SIC Saarland Informatics Campus

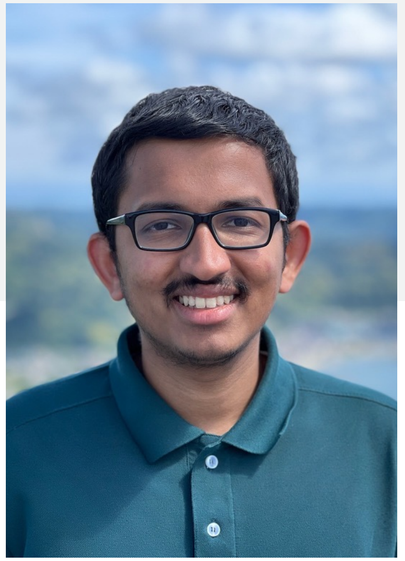
MAX PLANCK INSTITUTE FOR INFORMATICS

CVPR JUNE 3-7, 2026 DENVER COLORADO

Align Once to Explain: Feature Alignment for Scalable B-cosification of Foundational Vision Transformers

Raphael Maser

Siddhartha Gairola



Sukrut Rao

Bernt Schiele

Max Planck Institute for Informatics

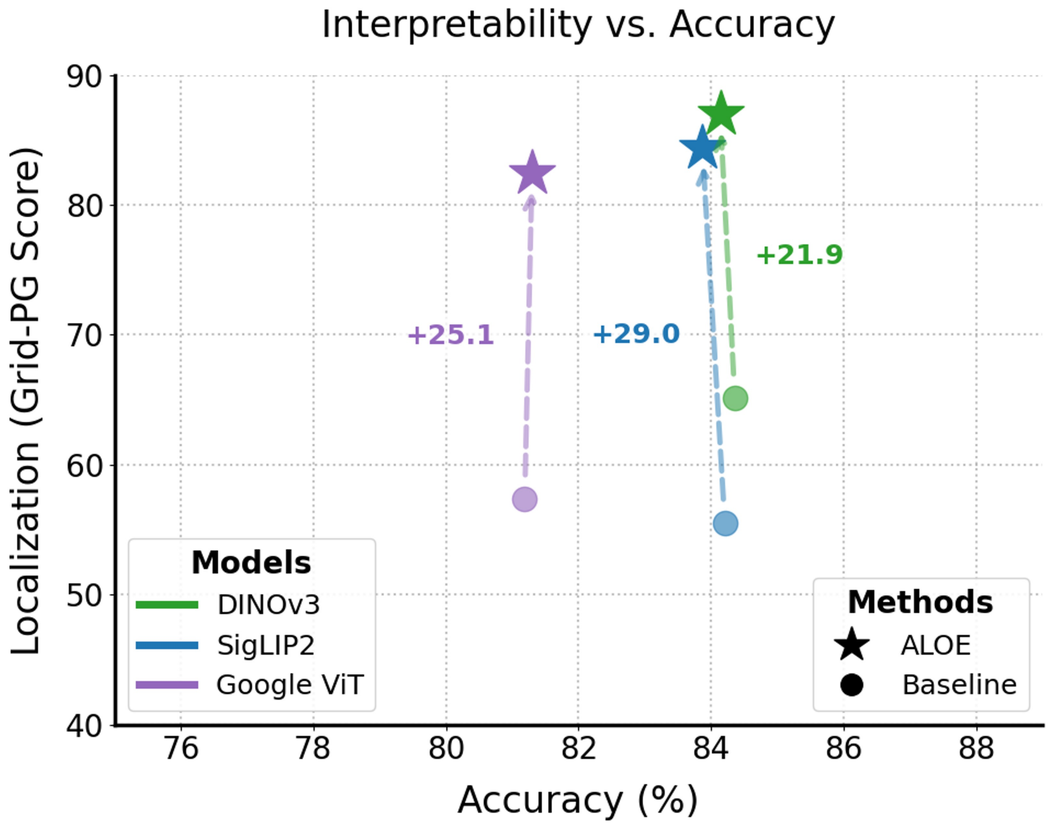
The Interpretability Gap

Opaque Foundation Models: incredible performance, remain opaque

Post-Hoc explanations: often unfaithful and compute intensive

Training Cost: re-training interpretable models is cost-prohibitive

→ **ALOE**: efficient distillation of foundation models into faithful and interpretable B-cos models

