Feature Alignment for Scalable B-cosification of Foundational ViTs

Raphael Maser*, Siddhartha Gairola*, Sukrut Rao, Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
{raphael.maser,siddhartha.gairola,sukrut.rao,schiele}@mpi-inf.mpg.de

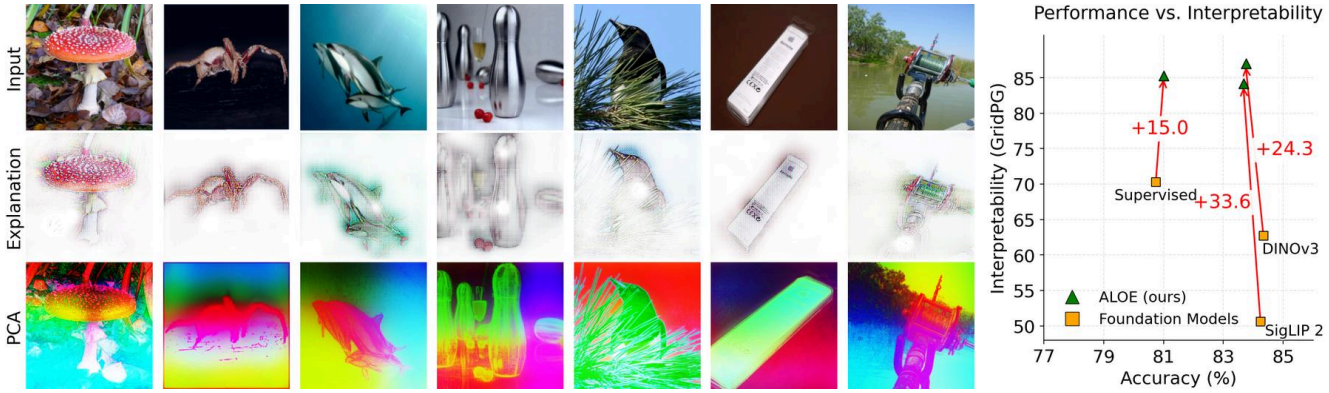

Figure 1. **ALOE: performant, preserves semantics, explains faithfully.**— *Left: Explanations and representations.* ALOE enables inherently interpretable B-cos: $\mathbf{W}(\mathbf{x})$ [10] attributions (**row. 2**) that are object-centric and class-specific. The accompanying PCA visualizations (**row. 3**; RGB = first three principal components of the final image representation for our ALOE aligned DINOv3 [67] model) preserves global feature geometry—indicating aligned semantics with improved explainability (*cf*. Fig. D1). *Right: Performance vs. interpretability.* Across ViT-B/16 backbones, ALOE substantially boosts localisation quality on GridPG [7] while maintaining competitive ImageNet accuracy relative to the corresponding foundation models (Supervised [22], DINOv3 [67], SigLIP2 [70]).

Abstract

Foundational vision models have become the de facto standard for many vision tasks due to their strong performance. However, they are notoriously opaque and remain hard to interpret. We present ***ALOE*** *(**AL**ign **O**nce to **E**xplain), a one-time, label-free feature alignment based approach that efficiently converts foundational vision models into inherently interpretable B-cos [10] variants. Once aligned, the B-cos backbone is used as a drop-in replacement across several downstream tasks—amortizing the cost of interpretability. ALOE is robust across pre-training paradigms (supervised, self-supervised, vision–language) and is* ***100–1000***$\times$ *more data-efficient than training from scratch. On classification, it outperforms fully-supervised B-cos models (e.g.,* ***+6.6 p.p.*** *top-1 on ImageNet for ViT-B/16), retains strong linear probing, k-NN, and zero-shot transfer performance competitive with foundational backbones (DINOv3 [67], SigLIP2 [70]) across diverse downstream datasets, while yielding well-localized and highly human interpretable explanations by design.*

1. Introduction

Vision foundation models, including self-supervised encoders [11, 51, 67] and contrastive vision-language models [55, 70, 74], are powerful feature extractors that achieve state-of-the-art performance across domains. However, their opaque decision-making hinders deployment in sensitive applications. Post-hoc explanation methods attempt to mitigate this, but often fail to faithfully reflect the model's true computations [3, 4, 56].

Inherently interpretable architectures, such as B-cos Networks [7, 10], offer faithful explanations by design. While recent "B-cosification" techniques [5] can efficiently retrofit existing CNNs into B-cos variants without full retraining, they provide only modest improvements for Vision Transformers (ViTs) [22]. Since most modern foundation models are ViT-based, this significantly limits their practical utility.

To bridge this gap, we propose **ALOE (ALign Once to Explain)** (Fig. 3), a scalable, label-free feature-alignment approach to efficiently convert ViT foundation models into performant B-cos variants. ALOE requires a single alignment phase, independent of the teacher's pre-training paradigm (supervised, self-supervised, contrastive). Once aligned, the backbone acts as a drop-in replacement for downstream tasks, amortizing the cost of interpretability.

ALOE is highly data- and compute-efficient. For instance, B-cosifying SigLIP2 [70] requires just $\approx$3M unlabeled images and 40 epochs—compared to its $\approx$10B image pre-training—while maintaining competitive ImageNet top-1 accuracy (83.67% vs. 84.24%; Fig. 1). Furthermore, ALOE consistently outperforms vanilla B-cosification [5] across pre-training paradigms by $>$ 4.9 p.p. (Fig. 2), de-

livering strong zero-shot performance and well-localized, faithful explanations.

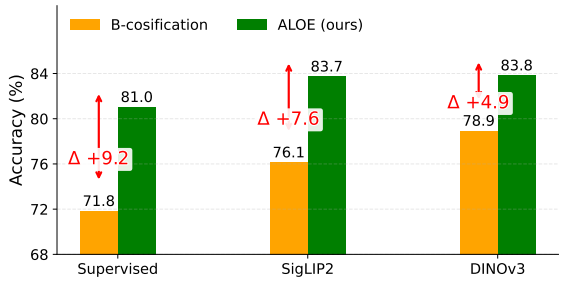

Figure 2. **ALOE vs. B-cosification [5].** ALOE (green) outperforms B-cosification (orange) on Supervised, SigLIP2, and DINOv3 paradigms. Arrows ↔ show absolute gains **Δ**.

Contributions & Findings:

- **Universal Alignment:** We introduce a one-time, label-free alignment recipe to transform foundational ViTs into interpretable B-cos models while preserving utility.
- **Design Insights:** We provide a comprehensive study of alignment targets, loss families (MSE, cosine, SigLIP, InfoNCE), model scales, and dataset sizes.
- **Generality & Efficiency:** ALOE approaches teacher-level generalization using $\sim$1000$\times$ fewer images than typical pre-training, significantly outperforming prior B-cosification methods on classification, linear probing, and zero-shot transfer.

Together, these contributions establish a practical path to scalable, faithful-by-design vision foundation models.

2. Background: B-cos networks

In this section we briefly review B-cos networks [10] (Sec. 2.1) and the B-cosification [5] (Sec. 2.2) recipe proposed for CNNs, before presenting our approach (Sec. 3).

2.1. B-cos networks

B-cos networks [10] are inherently interpretable architectures that provide faithful and human-interpretable explanations of model decisions. To do this, bias-free, dynamic-linear B-cos transforms are used in place of linear transforms that induce weight-input alignment across the model and provide faithful explanations that constitute an exact decomposition of the model's computation. The **B-cos transform** [7] is given by:

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

where B controls the alignment strength, cos is the cosine similarity between input $\mathbf{x}$ and weights $\mathbf{w}$, and $\widehat{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|_2$. Stacking such units builds a *dynamic-linear* network that produces an exact, faithful summary $\mathbf{y}(\mathbf{x}) = \mathbf{W}(\mathbf{x})\,\mathbf{x}$, where $\mathbf{W}(\mathbf{x})$ is the effective input-dependent dynamic linear weight from the full model. Increasing $B{>}1$ promotes weight–input alignment, making $\mathbf{W}(\mathbf{x})$ more task-relevant and interpretable. The cosine term raised to the power of $B{-}1$ provides the non-linearity needed for learning while preserving model expressivity. B-cos variants of standard CNNs (*e.g.*, B-cos ResNets [29]) can then be constructed—such models use B-cos transforms in place of convolutional and linear layers, remove activations, and remove all biases including in normalization layers. **B-cos ViTs** (Fig. 3) similarly replace linear layers in the patch embedding, MLP blocks, and projection heads with B-cos layers. Notably, self-attention, being already dynamic-linear [10], is left unchanged, as are positional embeddings.

Such models were shown to provide competitive performance while significantly improving interpretability [10], however, the need to train them from scratch remained a significant limitation to their adoption as compared to using already trained conventional models.

2.2. B-cosification recipe

To alleviate the need to retrain B-cos variants of existing models, 'B-cosification' [5] was proposed as a means to transform a pre-trained conventional model into its B-cos variant by making targeted architectural modifications followed by supervised fine-tuning for few epochs on the original task. Broadly, the transformation involves replacing linear transforms with corresponding B-cos transforms (Eq. (1)) with $B = 2$ and removing biases from all layers including normalization layers. In contrast to the original B-cos architecture [10], point-wise activation functions (ReLU/GELU) are left in and weights are not unit normalized, as it preserves the distribution of the learnt weights without harming interpretability. Following [10], 3-channel image inputs given by (r, g, b) are preprocessed to 6-channels $(r, g, b, 1{-}r, 1{-}g, 1{-}b)$ to allow visualizing explanations in color, with a transform to the first layer weights to maintain equivalence.

However, this recipe was designed with a focus on supervised CNNs, and was relatively ineffective for ViT-based architectures. In our work, we propose a scalable *feature-alignment* approach that is highly effective for modern ViT-based foundation models.

3. ALOE: ALign Once to Explain

In this section, we introduce **ALOE** (*ALign Once to Explain*), a *one-time, label-free feature-alignment* procedure that converts any ViT-based foundation encoder into a B-cos [10] counterpart and aligns it to a frozen teacher via simple representation matching (Fig. 3). We first describe the *teacher–student* setup and the architecture-preserving transformation used to initialize the student (Sec. 3.1). Next, we detail what we align and where in the network, including

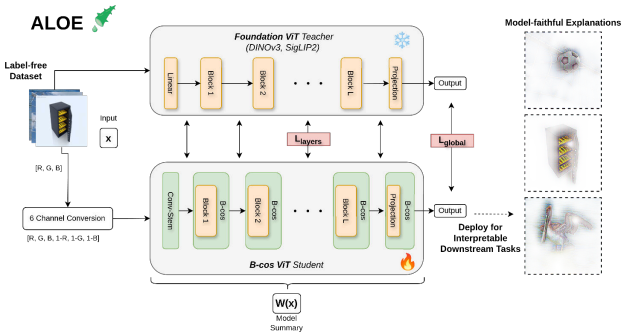

Figure 3. **Align Once to Explain.** *(1) Transform:* Convert a Foundation ViT into a bias-free, dynamic-linear B-cos backbone (Sec. 3.1). *(2) Align:* Label-free cosine feature alignment between frozen teacher and student. *(3) Deploy:* Freeze aligned backbone for downstream tasks. Faithful explanations emerge directly from $W(\mathbf{x})$, amortizing interpretability across tasks. (For high-resolution image see Fig. B2)

ViT-specific tokens (e.g., `[CLS]` and registers) and intermediate layers used to guide alignment (Sec. 3.2). Finally, we describe our setup (Sec. B.1) and show how the aligned, inherently interpretable backbone can be used as a drop-in component across downstream tasks, thereby amortizing the cost of interpretability (Sec. B.2).

