Selection-as-Nonlinearity: Bridging Attention and Activation via a Joint Game–Decision Lens for Interpretable, Discriminative Visual Representations

Supplementary Document

Sudong Cai[1,2] Shuai Yuan[2] Bingzhi Chen[2] Rui Mao[3] Bing Wang[1⋆]
[1]The Hong Kong Polytechnic University [2]BIT, Zhuhai [3]Shenzhen University
{sudong.cai,bing-w.wang}@polyu.edu.hk

A. Discussion and Proofs for Section 3

In this appendix, we further elaborate on the Selection-as-Nonlinearity (SaN) interpretation developed in Section 3.2 and provide detailed proofs of the main results. All discussions and proofs here inherit the assumptions, notations, and modeling choices established in the main paper.

At a high level, SaN consists of four tightly coupled pieces: (i) a context-gated primitive that reveals how directed selection induces effective nonlinearity and rules out an x-only linear surrogate; (ii) an interpretation of self-attention as an aggregation of such context-gated units, which leads to expressivity parity with group-gated FFNs at matched granularity; (iii) a budgeted cooperative game view that makes the row-wise simplex and shared values explicit; and (iv) a geometric analysis of the resulting row–column budgeting dilemma at first order, which clarifies why attention tends to exhibit weak-independence when used in isolation.

A.1. Context-gated selection as a primitive of nonlinearity

In this subsection, we clarify the rationale behind using the context-gated primitive $\Phi(\boldsymbol{x}, \boldsymbol{c}) = \rho(h(\boldsymbol{c}))\,\boldsymbol{x}$ as a starting point of SaN, and prove that it induces a genuine joint nonlinearity and admits no x-only linear surrogate once the gate depends non-trivially on the context. In the main text (main Equation (2)), we write this primitive more compactly as $\Phi(\boldsymbol{x}, \boldsymbol{c}) = \rho(\boldsymbol{c})\,\boldsymbol{x}$; throughout this appendix, we make the score map $h : \mathcal{C} \to \mathbb{R}$ explicit and regard $\rho(\boldsymbol{c})$ as a shorthand for $\rho(h(\boldsymbol{c}))$.

Motivation. Grounded in the decision-making perspective, we revisit the physical meaning of neural activation through the lens of *selection*. At a learning layer, we consider a feature vector $\boldsymbol{x} \in \mathbb{V} \subseteq \mathbb{R}^d$ and a context $\boldsymbol{c} \in \mathcal{C}$; the primitive

$$\Phi : \mathbb{V} \times \mathcal{C} \to \mathbb{V}, \qquad \Phi(\boldsymbol{x}, \boldsymbol{c}) = \rho(h(\boldsymbol{c}))\,\boldsymbol{x} \tag{1}$$

can be seen as a soft selection mechanism: the scalar gating weight $\rho(h(\boldsymbol{c})) \geqslant 0$ reweights how much of $\boldsymbol{x}$ is retained under context $\boldsymbol{c}$. Here $h : \mathcal{C} \to \mathbb{R}$ is a score function and $\rho : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ is a scalar gate applied to this score. When the score is input-driven (*e.g.*, $h(\boldsymbol{c}) = g(\boldsymbol{x})$), $\rho(g(\boldsymbol{x}))\,\boldsymbol{x}$ recovers a self-gated activation (a special case studied in several prior works from a decision-making perspective [1–3]), and the selection view reduces to the classical universal-approximation perspective on non-polynomial activations.

To link this selection-based intuition to nonlinearity, we adopt the isotonic bias in Definition 3.1 in the main text: higher scores $s = h(\boldsymbol{c})$ should not receive smaller retention. However, for the main result below we only require that the gate is *non-constant* across contexts for some nonzero $\boldsymbol{x}$.

Retrospect. For ease of reference, we restate Proposition 3.1 from the main text as follows.

Proposition A.1 (Context-induced joint nonlinearity and no x-only linear surrogate: Restatement of Proposition 3.1)**.** *Consider $\Phi : \mathbb{V} \times \mathcal{C} \to \mathbb{V}$ with $\Phi(\boldsymbol{x}, \boldsymbol{c}) = \rho(h(\boldsymbol{c}))\,\boldsymbol{x}$, where $\rho : \mathbb{R} \to \mathbb{R}$ and $h : \mathcal{C} \to \mathbb{R}$. Suppose there exist $\boldsymbol{x} \neq \mathbf{0}$ and $\boldsymbol{c}_1 \neq \boldsymbol{c}_2$ such that $\rho(h(\boldsymbol{c}_1)) \neq \rho(h(\boldsymbol{c}_2))$. Then:*

(i) Φ is nonlinear in the joint variable $(\boldsymbol{x}, \boldsymbol{c})$;

(ii) there is no linear $T : \mathbb{V} \to \mathbb{V}$ such that $T(\boldsymbol{x}) = \Phi(\boldsymbol{x}, \boldsymbol{c})$ holds simultaneously for both $\boldsymbol{c} = \boldsymbol{c}_1$ and $\boldsymbol{c} = \boldsymbol{c}_2$.

Proof. ***core insight.*** Any linear map on the product space $\mathbb{V} \times \mathcal{C}$ can be decomposed as

$$L(\boldsymbol{x}, \boldsymbol{c}) = A\boldsymbol{x} + B\boldsymbol{c} \tag{2}$$

for some linear operators $A : \mathbb{V} \to \mathbb{V}$ and $B : \mathcal{C} \to \mathbb{V}$. Matching $\Phi(\boldsymbol{x}, \boldsymbol{c}) = \rho(h(\boldsymbol{c}))\boldsymbol{x}$ at a few carefully chosen points forces $B = \mathbf{0}$, and then the remaining operator A is forced to map the *same* nonzero $\boldsymbol{x}$ to two distinct multiples when the context changes. This yields a contradiction whenever $\rho(h(\boldsymbol{c}_1)) \neq \rho(h(\boldsymbol{c}_2))$, and rules out both a globally linear surrogate in $(\boldsymbol{x}, \boldsymbol{c})$ and any x-only linear surrogate $T(\boldsymbol{x})$.

(ii) No x-only linear surrogate. We first prove the second claim, which is simpler and will be reused.

Assume for contradiction that there exists a linear $T : \mathbb{V} \to \mathbb{V}$ such that

$$T(\boldsymbol{x}) = \Phi(\boldsymbol{x}, \boldsymbol{c}_1) = \rho(h(\boldsymbol{c}_1))\, \boldsymbol{x} \quad \text{and} \quad T(\boldsymbol{x}) = \Phi(\boldsymbol{x}, \boldsymbol{c}_2) = \rho(h(\boldsymbol{c}_2))\, \boldsymbol{x} \tag{3}$$

for some fixed $\boldsymbol{x} \neq \mathbf{0}$ and distinct $\boldsymbol{c}_1, \boldsymbol{c}_2$. Then

$$\rho(h(\boldsymbol{c}_1))\, \boldsymbol{x} = \rho(h(\boldsymbol{c}_2))\, \boldsymbol{x} \quad \Rightarrow \quad \big(\rho(h(\boldsymbol{c}_1)) - \rho(h(\boldsymbol{c}_2))\big)\, \boldsymbol{x} = \mathbf{0}\,. \tag{4}$$

Since $\boldsymbol{x} \neq \mathbf{0}$ and by assumption $\rho(h(\boldsymbol{c}_1)) \neq \rho(h(\boldsymbol{c}_2))$, this is impossible. Thus no such T exists.

(i) Joint nonlinearity. Now assume, again for contradiction, that there exists a linear map $L : \mathbb{V} \times \mathcal{C} \to \mathbb{V}$ such that

$$L(\boldsymbol{x}, \boldsymbol{c}) = \Phi(\boldsymbol{x}, \boldsymbol{c}) = \rho(h(\boldsymbol{c}))\, \boldsymbol{x} \quad \text{for all } (\boldsymbol{x}, \boldsymbol{c}) \in \mathbb{V} \times \mathcal{C}\,. \tag{5}$$

Because $\mathbb{V} \times \mathcal{C}$ is a direct sum of vector spaces, any linear map L admits a unique decomposition

$$L(\boldsymbol{x}, \boldsymbol{c}) = A\boldsymbol{x} + B\boldsymbol{c} \tag{6}$$

for some linear maps $A : \mathbb{V} \to \mathbb{V}$ and $B : \mathcal{C} \to \mathbb{V}$.

