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, Saarland Informatics Campus, Saarbrücken, Germany
{raphael.maser,siddhartha.gairola,sukrut.rao,schiele}@mpi-inf.mpg.de

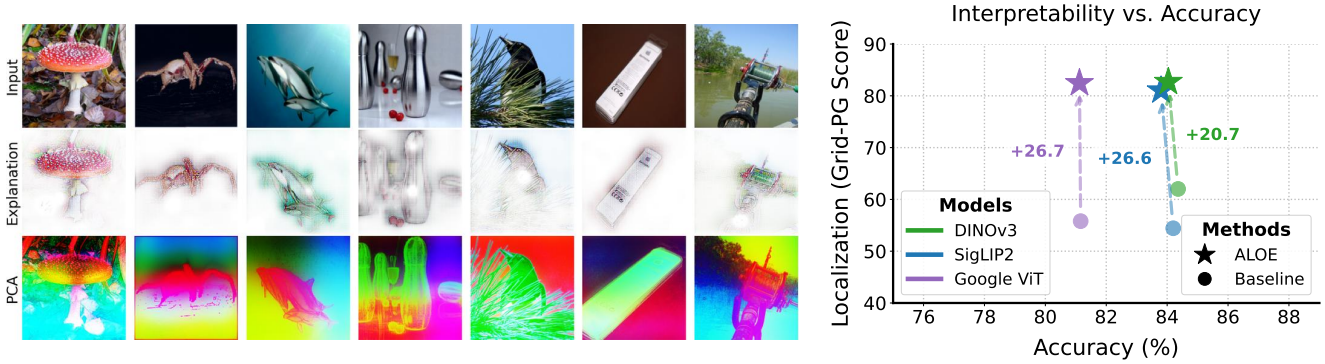


Figure 1. **ALOE: performant, preserves semantics, explains faithfully.**— *Left: Explanations and representations.* ALOE enables inherently interpretable B-cos: $\mathbf{W}(\mathbf{x})$ [11] attributions (**row. 2**) that are object-centric and class-specific. The accompanying PCA visualizations (**row. 3**; RGB = first three principal components of the final image representation for our ALOE aligned DINOv3 [70] model) preserves global feature geometry—indicating aligned semantics with improved explainability (*cf*. Fig. 5). *Right: Performance vs. interpretability.* Across ViT-B/16 backbones, ALOE substantially boosts localization quality on GridPG [6] while maintaining competitive ImageNet [63] accuracy relative to the corresponding foundation models (Supervised [25], DINOv3 [70], SigLIP2 [74]).

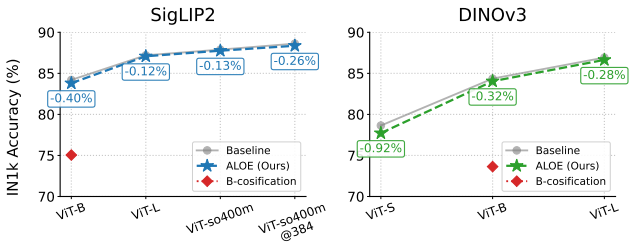
Abstract

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 approach that efficiently converts foundational vision models into inherently interpretable B-cos [11] 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****× more data-efficient than training from scratch. On classification, it strongly outperforms vanilla B-cosification (e.g.,* ***+9.2 p.p.*** *top-1 on ImageNet for supervised ViT-B/16), retains strong linear probing, k-NN, and zero-shot transfer competitive with foundational backbones (DINOv3 [70], SigLIP2 [74]) across diverse downstream datasets. It also preserves spatially structured features useful for dense prediction, while yielding well-localized and highly human-interpretable explanations by design. Code link https://www.github.com/rmaser/aloe.*

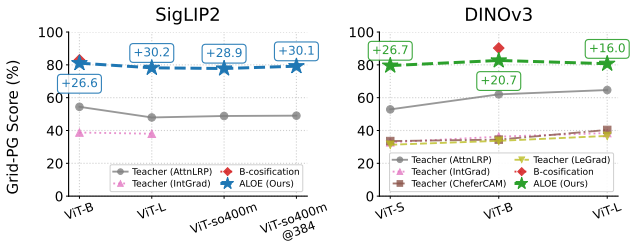
*Equal contribution.

1. Introduction

Large-scale vision foundation models, including self-supervised encoders in the DINO family [12, 54, 70] and contrastive vision–language models such as CLIP [58] and SigLIP [74, 78], are powerful and versatile feature extractors that achieve state-of-the-art performance across a diverse set of downstream tasks. Trained on billions of images and image–text pairs, these models serve as general-purpose building blocks across various domains, from fine-grained recognition to medical imaging, often replacing their fully supervised, domain-specific counterparts [70]. However, their decision-making processes remain largely opaque, which hinders their use in sensitive and safety-critical applications. Post-hoc explanation methods aim to mitigate this issue, but can be noisy and not always faithful to the underlying model being explained [2, 3, 59].



(a) ImageNet [63] top-1 accuracy (%).



(b) Grid-PG (localization) Score (%).

Figure 2. **ALOE vs. B-cosification [4].** ALOE consistently outperforms vanilla B-cosification on ImageNet [63] top-1 accuracy across model scales for both SigLIP2 [74] and DINOv3 [70] (see (a)). It reaches performance close to the original SigLIP2 and DINOv3 baselines while significantly improving localization scores (Grid-PG, see (b))—even compared against different SOTA post-hoc attribution methods. This trend also holds across diverse datasets (see Tab. 2).

An alternative line of work develops *inherently interpretable* architectures that yield model-faithful explanations by design. They introduce architectural constraints that lead to human-interpretable, faithful summaries of the model's computations; examples include prototype-based models [17, 24, 50], dynamic linear models [6, 10], and concept-bottleneck models [37, 53, 60]. A prominent example is the family of B-cos Networks [6, 11], which replace linear transforms with bias-free B-cos transforms that promote weight-input alignment to enhance interpretability and enable a faithful summary of the model's decision. To avoid training these models from scratch, recent work proposed *'B-cosification'* [4], an approach that efficiently transforms existing models to their B-cos variants post-hoc at a fraction of the full training cost. However, despite strong gains on fully-supervised convolutional backbones and competitive results on a CLIP ResNet [32] backbone, the B-cosification recipe provides only modest improvements for Vision Transformers (ViTs) [25], sometimes failing to match the performance of training from scratch. Since most modern foundation models are ViT-based, this significantly limits its practical utility and broader adoption.

In this work, we propose a scalable approach to efficiently transform existing foundation models into B-cos variants, that is *practically useful* for modern ViTs. Specifically, we explore feature-alignment methods for fine-grained matching between a pre-trained foundational teacher and its B-cosified student without any labeled supervision. We conduct a comprehensive study of key design choices, including which features/layers to align, alignment losses, model scales, and alignment dataset sizes. Unlike [4], our approach, **ALOE** ***(ALign Once to Explain)*** (Fig. 4), is a general-purpose recipe that requires only a single alignment of the backbone irrespective of the pre-training paradigm, such as fully-supervised, self-supervised, or contrastive vision-language training. Once aligned, the backbone can be used as a drop-in replacement across downstream tasks, amortizing the cost of interpretability.

Despite its simplicity, ALOE is surprisingly effective and highly data and compute efficient. For example, while training SigLIP2 [74] from scratch required ≈10B images and ≈12B alt-texts, ALOE B-cosifies such a model while maintaining its generalization with just ≈3M samples and 40 epochs of training. This small overhead obtains results comparable to the original model's performance (*e.g.*, 83.67% vs. 84.24% top-1 accuracy on ImageNet [63] for SigLIP2; see Fig. 1), making it clearly preferable to full retraining for achieving interpretability. Moreover, ALOE consistently outperforms vanilla B-cosification [4] across pre-training paradigms by a substantial margin (> 4.9 p.p.; see Fig. 2). Our models also deliver strong linear-probing and zero-shot performance across datasets, preserve spatially structured features useful for dense prediction, and provide faithful, well-localized explanations (Fig. 1).

In summary, our **contributions** are:

(1) We propose **ALOE** ***(ALign Once to Explain)***, a universal, one-time, label-free feature-alignment approach that enables efficient transformation of foundational ViT backbones into inherently interpretable B-cos variants while preserving downstream utility.

(2) We conduct a detailed study of the design choices and provide key insights into alignment targets (layer-wise vs. pooled; single vs. multi-layer), loss families (MSE, cosine, SigLIP, InfoNCE), model scales, and alignment dataset sizes.

We make the following **key empirical findings**:

(F1) Generality: ALOE B-cos ViTs outperform end-to-end trained B-cos and B-cosified [4] baselines in classification, retain strong linear probing, k-NN, and zero-shot transfer performance competitive with foundational backbones (DINOv3, SigLIP2), preserve spatially structured features for dense prediction, and provide faithful and well-localized explanations (e.g., $>$ **4.9 p.p.** accuracy over [4] on ImageNet [63]).

(F2) Data efficiency: feature alignment with as few as 3M images approaches teacher-level generalization—approximately $\mathbf{1000}\times$ fewer images than billion-scale pre-training (e.g., 3M vs. ∼10B), with comparable performance after ∼40 epochs.

With ***ALOE*** we establish a practical path to scalable trust-

worthy vision foundation models that are faithful by design, strong in transfer tasks, and easy to adopt.

2. 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 [12, 54, 70] and contrastive vision–language encoders such as CLIP [58] and SigLIP/SigLIP2 [74, 78]. Supervised ViTs pre-trained on large datasets (e.g., ImageNet21k/22k [39] or proprietary corpora) also serve as strong backbones for downstream usage [25, 39]. Such encoders are also used by large vision-language models and generative models together with a pre-trained LLM for multimodal understanding [21, 44]. In this work, we transform modern vision foundation models into an inherently interpretable B-cos backbone via a single alignment step, while preserving downstream utility.