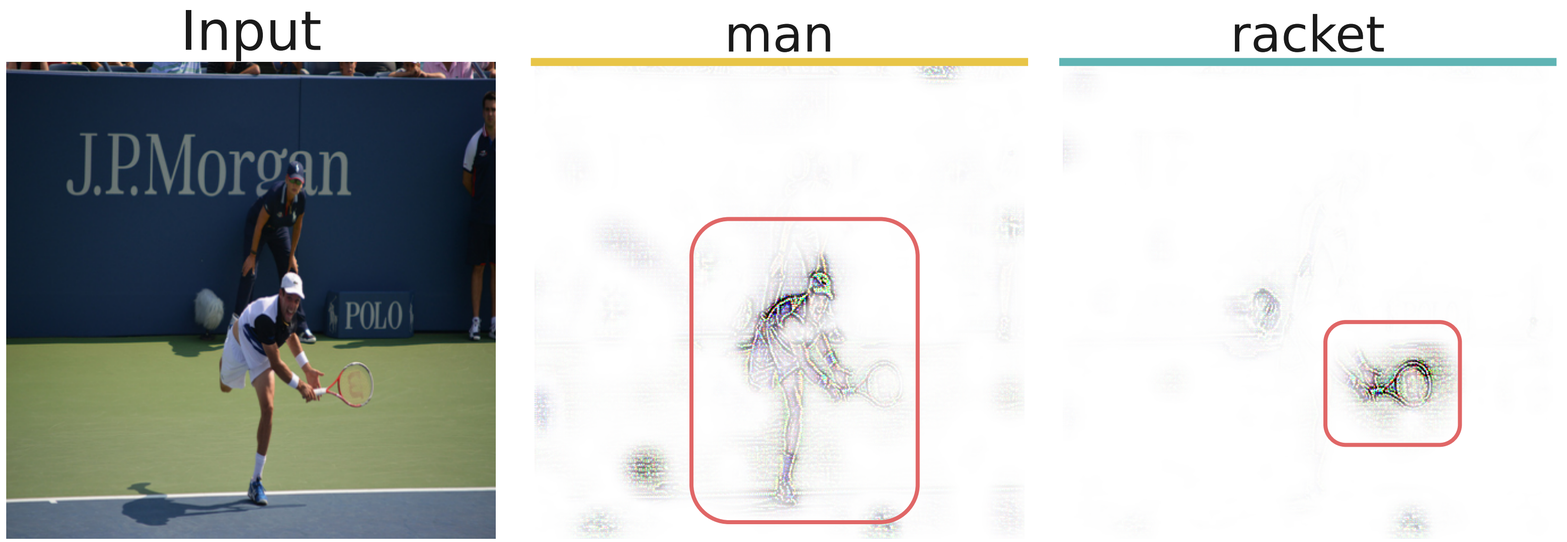


The Interpretability Gap

Understanding where models look can give us a better understanding about **failure modes**

→ Especially with **LVLMs**

[...] **man** in a white shirt playing tennis. He is in the middle of a swing, holding a tennis **racket** [...]



B-cos Models

B-cos models are **bias-free models** using **B-cos transformations**

→ hyperparameter **B steers alignment** with the input

$$\text{B-cos}(\mathbf{x}; \mathbf{w}) = \left(\left| \cos(\mathbf{x}, \mathbf{w}) \right|^{B-1} \times \widehat{\mathbf{w}} \right)^{\top} \mathbf{x} = \mathbf{w}(\mathbf{x})^{\top} \mathbf{x},$$

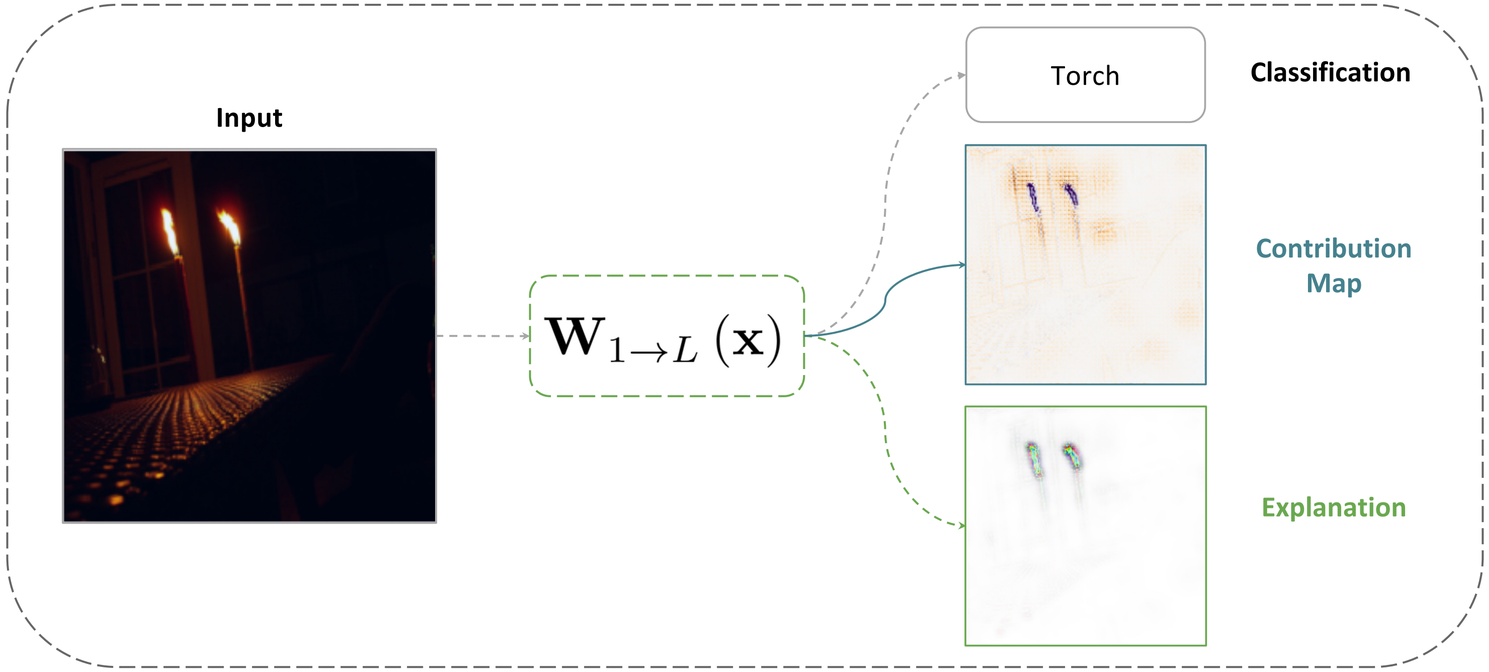
B-cos networks [Böhle et al., CVPR 2022]