Step 1: B must vanish. Set $\boldsymbol{x} = \mathbf{0}$ in equation 5. On the one hand,

$$L(\mathbf{0}, \boldsymbol{c}) = A\mathbf{0} + B\boldsymbol{c} = B\boldsymbol{c}\,. \tag{7}$$

On the other hand,

$$\Phi(\mathbf{0}, \boldsymbol{c}) = \rho(h(\boldsymbol{c}))\, \mathbf{0} = \mathbf{0}\,. \tag{8}$$

Thus $B\boldsymbol{c} = \mathbf{0}$ for all $\boldsymbol{c} \in \mathcal{C}$, which implies

$$B = \mathbf{0}\,. \tag{9}$$

Substituting equation 9 into equation 6, we obtain

$$L(\boldsymbol{x}, \boldsymbol{c}) = A\boldsymbol{x} \quad \text{for all } (\boldsymbol{x}, \boldsymbol{c})\,. \tag{10}$$

In particular, L cannot depend on $\boldsymbol{c}$ at all.

Step 2: Contradiction from two contexts. Fix the nonzero vector $\boldsymbol{x} \neq \mathbf{0}$ and two contexts $\boldsymbol{c}_1 \neq \boldsymbol{c}_2$ given in the proposition, with $\rho(h(\boldsymbol{c}_1)) \neq \rho(h(\boldsymbol{c}_2))$. Applying equation 5 and equation 10 at $(\boldsymbol{x}, \boldsymbol{c}_1)$ and $(\boldsymbol{x}, \boldsymbol{c}_2)$, we have

$$A\boldsymbol{x} = L(\boldsymbol{x}, \boldsymbol{c}_1) = \Phi(\boldsymbol{x}, \boldsymbol{c}_1) = \rho(h(\boldsymbol{c}_1))\, \boldsymbol{x}, \tag{11}$$

$$A\boldsymbol{x} = L(\boldsymbol{x}, \boldsymbol{c}_2) = \Phi(\boldsymbol{x}, \boldsymbol{c}_2) = \rho(h(\boldsymbol{c}_2))\, \boldsymbol{x}. \tag{12}$$

Comparing equation 11 and equation 12 yields

$$\rho(h(\boldsymbol{c}_1))\, \boldsymbol{x} = \rho(h(\boldsymbol{c}_2))\, \boldsymbol{x} \quad \Rightarrow \quad \big(\rho(h(\boldsymbol{c}_1)) - \rho(h(\boldsymbol{c}_2))\big)\, \boldsymbol{x} = \mathbf{0}\,, \tag{13}$$

which again contradicts $\boldsymbol{x} \neq \mathbf{0}$ and $\rho(h(\boldsymbol{c}_1)) \neq \rho(h(\boldsymbol{c}_2))$.

Thus the assumption that such a linear L exists leads to a contradiction, and no global linear map in $(\boldsymbol{x}, \boldsymbol{c})$ can represent Φ. Equivalently, Φ is nonlinear in the joint variable $(\boldsymbol{x}, \boldsymbol{c})$.

Combining the above, we conclude that both claims (i) and (ii) hold.

This completes the proof of Proposition A.1. □

Remark A.1. *Proposition A.1 formalizes the intuition that once a gate $\rho(h(\boldsymbol{c}))$ truly depends on the context, the resulting activation cannot be absorbed into an x-only linear transform. The nonlinearity arises from how the same $\boldsymbol{x}$ is* selectively reweighted *under different contexts, and this joint dependence on $(\boldsymbol{x}, \boldsymbol{c})$ constitutes the basic mechanism of SaN.*

A.2. Attention as aggregation of context-gated units and expressivity parity

We next clarify how self-attention fits the selection primitive, and why, at matched granularity and receptive field, a self-attention block is at least as expressive as a group-gated FFN.

Motivation. Mechanistically, each (i, j)-term in self-attention can be viewed as a context-gated unit: for query i and source j, the compatibility score is $c_{ij} = \boldsymbol{q}_i^\top \boldsymbol{k}_j$, and a row-wise normalizer R (*e.g.*, softmax) produces

$$\alpha_{ij} = R(c_{ij}; \{c_{ik}\}_k)\,, \tag{14}$$

which can be read as a context-dependent retention weight for value $\boldsymbol{v}_j$. Thus

$$\Phi_{i,j}(\boldsymbol{X}) = \alpha_{ij}\, \boldsymbol{v}_j \tag{15}$$

matches the primitive $\Phi(\boldsymbol{x}, \boldsymbol{c})$ once we identify the context with the local score c_{ij} and its competitors $\{c_{ik}\}_{k\neq j}$. Aggregating such units over j and heads yields the full attention output.

Retrospect: specialization to attention. We first restate Corollary 3.2.

Corollary A.2 (Specialization to attention: Restatement of Corollary 3.2)**.** *For a fixed row i and column j, hold $\{c_{ik}\}_{k\neq j}$ fixed and let*

$$\rho(s) = \frac{e^s}{\sum_k e^{c_{ik}}} \tag{16}$$

be the row-softmax gate. Then $s \mapsto \rho(s)$ is strictly increasing and non-constant in s. Consequently, if two inputs share $\boldsymbol{v}_j \neq \mathbf{0}$ and the same $\{c_{ik}\}_{k\neq j}$ but satisfy $c_{ij}^{(1)} \neq c_{ij}^{(2)}$, Proposition A.1 applies to the unit $\Phi_{i,j}$.

Proof. ***core insight.*** Row-softmax with a fixed set of competitor logits reduces to a one-dimensional logistic transform of s. The derivative of this logistic map is strictly positive everywhere, so it is strictly increasing and non-constant. Thus, changing c_{ij} while keeping the other logits fixed strictly changes the gate on a shared nonzero value vector $\boldsymbol{v}_j$, which is exactly the setup required to invoke Proposition A.1.

Detailed argument. Fix a row i and a column j, and define

$$Z_i^{(-j)} := \sum_{k\neq j} e^{c_{ik}} > 0\,, \tag{17}$$

which does not depend on $s = c_{ij}$. Then

$$\rho(s) = \frac{e^s}{e^s + Z_i^{(-j)}}\,. \tag{18}$$

Differentiating equation 18 with respect to s gives

$$\rho'(s) = \frac{e^s(e^s + Z_i^{(-j)}) - e^s(e^s)}{(e^s + Z_i^{(-j)})^2} = \frac{e^s Z_i^{(-j)}}{(e^s + Z_i^{(-j)})^2}\,.$$

Using equation 18, we can rewrite this as

$$\rho'(s) = \rho(s)\big(1 - \rho(s)\big)\,. \tag{19}$$

Since $Z_i^{(-j)} > 0$, we have $0 < \rho(s) < 1$ for all $s \in \mathbb{R}$, and therefore

$$\rho'(s) = \rho(s)\big(1 - \rho(s)\big) > 0 \quad \text{for all } s \in \mathbb{R}\,. \tag{20}$$

Thus $s \mapsto \rho(s)$ is strictly increasing and, in particular, non-constant: if $s_1 \neq s_2$, then $\rho(s_1) \neq \rho(s_2)$.

Now consider two inputs that share the same $\boldsymbol{v}_j \neq \mathbf{0}$ and the same $\{c_{ik}\}_{k\neq j}$ but satisfy $c_{ij}^{(1)} \neq c_{ij}^{(2)}$. By strict monotonicity, we obtain

$$\rho(c_{ij}^{(1)}) \neq \rho(c_{ij}^{(2)})\,. \tag{21}$$

Identifying $\boldsymbol{x}$ with $\boldsymbol{v}_j$ and contexts $\boldsymbol{c}_1, \boldsymbol{c}_2$ with score configurations that differ only in c_{ij}, the conditions of Proposition A.1 are met: $\boldsymbol{x} \neq \mathbf{0}$ and the gate takes different values under $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. Therefore, Proposition A.1 applies to the unit $\Phi_{i,j}(\boldsymbol{X}) = \rho(c_{ij})\, \boldsymbol{v}_j$.

This completes the proof. □

Canonical gated affine expert. To make the expressivity comparison precise, we next introduce a simple canonical template that captures both group-gated FFN experts and isolated attention gate units under receptive-field equivalence (RFE), matching the "gated affine expert" phrasing in the main text.