Inherently interpretable models. To understand deep neural networks (DNNs), post-hoc attribution methods spanning gradient- [5, 15, 68, 71] activation- [14, 36, 65, 75], and perturbation-based [57] 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 prototype-based models [17, 24, 50], dynamic-linear models [6, 10], and concept-bottleneck models [37, 53, 60]. B-cos networks [11] 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 [27, 55]. The cost of training such models from scratch at foundation scale motivated B-cosification [4], which transforms existing networks to B-cos variants post hoc. While B-cosification showed gains on supervised CNNs and for a 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 [33] which aligns output distributions between $\mathcal{T}$ and $\mathcal{S}$ has been extended with feature-level and attention-based alignment objectives [56, 62, 77], layer-wise and multi-stage schemes [80], and token-level distillation for ViTs [73]. Recent works [55] 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 [51, 67]. 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.

3. Background: B-cos networks

In this section we briefly review B-cos networks [11] (Sec. 3.1) and the B-cosification [4] (Sec. 3.2) recipe proposed for CNNs, before presenting our approach (Sec. 4).

3.1. B-cos networks

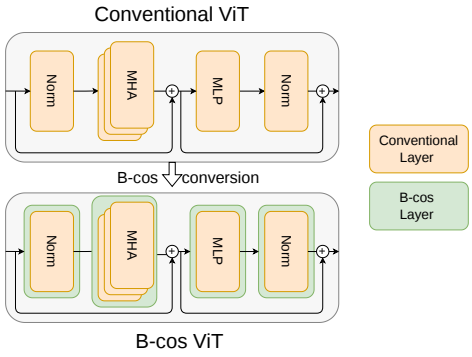


Figure 3. **Conventional vs. B-cos ViT.** We replace linear/MLP layers with *bias-free* B-cos transforms and uncentered normalization, keeping self-attention and residuals intact. This yields a dynamic-linear B-cos ViT with inherently faithful explanations via $\mathbf{W}(\mathbf{x})\mathbf{x}$ (Sec. 3.2).

B-cos networks [11] 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** [6] 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 [32]) 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 [11], is left unchanged, as are positional embeddings.

Such models were shown to provide competitive performance while significantly improving interpretability [11], however, the need to train them from scratch remained a significant limitation to their adoption as compared to using already trained conventional models.

3.2. B-cosification recipe

To alleviate the need to retrain B-cos variants of existing models, 'B-cosification' [4] 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 [11], 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 [11], 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.

4. 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 a ViT-based foundation encoder into a B-cos [11] counterpart and aligns it to a frozen teacher via representation matching (Fig. 4). We first describe the *teacher–student* setup and the architecture-preserving transformation used to initialize the student (Sec. 4.1). Next, we detail what is aligned and where in the network, including ViT-specific tokens (e.g., `[CLS]` and registers) and intermediate layers used to guide alignment (Sec. 4.2). Finally, we describe our setup (Sec. 4.3) 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. 4.4).

4.1. 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 [4](see Sec. 3.2), which preserves functional behavior where possible while requiring only 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 identical to those of 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 closely and is ready for label-free alignment.

4.2. Label-free alignment

After conversion (Sec. 4.1), 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. This motivates (i) a *global* guidance term to preserve the geometry of the final embedding space—critical for downstream tasks (classification, dense prediction, zero-shot transfer)—and (ii) *token-level, depth-wise* guidance to preserve intermediate computations and improve optimization stability. 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 depth-wise layer guidance:

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

$$\mathcal{L}_{\mathrm{global}} = \frac{1}{|\mathcal{B}|}\sum_{\mathbf{x}\in\mathcal{B}}\Big(1-\cos\big(E_{\mathcal{S}}(\mathbf{x}),\,E_{\mathcal{T}}(\mathbf{x})\big)\Big), \tag{3}$$

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

[1] SigLIP2 [74] image encoders do not use `[CLS]` or register tokens.

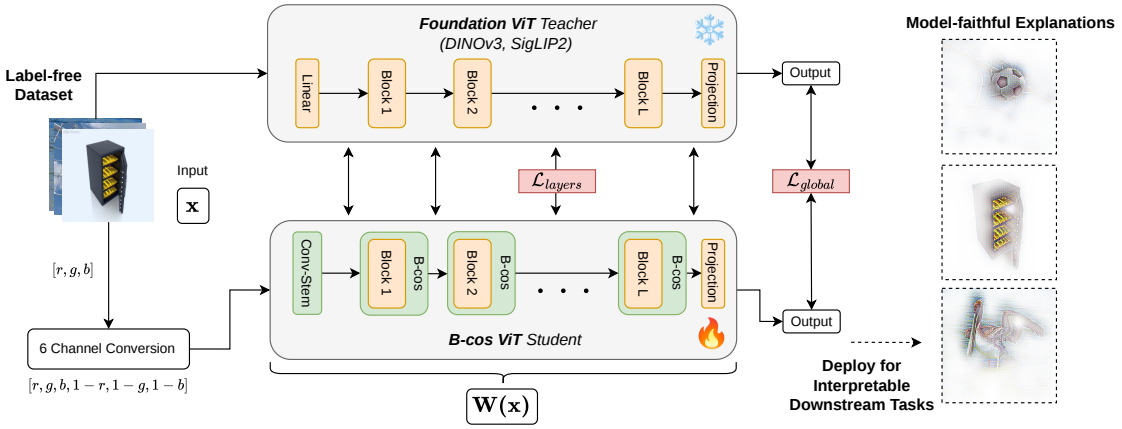


Figure 4. **Align Once to Explain.** *(1) Transform to B-cos*—convert a Foundation ViT encoder into a bias-free, dynamic-linear backbone (Sec. 4.1). *(2) Align once*—label-free feature alignment with cosine distance objective between foundational teacher and B-cos student, on unlabeled images. *(3) Deploy*—freeze the aligned B-cos backbone for linear probing and zero-shot transfer; faithful, well-localized explanations follow directly from $W(\mathbf{x})$ (*i.e.*, model summary), amortizing interpretability across tasks.

where $\mathcal{T}^{\ell}_{\text{tok}}(\mathbf{x})$ denotes the set of teacher tokens at depth ℓ.

Additionally, we include a regularization term that couples the student and teacher weight magnitudes. This is computed as the sum of layer-wise differences between the Frobenius norms $\|\cdot\|_F$ of the student ($\mathbf{W}^{(S)}_{\ell}$) and teacher ($\mathbf{W}^{(T)}_{\ell}$) weight matrices for all shared layers $\mathcal{P}$:

$$\mathcal{L}_{\text{reg}} = \alpha \sum_{l \in \mathcal{P}} \left(\|\mathbf{W}^{(T)}_{\ell}\|_F - \|\mathbf{W}^{(S)}_{\ell}\|_F \right)^2 \tag{5}$$

Consequently, this encourages the student to align the direction of weights (and not magnitude), relative to the teacher and avoids divergence due to exploding weight norms in longer training runs.

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 [70, 78]. 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. 5.2. 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 $\mathcal{L}_{\text{depth}} = \{\lfloor L/3 \rfloor, \lfloor 2L/3 \rfloor, L\}$ for a transformer of depth L (*i.e.*, $1/3$, $2/3$, and full depth). We align exactly the semantics-carrying tokens used by each teacher (*e.g.*, `[CLS]`/registers for DINOv3; attention pooling for SigLIP2; Tab. 1), ensuring one-to-one routing and preserving the teacher's computational pathways—crucial for faithful B-cos explanations.

Table 1. **Tokens used for alignment (Sec. 4.2) for each model.**

Model	Global Feature	Layer-wise Features
Supervised [25]	Final `[CLS]`	Image tokens, `[CLS]`
DINOv3 [70]	Final `[CLS]`	Image tokens, `[CLS]`, register tokens
SigLIP2 [74]	Attention-pooled final embedding	Image tokens

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, $(\lambda_{\text{g}}, \lambda_{\text{l}}) = (1, 1)$ works well across all models. We ablate the design choices—including the alignment objective and feature depths—to select the final configuration empirically (Sec. 5.2).

4.3. Implementation and training setup

We now describe the implementation and training setup for feature alignment.

Datasets. We use the unlabeled, web-scale image sets CC3M [66], CC12M [13], and YFCC15M [22] for alignment. The input resolution follows the teacher's default (typically 224×224). For our main results (Tab. 2), we report numbers for 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 [25]; (ii) *DINOv3* ViT-S/16, ViT-B/16, and ViT-L/16 [70]; (iii) *SigLIP2* ViT-B/16, ViT-L/16, and ViT-so400m/16 [74]. B-cos students $\mathcal{S}$ are constructed using the architecture-preserving conversion recipe (Sec. 3.2) and initialized to mirror $\mathcal{T}$ in width, depth, and tokenization, enabling one-to-one alignment (Sec. 4.2).

Augmentations. We apply random resized crops and horizontal flips as augmentations to prevent overfitting while preserving $\mathcal{T}$–$\mathcal{S}$ feature geometry during alignment.

Optimization. We freeze $\mathcal{T}$ and optimize $\mathcal{S}$ with AdamW [45] and a cosine learning-rate schedule, with mixed precision enabled. We follow an early-stopping criterion to train until the loss on a held-out validation set stops decreasing; we fix B=2, keep additive biases at zero, and use no explicit weight normalization (cf. Sec. 3.2). Gradient norm clipping of 1.0 (with no weight decay) is 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 (a 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. 5.2.

4.4. 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 $\mathbf{W}(\mathbf{x})$ [11]) and require no task-specific interpretability tuning.

Linear probing (LP). We evaluate the 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 the standard protocol [54] (see Sec. 5.1). 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 [12] and report accuracy (%) on the validation split(s).

Zero-shot evaluation. For the contrastive vision–language setting (SigLIP2), we use the ALOE-aligned B-cos image encoder as a *drop-in* replacement while keeping the *original* text encoder unchanged. Prompts, temperature, and normalization follow the default settings [58]. We evaluate zero-shot classification on standard benchmarks [58, 78] and report accuracy (%).

Dense prediction. We evaluate dense prediction by training a linear depth head on frozen features and reporting NYUv2 [42, 69] depth metrics.