3.1. Teacher–student setup

B-cos conversion. Given a frozen *teacher* encoder $\mathcal{T}$ (supervised, self-supervised, or vision–language) we construct a B-cos *student* $\mathcal{S}$ by applying the architecture-preserving transformation recipe ([5], see Sec. 2.2), which preserves functional behavior where possible with minimal edits needed for inherent interpretability.

Preserving special tokens. Unlike CNNs, modern ViTs use special tokens such as `[CLS]` and register tokens (*e.g.*, in DINOv3) that are crucial for performance. Given this, we keep these special-token pathways exactly as in the teacher so subsequent alignment can match tokens one-to-one and preserve the base model's computational routing[1].

These steps yield a bias-free, 6-input channel B-cos student that architecturally mirrors the teacher and is ready for label-free alignment.

3.2. Label-free alignment

After conversion (Sec. 3.1), we *align* the B-cos student $\mathcal{S}$ to the frozen teacher $\mathcal{T}$ using unlabeled images and a cosine-similarity objective. Our aim is to make the aligned B-cos backbone a *drop-in* substitute for the foundation teacher model. This motivates (i) a *global* guidance term to preserve the geometry of the final embedding space—critical for downstream tasks (classification, segmentation, zero-shot transfer) and (ii) *token-level, depth-wise* guidance to preserve intermediate computations and enhance stability during optimization. Therefore, we propose a multi-layer objective that is applied across selected depths of the two models during training.

Alignment objective. For an input $\mathbf{x}$, let $E_\star(\mathbf{x})$ denote the model's last-layer image representation, and let $h^\ell_{\star,t}(\mathbf{x})$ be the token-level features at transformer block ℓ for token t. We combine a global term with layer depth-wise guidance:

$$\mathcal{L} = \lambda_{\mathrm{g}}\,\mathcal{L}_{\mathrm{global}} \;+\; \lambda_{\mathrm{l}}\,\mathcal{L}_{\mathrm{layers}}\,, \tag{2}$$

$$\mathcal{L}_{\mathrm{global}} = \mathbb{E}_{\mathbf{x}\in\mathcal{B}}\Big[1 - \cos\big(E_{\mathcal{S}}(\mathbf{x}),\, E_{\mathcal{T}}(\mathbf{x})\big)\Big], \tag{3}$$

$$\mathcal{L}_{\mathrm{layers}} = \sum_{\ell\in\mathcal{L}_{\mathrm{depth}}} \mathbb{E}_{\mathbf{x}\in\mathcal{B},\, t\in\mathcal{T}^\ell_{\mathrm{tok}}(\mathbf{x})}\Big[1 - \cos\big(h^\ell_{\mathcal{S},t}(\mathbf{x}),\, h^\ell_{\mathcal{T},t}(\mathbf{x})\big)\Big], \tag{4}$$

where $\mathcal{T}^\ell_{\mathrm{tok}}(\mathbf{x})$ is the set of teacher tokens at depth ℓ.

We use cosine as our default alignment loss because it is scale-invariant and robust across teacher sizes and pre-training paradigms. Feature scales vary widely in practice; cosine normalizes this mismatch and directly optimizes angular agreement, which aligns well with objectives used during pre-training [67, 74]. We also experimented with compatible objectives (MSE, SigLIP, InfoNCE) and found cosine to be the most stable across pre-trained teachers; ablations are reported in Sec. C.3. In contrast, MSE is sensitive to absolute scale, and contrastive variants (InfoNCE/SigLIP) introduce batch negatives that likely distort the teacher's local geometry.

Targets and depth-wise guidance. We supervise at three evenly spaced depths $\mathcal{L}_{\mathrm{depth}} = \{\lfloor L/3\rfloor, \lfloor 2L/3\rfloor, L\}$ for a transformer of depth L (i.e., $1/3$, $2/3$, and full). We align exactly the semantics-carrying tokens used by each teacher (*e.g.*, `[CLS]`/registers for DINOv3; attention pooling for SigLIP2; Tab. B1), ensuring one-to-one routing and preserving the teacher's computational pathways—crucial for faithful B-cos explanations. .

This multi-scale supervision improves optimization stability and final alignment. Because the student $\mathcal{S}$ mirrors the teacher $\mathcal{T}$ in width, no projection heads are required. In practice, $(\lambda_{\mathrm{g}}, \lambda_{\mathrm{l}}) = (1, 1)$ works well across all models. We ablate the design choices—including the alignment objective and feature depths—to select the final configuration empirically (Sec. C.3).

4. Results

In the following we present our results. In Sec. 4.1, we report performance of ALOE as compared to baselines on the metrics described in Sec. B.2. We then provide detailed ablations on alignment objectives, feature depths, model scales and data size are in Sec. C.3.

[1] SigLIP2 [70] image encoders do not use `[CLS]` or register tokens.

Table 1. **Linear-probe accuracy.** Teachers in gray; best per block in **bold** (for B-cos models). ✓/✗: interpretable/not.

Feature	Arch	Inh. Int.	IN1k	Avg.
Fully Supervised Pre-Training				
Sup. [22]	ViT-B/16	✗	80.74	79.13
B-cosif. [5]	B-ViT-B/16	✓	71.76	66.99
ALOE (ours)	B-ViT-B/16	✓	**81.00**	**80.23**
			+9.24	+13.24
Vision Language Pre-training				
SigLIP2 [70]	ViT-B/16	✗	84.24	89.63
B-cosif. [5]	B-ViT-B/16	✓	75.84	80.86
ALOE (ours)	B-ViT-B/16	✓	**83.67**	**88.48**
			+7.83	+7.62
Self-Supervised Pre-training				
DINOv3 [67]	ViT-B/16	✗	84.34	90.25
B-cosif. [5]	B-ViT-B/16	✓	78.86	73.68
ALOE (ours)	B-ViT-B/16	✓	**83.75**	**89.50**
			+4.89	+15.82

Table 2. **Zero-shot IN-1k (SigLIP2 B/16).** ALOE substantially improves over vanilla B-cosification and closely matches the teacher.

Method	Top-1	Top-10
SigLIP2	69.2	92.1
B-cosif.	55.2	88.0
ALOE	**68.1**	**91.6**

4.1. Main results

Linear evaluation of frozen features. We compare ALOE to (i) the original foundational models, (ii) B-cos (from scratch) [10], and (iii) B-cosification [5]. All use the same linear protocol and input resolution (see Sec. B.2).

In Table 1, we report linear probing results on frozen features on IN-1k and averaged across 10 datasets ([16]). ALOE improves markedly over B-cos (from scratch) and B-cosification across ten datasets. For ViT-B/16, average gains vs. B-cosification are: **+13.24** p.p. (from 66.99% → 80.23%) for fully supervised; **+7.62** p.p. (80.86% → 88.48%) for SigLIP2; and **+15.82** p.p. (73.68% → 89.50%) for DINOv3. ALOE also closely matches its teachers (e.g., SigLIP2 average 88.48% vs. teacher 89.63%; DINOv3 89.50% vs. 90.25%). On IN-1k linear probing, we obtain state-of-the-art B-cos ViT results: **83.67%** for B-ViT-B/16, **85.65%** for B-ViT-L/16, and **87.00%** for B-ViT-so/16.

Zero-shot classification. ALOE preserves strong zero-shot transfer and outperforms B-cosification. On ImageNet-1k (zero-shot@1 setting) with ViT-B/16, ALOE achieves **68.12%** vs. B-cosification 55.16%; while competitive with the teacher SigLIP2 model, that gets **69.26%** (See Tab. 2).

Interpretability. We evaluate explanation quality for: (a) *model-inherent* B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$ from our ALOE-aligned models, and (b) *AttnLRP* [1] on conventional models. Following [10], we report GridPG (higher is better). Across supervised, SigLIP2, and DINOv3 teachers, ALOE yields higher localization scores than AttnLRP (see Fig. 1, right). For *e.g.*, for the SigLIP2 model, ALOE attains **84.2%** vs. teacher (AttnLRP) **50.6%**, and is competitive with from-scratch B-cos and B-cosification.

In Fig. 4, we show qualitative explanation maps, ALOE (B-cos: $W(x)$) and teacher (AttnLRP) for a DINOv3 ViT-B/16 model. The explanations for ALOE (**row 2**), appear to align with class-discriminant patterns in the input and resemble the class objects. This is a result of alignment pressure during optimisation (*cf*. [7]). In contrast the AttnLRP explanations are sparse and noisy (**row 3.**).

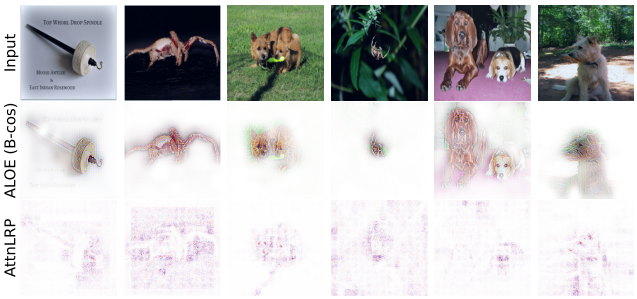

Figure 4. **Explanation quality.** Comparison between model inherent (B-cos: W(x) [10]) explanations after ALOE vs a popular strong method for visualizing vision transformers, AttnLRP [1].

In Fig. 5, we show the zero-shot explanations for ALOE ViT-B/16 models aligned with SigLIP2. We can then use our model as an explainable visual backbone, by combining it with the corresponding text encoder. We use prompts as "A photo of {class_name}"(*cf*. [5]), to compute similarity of text features with the visual features. The explanations are highly localized and class specific, thus we realise inherently-explainable vision language models.

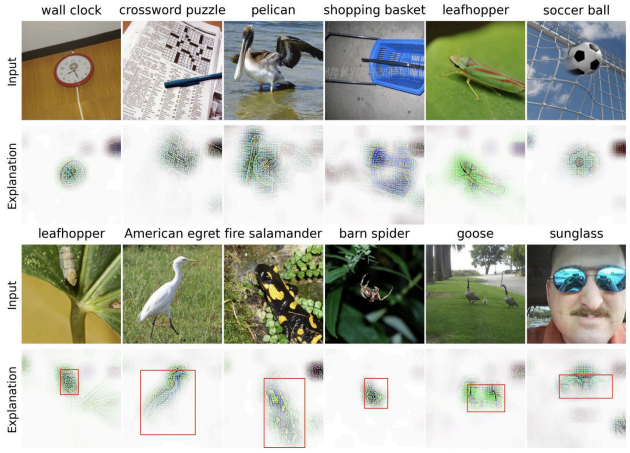

Figure 5. **VLM Zero-Shot Explanations.** ALOE aligned B-cos models (SigLIP2 teacher) yield zero-shot model-inherent explanations for VLMs, where the input prompt to the text encoder is "A photo of {class_name}"(*cf*. [5]). The explanations are visually well aligned with class-specific features (see bboxes in **row 4**).