Definition A.1 (Gated affine expert at a given granularity)**.** *Let $\boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{loc}}}$ be the local feature at a given granularity (e.g., a group of channels of a token). We call any map $E : \mathbb{R}^{d_{\mathrm{loc}}} \to \mathbb{R}^{d_{\mathrm{out}}}$ of the form*

$$E(\boldsymbol{x}) = W_{\mathrm{out}}\big(\gamma(\boldsymbol{x}) \cdot (W_{\mathrm{in}}\boldsymbol{x} + \boldsymbol{b}_{\mathrm{in}})\big) + \boldsymbol{b}_{\mathrm{out}} \tag{22}$$

a gated affine expert*, where:*

(i) $W_{\mathrm{in}} \in \mathbb{R}^{d_{\mathrm{mid}} \times d_{\mathrm{loc}}}$, $\boldsymbol{b}_{\mathrm{in}} \in \mathbb{R}^{d_{\mathrm{mid}}}$, $W_{\mathrm{out}} \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{mid}}}$, $\boldsymbol{b}_{\mathrm{out}} \in \mathbb{R}^{d_{\mathrm{out}}}$ *are learnable linear/affine parameters;*

(ii) $\gamma : \mathbb{R}^{d_{\mathrm{loc}}} \to \mathbb{R}$ *is a scalar gating function (not identically constant), and the scalar $\gamma(\boldsymbol{x})$ is broadcast and multiplied element-wise with $(W_{\mathrm{in}}\boldsymbol{x} + \boldsymbol{b}_{\mathrm{in}})$.*

This template matches the informal description used in the main text: a local feature is first mapped to some intermediate representation, then modulated by a scalar gate derived from the same local context, and finally passed through a linear readout.

Retrospect: parity statements. We now recall the expressivity statements from the main text.

Corollary A.3 (Unit-level expressivity parity: Restatement of Corollary 3.3)**.** *Under RFE and separate pre-/post-projections, an isolated attention gate unit composed with the block's projections is as expressive as a single group-gated FFN expert at the same granularity.*

Theorem A.4 (Expressivity parity at matched granularity (compact): Restatement of Theorem 3.4)**.** *Under RFE, matched per-token width d, and matched granularity H (heads versus group gates), a self-attention block with H heads and separate pre-/post-projections is at least as expressive as an H-group gated FFN on compact domains.*

Lemma A.5 (Both experts fit the gated affine template)**.** *Under RFE and separate pre-/post-projections, a single group-gated FFN expert and an isolated attention gate unit at the same granularity both realize maps of the form equation 22.*

Proof. ***FFN side.*** Under RFE, fix a particular group (or channel group) and let $\boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{loc}}}$ denote its local input feature. By construction in the main text, a single group-gated FFN expert at this granularity can be written as

$$E_{\mathrm{FFN}}(\boldsymbol{x}) = W_2\big(\gamma_{\mathrm{FFN}}(\boldsymbol{x}) \cdot \phi(\boldsymbol{x})\big) + \boldsymbol{b}_2\,, \tag{23}$$

where:

- $\phi(\boldsymbol{x})$ is the intermediate representation produced by the local FFN sub-block (e.g., a linear projection plus bias restricted to this group);
- $\gamma_{\mathrm{FFN}}(\boldsymbol{x})$ is a scalar gate produced by the FFN's gating sub-network for this group (e.g., a self-gated activation or group-wise gate);
- W_2 and $\boldsymbol{b}_2$ are the group's output projection and bias.

Writing $\phi(\boldsymbol{x}) = \tilde{W}_{\mathrm{in}}\boldsymbol{x} + \tilde{\boldsymbol{b}}_{\mathrm{in}}$ for some $\tilde{W}_{\mathrm{in}}, \tilde{\boldsymbol{b}}_{\mathrm{in}}$, and comparing with equation 22, we can identify

$$W_{\mathrm{in}} = \tilde{W}_{\mathrm{in}}, \quad \boldsymbol{b}_{\mathrm{in}} = \tilde{\boldsymbol{b}}_{\mathrm{in}}, \quad W_{\mathrm{out}} = W_2, \quad \boldsymbol{b}_{\mathrm{out}} = \boldsymbol{b}_2, \quad \gamma(\boldsymbol{x}) = \gamma_{\mathrm{FFN}}(\boldsymbol{x})\,. \tag{24}$$

Thus E_{FFN} is exactly a gated affine expert in the sense of Definition A.1.

Attention side. Under the same RFE, consider an isolated attention gate unit at the same granularity: for a fixed query index i and source index j, and for the local feature $\boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{loc}}}$ available after the block's pre-projection, we can write

$$c_{ij}(\boldsymbol{x}) = h_{\mathrm{att}}(\boldsymbol{x})\,, \tag{25}$$

$$\gamma_{\mathrm{att}}(\boldsymbol{x}) = \rho\big(c_{ij}(\boldsymbol{x})\big)\,, \tag{26}$$

$$\boldsymbol{v}_j(\boldsymbol{x}) = W_V\boldsymbol{x} + \boldsymbol{b}_V\,, \tag{27}$$

$$E_{\mathrm{att}}(\boldsymbol{x}) = W_O\big(\gamma_{\mathrm{att}}(\boldsymbol{x}) \cdot \boldsymbol{v}_j(\boldsymbol{x})\big) + \boldsymbol{b}_O\,, \tag{28}$$

where:

- h_{att} denotes the scalar compatibility function (e.g., the inner product between a query derived from $\boldsymbol{x}$ and a key derived from some source representation inside the same receptive field);

- $\rho(\cdot)$ is the row-normalizer (softmax or entmax) gate evaluated at c_{ij} as in Corollary A.2;
- $W_V, \boldsymbol{b}_V$ are the value projection and bias for this head;
- $W_O, \boldsymbol{b}_O$ are the head's output projection and bias.

By Corollary A.2, the scalar gate $\gamma_{\text{att}}(\boldsymbol{x}) = \rho(c_{ij}(\boldsymbol{x}))$ is a non-constant function of the same local context $\boldsymbol{x}$. Comparing with equation 22, we may set

$$W_{\text{in}} = W_V, \quad \boldsymbol{b}_{\text{in}} = \boldsymbol{b}_V, \quad W_{\text{out}} = W_O, \quad \boldsymbol{b}_{\text{out}} = \boldsymbol{b}_O, \quad \gamma(\boldsymbol{x}) = \gamma_{\text{att}}(\boldsymbol{x}), \tag{29}$$

and see that E_{att} is also a gated affine expert.

This completes the proof. □

We are now ready to prove the unit- and block-level parity results that correspond to Corollary 3.3 and Theorem 3.4.

Proof of Corollary A.3. ***core insight.*** Under RFE and separate pre-/post-projections, both a group-gated FFN expert and an isolated attention gate unit at a fixed granularity implement exactly the same canonical object: a gated affine expert in the sense of Definition A.1. As a consequence, the function classes realized by these two modules at this granularity coincide.

By Lemma A.5, both the group-gated FFN expert E_{FFN} and the isolated attention gate expert E_{att} admit representations of the form

$$E(\boldsymbol{x}) = W_{\text{out}}\big(\gamma(\boldsymbol{x}) \cdot (W_{\text{in}}\boldsymbol{x} + \boldsymbol{b}_{\text{in}})\big) + \boldsymbol{b}_{\text{out}}. \tag{30}$$

Under the modeling assumptions of SaN, we view $\gamma(\boldsymbol{x})$ as an arbitrary (non-constant) scalar gating function of the local context that is implementable by the corresponding gating sub-network (e.g., a small MLP or a qk-based gate satisfying Definition 3.1). Crucially, the canonical form equation 22 does not depend on the specific parameterization of γ; only the existence of such a scalar gate matters.

Given any group-gated FFN expert E_{FFN} with parameters $(W_{\text{in}}^{\text{FFN}}, \boldsymbol{b}_{\text{in}}^{\text{FFN}}, W_{\text{out}}^{\text{FFN}}, \boldsymbol{b}_{\text{out}}^{\text{FFN}}, \gamma_{\text{FFN}})$ in the template equation 22, we can construct an attention-gate expert E_{att} that realizes the *same* map by choosing

$$W_{\text{in}}^{\text{att}} = W_{\text{in}}^{\text{FFN}}, \quad \boldsymbol{b}_{\text{in}}^{\text{att}} = \boldsymbol{b}_{\text{in}}^{\text{FFN}}, \quad W_{\text{out}}^{\text{att}} = W_{\text{out}}^{\text{FFN}}, \quad \boldsymbol{b}_{\text{out}}^{\text{att}} = \boldsymbol{b}_{\text{out}}^{\text{FFN}}, \tag{31}$$

and configuring the attention gating sub-network (through its score function h_{att} and normalizer ρ) so that

$$\gamma_{\text{att}}(\boldsymbol{x}) = \gamma_{\text{FFN}}(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \tag{32}$$

within the local receptive-field domain under consideration.[1]

Under this identification, the resulting experts coincide pointwise:

$$E_{\text{att}}(\boldsymbol{x}) = E_{\text{FFN}}(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x}. \tag{33}$$

Symmetrically, any attention-gate expert can be matched by a group-gated FFN expert by the same argument. Thus, at the fixed granularity and under RFE, the two modules are equally expressive.