Evaluating explanations. We evaluate explanation quality using the Grid Pointing Game (GridPG) [6] for localization and pixel-deletion for faithfulness. For conventional models, we compare against several post-hoc methods (AttnLRP [1], Integrated Gradients [71], LeGrad [8], Chefer-CAM [16], and LIME [61]); we also report results on a human preference study. Implementation details and results are provided in the appendix Sec. A and B.2.

5. Results

In Sec. 5.1, we compare ALOE to baselines on the settings and metrics described in Sec. 4.4. Detailed ablations on alignment objectives, feature depths, model scales, and dataset size are provided in Sec. 5.2.

5.1. Main results

Linear evaluation of frozen features. We compare ALOE to (i) the original foundational models, (ii) B-cos (from scratch) [11], and (iii) B-cosification [4]. All use the same linear protocol and input resolution (see Sec. 4.4).

In Table 2, we report linear probing results on frozen features on ten datasets ([18]). 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 ImageNet-1k linear probing (SigLIP2 teachers), we obtain **83.67%** for B-ViT-B/16; larger variants reach **87.08%** and **87.76%** for B-ViT-L/16 and B-ViT-so400m/16 (Sec. B.1).

Similar trends hold for k-NN evaluation. In Table 3a we report accuracy on IN1k [63] for ALOE vs B-cosification; we get **+10.36** p.p. on IN1k and **+31.86** p.p. averaged across datasets (see appendix Sec. B.1 for additional results).

Zero-shot classification. ALOE preserves strong zero-shot transfer and outperforms B-cosification. On ImageNet-1k (zero-shot@1) with SigLIP2 ViT-B/16, ALOE achieves **77.20**% vs. B-cosification **61.01**%, close to the SigLIP2 teacher at **78.07**% (Tab. 3b).

Dense prediction. Beyond image-level recognition, ALOE also preserves spatially structured features useful for dense prediction. On monocular depth estimation with a linear probe (ViT-B/16), ALOE substantially outperforms B-cosification on both relative and absolute metrics, improving relative δ_1 from **0.83** to **0.94** and reducing RMSE from **0.46** to **0.30**, while approaching the DINOv3 teacher (**0.97/0.24**); see appendix (Sec B.1) for more dense results, including surface normals and multiview correspondence.

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. 5, ALOE preserves the teacher's global feature geometry (e.g., DINOv3 vs. ALOE), indicating aligned semantics while being inherently interpretable.

Table 2. **Linear-probe accuracy on base architectures (ViT-B/16).** ALOE substantially outperforms vanilla B-cosification while remaining competitive with the original foundation models. Teachers are shown in gray; best interpretable model in **bold**. *Note: For DINOv3 B-cosification we follow the supervised fine-tuning protocol [4] since the original work did not have a self-supervised baseline.*

Feature	Inh. Inter.	IN1k	Cal101	Flowers	Food	Aircr	DTD	Cars	SUN	C10	C100	Avg.
Sup. [25]	✗	80.74	100.00	99.35	86.00	39.51	73.83	55.19	73.54	97.07	86.05	79.13
B-cosif. [4]	✓	71.76	99.51	78.64	65.12	34.83	59.77	40.28	53.89	91.48	74.61	66.99
ALOE (ours)	✓	**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
SigLIP2 [74]	✗	84.24	99.93	98.31	94.43	75.48	85.21	95.43	81.67	96.89	84.66	89.63
B-cosif. [4]	✓	75.84	99.87	93.88	85.24	44.80	80.08	76.81	76.59	94.83	80.62	80.86
ALOE (ours)	✓	**83.67**	**99.90**	**99.09**	**92.60**	**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
DINOv3 [70]	✗	84.34	100.00	99.74	94.08	80.29	83.71	94.33	78.63	98.17	89.20	90.25
B-cosif. [4]	✓	78.86	99.80	82.29	74.33	44.32	71.48	54.70	61.41	92.71	76.87	73.68
ALOE (ours)	✓	**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

(a) k-NN Accuracy

Method	IN1k (%)	Avg (%)
DINOv3 [70]	82.27	82.95
B-cosif. [4]	71.03	50.77
ALOE	**81.39**	**82.63**

(b) Zero-shot Accuracy

Method	@1 (%)
SigLIP2 [74]	78.07
B-cosif. [4]	61.01
ALOE	**77.20**

Table 3. **Downstream and Zero-shot Evaluations. (a)** k-NN results (k=20). **(b)** Zero-shot top-1 accuracies. In both evaluation settings, ALOE substantially outperforms vanilla B-cosification and remains highly competitive with the original teacher models.

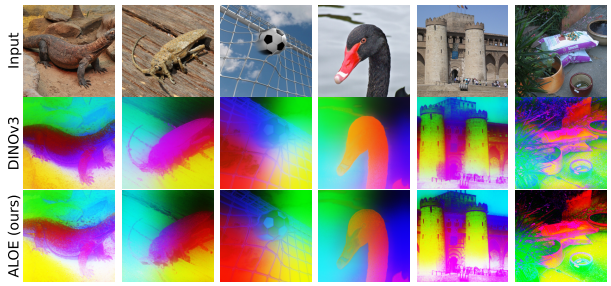


Figure 5. **PCA Visualization.** DINOv3 (**row 2.**) vs. ALOE (**row 3.**) last layer features visualized with 3 principal components and mapped to joint RGB space, preserves global feature geometry that indicates aligned semantics while being inherently interpretable.

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) post-hoc attributions on conventional models (AttnLRP [1] here). Following [11], we report GridPG (higher is better), and defer additional faithfulness metrics, human evaluation, and post-hoc comparisons to the appendix (Sec. B.2). 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) **54.4**%, and is competitive with B-cosification.

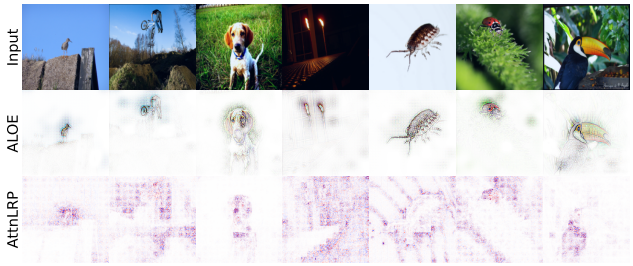


Figure 6. **Explanation quality.** Comparison between model inherent (B-cos: W(x) [11]) explanations after ALOE vs a popular strong method for visualizing vision transformers, AttnLRP [1].

In Fig. 6, 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 optimization (*cf*. [6]). In contrast the AttnLRP explanations are sparse and noisy (**row 3.**).

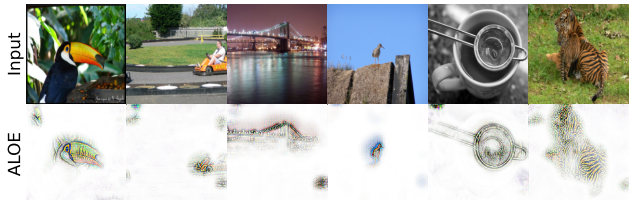


Figure 7. **VLM Zero-Shot Explanations.** ALOE aligned B-cos models (SigLIP2 teacher) yield zero-shot model-inherent explanations for VLMs, where the input prompt to the text encoder is "A photo of {class_name}"(*cf*. [4]). The explanations are visually well aligned with class-specific features.

In Fig. 7, we show zero-shot explanations for ALOE ViT-B/16 models aligned with SigLIP2. This lets us use the

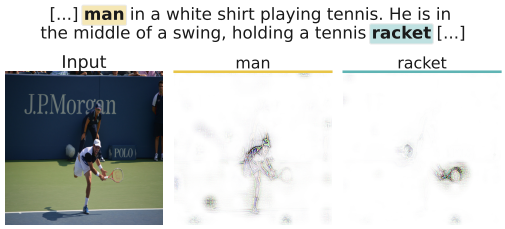


Figure 8. **Token-level Visual Grounding with ALOE.** By propagating relevance through the LLaVA-More [21] Gemma-9B [72] backbone (using AttnLRP) and extracting visual explanations from our B-cos SigLIP2 encoder, we achieve fine-grained localization for individual generated tokens. The relevance maps highlight well their corresponding semantic regions in the input image. Further examples and failure cases are shown in appendix Sec. C.

model as an explainable visual backbone combined with the corresponding text encoder. We use prompts as "A photo of {class_name}" (*cf*. [4]) to compute similarities between text and visual features. The resulting explanations are highly localized and class-specific, yielding inherently explainable vision–language models.

Multimodal LLMs. More broadly, our aligned SigLIP2 backbones can also serve as interpretable visual encoders within multimodal large language models (MLLMs). In preliminary experiments with a LLaVA-style Gemma-9B backbone [72], they enable fine-grained visual grounding for individual generated tokens (see Fig. 8), suggesting that ALOE can extend beyond pure vision tasks. Since the current setup still relies on AttnLRP [1] to propagate relevance through the language model, we defer a fuller treatment of end-to-end inherently interpretable MLLMs to future work.

5.2. Ablations

We ablate key design choices: *multi-layer feature alignment*, *model scale*, *data efficiency* and *alignment objectives*. For all ablations, we train for 40 epochs.

Table 4. **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. 4), accuracy improves with increasing depth until $\lfloor 2L/3 \rfloor$; supervising all layers yields no further gain. We thus use 3 evenly spaced depths.

Model scale. We see consistent gains in accuracy with increasing model size. For this we align DINOv3 and SigLIP2 teachers (and their B-cos students). Larger models close the gap to the teacher foundation model (see Fig. 2).

Data efficiency. Using only unlabeled YFCC15M [22], ALOE maintains nearly flat ImageNet-1k top-1 linear-probe accuracy when the alignment data shrinks from 100% to 1% (about 150k images), staying within $\sim$1 p.p.; see Fig. 9 for SigLIP2. This corresponds to using only $\approx 0.0015\%$ of a $\sim$10B-image pretraining corpus for SigLIP2 while delivering comparable downstream performance, highlighting the strong data efficiency of our one-time feature-alignment procedure.