Feature Alignment for Scalable B-cosification of Foundational ViTs

Appendix

Table of Contents

A. Related work

Large-scale vision foundation models trained on billion-scale datasets are de facto backbones for transfer learning and zero-shot tasks. These include self-supervised ViT encoders in the DINO family [11, 51, 67] and contrastive vision–language encoders such as CLIP [55] and SigLIP/SigLIP2 [70, 74]. Supervised ViTs pre-trained on large datasets (e.g., ImageNet21k/22k [35] or proprietary corpora) also serve as strong backbones for downstream usage [22, 35]. Such encoders are also used by large vision-language models and generative models together with a pre-trained LLM for multimodal understanding [19, 40]. In this work, we transform modern vision foundation models into an inherently interpretable B-cos backbone via a single alignment step, while preserving transfer learning and zero-shot utility.

Inherently interpretable models. To understand deep neural networks (DNNs), post-hoc attribution methods spanning gradient- [6, 14, 65, 68] activation- [13, 32, 62, 71], and perturbation-based [54] methods have been typically used, with the resulting explanation maps summarizing input regions contributing to the model's decision. In contrast, inherently interpretable architectures enforce structural constraints to give human-understandable, model-faithful summaries of the computation. These include, include prototype-based models [15, 21, 47], dynamic-linear models [7, 9], and concept-bottleneck models [33, 50, 57]. B-cos networks [10] replace linear layers with bias-free B-cos transforms that promote weight–input alignment, producing faithful, well-localized explanations by design and have gained recent prominence [24, 52]. The cost of training such models from scratch at foundation scale motivated B-cosification [5], which transforms existing networks to B-cos variants post hoc. While B-cosification showed gains on supervised CNNs and for CLIP ResNet model, the transformation recipe falls short in performance for ViTs, limiting applicability to foundation models. We therefore propose a scalable approach to transform ViT-based foundation backbones into inherently interpretable B-cos variants.

Knowledge distillation (KD) transfers behavior from a pre-trained model ('teacher' $\mathcal{T}$) to a smaller model ('student' $\mathcal{S}$) using soft targets or intermediate features, often for compression. Logit-based KD [30] which aligns output distributions between $\mathcal{T}$ and $\mathcal{S}$ has been extended with feature-level and attention-based alignment objectives [53, 59, 73], layer-wise and multi-stage schemes [76], and token-level distillation for ViTs [69]. Recent works [52] also leverage explanation map similarity objectives which improve faithfulness over logit-only KD, and improve robustness under distribution shifts. To reduce reliance on annotations, label-free KD techniques use unlabeled or synthetic data [48, 64]. In contrast to compression as a goal, we use a one-time, label-free *feature alignment* objective—akin to KD—to convert a frozen vision foundation model into an inherently interpretable B-cos counterpart. Our universal recipe retains the teacher's general-purpose features, provides model faithful explanations, and requires orders of magnitude ($\sim$100–1000$\times$) fewer images than full billion-scale pre-training.

B. Implementation Details

In this section we provide additional implementation details that complement the main paper (Sec. B.2), including details about the teacher checkpoints, datasets used for downstream evaluations, and finally the attribution methods along with the interpretability metrics we use.

B.1. Implementation and training setup

B-cos conversion An illustration of the B-cos conversion described in Sec. 2.2 and Sec. 3.1 is shown in Fig. B1. All layers are replaced by their bias-free B-cos equivalent, including linear projections, attention mechanism and norm layers.

Datasets. We use unlabeled, web-scale image sets CC3M [63], CC12M [12], YFCC15M [20] for alignment. The input resolution follows the teacher's default (typically 224$\times$224). For our main results(Tab. 1), we report numbers on models trained on YFCC15M.

Teachers and students. Teachers $\mathcal{T}$ are strong ViT-based foundation encoders spanning three pre-training paradigms: (i) *Supervised* ViT-B/16 [22]; (ii) *DINOv3* ViT-B/16 [67]; (iii) *SigLIP2* ViT-B/16, ViT-L/16, and ViT-so/16 [70]. B-cos students $\mathcal{S}$ are constructed by the architecture-preserving conversion recipe (Sec. 2.2) and initialized to mirror $\mathcal{T}$ in width, depth, and tokenization, enabling one-to-one alignment (Sec. 3.2).

Teacher checkpoints. For our feature-alignment step (Sec. 3.2) with pre-trained vision foundational models, we align to frozen teacher encoders of publicly available checkpoints and keep the associated text encoders for vision-language models unchanged. Table B2 lists the exact model IDs and native image resolutions we use.

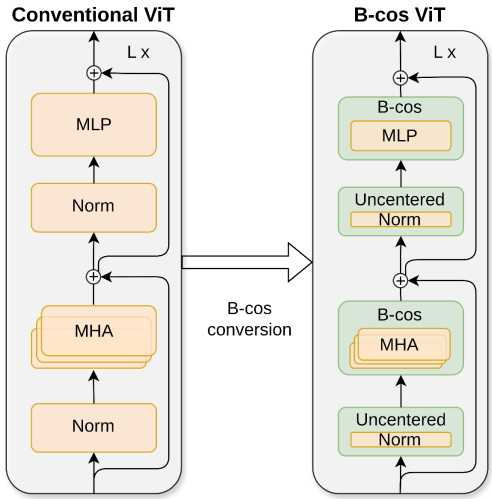

Figure B1. **Conventional vs. B-cos ViT.** We keep self-attention unchanged but replace all linear/MLP layers with *bias-free* B-cos transforms and use uncentered normalization, while preserving the residual topology. This architecture-preserving change yields a dynamic-linear B-cos ViT whose decisions admit model-intrinsic, faithful explanations via $\mathbf{W}(\mathbf{x})\mathbf{x}$ (Sec. 2.2).

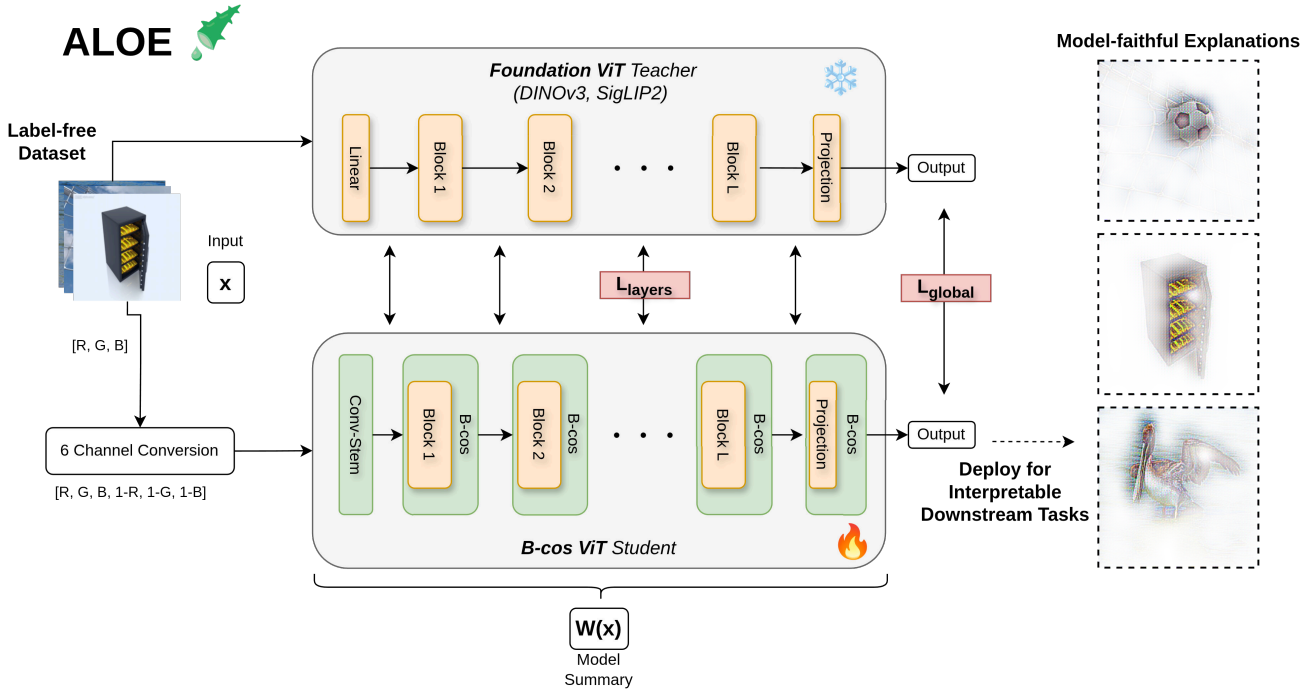

Figure B2. **Align Once to Explain.** High-resolution version of Fig. 3.

Tokens used for alignment Details about the tokes used for alignment are provided in Tab. B1

Table B1. **Tokens used for alignment (Sec. 3.2) for each model.**

Model	Global Feature	Layer-wise Features
Supervised [22]	Final `[CLS]`	Image tokens, `[CLS]`
DINOv3 [67]	Final `[CLS]`	Image tokens, `[CLS]`, register tokens
SigLIP2 [70]	Attention-pooled final embedding	Image tokens

Table B2. **Teacher encoders used for alignment.** We adopt each teacher's native evaluation resolution and keep their weights frozen during alignment. Identifiers are provided without URLs to comply with the submission policy for CVPR 2026.

Family	Architecture	Eval res.	Identifier
Fully Supervised Models [28]			
Supervised [22]	ViT-B/16	224×224	`google/vit-base-patch16-224`
Vision–Language Models [25–27]			
SigLIP2 [70]	ViT-B/16	224×224	`google/siglip2-base-patch16-224`
SigLIP2 [70]	ViT-L/16	256×256	`google/siglip2-large-patch16-256`
SigLIP2 [70]	ViT-so/16	256×256	`google/siglip2-so400m-patch16-256`
Self-Supervised Models [44–46]			
DINOv3 [67]	ViT-S/16	224×224	`facebook/dinov3-vits16-pretrain-lvd1689m`
DINOv3 [67]	ViT-B/16	224×224	`facebook/dinov3-vitb16-pretrain-lvd1689m`
DINOv3 [67]	ViT-L/16	224×224	`facebook/dinov3-vitl16-pretrain-lvd1689m`

Input resolutions. Alignment and evaluation follow the teacher's native resolution (Table B2): 224×224 for Supervised ViT-B/16 and all DINOv3 models; 224×224 for SigLIP2-B/16; 256×256 for SigLIP2-L/16 and SigLIP2-so/16.

Augmentations. We apply standard ViT augmentations (rand. resized crops, horizontal flips) to prevent overfitting while preserving $\mathcal{T}$-$\mathcal{S}$ feature geometry during alignment.