This completes the proof of Corollary A.3. □

Proof of Theorem A.4. ***core insight.*** At a fixed granularity (one head or one group gate), Corollary A.3 has already shown that attention and group-gated FFN experts implement the same class of gated affine maps. At the block level, both architectures simply aggregate H such experts that share the same receptive field and per-token width. Therefore, by matching experts head-wise/group-wise, any H-group FFN block can be realized by some H-head attention block (and vice versa). Since this equality holds globally, it also holds on any compact domain.

Step 1: Block-level form of an H-group gated FFN. Under RFE and matched per-token width d, fix a token (or local receptive field) and denote its local feature by $\boldsymbol{x} \in \mathbb{R}^d$. An H-group gated FFN block can be written as a sum of H group-wise experts:

$$F_{\text{FFN}}(\boldsymbol{x}) = \sum_{h=1}^{H} E_h^{\text{FFN}}(\boldsymbol{x}), \tag{34}$$

[1] Formally, we only require that the gating sub-networks on both sides are sufficiently expressive to realize the same family of scalar functions on the compact domains of interest, which is standard under non-polynomial activations.

where each $E_h^{\mathrm{FFN}} : \mathbb{R}^d \to \mathbb{R}^d$ operates on the same local feature $\boldsymbol{x}$ (RFE) but with its own parameters (gating sub-network, intermediate linear map, and output linear map) at granularity h. By Lemma A.5, each E_h^{FFN} is a gated affine expert.

Step 2: Block-level form of an H-head attention block. Under the same RFE and per-token width d, consider a self-attention block with H heads and separate pre-/post-projections. For each head $h \in \{1, \dots, H\}$ and the same local feature $\boldsymbol{x}$, we can write its contribution at this granularity as

$$E_h^{\mathrm{attn}}(\boldsymbol{x}) = W_O^{(h)} \Big(\gamma_h^{\mathrm{attn}}(\boldsymbol{x}) \cdot \big(W_V^{(h)} \boldsymbol{x} + \boldsymbol{b}_V^{(h)} \big) \Big) + \boldsymbol{b}_O^{(h)} , \tag{35}$$

where all symbols are as in Lemma A.5. Again, Lemma A.5 implies that each E_h^{attn} is a gated affine expert. The full attention block output at this granularity is obtained by summing the H head contributions (any concatenation and shared output projection can be absorbed into $W_O^{(h)}$ and $\boldsymbol{b}_O^{(h)}$):

$$F_{\mathrm{attn}}(\boldsymbol{x}) = \sum_{h=1}^{H} E_h^{\mathrm{attn}}(\boldsymbol{x}) . \tag{36}$$

Step 3: Matching group experts by head experts. Fix an H-group gated FFN block F_{FFN} of the form equation 34, and consider any compact set $K \subset \mathbb{R}^d$. By Corollary A.3, for each $h \in \{1, \dots, H\}$ there exists a choice of attention-head parameters such that

$$E_h^{\mathrm{attn}}(\boldsymbol{x}) = E_h^{\mathrm{FFN}}(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in K . \tag{37}$$

Substituting into equation 34 and equation 36, we obtain

$$F_{\mathrm{attn}}(\boldsymbol{x}) = \sum_{h=1}^{H} E_h^{\mathrm{attn}}(\boldsymbol{x}) = \sum_{h=1}^{H} E_h^{\mathrm{FFN}}(\boldsymbol{x}) = F_{\mathrm{FFN}}(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in K . \tag{38}$$

Thus, for any H-group gated FFN block and any compact domain K, there exists an H-head attention block (with separate pre-/post-projections and RFE) that realizes the *same* mapping on K. In particular, the function class represented by such attention blocks contains the function class of H-group gated FFN blocks under the stated conditions.

This proves that an H-head attention block is at least as expressive as an H-group gated FFN at matched granularity on compact domains, completing the proof of Theorem A.4. □

Remark A.2. *Corollary A.3 and Theorem A.4 formalize that, once we match receptive field, per-token width, and granularity H, an attention block is not inherently weaker than a group-gated FFN in terms of expressivity. Therefore, the empirically observed weak-independence of attention-only stacks Section 3.1) should be attributed to how this expressive mechanism is organized and optimized under shared budgets, rather than to a fundamental lack of representational power.*

A.3. Attention as a budgeted cooperative allocation game

We now discuss the game–decision interpretation of attention, and provide explicit gradient formulas that reveal how gates and values are coupled at first order.

Motivation. Row-softmax normalizes the scores of each query into a probability simplex, so that every query i distributes a unit-mass budget over the shared value bank $\{\boldsymbol{v}_j\}$. When the loss decomposes over queries (*e.g.*, standard token-wise or patch-wise objectives), this yields a natural view of attention as a cooperative allocation game: each row chooses a budget vector to minimize its local loss, while the shared parameters $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ couple these choices. The training loss then plays the role of a potential.

Retrospect. We restate the cooperative game proposition from the main text.

Proposition A.6 (Budgeted cooperative allocation view: Restatement of Proposition 3.5)**.** *With a row-wise normalizer R (*e.g.*, softmax or entmax) that yields nonnegative weights summing to one, each row i allocates a unit-mass budget $\boldsymbol{\alpha}_i \in \Delta^{N-1}$ over shared values $\{\boldsymbol{v}_j\}$, with $\alpha_{ij} = R(c_{ij}; \{c_{ik}\}_k)$ monotone in its own logit c_{ij}. Under an additive loss that decomposes over queries, joint updates of $(\{\boldsymbol{\alpha}_i\}_i, \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ form an exact potential game, where the training loss is the potential.*

Proof. ***core insight.*** Row-softmax (or entmax) enforces the budget constraints $\alpha_{ij} \geqslant 0$ and $\sum_j \alpha_{ij} = 1$ for every row, so each query i chooses a budget vector $\boldsymbol{\alpha}_i$ on the simplex. If the loss decomposes as a sum of row-wise terms, then the gradient of the global loss with respect to each $\boldsymbol{\alpha}_i$ coincides with the gradient of the corresponding row-wise term, which is the defining property of an exact potential game.

Let $c_{ij} = \boldsymbol{q}_i^\top \boldsymbol{k}_j$ and define

$$\alpha_{ij} = R(c_{ij}; \{c_{ik}\}_k) \quad \text{with} \quad \alpha_{ij} \geqslant 0,\ \sum_j \alpha_{ij} = 1\,. \tag{39}$$

Thus $\boldsymbol{\alpha}_i \in \Delta^{N-1}$ and the output is $\boldsymbol{y}_i = \sum_j \alpha_{ij} \boldsymbol{v}_j$. Assume the overall loss can be written as $\ell = \sum_i \ell_i(\boldsymbol{y}_i)$, *e.g.*, cross-entropy aggregated over positions. Then the gradient of ℓ with respect to $\boldsymbol{\alpha}_i$ is

$$\nabla_{\boldsymbol{\alpha}_i} \ell = \nabla_{\boldsymbol{\alpha}_i} \ell_i(\boldsymbol{y}_i)\,, \tag{40}$$

since other rows do not depend on $\boldsymbol{\alpha}_i$. If we interpret each row i as a player whose strategy is $\boldsymbol{\alpha}_i$ and whose payoff is $-\ell_i$, then the global potential $-\ell$ satisfies

$$\nabla_{\boldsymbol{\alpha}_i}(-\ell) = -\nabla_{\boldsymbol{\alpha}_i} \ell = -\nabla_{\boldsymbol{\alpha}_i} \ell_i\,, \tag{41}$$

which is exactly the definition of an (exact) potential game. This proves the proposition. □

We next formalize the first-order coupling between gates and values, as announced in Lemma 3.6 in the main text.

