CVPR JUNE 3-7, 2026 DENVER COLORADO

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

Sudong Cai[1,2], Shuai Yuan[2], Bingzhi Chen[2], Rui Mao[3], Bing Wang[1*]

[1]The Hong Kong Polytechnic University; [2]Beijing Institute of Technology, Zhuhai; [3]Shenzhen University

Presenter: Sudong Cai

Contact: bing-w.wang@polyu.edu.hk

Motivation: attention is expressive, but not independent

Result preview: CSaN turns the diagnosis into level-jump gains.

- **The puzzle**
  - Self-attention can be a universal approximator under mild conditions.
  - But attention-only stacks drop sharply when FFNs are removed.
- **Our answer**
  - SaN views attention as directed, budget-constrained selection.
  - CSaN relaxes binding budgets and adds a private value route.

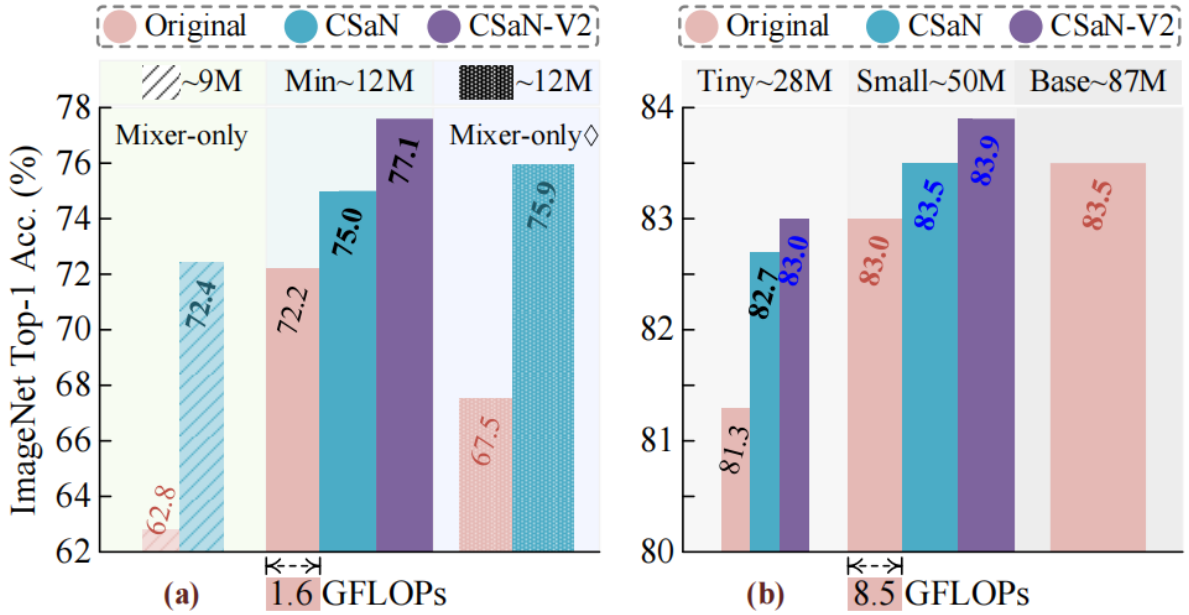


Fig. 1 shows two facts at once: weak-independence is real, and CSaN makes smaller models competitive with larger same-family models.

+5.5

Top-1 over original Swin-Min with CSaN-V2

~0.5x

Cost to rival much heavier variants

Weak-Independence: the empirical clue

Table 1 isolates the problem before introducing the remedy.

What Table 1 says

72.2 Original Attn-FFN

62.8 Attention-only

75.9 CSaN, size-matched

- **Key interpretation**

  - Replacing FFNs with more attention does not recover the baseline.
  - Width-matching helps, but still leaves a clear gap.
  - CSaN turns the attention-only skeleton into a stronger block.

Table 1. Empirical evidence for the *weak-independence* challenge of attention and the pronounced compensatory effect of **CSaN**.