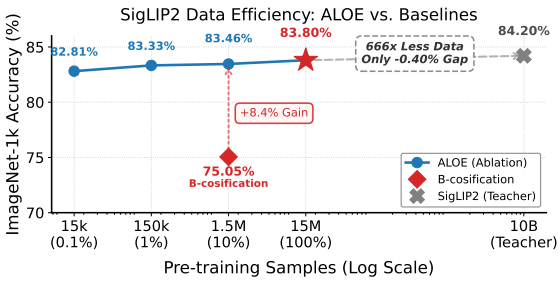


Figure 9. **Data efficiency of ALOE.** ImageNet top-1 linear-probe accuracy for ALOE aligned SigLIP2 (VIT-B/16) across varying subsets of YFCC15M (0.1%–100%). Accuracy plateaus at just 1% ($\approx$150k images), remaining within $\sim$1.0 p.p. of the SigLIP2 teacher (gray line) despite using only $\approx$0.0015% of its 10B pretraining corpus. At B-cosification's [4] data budget, ALOE still gains **+8.4** p.p. over it.

Alignment objectives. Under identical settings (Tab. B8), *cosine* and *SigLIP* are most consistent across models, *MSE* and *InfoNCE* showed inconsistencies. Given the ease of application and simplicity we adopt **cosine** by default.

6. Conclusion

We presented **ALOE** (ALign Once to Explain), a simple yet effective framework that converts ViT-based foundational encoders into inherently interpretable B-cos variants. By preserving special tokens and aligning both global embeddings and token-level features at a few depths with a cosine objective, ALOE yields backbones that achieve competitive performance as compared to conventional foundation models across a diverse set of downstream tasks, provide faithful, well-localized explanations by design, and scale across pre-training paradigms and model sizes with modest amounts of unlabeled data.

Compared to B-cosification [4], ALOE removes the need for supervised fine-tuning, addresses the ViT performance gap, and preserves contrastive zero-shot utility—offering a resource-efficient path to interpretable backbones at foundation scale. We show empirically that with our proposed recipe using $100 - 1000\times$ less images the model can already recover most of the teacher's generalization. Taken together, ALOE offers a practical path to scalable, trustworthy vision backbones: align once, reuse broadly, and amortize the cost of interpretability.

Acknowledgements

We thank Bart Pogodzinski and Wolfgang Boettcher for valuable feedback on the manuscript. Computations were performed on the HPC systems at the Max Planck Computing and Data Facility (MPCDF).

References

[1] 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 *ICML*, 2024. 6, 7, 8, 14, 15, 19, 26

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

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

[4] Shreyash Arya, Sukrut Rao, Moritz Böhle, and Bernt Schiele. B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable. In *NeurIPS*, 2024. 2, 3, 4, 6, 7, 8, 15, 16, 17

[5] 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), 2015. 3

[6] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos Networks: Alignment is All We Need for Interpretability. In *CVPR*, 2022. 1, 2, 3, 6, 7, 15

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*. Springer, 2014. 14

[8] Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. LeGrad: An Explainability Method for Vision Transformers via Feature Formation Sensitivity. In *ICCV*, 2025. 6, 15, 19

[9] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4), 1952. 20

[10] Moritz Böhle, Mario Fritz, and Bernt Schiele. Convolutional Dynamic Alignment Networks for Interpretable Classifications. In *CVPR*, 2021. 2, 3, 15

[11] Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers. *IEEE TPAMI*, 2024. 1, 2, 3, 4, 6, 7, 15, 19

[12] 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, 3, 6

[13] 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 *CVPR*, 2021. 5

[14] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In *WACV*, 2018. 3

[15] Hila Chefer, Shir Gur, and Lior Wolf. Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *ICCV*, 2021. 3

[16] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 6, 15, 19

[17] 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. 2, 3

[18] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 6

[19] 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*, 2023. 14, 16, 17

[20] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 14

[21] 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. In *ICCVW*, 2025. 3, 8, 14, 26

[22] 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. 5, 8

[23] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 14, 17, 18

[24] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable ProtoPNet: An Interpretable Image Classifier using Deformable Prototypes. In *CVPR*, 2022. 2, 3

[25] 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, 2, 3, 5, 7, 14, 15, 16, 19

[26] 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 *CVPR*, 2024. 14, 17, 18, 26

[27] Siddhartha Gairola, Moritz Böhle, Francesco Locatello, and Bernt Schiele. How to Probe: Simple Yet Effective Techniques for Improving Post-hoc Explanations. In *ICLR*, 2025. 3, 20

[28] Google Research. Siglip2 vision encoder checkpoint (vit-b/16, 224), 2025. Identifier: google/siglip2-base-patch16-224. 14

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

[30] Google Research. Siglip2 vision encoder checkpoint (so400m/16, 256), 2025. Identifier: google/siglip2-so400m-patch16-256. 14

[31] Google Research. Vit-b/16 supervised checkpoint (224), 2025. Identifier: google/vit-base-patch16-224. 14

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016. 2, 4

[33] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[34] 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. 14, 16, 17

[35] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. 14, 17, 18

[36] 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, 2021. 3

[37] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *ICML*, 2020. 2, 3

[38] 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. 15

[39] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*. Springer, 2020. 3

[40] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 14

[41] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. *Technical Report, Computer Science Department, University of Toronto*, 2009. 14

[42] L'ubor Ladickỳ, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*. Springer, 2014. 6, 14

[43] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022. 14

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[45] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 6

[46] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 14

[47] Meta AI. Dinov3 vit-b/16 pretrain checkpoint (224), 2025. Identifier: facebook/dinov3-vitb16-pretrain-lvd1689m. 14

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

[49] Meta AI. Dinov3 vit-s/16 pretrain checkpoint (224), 2025. Identifier: facebook/dinov3-vits16-pretrain-lvd1689m. 14

[50] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural Prototype Trees for Interpretable Fine-grained Image Recognition. In *CVPR*, 2021. 2, 3

[51] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-Shot Knowledge Distillation in Deep Networks. In *ICML*, 2019. 3

[52] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 14

[53] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-Free Concept Bottleneck Models. In *ICLR*, 2023. 2, 3

[54] 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. *TMLR*, 2024. Featured Certification. 1, 3, 6

[55] Amin Parchami-Araghi, Moritz Böhle, Sukrut Rao, and Bernt Schiele. Good Teachers Explain: Explanation-Enhanced Knowledge Distillation. In *ECCV*, 2024. 3

[56] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 3

[57] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *BMVC*, 2018. 3

[58] 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, 3, 6, 14

[59] Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards Better Understanding Attribution Methods. In *CVPR*, 2022. 1

[60] Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-Name: Task-Agnostic Concept Bottlenecks via Automated Concept Discovery. In *ECCV*, 2024. 2, 3

[61] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of any Classifier. In *KDD*, 2016. 6, 15, 19, 20

[62] 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. 3

[63] 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), 2015. 1, 2, 6, 14

[64] 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), 2016. 15

[65] 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*, 2017. 3

[66] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset for Automatic Image Captioning. In *ACL*, 2018. 5

[67] 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*, 2024. 3

[68] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features through Propagating Activation Differences. In *ICML*, 2017. 3

[69] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*. Springer, 2012. 6, 14, 26

[70] 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, 2, 3, 5, 7, 14, 15, 16, 17, 19

[71] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *ICML*, 2017. 3, 6, 15, 19

[72] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. 8, 14, 26

[73] 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*, 2021. 3

[74] 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, 2, 3, 4, 5, 7, 14, 15, 16, 17, 19

[75] 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 Neu-

ral Networks. In *CVPRW*, 2020. 3

[76] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 14

[77] 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. 3

[78] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, 2023. 1, 3, 5, 6

[79] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. *IJCV*, 126(10), 2018. 15

[80] Zikai Zhou, Yunhang Shen, Shitong Shao, Linrui Gong, and Shaohui Lin. Rethinking centered kernel alignment in knowledge distillation. *arXiv preprint arXiv:2401.11824*, 2024. 3

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

Appendix

In this appendix to our work on scalable B-cosification of foundational Vision Transformers, we provide:

A. Implementation Details

In this section we provide additional implementation details that complement the main paper (Sec. 4.4), including details about the teacher checkpoints, datasets used for downstream evaluations, and finally the attribution methods along with the interpretability metrics we use.

Teacher checkpoints. For our feature-alignment step (Sec. 4.2) with pre-trained vision foundational models ([25, 70, 74]), we align to frozen teacher encoders of publicly available checkpoints and keep the associated text encoders for vision-language models unchanged (for SigLIP2 [74] zero-shot). Table A1 lists the exact model IDs and native image resolutions we use.

Table A1. **Teacher encoders used for alignment.** We adopt each teacher's native evaluation resolution and keep their weights frozen during alignment.

Family	Architecture	Eval res.	Identifier
Fully Supervised Models [31]			
Supervised [25]	ViT-B/16	224×224	`google/vit-base-patch16-224`
Vision–Language Models [28–30]			
SigLIP2 [74]	ViT-B/16	224×224	`google/siglip2-base-patch16-224`
SigLIP2 [74]	ViT-L/16	256×256	`google/siglip2-large-patch16-256`
SigLIP2 [74]	ViT-so400m/16	256×256	`google/siglip2-so400m-patch16-256`
SigLIP2 [74]	ViT-so400m@384/14	384×384	`google/siglip2-so400m-patch16-384`
Self-Supervised Models [47–49]			
DINOv3 [70]	ViT-S/16	224×224	`facebook/dinov3-vits16-pretrain-lvd1689m`
DINOv3 [70]	ViT-B/16	224×224	`facebook/dinov3-vitb16-pretrain-lvd1689m`
DINOv3 [70]	ViT-L/16	224×224	`facebook/dinov3-vitl16-pretrain-lvd1689m`

Input resolutions. Alignment and evaluation follow the teacher's native resolution (Table A1): 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-so400m/16, and 384×384 for SigLIP2-so400m@384/14.

Evaluation datasets. Linear Probing (LP) and k-NN share the same 10 datasets: IN1K [63], CALTECH101 [43], FLOWERS102 [52], FOOD101 [7], FGVC-AIRCRAFT [46], DTD [20], STANFORD CARS [40], SUN397 [76], CIFAR-10 [41], CIFAR-100 [41]. Dense linear probing for depth uses NYUV2 [42, 69].

