Attention Transfer Is Not Universally Effective for Vision Transformers

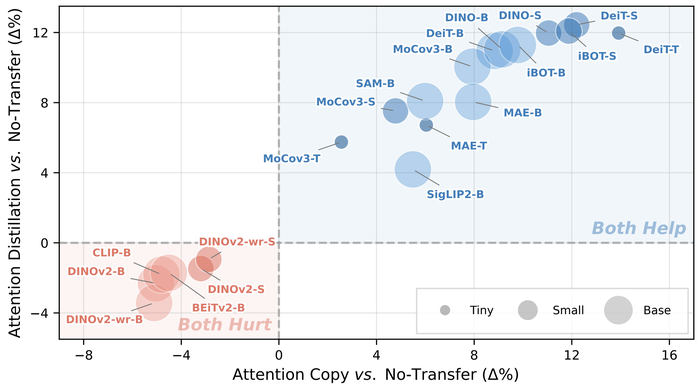
Huaiyuan Qin Muli Yang Gabriel James Goenawan Hongyuan Zhu
Institute of Advanced Intelligence and Computing (IAIC), A*STAR, Singapore

Is Attention Transfer Universally Effective?

1. Attention Transfer & the claim

- Transfer **only the attention patterns** from a pre-trained teacher ViT to a randomly initialized standard student: ***Attention Copy*** (directly copy the teacher's attention projection weights, kept frozen) or ***Attention Distillation*** (train the student to match the teacher's attention patterns).
- Li et al. (NeurIPS 2024): this alone is sufficient to **recover the full benefit** of the teacher's pre-trained weights — suggesting the essential transferable structure lies in the **inter-token routing patterns**.

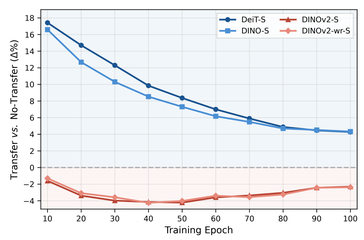
2. Our comprehensive benchmark answers: No ✗



- **20 pre-trained teachers, 11 well-known ViT families** across model sizes, under the default ImageNet-1K protocol, all results over 3 seeds.
- **7 families succeed** under both methods, with gains reaching **+13.9** (DeiT-T, Copy); **4 families consistently fail**: **DINOv2, DINOv2-wr, CLIP, BEiTv2**, falling up to **−5.1** below the from-scratch baseline (DINOv2-wr-B, Copy).
- All teachers evaluated by Li et al. fall in our **success** group.

3. The failure is robust

- **Not slow convergence**: the gap persists under extended training (up to 100 epochs).



- **Not the dataset or distribution**: the same success/failure split holds on iNaturalist and four OOD test sets.

Dataset robustness (Attention Distillation)

	NoT	DeiT	DINO	DINOv2	DINOv2-wr
IN-1K	63.1	75.6 +12.4	75.1 +12.0	61.7 −1.5	62.2 −0.9
iNat 2017	26.6	44.7 +18.1	46.0 +19.4	24.4 −2.2	24.5 −2.1
iNat 2018	35.0	50.2 +15.2	51.2 +16.3	33.1 −1.9	33.8 −1.2
iNat 2019	40.1	56.8 +16.7	55.7 +15.5	38.1 −2.1	38.7 −1.4

OOD robustness (Attention Distillation)

	NoT	DeiT	DINO	DINOv2	DINOv2-wr
IN-A	9.2	17.6 +8.4	17.2 +8.0	6.0 −3.2	5.8 −3.4
IN-R	35.9	42.3 +6.4	42.8 +6.9	32.8 −3.1	32.2 −3.7
IN-S	22.1	29.7 +7.6	29.4 +7.3	19.5 −2.6	19.0 −3.1
IN-V2	63.9	69.5 +5.5	69.9 +6.0	61.4 −2.5	61.6 −2.4

What Explains the Failure?

1. Generically bad weights? ✗ — it is the attention pathway

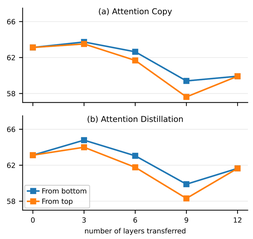
Component-level decomposition (fine-tuning accuracy)

Attn.	MLP	DeiT-S	DINOv2-S	CLIP-B
		66.0	65.3	70.3
✓		71.7 +5.7	63.7 −1.6	69.4 −0.9
	✓	77.7 +11.7	69.2 +3.9	74.7 +4.4
✓	✓	79.3 +13.3	71.0 +5.7	80.1 +9.8

- **MLP-only** initialization helps *all* families (even failures: +3.9, +4.4).
- **Attention-only** preserves the success/failure split ⇒ the failure is **consistently localized to the attention pathway**.

2. One bad depth? ✗

- Transferring top-k or bottom-k layers of DINOv2-S: **no subset** beats No-Transfer.
- Contrast: success teachers show monotonic gains as more layers are transferred (Li et al.).
- Interesting point: From-top is consistently worse than From-bottom.