Skeleton	Token-Mixer	#Stacking	#Params	FLOPs↓	Top-1(%)↑
Swin-Min	Swin-Original	Attn-FFN	11.8M	1.6G	72.2
		Attn-Only	8.6M	1.1G	62.8
		Attn-Only◊	11.8M	1.6G	67.5
	Swin-CSaN (ours)	Attn-Only	9.5M	1.3G	**72.4**
		Attn-Only◊	13.6M	1.8G	**75.9**

* The symbol ◊ denotes width-extended variants size-matched to their standard attention–FFN counterparts; "Attn" denotes "Attention."

Takeaway: the bottleneck is structural, not just insufficient parameter count.

This motivates a mechanism-level explanation of what FFNs compensate for.

Selection-as-Nonlinearity: attention as budgeted selection

SaN connects activation, gating and attention with one conceptual lens.

Activation = directed, context-conditioned selection

Context score	→	Gate / membership	→	Selected feature	→	Attention row
$s = h(\mathbf{c})$ or $\mathbf{q}_i^\top \mathbf{k}_j$		$\rho(s) \geq 0, \quad \rho(\cdot)$ is monotone		$\Phi(\mathbf{x}, \mathbf{c}) = \rho(\mathbf{c})\,\mathbf{x}$		$\mathbf{y}_i = \sum_j \alpha_{ij}\mathbf{v}_j, \quad \sum_j \alpha_{ij} = 1$

Three consequences

- Different contexts retain the same token differently, so the mapping becomes effectively nonlinear.
- Self-attention is an aggregation of context-gated activation units over a shared value bank.
- The row normalizer makes each query solve a unit-mass allocation game.

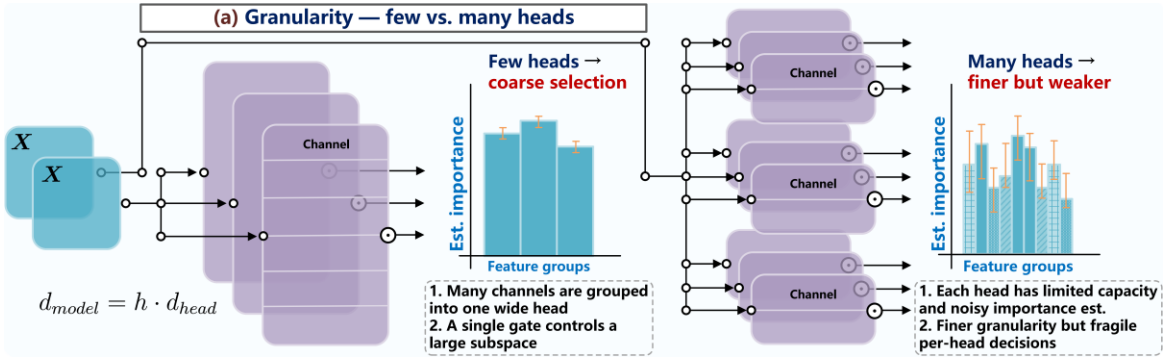
Why this matters

The same budget that gives attention meaningful selection also creates competition: rows compete over values, and values are shared across rows.

Two structural tensions behind weak-independence

SaN diagnoses why attention struggles when used alone.