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 [26] 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 [69] 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).

Surface normals estimation. We also evaluate surface normal estimation following the Probe3D [26] protocol by training a linear head on frozen features, and report δ_1, δ_2, δ_3, and RMSE.

Multiview correspondence. We evaluate multiview correspondence with Probe3D [26] on two different datasets: NAVI [35] and ScanNet [23]), reporting recall at standard error thresholds over view pairs from the same 3D scene.

Zero-shot (SigLIP2 [74]). We replace only the *image* encoder with its ALOE-aligned B-cos counterpart; the SigLIP2 *text* encoder, prompt templates, temperature, and normalization follow the originals [58, 74]. We report top-1 for a *single* class-name prompt ("a photo of a {class-name}") and the OpenCLIP 80-prompt template ([19, 34]) on ImageNet [63].

Multimodal large language model token grounding. For the multimodal large language model (MLLM) examples, we pair our ALOE-aligned SigLIP2 B-cos encoder with a LLaVA-More [21] Gemma-9B [72] language backbone. Relevance is propagated through the language model with AttnLRP [1], after which we extract model-inherent visual explanations from

the B-cos encoder to obtain token-level grounding maps.

Attribution methods visualized. For the inherently interpretable B-cos models [11], explanations are model-inherent $\mathbf{W}(\mathbf{x})\mathbf{x}$ [6]. For conventional teachers, we visualize AttnLRP [1], LeGrad [8] and CheferCAM [16] (using the authors' original implementation), and Integrated Gradients [71] as well as LIME [61] (from the `captum` library [38]). Where applicable, we use authors' recommended defaults.

Interpretability metrics. We evaluate model attributions with (i) the *Grid Pointing Game (GridPG)* [10, 79] for localization and (ii) *Pixel Deletion* [64] 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 most to least important and iteratively set the most important pixels to zero, plotting the target-class probability versus the removed-pixel fraction; *larger* drops (steeper curves) indicate attributions that are more consistent with model decisions.

B. Additional Quantitative Results

Table B1. **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 where a vanilla B-cosification baseline exists). ✓: denotes inherently interpretable models (vs. not ✗).

Feature	Arch	Inter.	IN1k	Cal101	Flowers	Food	Aircr	DTD	Cars	SUN	C10	C100	Avg.
Fully Supervised Pre-Training													
Sup. [25]	ViT-B/16	✗	81.16	97.65	99.74	86.16	41.28	74.69	57.06	74.26	97.12	86.54	79.57
B-cosif. [4]	B-ViT-B/16	✓	77.65	96.35	95.83	78.18	37.07	71.15	46.06	66.83	95.00	81.43	74.56
ALOE	B-ViT-B/16	✓	**81.12**	**97.78**	**99.87**	**86.57**	**44.17**	**75.00**	**60.29**	**74.62**	**97.42**	**87.21**	**80.41**
			+3.47	+1.43	+4.04	+8.39	+7.10	+3.85	+14.2	+7.79	+2.42	+5.78	+5.85
Vision Language Pre-training													
SigLIP2 [74]	ViT-B/16	✗	84.20	99.08	99.08	94.44	75.36	85.37	95.43	81.62	96.91	84.97	89.65
B-cosif. [4]	B-ViT-B/16	✓	75.05	97.52	94.14	83.74	45.04	79.18	73.70	75.68	94.44	81.11	79.96
ALOE	B-ViT-B/16	✓	**83.80**	**99.21**	**99.47**	**94.16**	**72.95**	**84.87**	**94.58**	**81.17**	**97.06**	**85.11**	**89.24**
			+8.75	+1.69	+5.33	+10.4	+27.9	+5.69	+20.9	+5.49	+2.62	+4.00	+9.28
Larger Architectures													
SigLIP2 [74]	ViT-L/16	✗	87.20	99.21	99.60	96.46	84.22	87.83	96.25	84.21	97.84	87.72	92.05
SigLIP2 [74]	ViT-so400m/16	✗	87.89	99.21	99.87	96.97	83.83	88.72	96.59	84.57	98.59	89.80	92.60
SigLIP2 [74]	ViT-so400m/14@384	✗	88.62	99.21	100.0	97.42	85.12	88.67	96.78	85.21	98.58	89.6	92.92
ALOE	B-ViT-L/16	✓	87.08	99.21	99.60	96.32	82.36	86.60	96.16	84.07	98.07	88.55	91.80
ALOE	B-ViT-so400m/16	✓	87.76	99.34	100.0	96.86	82.66	88.00	96.47	84.30	98.64	89.99	92.40
ALOE	B-ViT-so400m/16@432	✓	88.36	99.6	100.0	97.28	82.15	88.5	96.48	84.95	98.38	89.39	92.51
Self-Supervised Pre-training													
Smaller Architectures													
DINOv3 [70]	ViT-S/16	✗	78.64	98.43	99.74	89.62	73.25	80.8	91.59	74.50	96.24	85.00	86.78
ALOE	B-ViT-S/16	✓	77.72	98.30	99.74	87.10	71.39	79.68	91.04	73.64	95.34	84.21	85.97
Base Architecture													
DINOv3 [70]	ViT-B/16	✗	84.36	98.95	99.74	94.13	80.25	84.26	94.48	78.61	98.18	89.32	90.23
B-cosif. [4]	B-ViT-B/16	✓	73.64	95.18	82.68	68.24	41.67	68.02	50.90	58.25	92.04	76.44	70.71
ALOE	B-ViT-B/16	✓	**84.04**	**99.08**	**99.74**	**93.73**	**79.95**	**83.65**	**94.49**	**78.14**	**97.97**	**89.33**	**90.01**
			+10.4	+3.90	+17.1	+25.5	+38.3	+15.6	+43.6	+19.9	+5.93	+12.9	+19.3
Larger Architectures													
DINOv3 [70]	ViT-L/16	✗	86.92	98.95	99.74	95.86	80.64	86.94	94.68	80.76	99.13	93.24	91.69
ALOE	B-ViT-L/16	✓	86.64	98.95	99.74	95.58	80.37	85.99	94.88	80.25	99.16	92.57	91.41

In this section, we expand the quantitative evaluation along two axes: (1) *downstream performance*, and (2) *interpretability/faithfulness*. In Sec. B.1, for downstream performance, we report k-NN on frozen features (Tab. B2), zero-shot transfer for SigLIP2 (Tab. B3), dense prediction tasks (Tabs. B4 and B5, Fig. B1), and data-efficiency analyses (Fig. B2). In Sec. B.2, for interpretability and faithfulness, we quantify localization with GridPG (Tab. B6) and evaluate stability via pixel-deletion tests (Fig. B3). Additionally, we also report results on a human preference study. Unless noted otherwise, protocols match the main paper (see Sec. 4.4) and Sec. A.

Table B2. **k-NN accuracy on frozen features.** For the k-NN ($k = 20$) evaluation setting, ALOE B-cos ViTs again significantly outperform B-cosification [4] 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. 4.4). Teachers are shown in gray; best per block in **bold** (for B-cos models). ✓: denotes inherently interpretable models (vs. not ✗).

Feature	Arch	Inter.	IN1k	Cal101	Flowers	Food	Aircr	DTD	Cars	SUN	C10	C100	Avg.
Fully Supervised Pre-Training													
Sup. [25]	ViT-B/16	✗	80.72	92.44	79.55	78.88	22.17	63.17	29.76	68.62	96.41	82.30	69.40
B-cosif. [4]	B-ViT-B/16	✓	77.05	89.58	75.13	65.00	19.14	59.20	23.45	56.74	93.14	74.57	63.30
ALOE (ours)	B-ViT-B/16	✓	**80.77**	**92.05**	**79.42**	**79.58**	**23.16**	**63.00**	**30.82**	**69.24**	**96.80**	**83.34**	**69.82**
			+3.72	+2.47	+4.29	+14.58	+4.02	+3.80	+7.37	+12.50	+3.66	+8.77	+6.52
Vision Language Pre-training													
SigLIP2 [74]	ViT-B/16	✗	80.40	97.13	83.72	93.24	65.20	76.33	92.40	75.80	95.52	79.78	83.95
B-cosif. [4]	B-ViT-B/16	✓	68.42	94.14	74.87	76.16	25.45	72.43	49.57	68.46	91.74	72.63	69.39
ALOE (ours)	B-ViT-B/16	✓	**80.17**	**96.74**	**82.68**	**92.68**	**61.83**	**77.06**	**91.07**	**75.48**	**95.56**	**79.56**	**83.28**
			+11.75	+2.60	+7.81	+16.52	+36.38	+4.63	+41.50	+7.02	+3.82	+6.93	+13.89
Larger Architectures													
SigLIP2 [74]	ViT-L/16	✗	83.78	97.91	85.80	95.55	73.82	77.73	93.42	77.18	96.54	82.48	86.42
SigLIP2 [74]	ViT-so400m/16	✗	84.51	98.56	85.16	96.13	73.97	78.79	93.41	77.66	97.78	84.81	87.08
SigLIP2 [74]	ViT-so400m/14@384	✗	85.06	98.43	84.24	96.60	72.89	78.51	93.83	77.52	97.72	84.81	86.96
ALOE (ours)	B-ViT-L/16	✓	83.92	97.65	83.85	95.57	71.72	76.95	93.52	77.92	96.87	83.94	86.19
ALOE (ours)	B-ViT-so400m/16	✓	84.62	97.78	83.33	96.15	73.01	78.40	93.49	77.89	97.80	85.38	86.78
ALOE (ours)	B-ViT-so400m/16@432	✓	85.17	97.13	82.42	96.48	71.45	77.98	93.24	77.64	97.51	84.92	86.39
Self-Supervised Pre-training													
Smaller Architectures													
DINOv3 [70]	ViT-S/16	✗	76.91	93.75	87.50	85.85	51.89	74.88	83.56	67.96	95.76	81.66	79.97
ALOE (ours)	B-ViT-S/16	✓	75.70	94.21	87.10	84.40	50.48	73.71	82.71	67.05	94.89	79.99	79.02
Base Architecture													
DINOv3 [70]	ViT-B/16	✗	82.27	95.05	80.46	91.30	58.65	77.51	88.92	72.40	97.31	85.66	82.95
B-cosif. [4]	B-ViT-B/16	✓	71.03	83.33	41.27	44.97	16.25	44.64	22.90	34.86	87.32	61.18	50.77
ALOE (ours)	B-ViT-B/16	✓	**81.39**	**95.18**	**80.46**	**90.57**	**58.17**	**77.12**	**88.36**	**72.20**	**97.42**	**85.41**	**82.63**
			+10.36	+11.85	+39.19	+45.60	+41.92	+32.48	+65.46	+37.34	+10.10	+24.23	+31.86
Larger Architectures													
DINOv3 [70]	ViT-L/16	✗	84.73	94.92	73.69	93.74	57.27	77.06	90.49	74.52	98.48	90.03	83.49
ALOE (ours)	B-ViT-L/16	✓	84.35	94.01	71.61	93.23	56.07	77.62	90.33	74.77	98.61	89.21	82.98