Optimization. We freeze $\mathcal{T}$ and optimize $\mathcal{S}$ with AdamW [41] and a cosine learning-rate schedule (mixed precision enabled). We follow the early stopping criterion to train until the loss on a held out validation set stops decreasing; fix $B{=}2$, additive biases remain zero, and use no explicit weight normalization (cf. Sec. 2.2). Gradient norm clipping of 1.0 and weight decay of $1{\times}10^{-2}$ are used to stabilize large models. We sweep learning rates over $\{3e^{-3}, 1e^{-3}, 5e^{-4}\}$ and select the model with the lowest alignment loss on the held-out split (randomly sampled 30k subset of images from the training set). We use a batch-size of 1024 for all experiments.

Loss arguments. Unless specified: $(\lambda_g, \lambda_l) = (1, 1)$; cosine-similarity alignment loss; depth supervision at $\{\lfloor L/3 \rfloor, \lfloor 2L/3 \rfloor, L\}$. We ablate alignment objectives, feature depths, model sizes and dataset scales in Sec. C.3.

B.2. Evaluation protocols and downstream usage

Train minimally, explain everywhere. Once aligned, the B-cos backbone is used as a drop-in encoder across tasks; explanations are model-inherent (derived from $\mathbf{W}(\mathbf{x})$ [10]), requiring no task-specific interpretability tuning.

Evaluation datasets. Linear Probing (LP) and k-NN share the same 10 datasets: IN1K [60], CALTECH101 [39], FLOWERS102 [49], FOOD101 [8], FGVC-AIRCRAFT [43], DTD [18], STANFORD CARS [36], SUN397 [72], CIFAR-10 [37], CIFAR-100 [37]. Dense linear probing for depth uses NYUV2 [38, 66].

Linear probing (LP). We evaluate aligned representation quality of $\mathcal{S}$ by training a linear head on frozen features, reporting top-1 accuracy on the validation sets of 10 downstream classification datasets (including IN1k) following standard protocol [51] (see Sec. 4.1). To speed up evaluation we do *not* apply augmentations during probing.

k-NN evaluation. To assess feature quality without additional training, we use a weighted k-NN classifier on frozen embeddings. We pre-compute features on the training split and use $k{=}20$, a robust choice across datasets [11] and report accuracy (%) on the validation split(s).

Dense linear probing for depth. We train a *linear* depth head on frozen features, and optimize with an L1 objective on inverse depth plus a scale-invariant gradient prior, following the Probe3D [23] protocol. Inputs are resized to the teacher's native resolution and center-cropped; no test-time augmentation is used. We evaluate on the standard NYUV2 [66] split and report both **relative** and **absolute** metrics: $\delta_1 \uparrow$ (fraction of pixels for which the ratio of prediction to ground truth is < 1.25; higher is better) and RMSE$\downarrow$ (lower is better).

Zero-shot evaluation (SigLIP2 [70]). For the contrastive vision–language setting (SigLIP2), we replace only the *image* encoder with its ALOE-aligned B-cos counterpart as a *drop-in* replacement; the SigLIP2 *text* encoder, prompt templates, temperature, and normalization follow the originals [55, 70]. We report top-1 for a *single* class-name prompt ("a photo of a

{class-name}") and the OpenCLIP 80-prompt template ([17, 31]) on ImageNet [60], as well as standard benchmarks [55, 74].

Attribution methods visualized. For the inherently interpretable B-cos models [10], explanations are model-inherent $\mathbf{W}(\mathbf{x})\mathbf{x}$ [7]. For conventional teachers, we visualize AttnLRP [2] (using the original authors' implementation) and Input×Gradient [65], Integrated Gradients [68], GradSHAP [42] as well as LIME [58] (from the `captum` library [34]). Where applicable, we use authors' recommended defaults.

Evaluating explanations & Interpretability metrics. We evaluate model attributions with (i) the *Grid Pointing Game (GridPG)* [9, 75] for localization and (ii) *Pixel Deletion* [61] for faithfulness, following standard protocol from prior work.

GridPG. We build $N \times N$ grids (we use 2×2) from images of *distinct* classes that are individually and confidently correctly classified. For each class i, we measure the fraction of *positive* attribution mass inside its corresponding grid cell. Let $A(p)$ be the attribution at pixel p and $A^+(p) = \max(A(p), 0)$ its positive part; the localization score for cell i is

$$L_i = \frac{\sum_{p \in \text{cell}_i} A^+(p)}{\sum_{j=1}^{N^2} \sum_{p \in \text{cell}_j} A^+(p)},$$

and the GridPG score is the average of L_i over several grids (grids with zero total positive mass are discarded).

Pixel Deletion. We rank pixels by attribution scores from least to most important and iteratively set the least important pixels to zero, plotting the target-class probability versus the removed-pixel fraction; *smaller* drops (flatter curves) indicate attributions that are more consistent with model decisions.

Note. To comply with the CVPR 2026 submission policy, we omit all external URLs (including model hubs and code). Identifiers are provided for clarity; full links and artifacts will be added in the final version.

C. Additional Quantitative Results

In this section, we expand the quantitative evaluation along two axes: (1) *downstream performance*, and (2) *interpretability/faithfulness*. In Sec. C.1, for downstream performance, we report k-NN on frozen features (Tab. C2), scaling trends across model sizes (Fig. C1), zero-shot transfer for SigLIP2 (Tab. C3), dense linear probing for monocular depth (Tab. C4), and data-efficiency analyses (Fig. C2). In Sec. C.2, for interpretability and faithfulness, we quantify localization with GridPG (Tab. C5) and evaluate stability via pixel-deletion tests (Fig. C3). Unless noted otherwise, protocols match the main paper (see Sec. B.2) and Sec. B.

C.1. Downstream performance

LP on frozen features. Additionally to the average LP accuracy shown in Tab. 1 we provide the performance on all ten datasets in Tab. C1. Across all datasets ALOE substantially outperforms B-cosification while maintaining most of the original teacher's capabilities.

k-NN on frozen features. Across ten datasets, ALOE B-ViTs substantially outperform B-cosification while remaining competitive with their teachers (Tab. C2). Gains are especially pronounced on fine-grained or texture-heavy benchmarks (e.g., CARS, AIRCR, DTD, FOOD), indicating that alignment preserves discriminative structure in feature space without additional fine-tuning.

Model scaling. Performance improves monotonically with increase in model size (number of parameters) from ViT-S/16 to ViT-so/16 (Fig. C1). Both linear-probe (solid) and k-NN (dashed) curves rise with model scale under DINOv3 [67] and SigLIP2 [70] teachers, and the gap to the teacher narrows for larger models (also see Tab. C5), which suggests that alignment benefits improve with an increased model capacity.

Zero-shot (SigLIP2 [70]). Replacing the SigLIP2 image encoder with its ALOE B-cos counterpart preserves strong zero-shot classification performance and markedly outperforms B-cosification for both single-prompt ("A photo of a {class-name}") and OpenCLIP 80-prompt settings ([17, 31]), while staying close to teacher performance (Tab. C3).

Dense linear probing (depth estimation). On monocular depth estimation with ViT-B/16, ALOE approaches the teacher and surpasses B-cosification on both relative and absolute metrics (Tab. C4) by a large margin. This indicates that aligned B-cos features remain useful for dense prediction, not only global image classification.

Table C1. **Linear-probe accuracy on frozen features.** ALOE B-cos ViTs substantially outperform B-cosification while remaining competitive with the original foundation models on ImageNet-1k and on the 10-dataset average. All models use the same protocol and resolution. Teachers are shown in gray; best per block in **bold** (for B-cos models). ✓: denotes inherently interpretable models (vs. not ✗).

Feature	Arch	Inh.	Inter.	IN1k	Cal101	Flowers	Food	Aircr	DTD	Cars	SUN	C10	C100	Avg.
Fully Supervised Pre-Training														
Sup. [22]	ViT-B/16	✗		80.74	100.00	99.35	86.00	39.51	73.83	55.19	73.54	97.07	86.05	79.13
B-cosif. [5]	B-ViT-B/16	✓		71.76	99.51	78.64	65.12	34.83	59.77	40.28	53.89	91.48	74.61	66.99
ALOE (ours)	B-ViT-B/16	✓		**81.00**	**99.80**	**98.95**	**86.52**	**42.42**	**73.54**	**59.75**	**75.19**	**97.62**	**97.61**	**80.23**
				+9.24	+0.29	+20.31	+21.4	+7.59	+13.77	+19.47	+21.30	+6.14	+23.00	+13.24
Vision Language Pre-training														
SigLIP2 [70]	ViT-B/16	✗		84.24	99.93	98.31	94.43	75.48	85.21	95.43	81.67	96.89	84.66	89.63
B-cosif. [5]	B-ViT-B/16	✓		75.84	99.87	93.88	85.24	44.80	80.08	76.81	76.59	94.83	80.62	80.86
ALOE (ours)	B-ViT-B/16	✓		**83.67**	**99.90**	**99.09**	**92.6**	**70.05**	**82.52**	**94.36**	**81.33**	**96.77**	**84.54**	**88.48**
				+7.83	+0.03	+5.21	+7.36	+25.25	+2.44	+17.55	+4.74	+1.94	+3.92	+7.62
Larger Architectures														
SigLIP2 [70]	ViT-L/16	✗		87.15	100.00	99.22	96.41	83.01	86.33	96.33	84.39	97.80	87.57	91.19
SigLIP2 [70]	ViT-so/16	✗		87.85	100.00	99.74	96.90	82.78	86.13	96.75	84.75	98.55	89.34	91.65
ALOE (ours)	B-ViT-L/16	✓		85.65	99.90	99.48	95.24	75.42	82.52	95.52	83.29	98.24	89.31	90.46
ALOE (ours)	B-ViT-so/16	✓		**87.00**	**99.90**	**99.61**	**96.31**	**79.39**	**84.08**	**95.84**	**84.16**	**98.59**	**90.66**	**91.55**
Self-Supervised Pre-training														
DINOv3 [67]	ViT-B/16	✗		84.34	100.00	99.74	94.08	80.29	83.71	94.33	78.63	98.17	89.20	90.25
B-cosif. [5]	B-ViT-B/16	✓		78.86	99.80	82.29	74.33	44.32	71.48	54.70	61.41	92.71	76.87	73.68
ALOE (ours)	B-ViT-B/16	✓		**83.75**	**99.90**	**99.74**	**93.30**	**77.15**	**82.03**	**93.97**	**77.98**	**98.08**	**89.14**	**89.50**
				+4.89	+0.10	+17.45	+18.97	+32.83	+10.55	+39.27	+16.57	+5.37	+12.27	+15.82

Table C2. **k-NN accuracy on frozen features.** For the k-NN ($k = 20$) evaluation setting, ALOE B-cos ViTs again significantly outperform B-cosification [5] while remaining competitive with the original foundation models on ImageNet-1k and on the 10-dataset average. All models use the same protocol and resolution (see Sec. B.2). Teachers are shown in gray; best per block in **bold** (for B-cos models). ✓: denotes inherently interpretable models (vs. not ✗).