Tension I: granularity vs. reliability

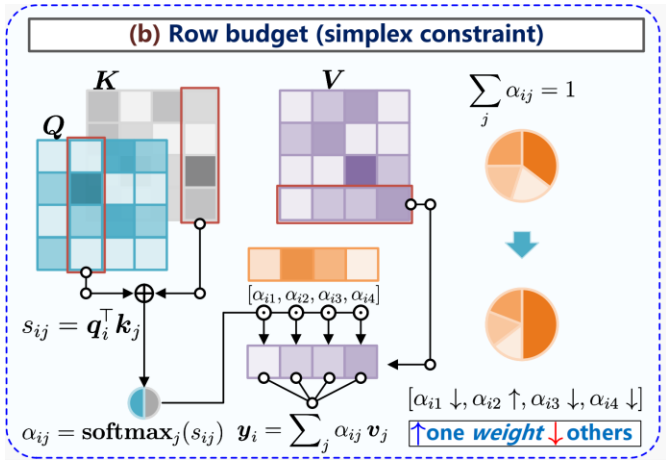


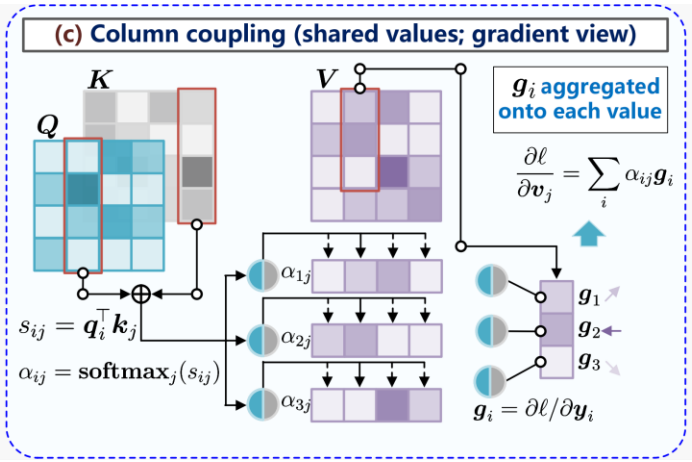
- Few heads give coarse selection;
- many heads give fine but fragile gates.

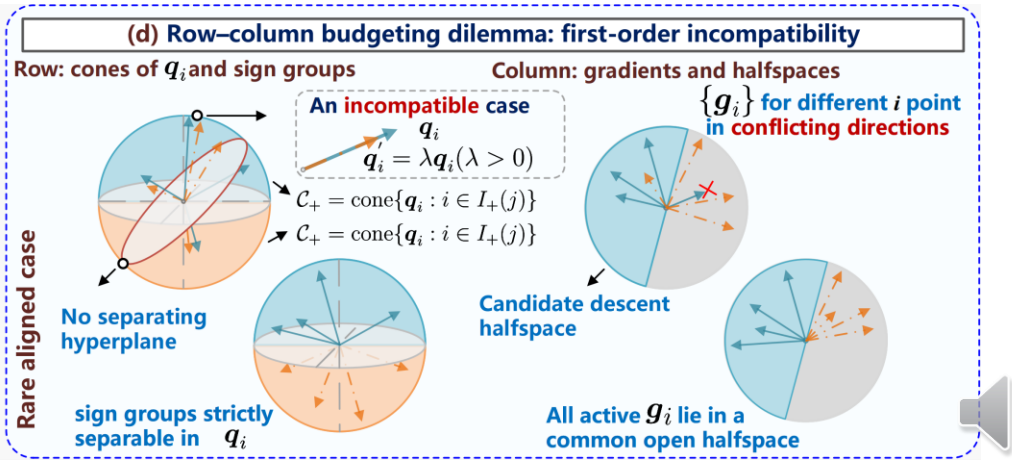
- **A single shared value update must satisfy many row-wise descent demands; without rare directional alignment, helping one row can hurt another.**

CSaN targets both tensions: finer readout, relaxed budgets, and a private value route.

Tension II: row-column budgeting dilemma







CSaN: An Interpretable Remedy

A drop-in attention compensation mechanism guided by SaN

Component 1

Hierarchical budget calibration

- Row temperature tau: reweights inter-query budgets.
- Column magnifier mu: eases column-side crowding.
- Softmax is preserved, so selection semantics remain interpretable.

Component 2

Public-private cooperation

- Public path: normal attention for cross-token mixing.
- Private value path: per-token route that bypasses shared-budget conflict.
- Unit-wise gains improve granularity without adding many heads.