B.1. Downstream performance

LP on frozen features. Across ten datasets, ALOE B-ViTs substantially outperform B-cosification while remaining competitive with their teachers (Tab. B1). Gains are especially pronounced on fine-grained or texture-heavy benchmarks (e.g., CARS, AIRCR, DTD, FOOD).

k-NN on frozen features. Similar to the LP performance Tab. B1 our models substantially outperform B-cosification while preserving most of the baseline's capabilities, indicating that alignment preserves discriminative structure in feature space without additional fine-tuning (see Tab. B2).

Zero-shot (SigLIP2 [74]). Replacing the SigLIP2 image encoder with its ALOE 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 ([19, 34]), while staying close to teacher performance (Tab. B3).

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. B4) by a large margin. This indicates that aligned B-cos features remain useful for dense prediction, not only global image classification.

Table B3. **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 the OpenCLIP 80-prompt template [19, 34]. 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.	Zero-shot Acc.
Base architecture		
SigLIP2 [74] — ViT-B/16	✗	78.07
B-cosif. [4] — B-ViT-B/16	✓	61.01
ALOE (ours) — B-ViT-B/16	✓	**77.20**
Δ *(ALOE vs. B-cosif.)*		+16.19
Larger architectures		
SigLIP2 [74] — ViT-L/16	✗	82.28
SigLIP2 [74] — ViT-so400m/16	✗	82.56
SigLIP2 [74] — ViT-so400m/14@384px	✗	83.53
ALOE (ours) — B-ViT-L/16	✓	**81.97**
ALOE (ours) — B-ViT-so400m/16	✓	**82.39**
ALOE (ours) — B-ViT-so400m/16@432	✓	**83.13**

Table B4. **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 [70]	✗	0.9669	0.2464	0.8706	0.3957
B-cosif. [4]	✓	0.8311	0.4604	0.6503	0.6804
ALOE (ours)	✓	**0.9416**	**0.2988**	**0.7850**	**0.4850**
Δ *(ALOE vs. B-cosif.)*		+0.1105	-0.1616	+0.1347	-0.1954

Table B5. **Surface Normals Estimation.** We report angular error metrics (δ_n) and RMSE using Probe3D. Higher is better for δ metrics; lower is better for RMSE. The Δ (**row 4.**) reports ALOE minus B-cosification.

Method	Inh. Inter.	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	RMSE$\downarrow$
Baseline	✗	0.58	0.77	0.83	23.7
B-cosif.	✓	0.38	0.61	0.70	31.7
ALOE (ours)	✓	**0.54**	**0.75**	**0.82**	**24.9**
Δ *(ALOE vs. B-cosif.)*		+0.16	+0.14	+0.12	-6.8

Surface normal estimation. On surface normal estimation, ALOE again approaches the teacher and clearly outperforms B-cosification across all angular accuracy metrics and RMSE (Tab. B5). In particular, ALOE improves over B-cosification by **+0.16**, **+0.14**, and **+0.12** on δ_1, δ_2, and δ_3, respectively, while reducing RMSE by **6.8**. This further indicates that aligned B-cos features preserve spatial structure useful for dense prediction beyond image-level recognition.

Multiview correspondence. Multiview correspondence attempts to match two images with different viewing angles from the same 3D scene. Using the Probe3D framework [26], we also evaluated the multiview correspondence on two different datasets (NAVI [35] and ScanNet [23]). The results are averaged over the different viewing angles and are shown for predefined error margins (1 cm, 2 cm, and 5 cm in our case). Figure B1 shows the evaluation results for this experiment.

Data efficiency. In addition to SigLIP2 (as demonstrated in Fig. 9 of the main paper), we report data-scaling results for DINOv3 that shows a similar trend (Fig. B2), although this ablation was performed on shorter training schedules.

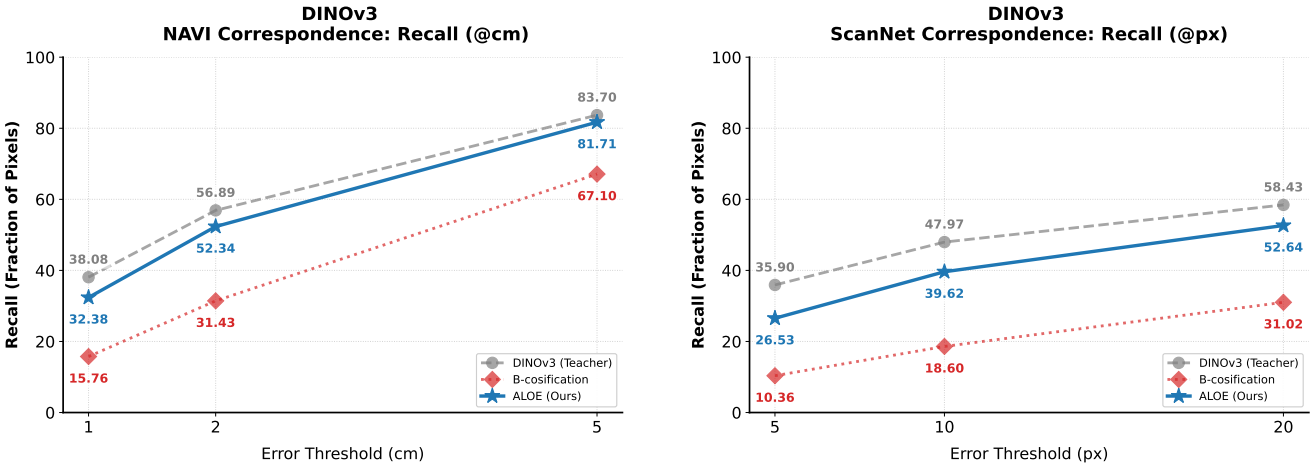


Figure B1. **Multiview correspondence.** Evaluated on two different datasets (left: NAVI [35], right: ScanNet [23]) using Probe3D [26]. ALOE increases performance substantially over B-cosification and closes the gap to the teacher model.

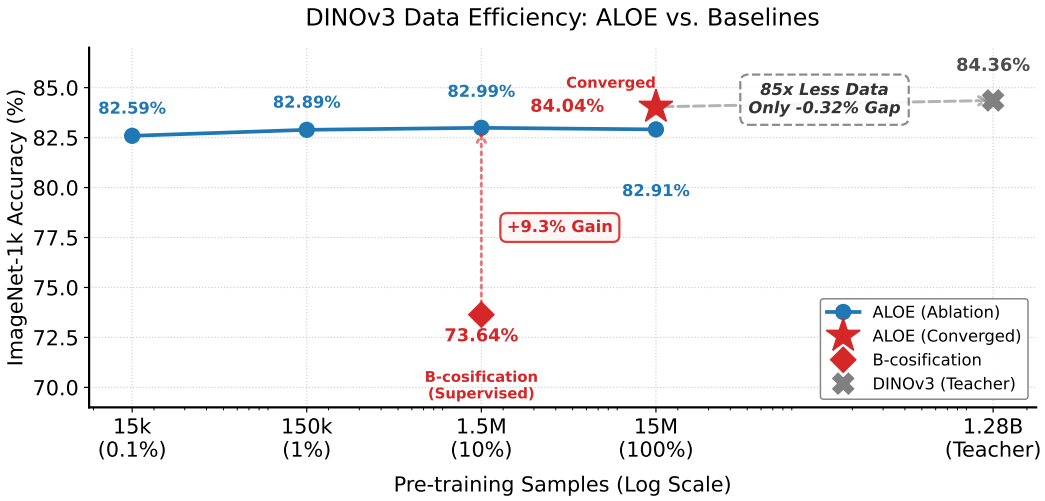


Figure B2. **Data efficiency of ALOE for DINOv3 on YFCC15M (ViT-B/16).** We vary the amount of unlabeled YFCC15M data used for label-free alignment from 0.1% (15k images) to 100% (15M images) and report ImageNet-1k linear-probe accuracy. Across these subsets, ALOE remains nearly flat (82.59%–82.99%), already substantially outperforming B-cosification (+9.3 p.p. at 1.5M images). With longer training to convergence, ALOE reaches 84.04%, leaving only a 0.32 p.p. gap to the DINOv3 teacher (84.36%).