Feature	Arch	Inh.	Inter.	IN1k	Cal101	Flowers	Food	Aircr	DTD	Cars	SUN	C10	C100	Avg.
Fully Supervised Pre-Training														
Sup. [22]	ViT-B/16	✗		80.74	93.29	77.21	78.92	22.18	61.94	29.70	68.49	96.41	82.42	69.13
B-cosif. [5]	B-ViT-B/16	✓		71.18	83.50	36.58	36.09	12.14	37.21	12.23	31.08	82.15	56.92	45.91
ALOE (ours)	B-ViT-B/16	✓		**79.85**	**93.95**	**76.17**	**80.13**	**22.49**	**61.13**	**29.85**	**68.47**	**96.61**	**82.24**	**69.09**
				+8.67	+10.45	+39.59	+44.04	+10.35	+23.92	+17.62	+37.39	+14.46	+25.32	+23.18
Vision Language Pre-training														
SigLIP2 [70]	ViT-B/16	✗		80.41	98.44	79.95	93.22	65.23	77.06	92.47	75.91	95.18	79.46	83.73
B-cosif. [5]	B-ViT-B/16	✓		69.63	95.57	70.18	78.03	27.37	72.54	54.10	69.42	92.41	72.31	70.16
ALOE (ours)	B-ViT-B/16	✓		**79.69**	**98.14**	**76.69**	**90.53**	**57.68**	**74.12**	**90.33**	**75.48**	**95.00**	**78.07**	**81.57**
				+10.06	+2.57	+6.51	+12.50	+30.31	+1.58	+36.23	+6.06	+2.59	+5.76	+11.42
Self-Supervised Pre-training														
DINOv3 [67]	ViT-B/16	✗		82.29	95.70	79.69	91.3	58.71	77.68	88.96	72.42	97.35	85.61	82.97
B-cosif. [5]	B-ViT-B/16	✓		78.69	90.76	52.60	59.22	21.63	62.17	29.73	50.28	90.33	69.14	60.46
ALOE (ours)	B-ViT-B/16	✓		**81.22**	**95.90**	**80.99**	**89.87**	**54.04**	**76.07**	**87.75**	**72.00**	**97.46**	**86.22**	**82.15**
				+2.53	+5.14	+28.39	+30.65	+32.41	+13.90	+58.02	+21.72	+7.13	+17.08	+21.70

Data efficiency. Using only un-labeled CC3M [63], ALOE maintains nearly flat ImageNet top-1 linear-probe accuracy when the alignment data shrinks from 100% to 0.5% (about 15k images) , staying within $\sim$1 p.p.; see Fig. C2. This corresponds to using $\approx 0.0009\%$ of a $\sim$1.6B pretraining corpus (for DINOv3) while delivering comparable downstream performance, underscoring the strong data efficiency of our proposed one-time feature alignment procedure.

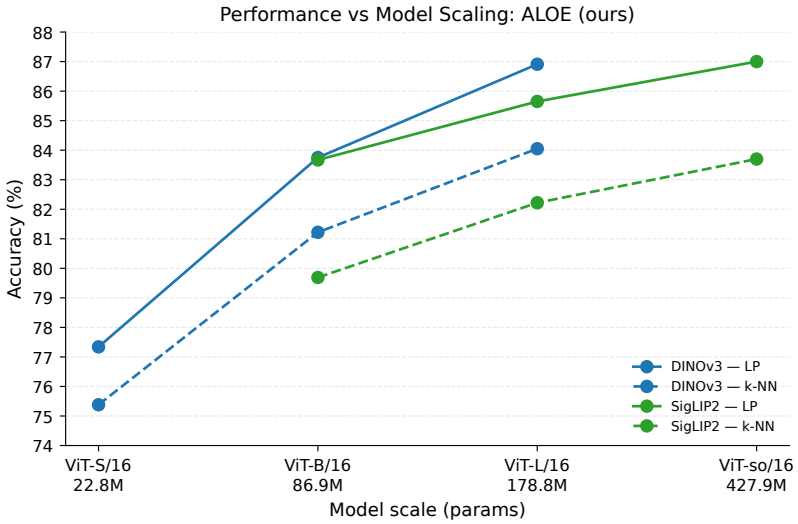

Figure C1. **Performance vs. model scale.** Linear-probe (LP, solid) and k-NN (dashed) ImageNet top-1 accuracy for ALOE-aligned B-cos ViTs under DINOv3 (blue) and SigLIP2 (green) teachers. Accuracy improves with model size from ViT-S/16 (22.8M) to ViT-so/16 (427.9M), and the gap to the teacher narrows at larger scales (also see Tab. C5).

Table C3. **Zero-shot ImageNet-1k with SigLIP2 prompts.** We replace the SigLIP2 image encoder with the ALOE-aligned B-cos counterpart and evaluate zero-shot classification with either a single class-name prompt ("A photo of a {class-name}") or the OpenCLIP 80-prompt template ([17, 31]). Values are ImageNet top-1 accuracy (%). Teachers are shown in gray; ✓ denotes inherently interpretable B-cos models (✗ are not). For ViT-B/16, ALOE substantially outperforms B-cosification and remains competitive with the teacher; similar trends hold for larger models. The Δ (**row 4.**) reports ALOE minus B-cosification.

	Architecture	Inh. Inter.	Single prompt	OpenCLIP (80 prompts)
	Base architecture			
	SigLIP2 [70] — ViT-B/16	✗	69.26	78.06
	B-cosif. [5] — B-ViT-B/16	✓	55.16	61.01
	ALOE (ours) — B-ViT-B/16	✓	**68.12**	**76.98**
	Δ *(ALOE vs. B-cosif.)*		+12.96	+15.97
	Larger architectures			
	SigLIP2 [70] — ViT-L/16	✗	72.65	82.27
	SigLIP2 [70] — ViT-so/16	✗	73.81	82.58
	ALOE (ours) — B-ViT-L/16	✓	70.55	79.93
	ALOE (ours) — B-ViT-so/16	✓	72.98	81.40

Table C4. **Dense linear probing for monocular depth (ViT-B/16).** We report relative and absolute depth metrics; higher is better for δ_1 and lower is better for RMSE. The Δ (**row 4.**) reports ALOE minus B-cosification.

Method	Inh. Inter.	Relative		Absolute	
		$\delta_1 \uparrow$	RMSE$\downarrow$	$\delta_1 \uparrow$	RMSE$\downarrow$
DINOv3 [67]	✗	0.9545	0.2776	0.8229	0.4534
B-cosif. [5]	✓	0.8311	0.4604	0.6503	0.6804
ALOE (ours)	✓	**0.9341**	**0.3222**	**0.7709**	**0.5071**
Δ *(ALOE vs. B-cosif.)*		+0.1030	-0.1382	+0.1206	-0.1733

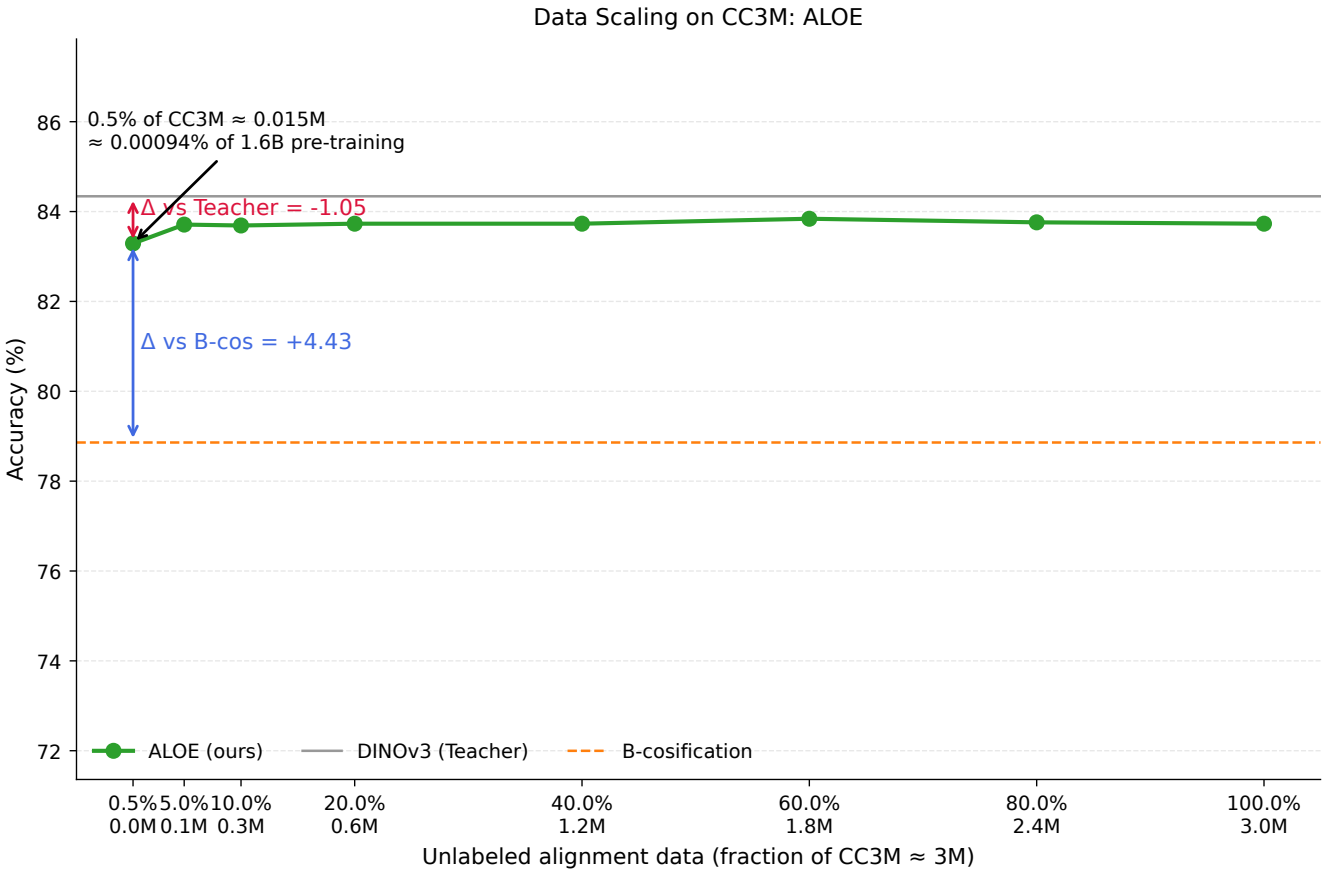

Figure C2. **Data efficiency of ALOE on CC3M for ViT-B/16.** ALOE's (green) linear-probe ImageNet top-1 accuracy as we vary the fraction of CC3M used for label-free alignment from 0.5% to 100% for DINOv3 teacher (shown in gray horizontal line). Performance is essentially flat from 0.5% ($\approx$15k images) $\rightarrow$100%, staying within $\sim$1.0 p.p., while using only $\approx$0.0009% of a $\sim$1.6 B-image pre-training image corpus used by DINOv3; the black arrow in the plot highlights the 0.5% point. For the same point ALOE gains **+4.43** p.p. over vanilla B-cosification [5] (orange).