B-cos Models

B-cos models are **bias-free models** using **B-cos transformations**

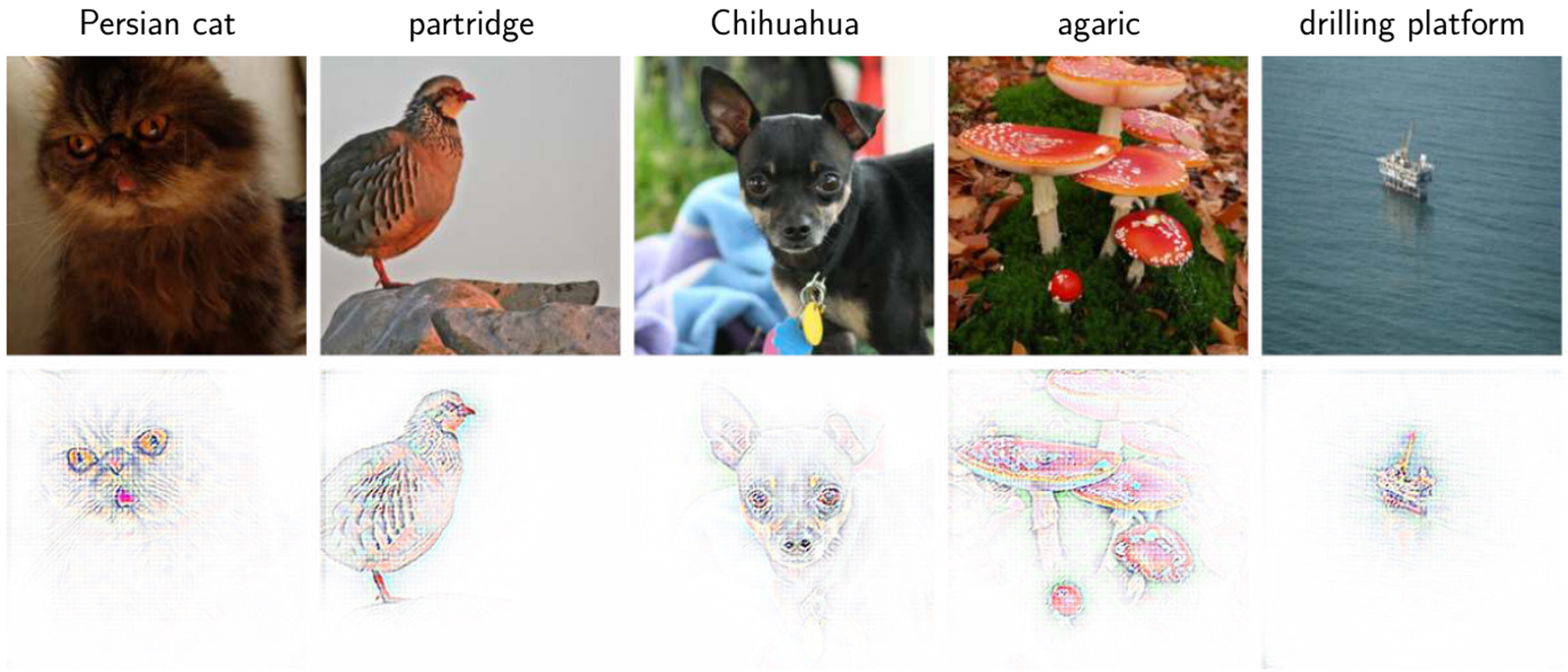
→ they can be reformulated as **dynamic linear** functions (=input-dependent linear functions)

$$\mathbf{f}^*(\mathbf{x};\theta) = \mathbf{W}_{1\to L}(\mathbf{x})\,\mathbf{x}\,,$$
$$\mathbf{W}_{1\to L}(\mathbf{x}) = \prod_{j=1}^{L} \widetilde{\mathbf{W}}_j(\mathbf{a}_j)$$



B-cos networks [Böhle et al., CVPR 2022]

B-cos Models: Interpretable Explanations



→ B-cos models produce **human-interpretable**, **faithful** **explanations**

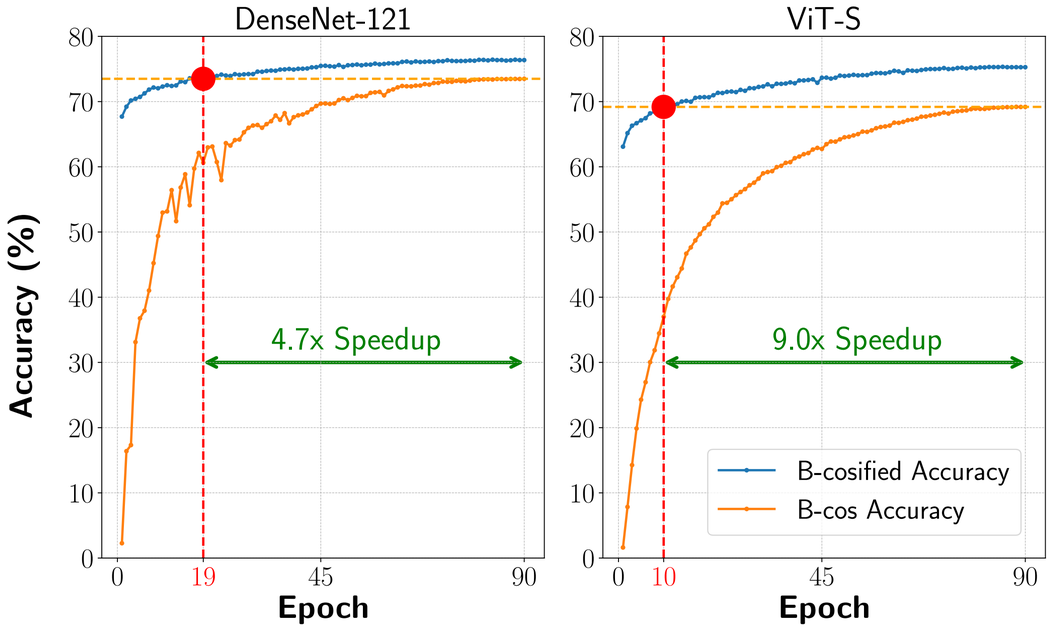
B-cos networks [Böhle et al., CVPR 2022]

B-cosification

B-cosification proposes **conversion of standard neural networks to B-cos models**

→ re-using trained weights shortens training time

→ **caveat**: **not applicable to most foundation models**



B-cosification [Arya et al., NeurIPS 2024]

ALign Once to Explain (ALOE)

We extend the approach of B-cosification with a layer-wise distillation loss, creating a