Model	Lin. Probe ↑		k-NN ↑		Grid-PG ↑					Δ_{GridPG}
	Teacher (✗)	ALOE (✓)	Teacher (✗)	ALOE (✓)	LRP	IG	LeG	Chf	ALOE (✓)	
Vision–language teacher: SigLIP2 [74]										
ViT-B/16	84.20	83.80	80.40	80.17	54.43	38.76	-	-	**81.04**	+26.61
ViT-L/16	87.20	87.08	83.78	83.92	47.95	38.03	-	-	**78.20**	+30.25
ViT-so/16	87.89	87.76	84.51	84.62	48.84	*	-	-	**77.77**	+28.93
ViT-so/14@384	88.62	88.36	85.06	85.17	49.04	*	-	-	**79.19**	+30.15
Self-supervised teacher: DINOv3 [70]										
ViT-S/16	78.64	77.72	76.91	75.70	52.86	32.80	31.32	33.56	**79.55**	+26.69
ViT-B/16	84.36	84.04	82.27	81.39	62.02	36.24	33.69	34.44	**82.69**	+20.67
ViT-L/16	86.92	86.64	84.73	84.35	64.66	38.44	36.69	40.35	**80.69**	+16.03
Supervised teacher: ViT [25]										
ViT-B/16	81.16	81.12	80.72	80.77	55.80	43.06	-	-	**82.45**	+26.65

Table B6. **Localization (Grid-PG) vs. recognition.** ALOE improves Grid-PG substantially across backbones and teachers while maintaining competitive linear-probe and k-NN accuracy on Imagenet-1k. Grid-PG is reported for Teacher using AttnLRP (LRP) [1], Integrated Gradients (IG) [71], LeGrad (LeG) [8], and CheferCAM (Chf) [16], while ALOE uses model inherent B-cos attributions. Δ_{GridPG} is ALOE minus the best teacher baseline (LRP). ✓: denotes inherently interpretable models. For fields marked with - it was too expensive to compute the attribution method for the given model while fields marked with * indicate missing model-specific implementations.

B.2. Interpretability and faithfulness

Localization (GridPG). ALOE aligned models yield substantial improvements in GridPG localization across pre-training paradigms and backbones while maintaining competitive performance in both recognition and dense prediction tasks. Compared to conventional teacher models explained with post-hoc attribution methods (AttnLRP [1], Integrated Gradients [71], LeGrad [8], and CheferCAM [16]), inherently interpretable B-cos explanations from ALOE achieve consistently higher localization scores. The gains are substantial across SigLIP2, DINOv3, and supervised ViT backbones, with Δ_{GridPG} ranging from **+16.03** to **+30.25**, showing that aligned B-cos models provide more localized, class-specific attributions without sacrificing downstream accuracy.

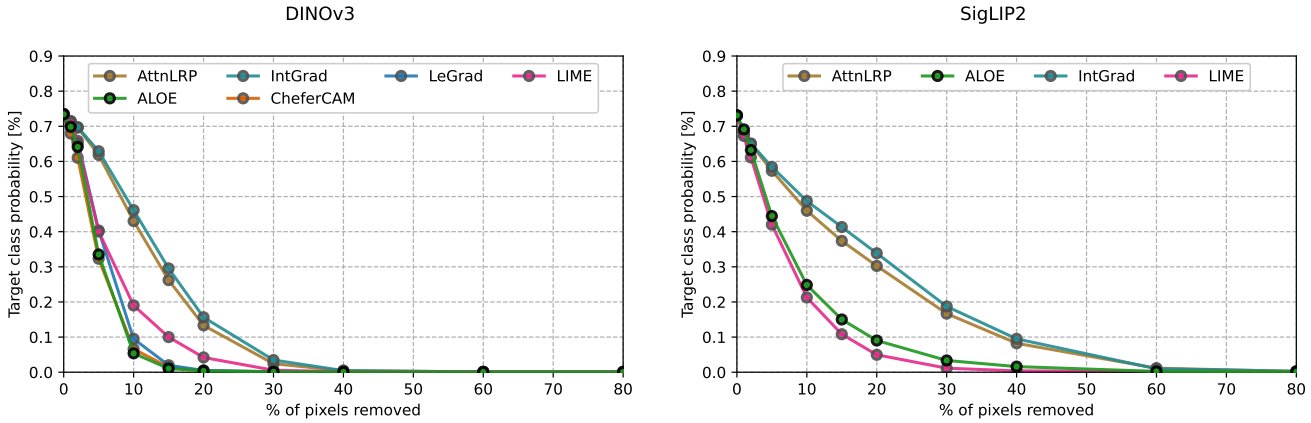


Figure B3. **Perturbation stability of explanations on ViT-B/16.** Target-class probability vs. percentage of top-attributed 16×16 pixel blocks removed (lower curves are better). Across both teachers—**DINOv3** (a) and **SigLIP2** (b)—**ALOE (ours)** using model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$ [11] removing most the blocks with the highest attribution leads to very fast drop in model confidence for ALOE, indicating stable and faithful localization. Interestingly, LIME [61] attributions computed for the SigLIP2 model perform the best, while ALOE still remains competitive.

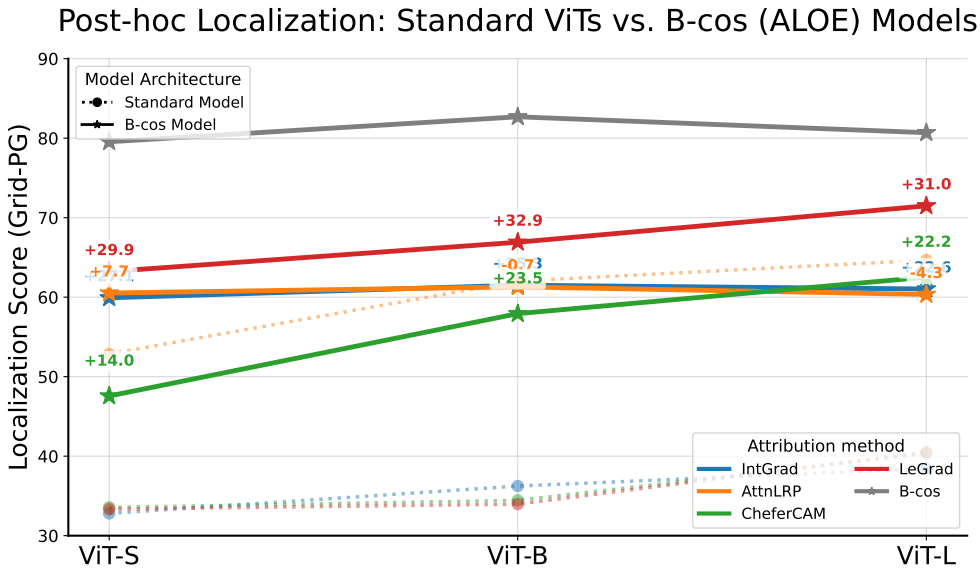


Figure B4. **GridPG on baseline and ALOE.** We computed the GridPG score on an DINOv3 ViT-B/16 model using different post-hoc attribution methods on both the teacher and distilled ALOE model. Applying post-hoc methods which depend on the attention scores and its gradient to ALOE models increase the score strongly, closing the gap to inherent explanations substantially.

Faithfulness under pixel perturbation. In pixel-deletion tests on ViT-B/16 (Fig. B3), target-class probability for ALOE (using model-inherent $\mathbf{W}(\mathbf{x})\mathbf{x}$) drops very fast for most important when 16x16 blocks with the highest attribution are removed, outperforming AttnLRP and Integrated Gradients for both DINOv3 and SigLIP2. Interestingly, LIME [61] attributions computed for the SigLIP2 model perform the best, while ALOE still remains competitive. The faster decay indicates more faithful attributions that better reflect the model's decision computations. The usage of blocks instead of single pixels is necessitated by the inclusion of CheferCAM, LIME and LeGrad which work internally at token-granularity; thus providing less fine-grained attribution maps compared to B-cos, AttnLRP or IntGrad.

Post-hoc method increase localization score. Applying the same post-hoc methods to the aligned ALOE models (see Fig. B4) substantially increases the GridPG score by up to $\sim 30\%$, which comes surprisingly close to the score achieved by the inherent explanation mechanism. This works for CheferCAM, LeGrad, and IntGrad but does not increase the score of AttnLRP. This suggests that B-cos models generally increase the localization of objects, independent of the attribution method (similar findings were also presented in [27]). AttnLRP might not benefit because it already achieves relatively high scores, at least compared to other post-hoc methods.

Human user study. We also performed a human user study in which the subjective preferences of different attribution methods for DINOv3 ViT-B/16 were evaluated. 41 users on Amazons Mechanical Turk performed 50 trials, each consisting of a random image with 2 attribution maps from randomly drawn attribution methods (AttnLRP, IntGrad, Chefer, LeGrad, LIME, B-cos). An example for one trial is shown in Fig. B5. The predicted class of the image was provided, and the user had to choose the explanation that best helped them understand the model's prediction. We fitted a Bradley–Terry (BT) model [9] to the outcome of the study in order to derive the pairwise win-probabilities between attribution methods. BT is similar to logistic regression on the one-hot encoded study data, where the winning class receives the score $+1$ and the negative -1 (with 1 as the target value). The derived BT weights indicate the overall performance relative to the other models. By computing the sigmoid of the difference $\beta_1 - \beta_2$ between the weights of two attribution methods, we can derive the probability of method 1 winning against method 2. The probabilities, together with the BT-scores, are shown in Tab. B7. ALOE is preferred over all post-hoc baselines ($> 50\%$ win rate) while being the most compute-efficient method.

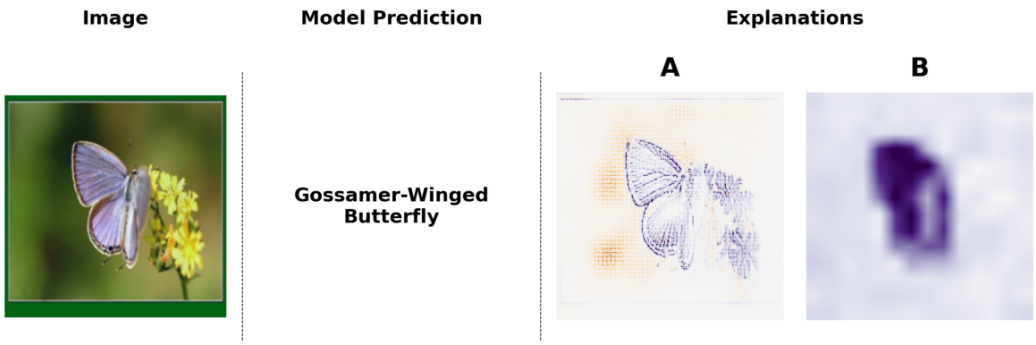


Figure B5. **Example for the User Study.** Users were given two attribution maps from randomly drawn methods and had to decide which attribution map helps them best to understand the model's prediction. Blue shows positive attributions while yellow shows negative ones.