Model	Linear probe ↑		k-NN ↑		Grid-PG ↑		Δ_{GridPG}
	Teacher (✗)	ALOE (✓)	Teacher (✗)	ALOE (✓)	Teacher (✗)	ALOE (✓)	
Vision–language teacher: SigLIP2 [70]							
ViT-B/16	84.24	83.67	80.41	79.69	50.60	**84.16**	+33.56
ViT-L/16	87.15	85.65	83.78	82.22	39.47	**78.84**	+39.37
ViT-so/16	87.85	87.00	84.48	83.70	38.75	**78.92**	+40.17
Self-supervised teacher: DINOv3 [67]							
ViT-S/16	78.63	77.34	76.93	75.38	54.10	**73.62**	+19.52
ViT-B/16	84.34	83.75	82.29	81.22	62.70	**87.01**	+24.31
ViT-L/16	86.99	86.91	84.75	84.05	64.97	**80.38**	+15.41

Table C5. **Localization (Grid-PG) vs. recognition.** ALOE improves Grid-PG substantially across backbones and teachers while maintaining competitive linear-probe and k-NN accuracy. Δ_{GridPG} is ALOE minus teacher (percentage points). ✓: denotes inherently interpretable models (vs. not ✗).

C.2. Interpretability and faithfulness

Localization (GridPG). ALOE aligned models yield large improvements in GridPG localization across backbones and teachers (SigLIP2, DINOv3) while keeping classification performance (Linear Probe, k-NN) competitive (Tab. C5). Relative to teachers (when using AttnLRP [2]) vs inherently interpretable B-cos explanations [10], we observe substantial GridPG gains (Δ_{GridPG}), thus demonstrating that aligned B-cos models provide more localized, class-specific attributions while not sacrificing downstream accuracy.

Faithfulness under pixel perturbation. In pixel-deletion tests on ViT-B/16 (Fig. C3), target-class probability for ALOE (using model-inherent $\mathbf{W}(\mathbf{x})\mathbf{x}$) degrades slowest as least important (lowest attribution scores) pixels are removed, outperforming AttnLRP, Integrated Gradients, Input×Gradient, and GradSHAP for both DINOv3 and SigLIP2. The flatter decay indicates more stable, faithful attributions that better reflect the model's decision computations.

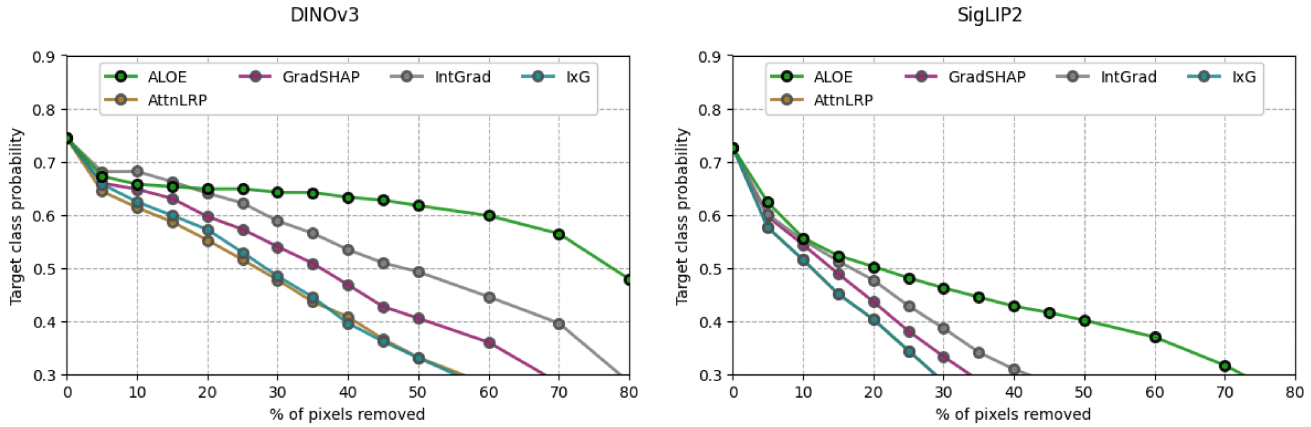

Figure C3. **Perturbation stability of explanations on ViT-B/16.** Target-class probability vs. percentage of top-attributed pixels removed (higher curves are better). Across both teachers—**DINOv3** (left) and **SigLIP2** (right)—**ALOE (ours)** using model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$ [10] degrades slowest and stays consistently above popular post-hoc methods (AttnLRP, Integrated Gradients, Input×Gradient, GradSHAP), indicating more stable and faithful localization.

C.3. Ablations

We ablate components: *alignment objectives*, *multi-layer feature alignment*, *model scale* and *dataset size*. For all the ablations, we only run the training for 40 epochs.

Alignment objectives. Under identical settings (Tab. C6), *cosine* and *SigLIP* are most consistent across models, *MSE* and

Table C6. **Alignment objective ablation.** IN1k top-1 (%).

Loss	**MSE**	**Cosine**	**SigLIP**	**InfoNCE**
DINOv3	83.9	**84.0**	83.7	83.5
SigLIP2	75.5	75.8	**76.0**	75.7
Google ViT	81.0	**81.1**	**81.1**	80.9

InfoNCE showed inconsistencies. Given the ease of application and simplicity we adopt **cosine** by default.

Table C7. **Feature depth ablation.** Avg. (LP) acc. for SigLIP2.

Pool	$+L$	$+\{2/3,\ L\}$	$+\{1/3,\ 2/3,\ L\}$	+All
75.51	77.85	85.24	**85.42**	84.93

Feature Depth. We ablate loss placement across depth(s): *global*-only; *global*$+L$; *global*$+\{\lfloor 2L/3 \rfloor,\ L\}$; *global*$+\{\lfloor L/3 \rfloor,\ \lfloor 2L/3 \rfloor,\ L\}$; and *global*+all. For SigLIP2 ViT-B/16 (Tab. C7), accuracy improves with increasing depth until $\lfloor 2L/3 \rfloor$; supervising all layers yields no further gain. We thus use 3 evenly spaced depths by default.

Table C8. **Model scale ablation.** IN1k and average (LP) accuracy.

Model	IN1k	Avg
B-cos ViT-B/16	83.67	88.48
B-cos ViT-L/16	85.65	90.46
B-cos ViT-so/16	**87.00**	**91.55**

Model scale. We align SigLIP2 ViT-B/16, ViT-L/16, and ViT-So400m/16 teachers (and their B-cos students). We see consistent gains in accuracy with increasing model size(Fig. C8). Larger models close the gap to the teacher foundation model.

Table C9. **Dataset scale ablation.** IN1k acc. for SigLIP2.

CC3M	CC12M	YFCC15M
83.42	83.49	**83.50**

Dataset size scaling. We vary alignment data from **3M** $\rightarrow$ **15M** samples (CC3M, CC12M, YFCC15M). For SigLIP2 ViT-B/16, ImageNet-1k accuracy increases minimally beyond CC12M (Tab. C9). Since we get highest accuracy on YFCC15M, we use that for our main experiments.

D. Additional Qualitative Results

In this section, we complement the quantitative results with qualitative evidence across three settings: (i) *zero-shot, model-inherent* explanations (Figure D2); (ii) side-by-side comparisons against popular post-hoc attribution methods (Figures D3 to D5); and (iii) dense predictions from a linear depth probe (Figure D6). We also compare PCA between DINOv3 and ALOE to demonstrate that the alignment process retains the feature geometry of the teacher model.

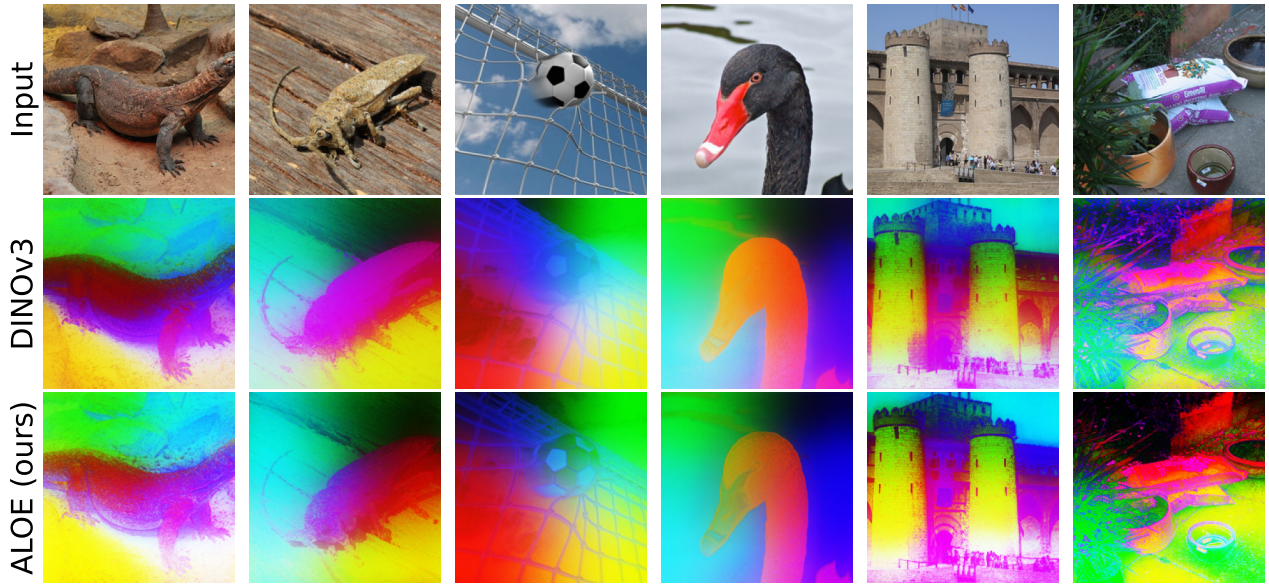

Figure D1. **PCA Visualization.** DINOv3 (**row 2.**) vs. ALOE (**row 3.**) last layer features visualized with 3 principal components and mapped to joint RGB space, preserves global feature geometry that indicates aligned semantics while being inherently interpretable.

PCA Visualization. To visualize alignment quality, we project last-layer features to RGB using the first three principal components with shared scaling per teacher–student pair. As shown in Fig. D1, ALOE preserves the teacher's global feature geometry (e.g., DINOv3 vs. ALOE), indicating aligned semantics while being inherently interpretable.

Zero-shot, model-inherent VLM explanations. In Figure D2, we visualize zero-shot predictions by swapping the SigLIP2 image encoder with its ALOE-aligned B-cos counterpart while keeping the original text encoder and prompts. The resulting *inherent* explanations localize the class-relevant regions (e.g., discriminative parts, textures) without any additional tuning. Notably, maps remain well-aligned and class-specific, consistent with our zero-shot accuracy in Tab. C3.

Comparisons with popular attribution methods. Figures D3 to D5 contrasts **ALOE (ours)**—which uses model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$—with AttnLRP, Input×Gradient, Integrated Gradients, LIME, and GradSHAP. Across diverse categories, ALOE produces more object-centric explanations with sharper boundaries and less background noise. The six-channel encoding preserves color semantics, yielding explanations that align with class-specific parts and textures. These trends also mirror our quantitative gains in GridPG and pixel-perturbation stability (Tab. C5 and Fig. C3). *All examples for this use DINOv3-based backbones (ViT-B/16) with linear probes trained on ImageNet-1k.*

Depth maps (dense linear probing). Figure D6 shows monocular depth outputs from a shallow linear head trained on frozen features (ViT-B/16). ALOE-aligned features yield depth maps with coherent geometry that visually are very similar to the DINOv3 teacher. These visuals complement the relative/absolute depth metrics in Tab. C4, underscoring that the aligned B-cos backbone provides useful *dense* representations—not only global classification signals—while retaining inherent interpretability. The depth maps from the vanilla B-cosification model seem to be more noisy and blurred.

