioned bilinear refinement, we reconstruct the unscaled correction

$$\Delta_0 = A\left[(B^{\top} S) \odot C\right], \tag{119}$$

and the refined state

$$\widetilde{S} = S + \gamma\, \Delta_0, \qquad \gamma = \gamma_{\text{impl}}\, s_{\text{cross}}, \tag{120}$$

where A, B, and C are the learned low-rank bilinear factors, γ_{impl} is the learned branch coefficient, and $s_{\text{cross}} = 1/\sqrt{R}$ is the fixed rank-dependent scale. The effective coefficient γ therefore already absorbs s_{cross}, consistent with the convention used in Sec. D and Sec. E. The associated global outputs are

$$Y_{\text{base}} = (Q_{\text{rope}} S) \odot z, \qquad Y_{\text{bil}} = (Q_{\text{rope}} \widetilde{S}) \odot z, \tag{121}$$

where z denotes the linear-attention normalizer.

We apply numerical rank, effective rank, stable rank, and singular-spectrum analysis separately to $(S, \widetilde{S})$ and $(Y_{\text{base}}, Y_{\text{bil}})$. This controlled comparison tests the mechanism at the exact point where rank enrichment is expected to occur, rather than attributing changes observed after subsequent nonlinear or residual operations to the bilinear refinement. The comparison holds the learned parameters fixed and therefore isolates the immediate effect of the correction within a single forward pass; it is not a retraining ablation.

Interpretation. Evidence for successful spectral enrichment consists of **higher effective and stable rank and a broader singular-value spectrum for $\widetilde{S}$ relative to S, with the same tendency propagating from Y_{base} to Y_{bil}**. Differences in numerical rank are also informative when the base representation is rank-deficient, but rank_τ depends on the tolerance τ and can saturate at its maximum; an increase indicates that additional singular directions exceed the evaluation threshold rather than establishing an increase in exact algebraic rank. This direct state-level analysis provides a mechanism-specific complement to the downstream accuracy results.

H.11 Qualitative Representation and Attention Visualization

Finally, we complement the scalar diagnostics with representation and attention visualizations. For the pre-projection feature maps, raw feature values are displayed under a shared symmetric color scale within each architectural setting, enabling direct comparison of activation magnitudes among models in the same linear-only or hybrid group. This is essential because separately normalizing each model would conceal differences in feature magnitude and could make weak and strong activations appear visually equivalent. The heatmaps expose token–channel structure and provide a qualitative counterpart to rank, spectral, and correlation analyses.

For spatial attention visualization, the normalized interaction magnitude associated with a selected query token is reshaped to the spatial token grid and overlaid on the original image. We use the same visualization normalization across the compared models so that concentration patterns remain directly comparable. Such visualizations are intended to answer whether strong interactions correspond to semantically meaningful image regions rather than simply whether the numerical distribution is sparse. Similar attention-map analyses have been widely used to investigate the spatial behavior of vision Transformers, including recent studies of feature and attention-map structure (Darcet et al., 2024).

Unlike rank or entropy, these visualizations have no universal scalar "higher-is-better" interpretation. Their role is complementary: a desirable model should exhibit structured, query-dependent attention over relevant visual content while the quantitative metrics simultaneously indicate sufficient dimensionality, limited redundancy, and effective attention concentration.