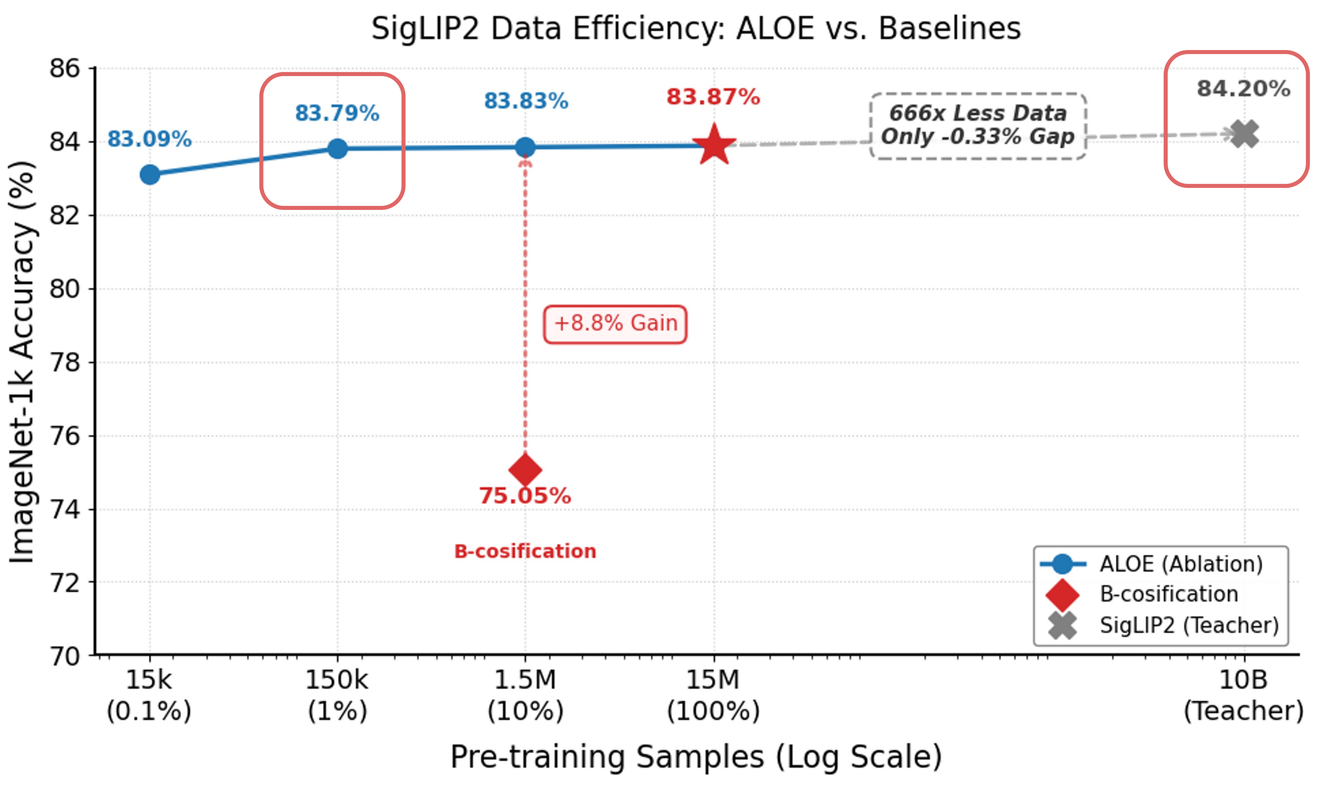
1. **Label-free**
2. **Model-agnostic**
3. **Efficient**