Algorithm 1: CSaN: descriptor → hierarchical budgets → public/private readout (rectangular $M \times N$)

Inputs: logits $\mathbf{S} \in \mathbb{R}^{M \times N \times H}$, values $\mathbf{V} \in \mathbb{R}^{N \times H \times d_h}$, head template $\mathbf{U} = f_u(\mathbf{X})$

▷ w/ reduction ratio: r_u

Output: $\mathbf{X} \in \mathbb{R}^{M \times H \times d_h}$

if $M \neq N$ **then**

  $\hat{\mathbf{U}} \leftarrow \Pi_{N \to M}(\mathbf{U})$ ▷ align N-indexed template to M (pool/interp)

Descriptors: $\iota_{ih} \leftarrow \frac{1}{N} \sum_j \phi_\beta(s_{ijh})$;

$\iota_{ih} \leftarrow \sigma(u_{ih}) \cdot \iota_{ih}$; $\boldsymbol{z}_i \leftarrow \psi \circ f_\iota(\iota_{i:})$

▷ nonneg., isotonic

Budgets: $\tau_{ih} \leftarrow 1 + f_\tau(\boldsymbol{z}_i^{\text{row}})$, $\mu_{jh} \leftarrow 1 + f_\mu(\boldsymbol{z}_j^{\text{col}})$

▷ row temp., column magnifier, w/ reduction ratio: r_b

Calibrated attention:

$\tilde{\boldsymbol{A}}_{ijh} \leftarrow \text{softmax}_j(\tau_{ih} s_{ijh}) \cdot \mu_{jh}$ ▷ keep row normalizer R

Readout: $\boldsymbol{\kappa}_{ih}^{\text{pub}} \leftarrow 1 + f_a(\boldsymbol{z}_i^{\text{row}})$, $\boldsymbol{\kappa}_{ih}^{\text{priv}} \leftarrow f_v(\boldsymbol{z}_i^{\text{row}})$;

$\boldsymbol{x}_{ih} \leftarrow (\sum_j \tilde{\boldsymbol{A}}_{ijh} \boldsymbol{v}_{jh}) \odot \boldsymbol{\kappa}_{ih}^{\text{pub}} \oplus \boldsymbol{V}_{i,h} \odot \boldsymbol{\kappa}_{ih}^{\text{priv}}$

▷ Defaults: $\beta = 0.25$; $f_u, f_\iota, f_\tau, f_\mu, f_a, f_v$ are linear layers; $r_u = 4$, $r_b = 8$.

Result I: generalization across backbones and tasks

ImageNet and COCO show that CSaN is not tied to one attention design.

Table 2. Comprehensive evaluation of CSaN's effectiveness and generalization on ImageNet(-1K) across three Vision Transformer families (Swin [30], ViT [16] (DeiT version [42]), Hiera [38]) and multiple model scales. With negligible overhead (∼5% in parameters and FLOPs), CSaN consistently improves the baselines, enabling lighter variants to rival same-family much heavier counterparts that are ∼ 2× as large.