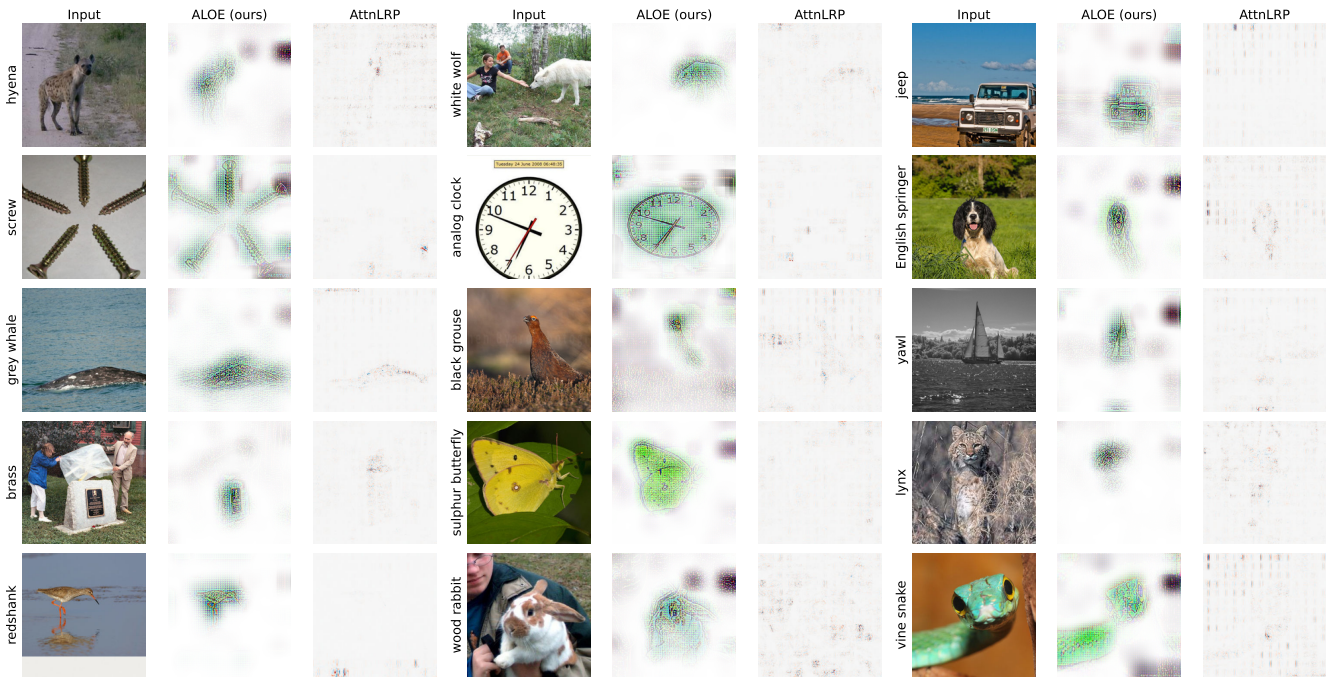

Figure D2. **Zero-shot, model-inherent VLM explanations.** Qualitative comparison of **ALOE (ours)** using $\mathbf{W}(\mathbf{x})\mathbf{x}$ attributions vs. AttnLRP, given the fixed text prompt *"A photo of a {class-name}"*. Our explanations are sharply localized on class-relevant regions, whereas AttnLRP appears diffuse and noisy.

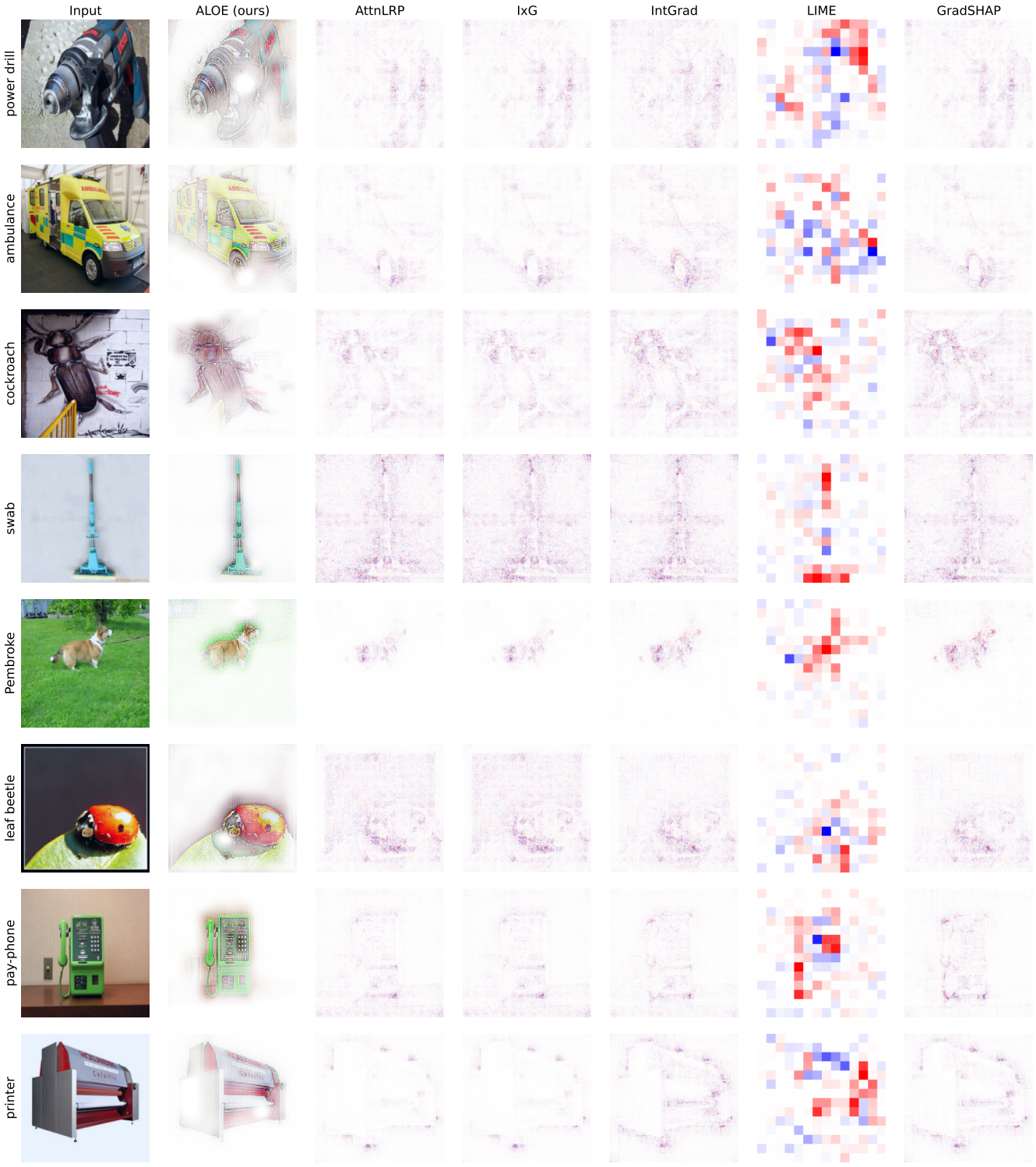

Figure D3. **Qualitative attribution comparisons.** Example 1: Visualizations for **ALOE (ours)**—using model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$—versus popular post-hoc methods (AttnLRP, Input×Gradient, Integrated Gradients, LIME, GradSHAP). ALOE produces sharper, better-localized, and color-faithful highlight maps with less background noise, focusing on class-relevant object regions consistently across examples.

Figure D4. **Qualitative attribution comparisons.** Similar to Figures D3, D5

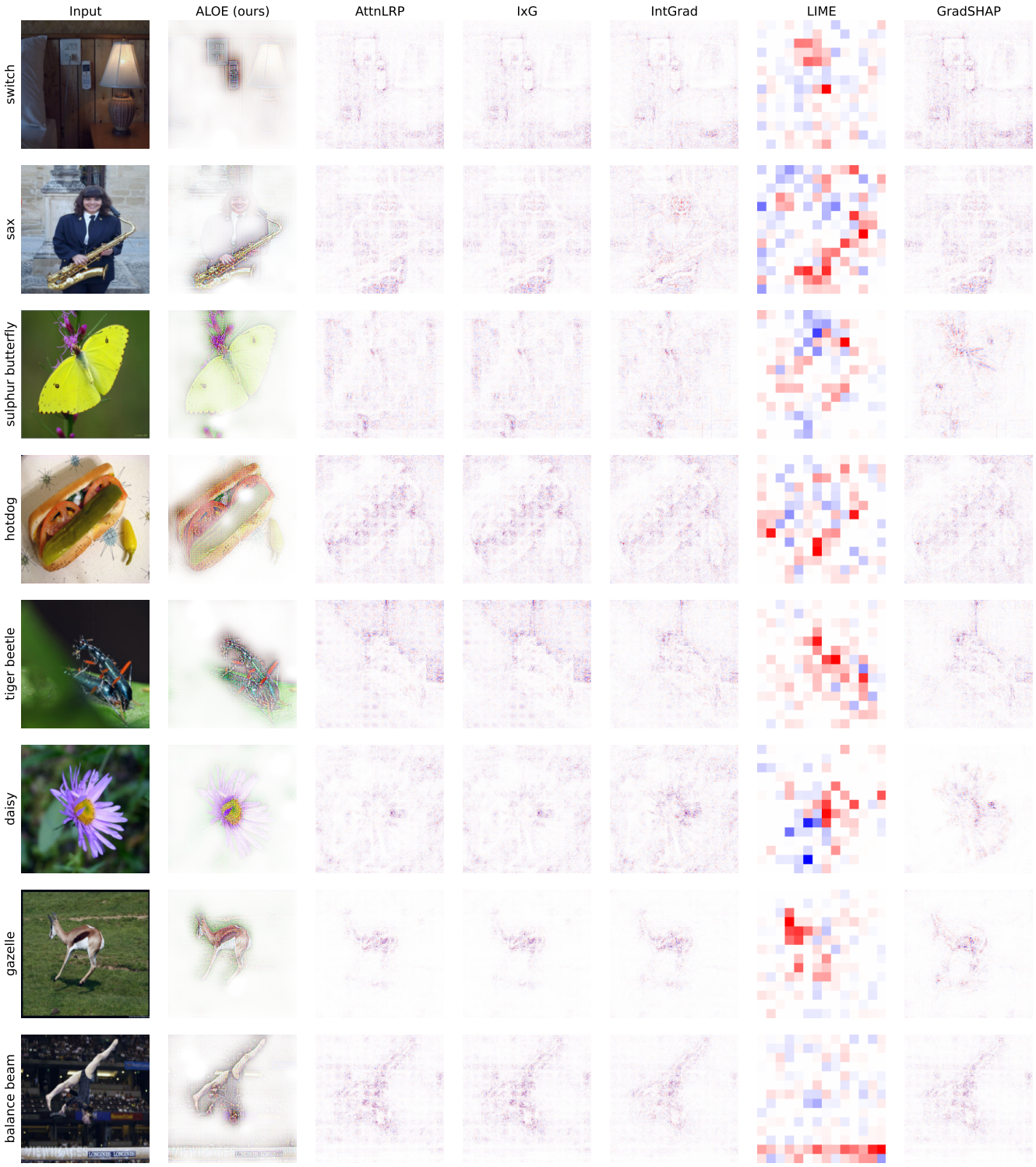

Figure D5. **Qualitative attribution comparisons.** Similar to Figures D3, D4.