framework for **translating existing foundation models to B-cos models**

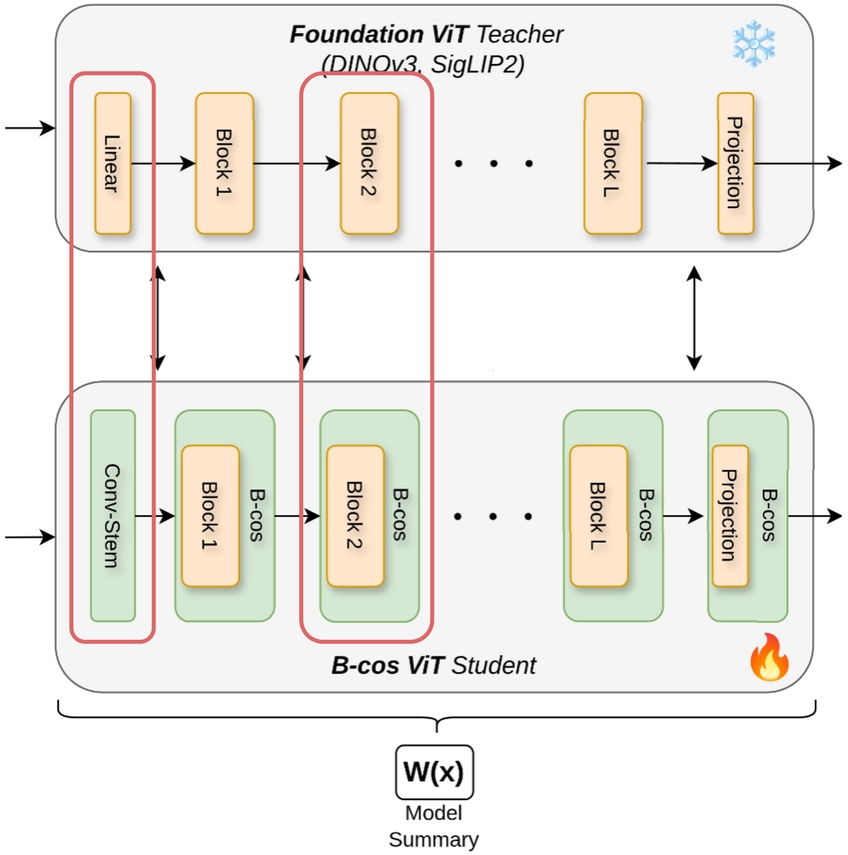
0.0015% of SigLIP2 pre-training data

=

<0.5% gap

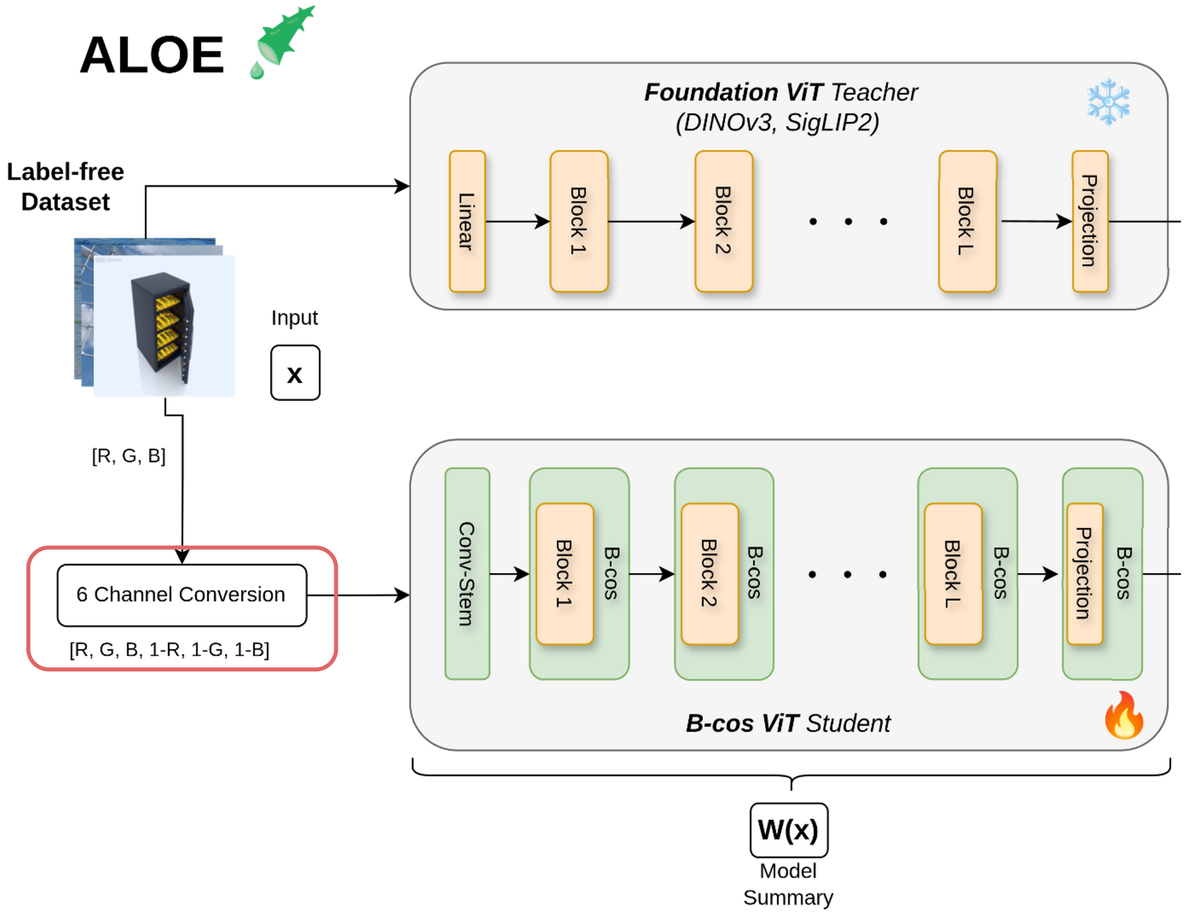


Step 1: Conversion

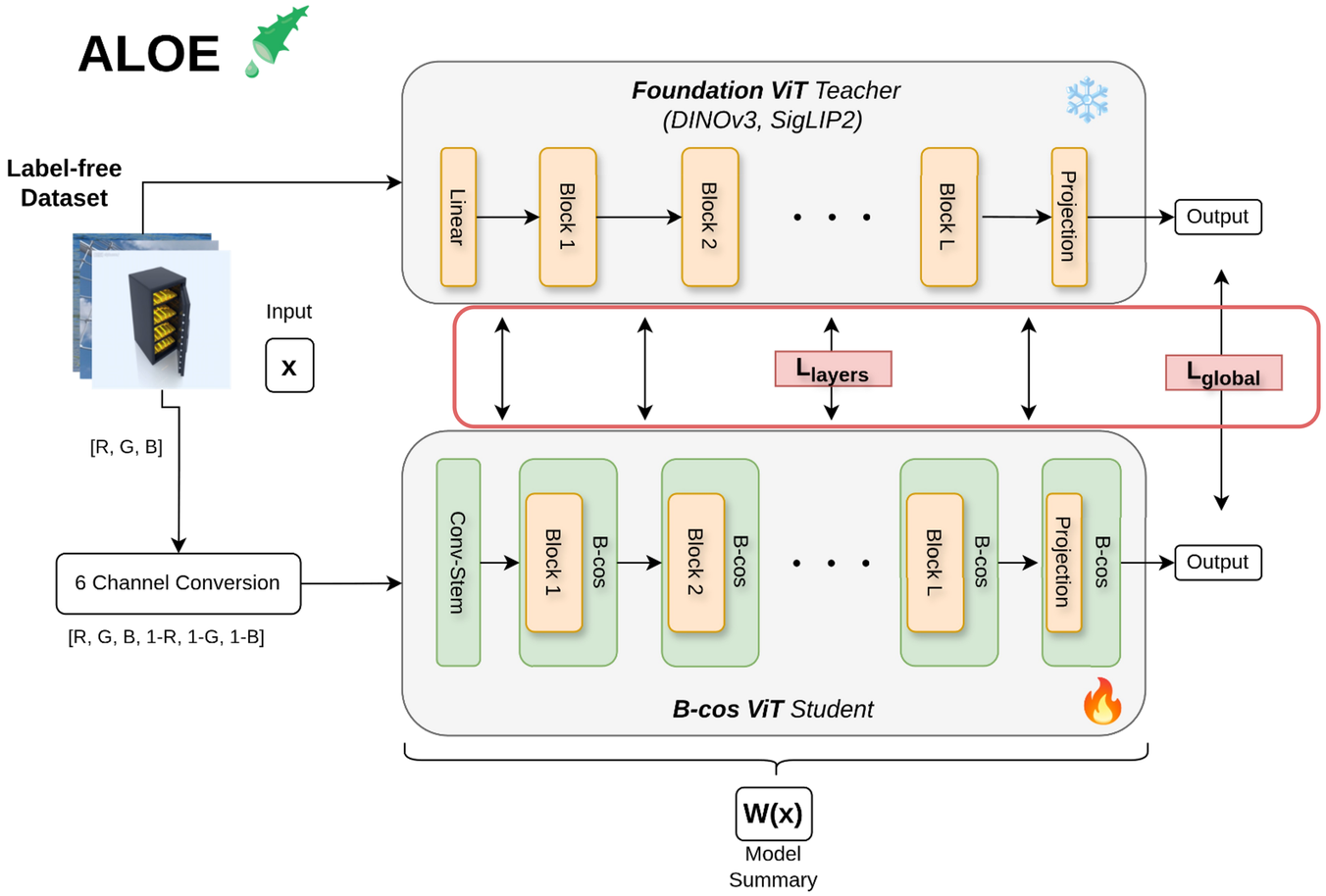


1. Replace **linear projection** by convolutional stem
2. Replace **linear layers** by B-cos layers
3. **Replace norms** by bias-free variant