Lemma A.7 (Row-softmax gradients: coupling of qk and values: Restatement of Lemma 3.6)**.** *Let* $s_{ij} = \boldsymbol{q}_i^\top \boldsymbol{k}_j$, $\alpha_{ij} = \mathrm{softmax}_j(s_{ij})$, $\boldsymbol{y}_i = \sum_j \alpha_{ij} \boldsymbol{v}_j$, *and* $\boldsymbol{g}_i = \partial \ell / \partial \boldsymbol{y}_i$. *Then*

$$\frac{\partial \ell}{\partial \boldsymbol{v}_j} = \sum_i \alpha_{ij}\, \boldsymbol{g}_i\,, \tag{42}$$

$$\frac{\partial \ell}{\partial s_{ij}} = \alpha_{ij}\, \boldsymbol{g}_i^\top (\boldsymbol{v}_j - \boldsymbol{y}_i)\,, \tag{43}$$

$$\frac{\partial \ell}{\partial \boldsymbol{k}_j} = \sum_i \alpha_{ij} \big(\boldsymbol{g}_i^\top (\boldsymbol{v}_j - \boldsymbol{y}_i)\big) \boldsymbol{q}_i\,, \tag{44}$$

$$\frac{\partial \ell}{\partial \boldsymbol{q}_i} = \sum_j \alpha_{ij} \big(\boldsymbol{g}_i^\top (\boldsymbol{v}_j - \boldsymbol{y}_i)\big) \boldsymbol{k}_j\,. \tag{45}$$

Proof. ***core insight.*** The gradient with respect to each value $\boldsymbol{v}_j$ aggregates row-wise demands weighted by α_{ij}; the gradient with respect to each logit s_{ij} is proportional to how much row i prefers $\boldsymbol{v}_j$ over its current mixture $\boldsymbol{y}_i$. Chaining these derivatives through $s_{ij} = \boldsymbol{q}_i^\top \boldsymbol{k}_j$ yields the coupling between $\boldsymbol{q}_i, \boldsymbol{k}_j$ and the values.

Derivatives w.r.t. values. By definition, $\boldsymbol{y}_i = \sum_k \alpha_{ik} \boldsymbol{v}_k$ and $\ell = \sum_i \ell_i(\boldsymbol{y}_i)$. Thus, for a fixed j,

$$\frac{\partial \ell}{\partial \boldsymbol{v}_j} = \sum_i \frac{\partial \ell_i}{\partial\, \boldsymbol{y}_i} \frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{v}_j} = \sum_i \boldsymbol{g}_i\, \alpha_{ij}\,, \tag{46}$$

which is equation 42.

Derivatives w.r.t. logits. For a fixed (i, j), the Jacobian of row-softmax satisfies

$$\frac{\partial \alpha_{ik}}{\partial s_{ij}} = \alpha_{ik}(\delta_{kj} - \alpha_{ij})\,, \tag{47}$$

where δ_{kj} is the Kronecker delta. Hence

$$\begin{aligned} \frac{\partial \boldsymbol{y}_i}{\partial s_{ij}} &= \sum_k \frac{\partial \alpha_{ik}}{\partial s_{ij}} \boldsymbol{v}_k = \sum_k \alpha_{ik}(\delta_{kj} - \alpha_{ij}) \boldsymbol{v}_k \\ &= \alpha_{ij} \boldsymbol{v}_j - \alpha_{ij} \sum_k \alpha_{ik} \boldsymbol{v}_k = \alpha_{ij}(\boldsymbol{v}_j - \boldsymbol{y}_i). \end{aligned}$$

Applying the chain rule, we obtain

$$\frac{\partial \ell}{\partial s_{ij}} = \boldsymbol{g}_i^\top \frac{\partial \boldsymbol{y}_i}{\partial s_{ij}} = \alpha_{ij}\, \boldsymbol{g}_i^\top (\boldsymbol{v}_j - \boldsymbol{y}_i)\,, \tag{48}$$

which gives equation 43.

Derivatives w.r.t. keys and queries. Finally, $s_{ij} = \boldsymbol{q}_i^\top \boldsymbol{k}_j$ implies

$$\frac{\partial s_{ij}}{\partial \boldsymbol{k}_j} = \boldsymbol{q}_i\,, \qquad \frac{\partial s_{ij}}{\partial \boldsymbol{q}_i} = \boldsymbol{k}_j\,. \tag{49}$$

Using equation 43 and summing over i (or j) yields

$$\frac{\partial \ell}{\partial \boldsymbol{k}_j} = \sum_i \frac{\partial \ell}{\partial s_{ij}} \frac{\partial s_{ij}}{\partial \boldsymbol{k}_j} = \sum_i \alpha_{ij} \big(\boldsymbol{g}_i^\top (\boldsymbol{v}_j - \boldsymbol{y}_i)\big) \boldsymbol{q}_i\,, \tag{50}$$

and similarly for $\partial \ell / \partial \boldsymbol{q}_i$, which proves Equations (44) and (45).

This completes the proof. □

Remark A.3. *Lemma A.7 shows that increasing a logit s_{ij} helps row i exactly when the local descent direction $\boldsymbol{g}_i$ prefers $\boldsymbol{v}_j$ over the current mixture $\boldsymbol{y}_i$, while each value $\boldsymbol{v}_j$ aggregates multi-row demands $\{\alpha_{ij}\boldsymbol{g}_i\}_i$. This gradient-level coupling underlies the row–column budgeting dilemma analyzed next.*

A.4. Row–column budgeting dilemma: KKT view and geometric incompatibility

We finally discuss how the row-wise simplex and shared values jointly induce a structural tension—the row–column budgeting dilemma—that helps explain the weak-independence behavior of attention-only stacks.

Motivation. Horizontally (row side), each query i would like to concentrate its unit-mass budget on a small number of values that align best with its descent direction. Vertically (column side), each value $\boldsymbol{v}_j$ must simultaneously support many rows. We formalize the row-side preference via a simple KKT analysis of the best response on the simplex, and the column-side coupling via the cones spanned by query and gradient vectors. The key question is: *when does there exist a single update direction for $\boldsymbol{k}_j$ or $\boldsymbol{v}_j$ that strictly decreases the loss for all involved rows at first order?*

Retrospect. We restate the relevant results from the main text.

Proposition A.8 (Row-side resource competition: a KKT view: Restatement of Proposition 3.7)**.** *Fix a row i and a target direction $\boldsymbol{t}_i$ (*e.g.*, $\boldsymbol{t}_i = -\boldsymbol{g}_i$). Consider the optimization problem*

$$\max_{\alpha_{ij}\geqslant 0,\ \sum_j \alpha_{ij}=1} \Big\langle \sum_j \alpha_{ij} \boldsymbol{v}_j,\ \boldsymbol{t}_i \Big\rangle\,. \tag{51}$$

At any optimum, there exists a scalar λ_i such that

$$\boldsymbol{v}_j^\top \boldsymbol{t}_i - \lambda_i \begin{cases} = 0, & \alpha_{ij} > 0, \\ \leqslant 0, & \alpha_{ij} = 0, \end{cases} \tag{52}$$

so sparse optima cut off many columns from row-i credit.

Proof. ***core insight.*** Problem equation 51 is a finite-dimensional linear program over the probability simplex. The KKT conditions indicate that only values with maximal alignment $\boldsymbol{v}_j^\top \boldsymbol{t}_i$ receive positive mass at optimum; all others are driven to zero. This formalizes the tendency for row-wise budgets to concentrate on a few values.

The Lagrangian for equation 51 is

$$\mathcal{L}(\boldsymbol{\alpha}_i, \lambda_i, \boldsymbol{\mu}_i) = \Big\langle \sum_j \alpha_{ij} \boldsymbol{v}_j,\ \boldsymbol{t}_i \Big\rangle + \lambda_i \Big(1 - \sum_j \alpha_{ij}\Big) + \sum_j \mu_{ij} \alpha_{ij}\,, \tag{53}$$

with multipliers $\mu_{ij} \geqslant 0$. The KKT conditions are:

(a) Primal feasibility: $\alpha_{ij} \geqslant 0$, $\sum_j \alpha_{ij} = 1$;

(b) Dual feasibility: $\mu_{ij} \geqslant 0$;
(c) Complementary slackness: $\mu_{ij}\alpha_{ij} = 0$;
(d) Stationarity: $\partial\mathcal{L}/\partial\alpha_{ij} = \boldsymbol{v}_j^\top \boldsymbol{t}_i - \lambda_i + \mu_{ij} = 0$.
Rearranging the stationarity condition gives

$$\boldsymbol{v}_j^\top \boldsymbol{t}_i - \lambda_i = -\mu_{ij}\,. \tag{54}$$

If $\alpha_{ij} > 0$, complementary slackness implies $\mu_{ij} = 0$, so $\boldsymbol{v}_j^\top \boldsymbol{t}_i - \lambda_i = 0$. If $\alpha_{ij} = 0$, then $\mu_{ij} \geqslant 0$ and hence $\boldsymbol{v}_j^\top \boldsymbol{t}_i - \lambda_i = -\mu_{ij} \leqslant 0$. This is exactly equation 52, and immediately shows that values with strictly suboptimal alignment are assigned zero budget in any optimal solution. □

We now turn to the joint row–column dilemma.