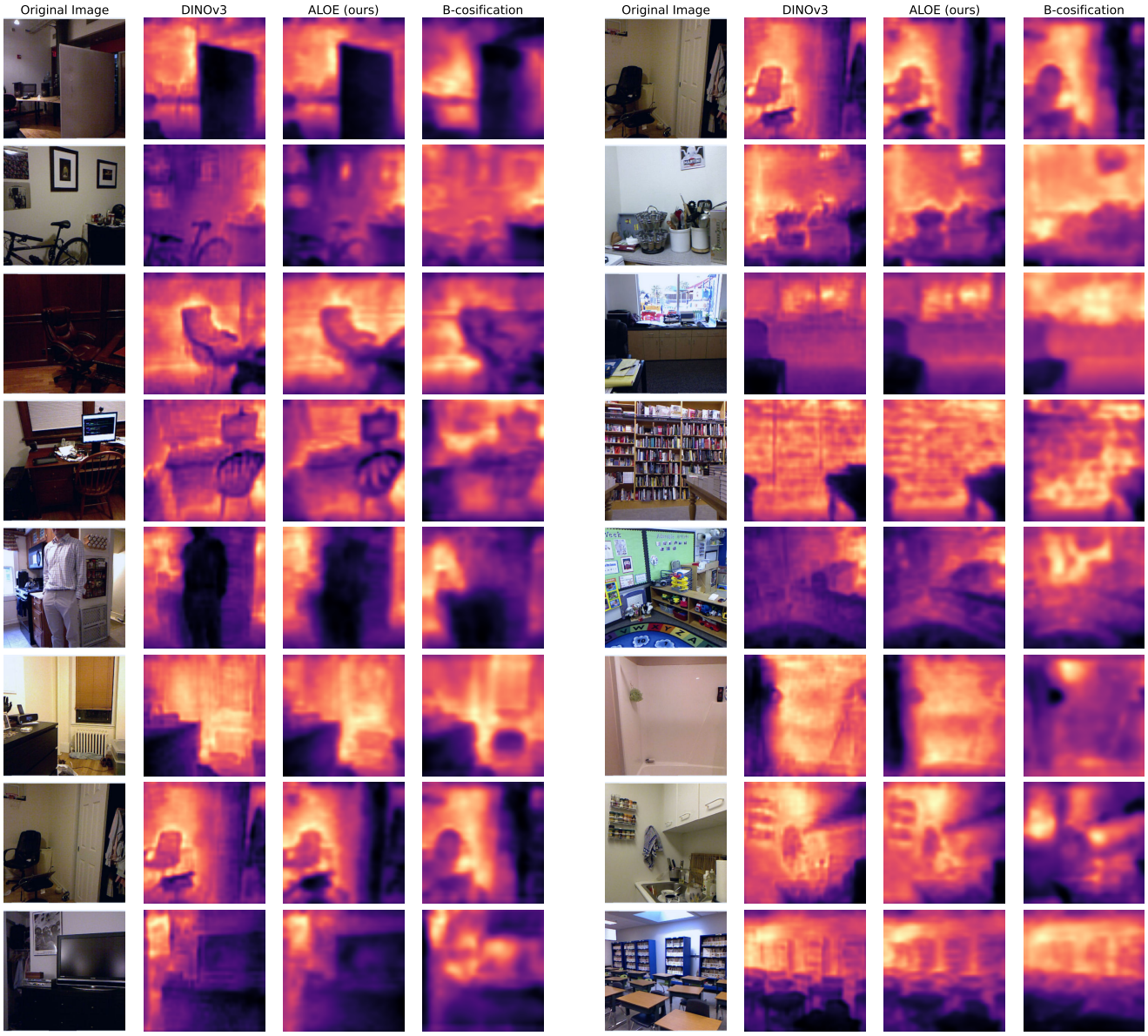

Figure D6. **Depth Estimation.** Visualization of predicted depth maps for **ALOE (ours)**, DINOv3 and B-cosification on the NYUv2 [66] depth-estimation dataset using the Probe3D [23] protocol. ALOE-aligned features yield depth maps with coherent geometry that visually are very similar to the DINOv3 teacher model. The depth maps from the vanilla B-cosification model seem to be more noisy and blurred.

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From Attribution Maps to Human-Understandable Explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. 4

[2] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 9, 13

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *NeurIPS*, 2018. 1

[4] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post Hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation. In *ICLR*, 2021. 1

[5] Shreyash Arya, Sukrut Rao, Moritz Böhle, and Bernt Schiele. B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable. *NeurIPS*, 37:62756–62786, 2024. 1, 2, 3, 4, 6, 10, 11, 12

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation. *PloS one*, 10(7):e0130140, 2015. 6

[7] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos Networks: Alignment is All We Need for Interpretability. In *CVPR*, pages 10329–10338, 2022. 1, 2, 4, 6, 9

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 8

[9] Moritz Böhle, Mario Fritz, and Bernt Schiele. Convolutional Dynamic Alignment Networks for Interpretable Classifications. In *CVPR*, 2021. 6, 9

[10] Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers. *IEEE TPAMI*, 2024. 1, 2, 4, 6, 8, 9, 13

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, 2021. 1, 6, 8

[12] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 6

[13] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In *WACV*, pages 839–847, 2018. 6

[14] Hila Chefer, Shir Gur, and Lior Wolf. Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *ICCV*, 2021. 6

[15] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *NeurIPS*, 2019. 6

[16] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 4

[17] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *CVPR*, pages 2818–2829, 2023. 9, 11

[18] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 8

[19] Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Llava-more: A comparative study of llms and visual backbones for enhanced visual instruction tuning. *arXiv preprint arXiv:2503.15621*, 2025. 6

[20] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. *arXiv preprint arXiv:2203.05796*, 2022. 6

[21] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable ProtoPNet: An Interpretable Image Classifier using Deformable Prototypes. In *CVPR*, 2022. 6

[22] 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 *ICLR*, 2021. 1, 4, 6, 7, 8, 10

[23] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 8, 20

[24] Siddhartha Gairola, Moritz Böhle, Francesco Locatello, and Bernt Schiele. How to Probe: Simple Yet Effective Techniques for Improving Post-hoc Explanations. In *ICLR*, 2025. 6

[25] Google Research. Siglip2 vision encoder checkpoint (vit-b/16, 224), 2025. Identifier: google/siglip2-base-patch16-224. 8

[26] Google Research. Siglip2 vision encoder checkpoint (vit-l/16, 256), 2025. Identifier: google/siglip2-large-patch16-256.

[27] Google Research. Siglip2 vision encoder checkpoint (so400m/16, 256), 2025. Identifier: google/siglip2-so400m-patch16-256. 8

[28] Google Research. Vit-b/16 supervised checkpoint (224), 2025. Identifier: google/vit-base-patch16-224. 8

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016. 2

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 6

[31] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. If you use this software, please cite it as below. 9, 11

[32] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE TIP*, 30:5875–5888, 2021. 6

[33] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *ICML*, pages 5338–5348, 2020. 6

[34] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A Unified and Generic Model Interpretability Library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020. 9

[35] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020. 6

[36] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 8

[37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. *Technical Report, Computer Science Department, University of Toronto*, 2009. 8

[38] L'ubor Ladickỳ, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *European conference on computer vision*, pages 468–484. Springer, 2014. 8

[39] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022. 8

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 6

[41] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 8

[42] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *NeurIPS*, 30, 2017. 9

[43] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 8

[44] Meta AI. Dinov3 vit-b/16 pretrain checkpoint (224), 2025. Identifier: facebook/dinov3-vitb16-pretrain-lvd1689m. 8

[45] Meta AI. Dinov3 vit-l/16 pretrain checkpoint (224), 2025. Identifier: facebook/dinov3-vitl16-pretrain-lvd1689m.

[46] Meta AI. Dinov3 vit-s/16 pretrain checkpoint (224), 2025. Identifier: facebook/dinov3-vits16-pretrain-lvd1689m. 8

[47] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural Prototype Trees for Interpretable Fine-grained Image Recognition. In *CVPR*, pages 14933–14943, 2021. 6

[48] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-Shot Knowledge Distillation in Deep Networks. In *ICML*, pages 4743–4751. PMLR, 2019. 6

[49] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 8

[50] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-Free Concept Bottleneck Models. In *ICLR*, 2023. 6

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 1, 6, 8

[52] Amin Parchami-Araghi, Moritz Böhle, Sukrut Rao, and Bernt Schiele. Good Teachers Explain: Explanation-Enhanced Knowledge Distillation. In *ECCV*, 2024. 6

[53] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 6

[54] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *BMVC*, 2018. 6

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, pages 8748–8763, 2021. 1, 6, 8, 9

[56] Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards Better Understanding Attribution Methods. In *CVPR*, pages 10213–10222, 2022. 1

[57] Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-Name: Task-Agnostic Concept Bottlenecks via Automated Concept Discovery. In *ECCV*, 2024. 6

[58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of any Classifier. In *KDD*, pages 1135–1144, 2016. 9

[59] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arxiv 2014. *arXiv preprint arXiv:1412.6550*, 2014. 6

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3): 211–252, 2015. 8, 9

[61] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the Visualization of What a Deep Neural Network has Learned. *IEEE Trans. Neural Netw. Learn. Syst.*, 28(11):2660–2673, 2016. 9

[62] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*, pages 618–626, 2017. 6

[63] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset for Automatic Image Captioning. In *ACL*, pages 2556–2565, 2018. 6, 11

[64] Hyunjune Shin and Dong-Wan Choi. Teacher as a lenient expert: Teacher-agnostic data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14991–14999, 2024. 6

[65] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features through Propagating Activation Differences. In *ICML*, pages 3145–3153, 2017. 6, 9

[66] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 8, 20

[67] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 3, 4, 6, 7, 8, 9, 10, 12, 13

[68] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Metworks. In *ICML*, pages 3319–3328. PMLR, 2017. 6, 9

[69] 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*, pages 10347–10357. PMLR, 2021. 6

[70] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1, 3, 4, 6, 7, 8, 9, 10, 11, 13

[71] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *CVPRW*, pages 111–119, 2020. 6

[72] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 8

[73] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 6

[74] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, pages 11975–11986, 2023. 1, 3, 6, 9

[75] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. *IJCV*, 126(10):1084–1102, 2018. 9

[76] Zikai Zhou, Yunhang Shen, Shitong Shao, Linrui Gong, and Shaohui Lin. Rethinking centered kernel alignment in knowledge distillation. *arXiv preprint arXiv:2401.11824*, 2024. 6