Step 2: Distillation

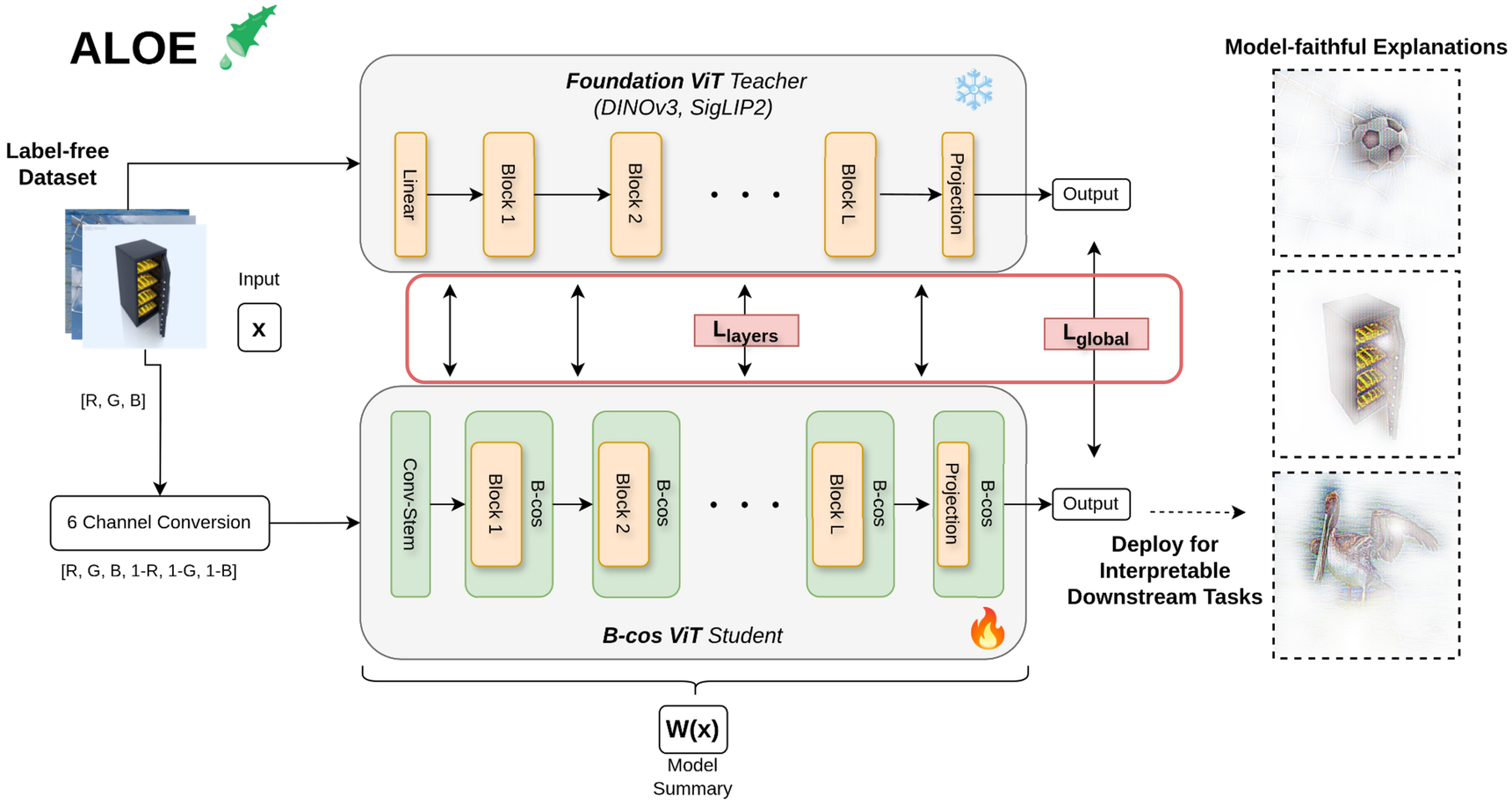


Step 2: Distillation



→ **minimize** mean **cosine distance** between feature embeddings

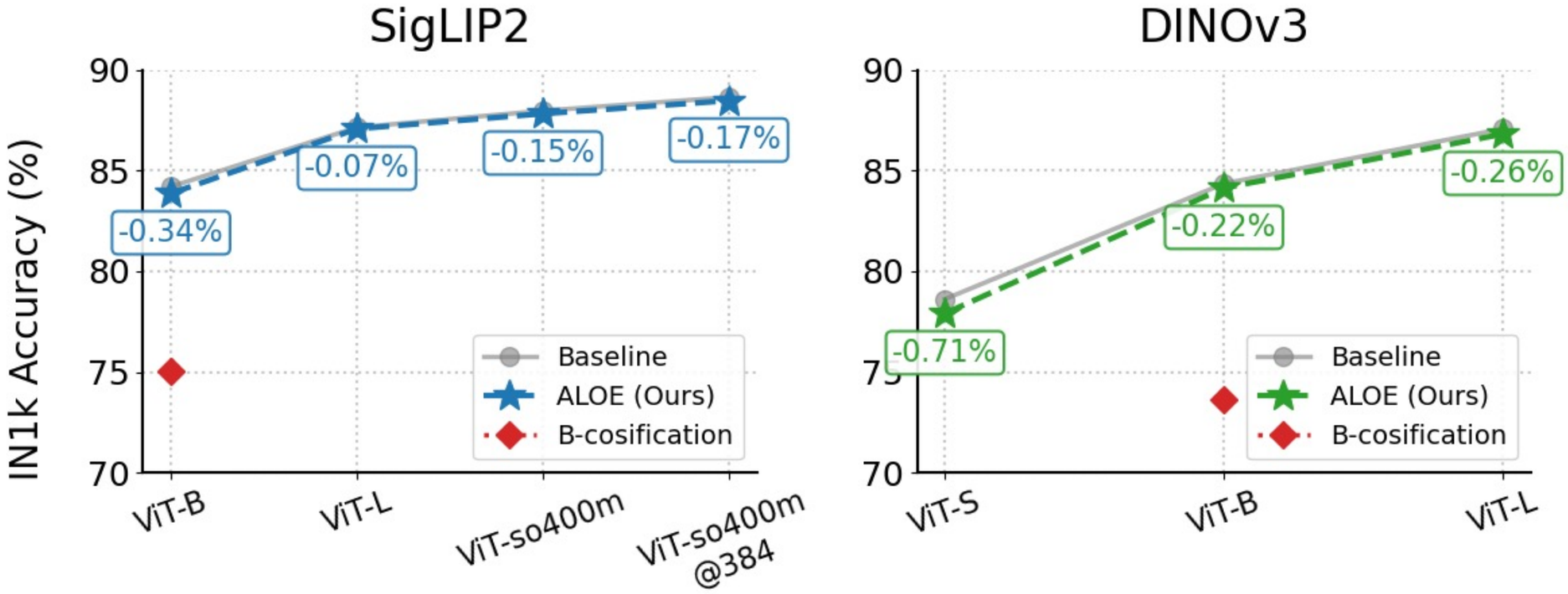