Theorem A.9 (Row–column budgeting dilemma: first-order incompatibility: Restatement of Theorem 3.8)**.** *Let* $c_i(j) = \alpha_{ij}\, \boldsymbol{g}_i^\top(\boldsymbol{v}_j - \boldsymbol{y}_i)$. ***Row side (via*** $\boldsymbol{k}_j$***).*** *Partition* $I_\pm(j) = \{\, i : \pm c_i(j) > 0 \,\}$ *and let* $\mathcal{C}_\pm(j) = \mathrm{cone}\{\, \boldsymbol{q}_i : i \in I_\pm(j) \,\}$. *There exists* $\Delta\boldsymbol{k}_j$ *that makes all involved rows strictly descend at first order if and only if* $\mathcal{C}_+(j)$ *and* $\mathcal{C}_-(j)$ *are strictly separable by a hyperplane through the origin.* ***Column side (via*** $\boldsymbol{v}_j$***).*** *Let* $A_j = \{\, i : \alpha_{ij} > 0 \,\}$. *There exists* $\Delta\boldsymbol{v}_j$ *such that* $\langle \boldsymbol{g}_i, \alpha_{ij}\Delta\boldsymbol{v}_j \rangle < 0$ *for all* $i \in A_j$ *if and only if* $\{\boldsymbol{g}_i : i \in A_j\}$ *lies in a common open halfspace (equivalently,* $\mathbf{0} \notin \mathrm{cone}\{\boldsymbol{g}_i : i \in A_j\}$*).*

Proof. ***core insight.*** At first order, the effect of a small update is captured by a collection of inner products between the update direction and row-specific gradient vectors. On the row side, these vectors are $\{c_i(j)\boldsymbol{q}_i\}_i$; requiring all inner products to be negative is precisely the condition that a separating hyperplane exists between the cones generated by $\{\boldsymbol{q}_i : c_i(j) > 0\}$ and $\{\boldsymbol{q}_i : c_i(j) < 0\}$. On the column side, requiring a single $\Delta\boldsymbol{v}_j$ to simultaneously yield negative inner products with all active gradients $\{\alpha_{ij}\boldsymbol{g}_i\}_{i\in A_j}$ is equivalent to $\mathbf{0}$ not belonging to their conic hull. Both equivalences follow from standard hyperplane separation results for finite cones.

Row side. By Lemma A.7,

$$\frac{\partial \ell}{\partial \boldsymbol{k}_j} = \sum_i \alpha_{ij}\big(\boldsymbol{g}_i^\top(\boldsymbol{v}_j - \boldsymbol{y}_i)\big)\boldsymbol{q}_i = \sum_i c_i(j)\, \boldsymbol{q}_i\,. \tag{55}$$

Decompose $c_i(j)$ by sign: for $i \in I_+(j)$, $c_i(j) > 0$; for $i \in I_-(j)$, $c_i(j) < 0$. The first-order change in the loss contributed by column j under an update $\Delta\boldsymbol{k}_j$ is

$$\Delta\ell_j \approx \Big\langle \frac{\partial \ell}{\partial \boldsymbol{k}_j},\ \Delta\boldsymbol{k}_j \Big\rangle = \sum_i c_i(j)\, \boldsymbol{q}_i^\top \Delta\boldsymbol{k}_j\,. \tag{56}$$

To make all involved rows strictly descend at first order, we require

$$c_i(j)\, \boldsymbol{q}_i^\top \Delta\boldsymbol{k}_j < 0 \quad \text{for all } i \in I_+(j) \cup I_-(j)\,. \tag{57}$$

Equivalently,

$$\boldsymbol{q}_i^\top \Delta\boldsymbol{k}_j < 0 \quad (i \in I_+(j))\,, \qquad \boldsymbol{q}_i^\top \Delta\boldsymbol{k}_j > 0 \quad (i \in I_-(j))\,. \tag{58}$$

This is precisely the existence of a vector $\Delta\boldsymbol{k}_j$ that strictly separates the two conic sets $\mathcal{C}_+(j) = \mathrm{cone}\{\boldsymbol{q}_i : i \in I_+(j)\}$ and $\mathcal{C}_-(j) = \mathrm{cone}\{\boldsymbol{q}_i : i \in I_-(j)\}$ by a hyperplane through the origin. By the standard hyperplane separation theorem for closed convex cones generated by finite sets, such a vector exists if and only if the two cones are strictly separable. This proves the row-side statement.

Column side. Similarly, by equation 42,

$$\frac{\partial \ell}{\partial \boldsymbol{v}_j} = \sum_i \alpha_{ij}\, \boldsymbol{g}_i\,. \tag{59}$$

The first-order change under an update $\Delta\boldsymbol{v}_j$ is

$$\Delta\ell_j \approx \Big\langle \frac{\partial \ell}{\partial \boldsymbol{v}_j},\ \Delta\boldsymbol{v}_j \Big\rangle = \sum_{i\in A_j} \alpha_{ij}\, \boldsymbol{g}_i^\top \Delta\boldsymbol{v}_j\,, \tag{60}$$

where we may restrict to A_j since $\alpha_{ij} = 0$ otherwise. To make each active row $i \in A_j$ strictly descend at first order via the same $\Delta\boldsymbol{v}_j$, we need

$$\boldsymbol{g}_i^\top \Delta\boldsymbol{v}_j < 0 \quad \text{for all } i \in A_j\,. \tag{61}$$

This is equivalent to finding a vector $\Delta \boldsymbol{v}_j$ in the intersection of the open halfspaces $\{u : \boldsymbol{g}_i^\top u < 0\}$ for all $i \in A_j$. By the finite-dimensional hyperplane separation theorem, such a vector exists if and only if $\mathbf{0}$ is not contained in the conic hull $\mathrm{cone}\{\boldsymbol{g}_i : i \in A_j\}$. Equivalently, the set $\{\boldsymbol{g}_i : i \in A_j\}$ lies in a common open halfspace. This proves the column-side statement.

This completes the proof. □

We finally restate the corollaries that highlight simple sufficient conditions.

Corollary A.10 (Simple sufficient conditions for incompatibility: Restatement of Corollary 3.9)**.** *If $c_i(j)c_{i'}(j) < 0$ and $\boldsymbol{q}_{i'} = \lambda \boldsymbol{q}_i$ for some $\lambda > 0$, then no $\Delta \boldsymbol{k}_j$ can decrease both rows at first order. Likewise, if A_j contains $i \neq i'$ with $\boldsymbol{g}_{i'} = -\mu \boldsymbol{g}_i$ for some $\mu > 0$, then no single $\Delta \boldsymbol{v}_j$ decreases both rows at first order.*

Proof. ***core insight.*** The special cases in Corollary A.10 correspond to maximally conflicting demands: two rows either request opposite updates along exactly the same query direction (row side), or along exactly opposite gradient directions (column side).

For the row side, if $c_i(j)c_{i'}(j) < 0$ and $\boldsymbol{q}_{i'} = \lambda \boldsymbol{q}_i$ with $\lambda > 0$, then the cones $\mathcal{C}_+(j)$ and $\mathcal{C}_-(j)$ share a common ray. By Theorem A.9, they cannot be strictly separable by a hyperplane through the origin, so no $\Delta \boldsymbol{k}_j$ can make both rows strictly descend.

For the column side, if $\boldsymbol{g}_{i'} = -\mu \boldsymbol{g}_i$ with $\mu > 0$, then $\mathbf{0} \in \mathrm{cone}\{\boldsymbol{g}_i, \boldsymbol{g}_{i'}\}$. Thus the conic hull of $\{\boldsymbol{g}_k : k \in A_j\}$ contains the origin, and again by Theorem A.9, there is no single $\Delta \boldsymbol{v}_j$ that decreases the loss of both rows at first order. □

Corollary A.11 (When the dilemma disappears: alignment conditions: Restatement of Corollary 3.10)**.** *Fix a column j. The incompatibility in Theorem A.9 disappears if and only if either holds:*

(i) (Row alignment) *the sign-group cones $\mathcal{C}_+(j)$ and $\mathcal{C}_-(j)$ are strictly separable by a hyperplane through the origin;*

(ii) (Column alignment) *the active gradients $\{\boldsymbol{g}_i : i \in A_j\}$ lie in a common open halfspace.*

Proof. This is an immediate restatement of the two "if and only if" conditions in Theorem A.9. □

Remark A.4. *Taken together, Proposition A.8, Theorem A.9, and Corollaries A.10 and A.11 show that, under a unit-mass row budget and a shared value bank, achieving simultaneous first-order improvement for all rows via a single key or value update requires rare geometric alignment in either query space or gradient space. Otherwise, any update that helps some rows will necessarily hurt others. This structural tension, coupled with the granularity–reliability trade-off discussed in the main text, provides a mechanistic explanation for the observed weak-independence of attention when FFNs are removed.*

B. Complementary Ablation Studies for Section 5.4

B.1. Query-importance measure

We model cross-query importance with an *omnidirectional key-calibrated statistic*, instantiated via a quasi-linear mapping to impose smooth non-negativity. A simpler alternative is the naive mean responses of the keys. Our analysis expects that naive mean conflates prevalence with salience, thereby confusing query-importance estimation and degrading discriminability. We test this by comparing **CSaN** with **CSaN–Naive-Mean (NM)**. As shown in Table 1, substituting *naive-mean* for the *omnidirectional key re-calibration* yields a **clear accuracy degradation**. Notably, CSaN–NM still outperforms the original baseline: even so, the prevalence–salience conflation clearly caps the discriminability. This validates our insight.