3. One bad Q/K/V term? ✗

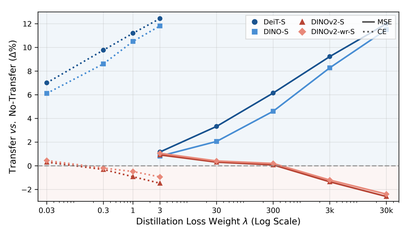
- Transferring any of Q, K, or V alone remains beneficial for DeiT-S, but is **harmful** for DINOv2-S (V: largest drop, −15.9) ⇒ the failure is not confined to one subcomponent, but **distributed broadly throughout the overall attention pathway**.

QKV decomposition (Δ vs. NoT = 63.1)

	DeiT-S		DINOv2-S	
	Copy	Distill	Copy	Distill
Q	78.1 +15.0	77.0 +13.9	57.9 −5.2	59.0 −4.1
K	78.7 +15.6	76.7 +13.6	57.4 −5.7	58.7 −4.4
V	78.4 +15.3	78.0 +14.9	47.3 −15.9	54.9 −8.3
Attn.	75.3 +12.2	75.6 +12.4	59.9 −3.2	61.7 −1.5

4. The transfer loss? ✗

- CE vs. MSE with λ from 0.03 to 30k: **same family-level trends** — stronger transfer signal helps success families, but *amplifies* harm for failure families.



5. The pre-training recipe? ✗

No single-factor recipe property predicts the family-level failure boundary

Hypothesis	Failure	Counter-example
Pre-training signal	DINOv2 (Self-Distillation) CLIP (Contrastive) BEiTv2 (Masked Image Modeling)	DINO (Self-Distillation) SigLIP2 (Contrastive) iBOT/MAE (Masked Image Modeling)
Data source	DINOv2 (Non-ImageNet Data) CLIP (Multi-Modal Data)	SAM (Non-ImageNet Data) SigLIP2 (Multi-Modal Data)
Special attribute	DINOv2 (with Attention Sink) DINOv2 (patch 14 to 16 to fit standard-arch)	DeiT (with Attention Sink) CLIP (native patch 16, also fails)

Architectural Compatibility Is the Key

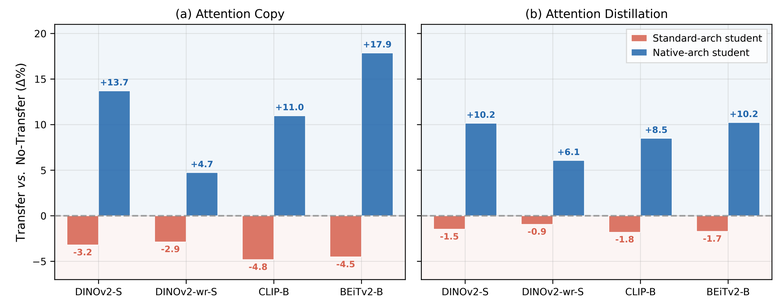
1. The hidden mismatch

- Each failure family contains one or more components that are **native to the teacher but absent from the standard student architecture**:

Failure family	Missing native component
DINOv2, DINOv2-wr	`LayerScale`
CLIP	`PreLayerNorm`
BEiTv2	`LayerScale`, `RelativePositionBias`

- These native components define the **scaling, normalization, and positional context** under which the teacher's attention was learned.

2. Native-architecture rescue ✓



- Add **only** the teacher-native missing components — initialized **randomly** and trained jointly, with no pre-trained weights loaded.
- The failure is **completely reversed for all 4 families**: standard-arch transfer remains harmful, while native-arch transfer consistently **flips to positive gains** under both Attention Transfer methods.

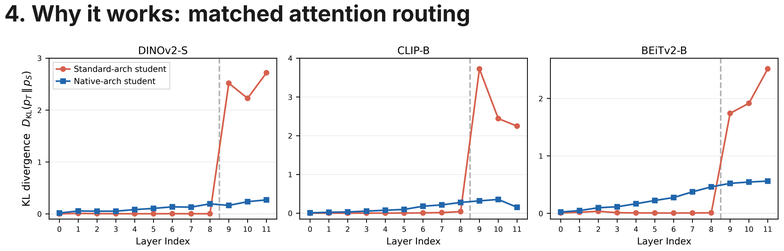
3. Control: the components alone do not help

- Native-arch NoT baselines are **worse** than standard-arch ones: the added components **do not by themselves improve performance** ⇒ the rescue comes specifically from **restoring architectural compatibility**.

From-scratch (NoT) baselines

	DINOv2-S	CLIP-B	BEiTv2-B
Standard-arch	63.1	69.0	69.0
Native-arch	60.1 −3.0	66.7 −2.3	66.3 −2.7

4. Why it works: matched attention routing



- The standard-arch student matches the early layers almost perfectly with **near-zero KL**, but exhibits a **sharp divergence spike in the last three layers**.
- The native-arch student remains **consistently low and smooth across all layers** ⇒ resolving this compatibility issue recovers the transfer gains.

Our findings refine the prevailing understanding of attention in ViT representations:

Attention is sufficient *only* when the student architecture matches the teacher.

More details & Full paper: arXiv:2605.07191