Step 2: Distillation



→ **minimize** mean **cosine distance** between feature embeddings

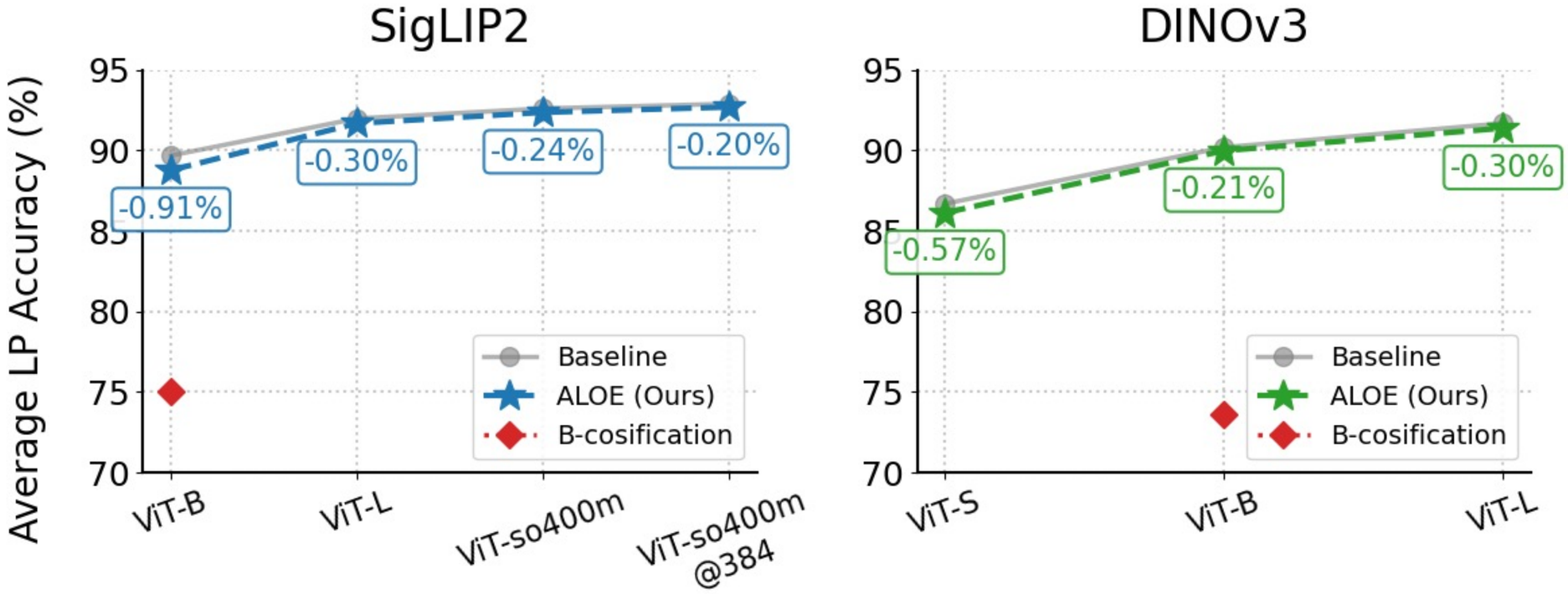
Model-agnostic and **label-free** distillation

Linear Probing (IN1k)