Table 1. Ablation study of query-importance measure.

Token-Mixer	Backbone	#Params	FLOPs↓	Top-1(%)↑
Swin-Original	Swin-Tiny [8]	28.3M	4.4G	81.3
Swin-CSaN-NM		29.5M	4.6G	82.2
Swin-CSaN		29.5M	4.6G	**82.7**

B.2. Quasi-gating private value pathway

We introduce private value pathway (*i.e.*, $\boldsymbol{\kappa}_{[v]} \odot v$) as a per-token, unit-wise gated copy of the value vector that lies outside the shared attention budget, acting as a private endowment that preserves self-information and relaxes the row–column coupling behind weak-independence. Here, we investigate its empirical contribution. Table 2 reports that removing the private pathway (**CSaN-rPriv**) leads to a pronounced accuracy drop, reflecting the pathway's critical role as an effective *re-calibration anchor*, as well as an optimization aid (less cross-token contention, better-conditioned gradients), rather than mere capacity. These findings validate our insight.

Table 2. Validation of the private pathway's contribution.

Token-Mixer	Backbone	#Params	FLOPs↓	Top-1(%)↑
Swin-Original	Swin-Tiny [8]	28.3M	4.4G	81.3
Swin-CSaN-rPriv		29.2M	4.5G	**82.1**
Swin-CSaN		29.5M	4.6G	**82.7**

B.3. Ratios: expressivity-overhead trade-off

In configuring CSaN, we balance expressivity against overhead by two reduction ratios associated with the weight-budget relaxation learners. Let r_u denote the reduction ratio for the gated learner mixing query scores and raw input, and r_b the reduction ratio for the terminal relaxation learners (*i.e.*, unit-wise re-calibrators). We set r_u=4 and r_b=8 by default. Below we justify these choices and examine performance under alternative settings. As shown in Table 3, we observe:

1. Decreasing r_u toward 1 (*i.e.*, no compression) yields monotonic but modest accuracy gains.
2. Increasing r_u up to 16 (strong compression) gradually reduces the gains, yet results remain clearly above the original Swin-Min.
3. Decreasing r_b toward 1 first improves accuracy slightly and then plateaus, indicating representational redundancy at low compression.
4. Increasing r_b up to 16 diminishes the gains marginally, but they still exceed the original Swin-Min by a notable margin.

Taken together, these results indicate that (i) r_u is marginally more influential than r_b, yet CSaN is insensitive to these hyper-parameters within a reasonable range; and (ii) setting $r_u = 4$ and $r_b = 8$ achieves favorable performance with only slight overhead, so we adopt this configuration as the default.

Table 3. Ablation study of expressivity–overhead trade-off.

Token-Mixer	Backbone	r_u	r_b	#Params	FLOPs	Top-1(%)↑
Swin-Original	Swin-Min [8]	—	—	11.8M	1.6G	72.2
Swin-CSaN	Swin-Min [8]	✓	✓	12.2M	1.7G	75.0
		16	✓	12.0M	1.6G	74.6
		8	✓	12.1M	1.6G	74.7
		2	✓	12.4M	1.7G	75.1
		1	✓	12.9M	1.8G	75.1
		✓	16	12.1M	1.6G	74.7
		✓	4	12.5M	1.7G	75.0
		✓	2	13.1M	1.8G	75.1
		✓	1	14.6M	2.0G	75.1

* "✓" denotes the ***default setting(s)***, where $r_u = 4$ and $r_b = 8$.

B.4. Row temperatures and column magnifiers

In configuring CSaN, we learn per-head row temperatures and column magnifiers and use them to modulate the attention map before normalization (*i.e.*, $\mathrm{softmax}$ in our case), injecting finer relaxation signals. Below, we ablate this design on ImageNet with a Swin-Tiny [8] backbone.

Table 4 reports the comparative results, where we observe that row temperatures (τ) and column magnifiers (μ) yield similarly improvements. Given that both are very light computationally, we recommend including these components in CSaN by default.

Table 4. Ablation study of *row temperatures* (τ) and *column magnifiers* (μ).

Token-Mixer	Backbone	τ	μ	#Params	FLOPs	Top-1(%)↑
Swin-Original	Swin-Tiny [8]	—	—	28.3M	4.4G	81.3
Swin-CSaN	Swin-Tiny [8]	✓	✗	29.5M	4.6G	82.5
		✗	✓	29.5M	4.6G	82.5
		✓	✓	29.5M	4.6G	**82.7**

* "✓" indicates that the component is present, whereas "✗" indicates that it is absent (note that default CSaN applies both). "—" denotes non-applicable.

C. Implementation Recipe for ImageNet Experiment

For fair comparisons, we follow the standard training–evaluation recipe [8, 11] to train implemented variants of the Swin [8], ViT [4], and Hiera [10] families. Specifically, we adopt the standard data augmentations recommended in [8, 11] and the widely adopted AdamW optimizer [9], training each model with a cosine learning-rate scheduler for 300 epochs, including 20 epochs of linear warm-up (*i.e.*, DeiT's recommended recipe omitting distillation-related augmentations). The learning rate starts from 1×10^{-3} with an effective batch size of 1024 by default and decays to 2.5×10^{-6}, smoothly. The weight decay is set to 0.05 and label-smoothing of 0.1. Following the common practice, we (1) train and test all models with an image size of 224×224; (2) report the results of our models and the official Top-1 accuracies for the original baselines (if publicly available), rounded to one decimal place.

D. Robustness Under Long-Tailed Distributions

Setup. We validate CSaN on ImageNet-LT [7] (long-tailed distribution) using Swin-Min [8] and ViT-Tiny [4]. To isolate the intrinsic effect of the original/CSaN-enhanced attention paradigms, we *do not* apply any specialized long-tail techniques (*e.g.*, re-weighting/re-sampling, deferred re-balancing). Training follows common long-tail practice: 300 epochs, 20 warm-up epochs, cosine LR, base LR 5×10^{-4}, AdamW, weight decay 0.05, label-smoothing of 0.1.

Table 5. Comparative evaluation on ImageNet-LT.

Token-Mixer	Backbone	#Params	FLOPs	Top-1(%)↑
Swin-Original	Swin-Min [8]	11.8M	1.6G	31.3
Swin-CSaN		12.2M	1.7G	**32.1**
ViT-Original	ViT-Tiny [4]	5.7M	1.1G	27.4
ViT-CSaN		6.0M	1.1G	**28.6**

Observation. As reported in Table 5, **CSaN** improves the original transformer counterparts under class-imbalance, indicating stronger adaptability to long-tailed distributions even without any imbalance-specific heuristics.

E. Convergence Attribute

We present the convergence curves of the original Swin-Min [8] and the CSaN-enhanced counterpart. Figure 1 depicts the convergence trends in Top-1 accuracy (the higher the better) and training loss (the lower the better), respectively. Notably, the CSaN-enhanced model achieve higher Top-1 accuracy and lower loss in epochs after the mid-training stage, once the model are sufficiently warmed up. This finding validates the favorable convergence attributes of CSaN.