Backbone	Token-Mixing Paradigm	Layer Setting	Embed. Dimension	Num. Heads	Throughput (images/sec.)	#Params (M)	FLOPs↓ (G)	Top-1↑ (%)
SWIN-TRANSFORMER FAMILY (shifted-window attention) [30] w/ Resolution @ 224								
Swin-Min	Swin-Original	[1, 1, 1, 1]	[96, 192, 384, 768]	[3, 6, 12, 24]	4207.2	11.8	1.6	72.2
	Swin-CSaN (ours)				3555.7	12.2	1.7	**75.0**
Swin-Tiny	Swin-Original	[2, 2, 6, 2]	[96, 192, 384, 768]	[3, 6, 12, 24]	1622.5	28.3	4.4	81.3
	Swin-CSaN (ours)				1384.1	29.5	4.6	**82.7**
Swin-Small	Swin-Original	[2, 2, 18, 2]	[96, 192, 384, 768]	[3, 6, 12, 24]	906.9	49.6	8.5	83.0
Swin-Base	Swin-Original		[128, 256, 512, 1024]	[4, 8, 16, 32]	646.9	87.8	15.1	**83.5**
Swin-Small	**Swin-CSaN (ours)**		[96, 192, 384, 768]	[3, 6, 12, 24]	789.6	51.8	8.9	**83.5**
VISION TRANSFORMER FAMILY (global attention over patch tokens) [42] w/ Resolution @ 224								
ViT-Tiny/16	ViT (DeiT version)	[12]	192	3	5228.9	5.7	1.1	72.2
	ViT-CSaN (ours)				4638.2	6.0	1.1	**74.9**
ViT-Small/16	ViT (DeiT version)	[12]	384	6	2089.2	22.1	4.2	79.8
	ViT-CSaN (ours)				1885.4	23.0	4.4	**81.3**
ViT-Base-Slim/16	ViT (DeiT version)		512	8	1376.3	38.8	7.5	80.7
ViT-Base/16	ViT (DeiT version)	[12]	768	12	714.9	86.6	16.9	**81.8**
ViT-Base-Slim/16	**ViT-CSaN (ours)**		512	8	1258.7	40.6	7.9	**82.1**
HIERA TRANSFORMER FAMILY (local-then-global, multi-scale attention) [38] w/ Resolution @ 224								
Hiera-Tiny	Hiera-Original	[1, 2, 7, 2]	[96, 192, 384, 768]	[1, 2, 4, 8]	1791.5	27.9	4.6	80.9
	Hiera-CSaN (ours)				1547.8	29.1	4.9	**81.8**
Hiera-Small	Hiera-Original	[1, 2, 11, 2]	[96, 192, 384, 768]	[1, 2, 4, 8]	1415.2	35.0	6.0	81.3
	Hiera-CSaN (ours)				1236.0	36.5	6.4	**82.5**
Hiera-Tiny-Plus	Hiera-Original	[1, 2, 7, 2]	[64, 128, 256, 512]	[3, 6, 12, 24]	1627.6	27.9	4.6	81.5
Hiera-Base	Hiera-Original	[2, 3, 16, 3]	[96, 192, 384, 768]	[1, 2, 4, 8]	957.7	51.5	8.8	**82.4**
Hiera-Tiny-Plus	**Hiera-CSaN (ours)**	[1, 2, 7, 2]	[64, 128, 256, 512]	[3, 6, 12, 24]	1401.3	29.1	4.9	**82.4**

Table 3. Comparative evaluation on MS COCO [26]. Both variants are initialized from the corresponding baseline pre-trained weights.

Token-Mixer	Encoder	Head	mAP (%)↑	AP_{50}(%)↑	AP_{75}(%)↑	AP_S(%)↑	AP_M(%)↑	AP_L(%)↑
Swin-Original	Swin-Tiny [30]	RetinaNet [27]	37.3	57.6	39.8	22.1	40.7	49.7
Swin-CSaN (Ours)			**38.0**	**58.3**	**40.3**	**22.2**	**41.5**	**51.3**

ImageNet-1K

- Consistent gains on Swin, ViT and Hiera.
- Lighter models rival much heavier same-family baselines.
- Only slight overhead: about 5% in parameters/FLOPs.

COCO detection

- The downstream gain remains even when both models start from the same original Swin-Tiny pretrained weights.

Result II: extensibility, ablation, and dynamics

CSaN is a flexible basis, and its components match the SaN diagnosis.

Extensible remedy

Table 4. Validation of CSaN's extensibility. CSaN-V2 is compared with CSaN and the original Swin models across multiple sizes.

Backbone	Token-Mixing Paradigm	#Params (M)	FLOPs↓ (G)	Top-1↑ (%)
Swin-Min [30]	Swin-Original	11.8	1.6	72.2
	Swin-CSaN	12.2	1.7	**75.0**
	Swin-CSaN-V2	12.3	1.7	**77.1**
Swin-Tiny [30]	Swin-Original	28.3	4.4	81.3
	Swin-CSaN	29.5	4.6	**82.7**
	Swin-CSaN-V2	29.7	4.6	**83.0**
Swin-Small [30]	Swin-Original	49.6	8.5	83.0
Swin-Base [30]	Swin-Original	87.8	15.1	**83.5**
Swin-Small [30]	**Swin-CSaN**	51.8	8.9	**83.5**
Swin-Small [30]	**Swin-CSaN-V2**	52.3	9.0	**83.9**