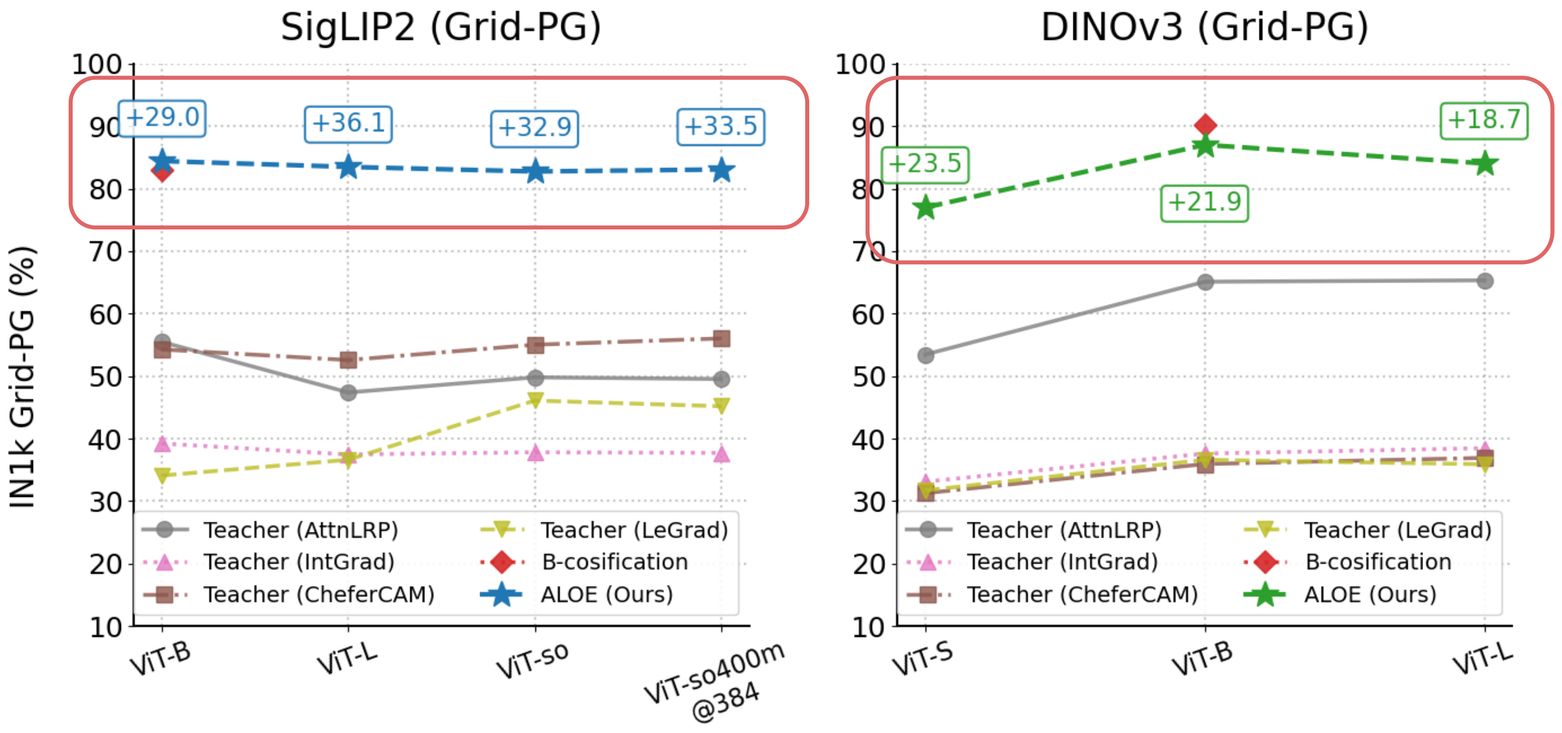
→ ALOE **nearly matches teacher performance**

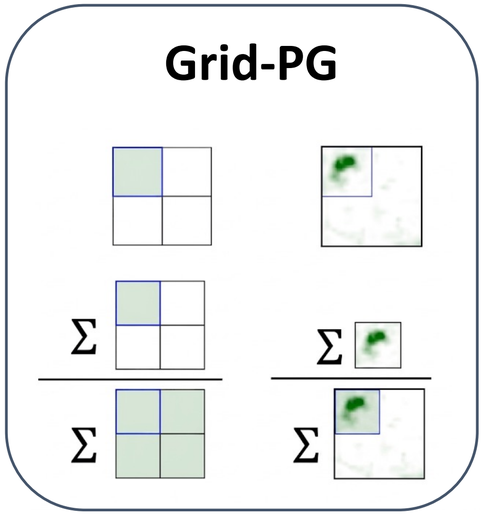
Linear Probing (all 10 datasets)



→ ALOE **nearly matches teacher performance**

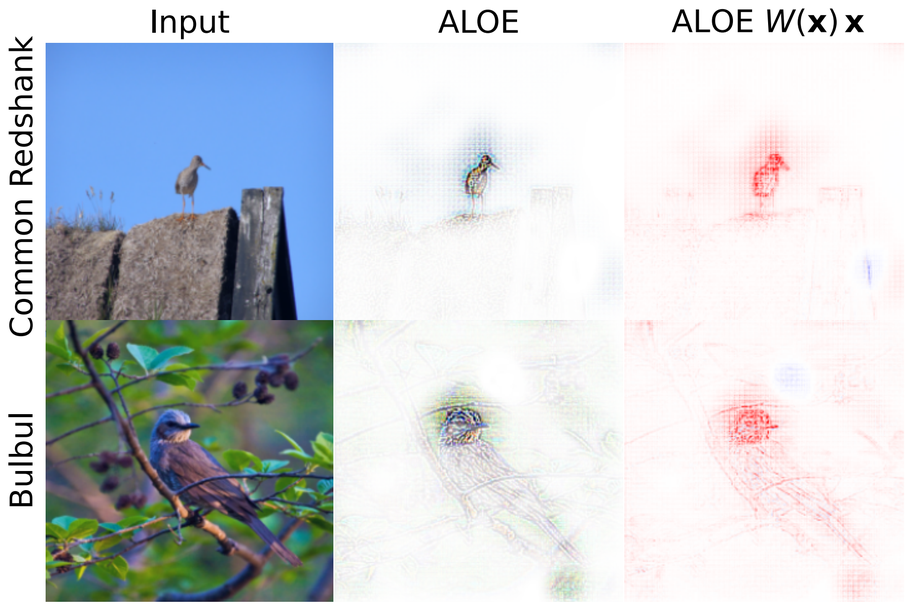
Localization: Grid-PG





→ ALOE **increases Grid-PG score substantially**

Qualitative Results: DINOv3



Qualitative attribution comparisons. Visualizations for ALOE (ours)—using model-inherent B-cos attributions W(x)x—versus popular post-hoc methods. **Positive contributions are red, negative ones blue.**

Takeaways

ALOE is a

1. **Label-free**
2. **Model-agnostic**
3. **Efficient**

framework for **translating existing foundation models to B-cos models**

Code

https://github.com/rmaser/ALOE

Models

https://huggingface.co/collections/rmaser/aloe