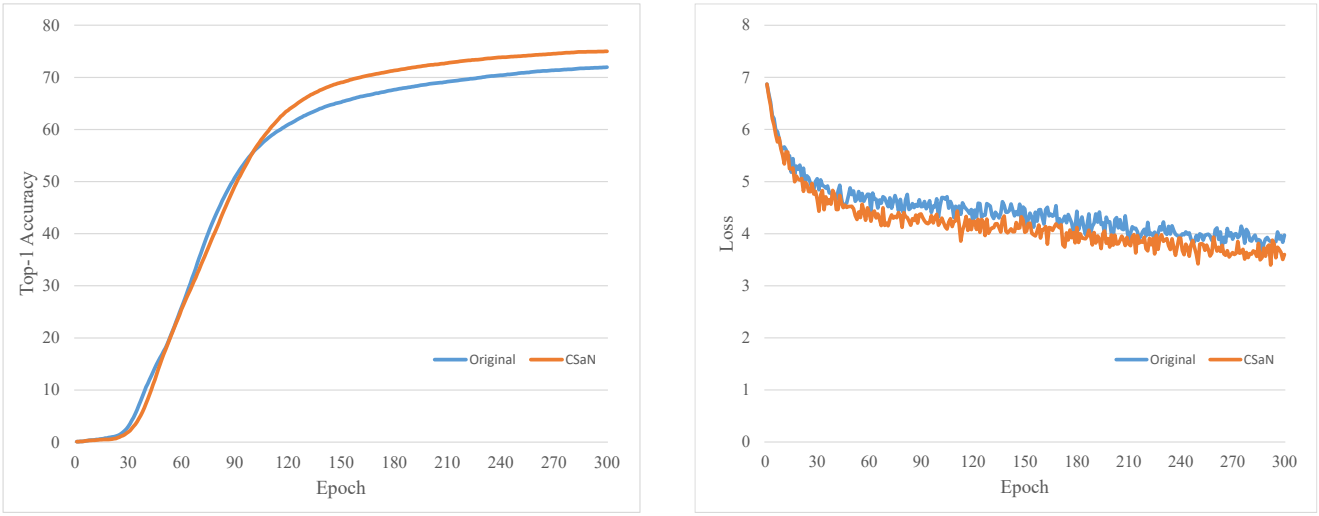

Figure 1. The accuracy curves (left) and loss curves (right) on Swin-Min [8] backbone with the original and CSaN-enhanced paradigms.

F. Public–Private Interaction Dynamics

The learned public–private dynamics (Figure 2) suggest that the two routes are not simply fused to accumulate their respective information; rather, CSaN decomposes attention into a shared *public* selector and a corrective *private* self-anchor. Throughout training, the stage-wise mean of $\boldsymbol{\kappa}^{\mathrm{pub}}$ over all tokens remains positive across all stages, indicating that the public pathway consistently preserves its role as the carrier of consensus-like mutual scoring and dynamic budgeting. In contrast, the stage-wise mean of $\boldsymbol{\kappa}^{\mathrm{priv}}$ over all tokens is learned from near zero to stable negative values, with larger magnitudes in deeper stages, showing that the private pathway is (more likely) recruited as an oppositional self-evaluation term. Mechanistically, public scoring determines which candidates are globally favored, while private self-anchoring can veto, preserve, or restore candidates that would otherwise be over-suppressed by mutual evaluation alone. From the SaN perspective, this highlights the significance of CSaN: effective attention nonlinearity is realized not by public mutual scoring alone, but by a principled coupling of *public agreement* and *private correction*, which stabilizes selection and mitigates weak-independence.

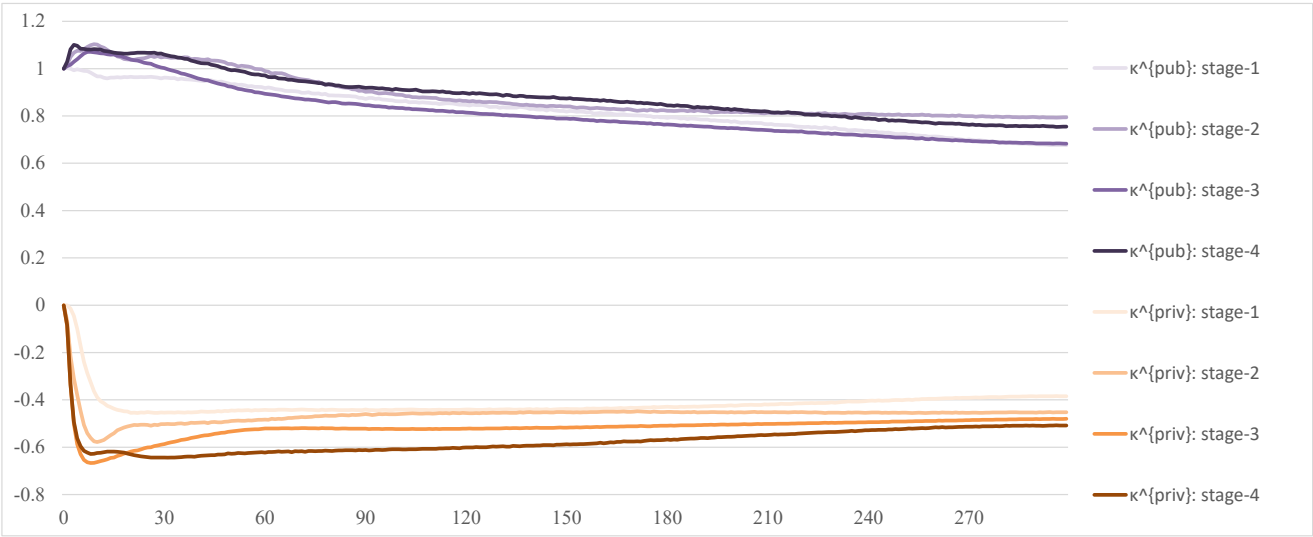

Figure 2. Illustration of the interaction dynamics between the public and private pathways.

G. CIFAR-100 Evaluation

Implementation details. We further validate CSaN on CIFAR-100 [5], using Swin-Tiny as the backbone. CIFAR-Swin-Tiny is a modified version of the original Swin-Tiny [8] for ImageNet and downstream tasks. Specifically, the base embedding dimension is reduced from 96 to 24 to prevent redundant parameters, as CIFAR-100 contains far fewer images than ImageNet and has significantly lower image resolution.

To ensure fair comparisons, all models are trained from scratch using the same standard training-evaluation recipe: we adopt most of the training protocols and data augmentations suggested in [6], with slight modifications to fit the Transformer-based backbone. Specifically, each model is train for 350 epochs with a batch-size of 256, by an AdamW optimizer with a weight-decay of 0.05. The learning rate starts from 1^{-3} and decreases to 1^{-6} by following the standard cosine learning rate schedule. All the input images are fixed to the size of 32×32 by following the common practice.

Experimental results. Comparative results for the mean and standard deviation (over 8 runs) of Top-1 accuracy are reported in Tab. 6, where the CSaN-enhanced model improves upon the original model by a clear margin. This trend is consistent with our findings on ImageNet, further supporting the adaptability of CSaN across datasets of different scales.

Table 6. Comparative evaluation on CIFAR-100 benchmark dataset.

Token-Mixer	Backbone	#Params	Top-1(%)↑
Swin-Original	CIFAR-Swin-Tiny	1.8M	66.7±0.3
Swin-CSaN		2.0M	**67.6±0.4**

H. Other Details

H.1. Quasi-linear function in Section 4

We define a quasi-linear map ϕ_β to impose soft non-negative constraints on the attention map to obtain the omnidirectional query scores, where ϕ_β is defined as:

$$s = \begin{cases} s, & \text{if } s \geqslant \beta\,, \\ \beta \cdot \exp\left(\dfrac{s}{\beta} - 1\right), & \text{if } s < \beta\,, \end{cases} \tag{62}$$

where β is a small threshold (0.25 by default in our experiments). Note that ϕ_β defines a differentiable function on $\mathbb{R}$.

H.2. LLM usage

ChatGPT was used to aid in polishing the writing. Specifically, it was employed to correct grammar, improve readability, and refine the clarity of sentences.

References

[1] Sudong Cai. Iieu: Rethinking neural feature activation from decision-making. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 5796–5806, 2023. 1

[2] Sudong Cai. Adashift: Learning discriminative self-gated neural feature activation with an adaptive shift factor. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5947–5956, 2024.

[3] Sudong Cai, Shuyuan Zheng, Bingzhi Chen, Shuai Yuan, Chuan Xiao, Jianbin Qin, and Bing WANG. Toward principled flexible scaling for self-gated neural activation. In *International Conference on Learning Representations (ICLR)*, 2026. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 12

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009. 14

[6] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective Kernel Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–519, 2019. 14

[7] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 12

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 10, 11, 12, 13, 14

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 12

[10] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *Proc. International Conference on Machine Learning (ICML)*, pages 29441–29454. PMLR, 2023. 12

[11] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training Data-Efficient Image Transformers & Distillation Through Attention. In *Proc. International Conference on Machine Learning (ICML)*, 2021. 12