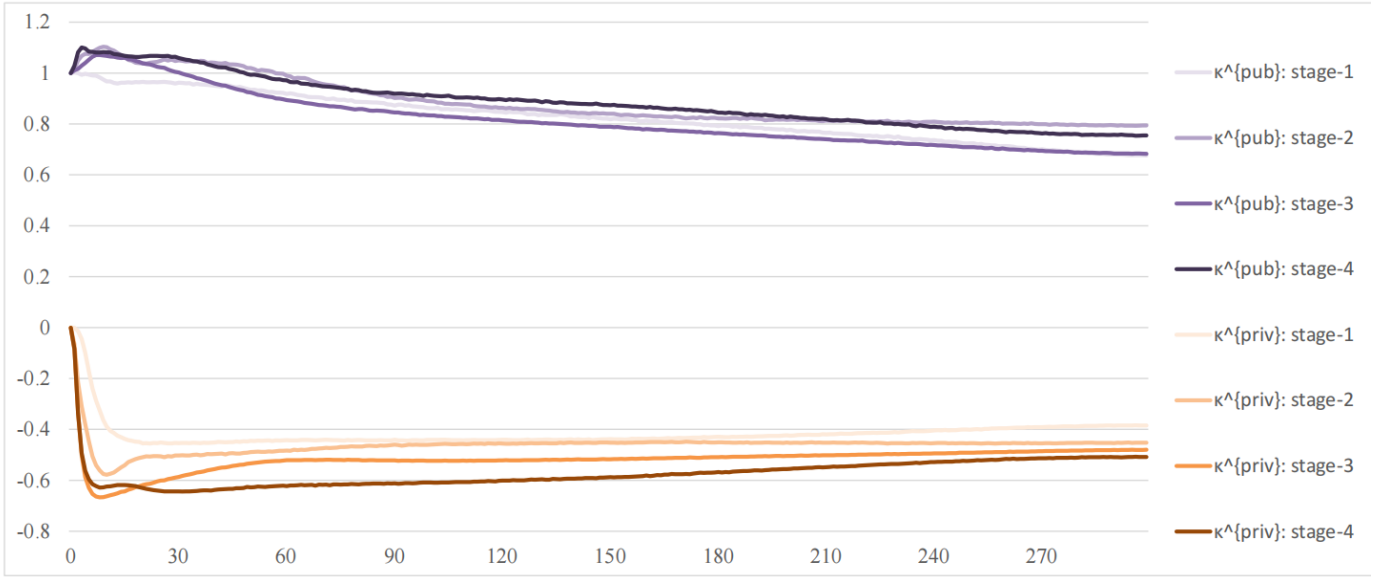
* With slight overhead, CSaN-V2 boosts the lighter variants by a level-jump improvements over same-family much heavier counterparts.

Granularity ablation

Table 5. Validation of granularity in weight-budget relaxation.

Token-Mixer	Backbone	#Params	FLOPs↓	Top-1(%)↑
Swin-Original	Swin-Tiny [30]	28.3M	4.4G	81.3
Swin-CSaN-HdW		29.0M	4.5G	82.5
Swin-CSaN		29.5M	4.6G	**82.7**

Public-private dynamics



- CSaN-V2 adds structure-aware integration and further improves CSaN.
- Unit-wise relaxation outperforms head-wise relaxation, supporting the granularity diagnosis.
- Public pathway remains positive; private pathway learns corrective self-anchoring.

Conclusion

Attention is not just a linear mixer. It is budget-constrained selection.

- **SaN explains weak-independence as two structural tensions:**
  - Granularity-reliability trade-off
  - Row-column budgeting dilemma

- **CSaN operationalizes the lens:**
  - keep the normalizer
  - relax the budgets
  - add public-private cooperation

Result: interpretable attention compensation with strong gains and negligible overhead.

Thanks!

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

Sudong Cai[1,2], Shuai Yuan[2], Bingzhi Chen[2], Rui Mao[3], Bing Wang[1*]
[1]The Hong Kong Polytechnic University; [2]Beijing Institute of Technology, Zhuhai; [3]Shenzhen University