Table B7. **Human Evaluation.** ALOE achieves the highest Bradley–Terry (BT) score and outperforms all baselines in pairwise comparisons (Win Rate $> 50\%$).

Method	BT Score ↑	ALOE Win Rate ↑
ALOE	**0.97**	–
IntGrad	0.41	63.5%
AttnLRP	0.11	70.2%
CheferCAM	-0.11	74.5%
LeGrad	-0.59	82.6%
LIME	-0.79	85.3%

B.3. Ablations

We present the table on ablation for *alignment objectives* as reported in Sec. 5.2 of the main paper.

Table B8. **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

Alignment objectives. Under identical settings (Tab. B8), *cosine* and *SigLIP* are most consistent across models, *MSE* and *InfoNCE* showed inconsistencies. Given the ease of application and simplicity we adopt **cosine** by default.

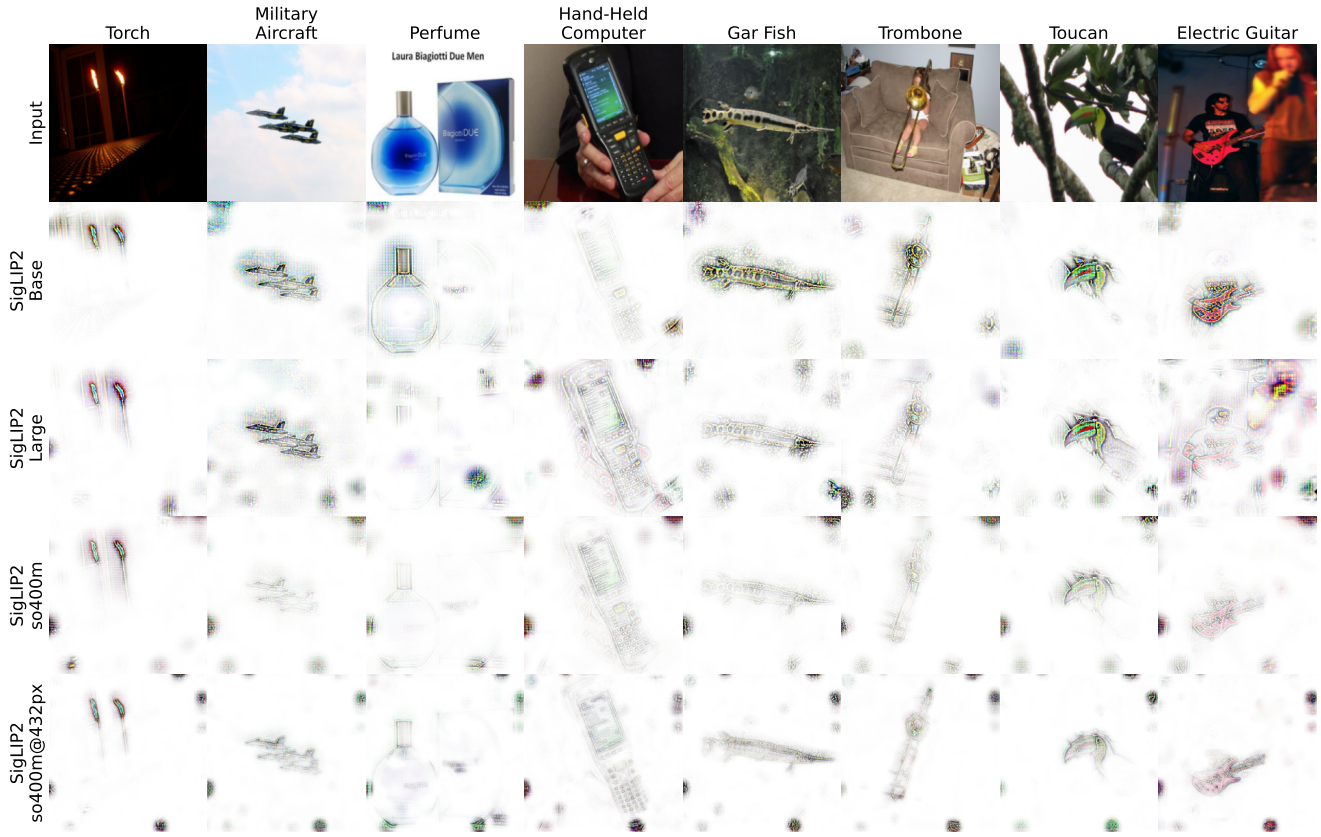


Figure C1. **Zero-shot, model-inherent VLM explanations.** Qualitative comparison of **ALOE (ours)** using $\mathbf{W}(\mathbf{x})\mathbf{x}$ attributions vs. AttnLRP, given the fixed text prompt *"A photo of a* $\{$*class-name*$\}$*"*. Our explanations are sharply localized on class-relevant regions, whereas AttnLRP appears diffuse and noisy.

C. Additional Qualitative Results

In this section, we complement the quantitative results with qualitative evidence across three settings: (i) *zero-shot, model-inherent* explanations (Figure C1), (ii) side-by-side comparisons against popular post-hoc attribution methods (Figures C2 to C4), (iii) dense predictions from a linear depth probe (Figure C5), and multimodal large language model explanations (Figs. C6 to C8).

Zero-shot, model-inherent VLM explanations. In Figure C1, 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. B3.

Comparisons with popular attribution methods. Figures C2 to C4 contrasts **ALOE (ours)**—which uses model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$—with AttnLRP, Integrated Gradients, LeGrad, CheferCAM, and LIME. 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. B6 and Fig. B3). *All examples for this use DINOv3-based backbones (ViT-B/16) with linear probes trained on ImageNet-1k.*

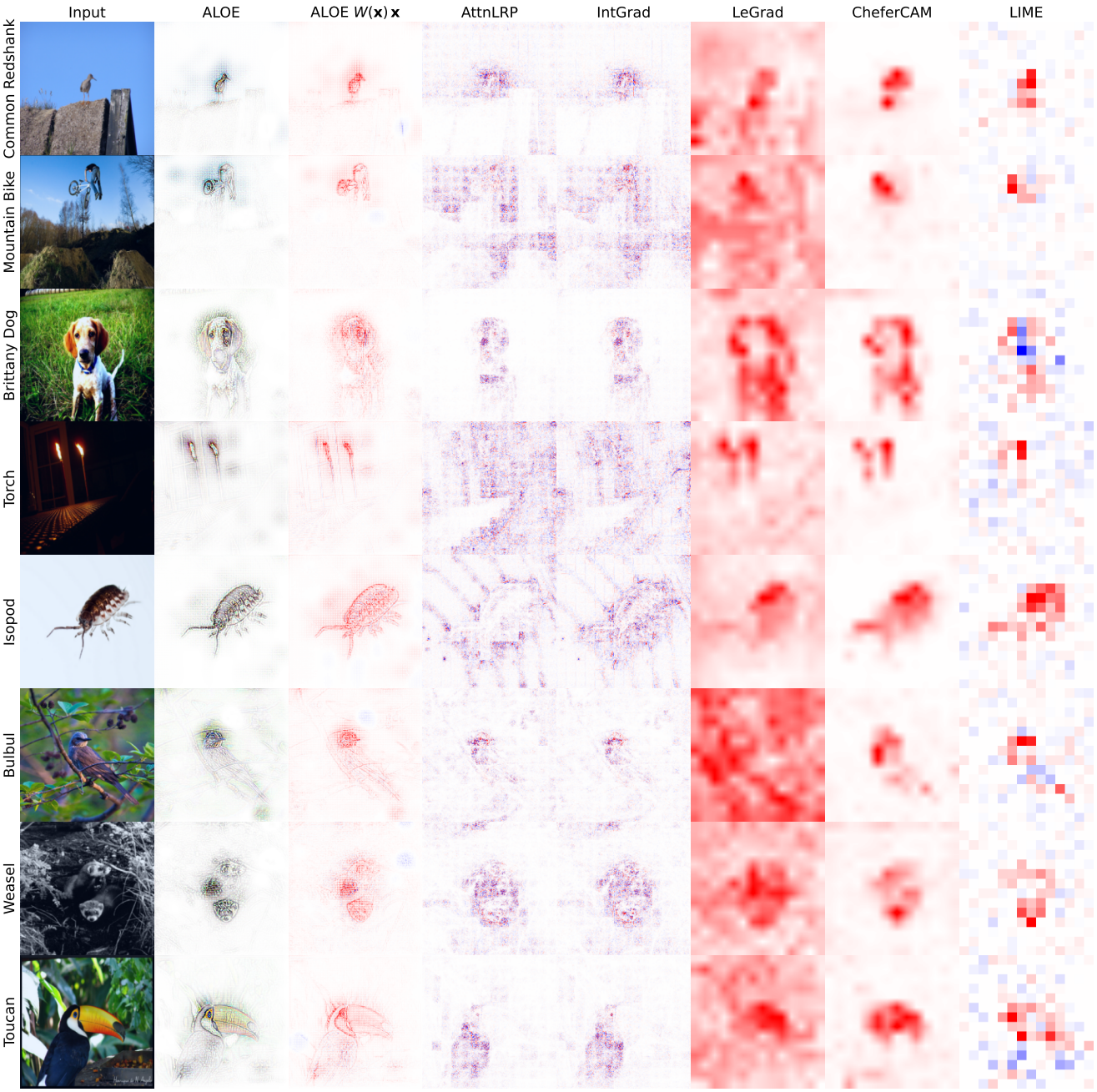


Figure C2. **Qualitative attribution comparisons.** Example 1: Visualizations for **ALOE (ours)**—using model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$—versus popular post-hoc methods (AttnLRP, Integrated Gradients (IntGrad), LeGrad, CheferCAM, and LIME). ALOE produces sharper, better-localized, and color-faithful highlight maps with less background noise, focusing on class-relevant object regions consistently across examples.

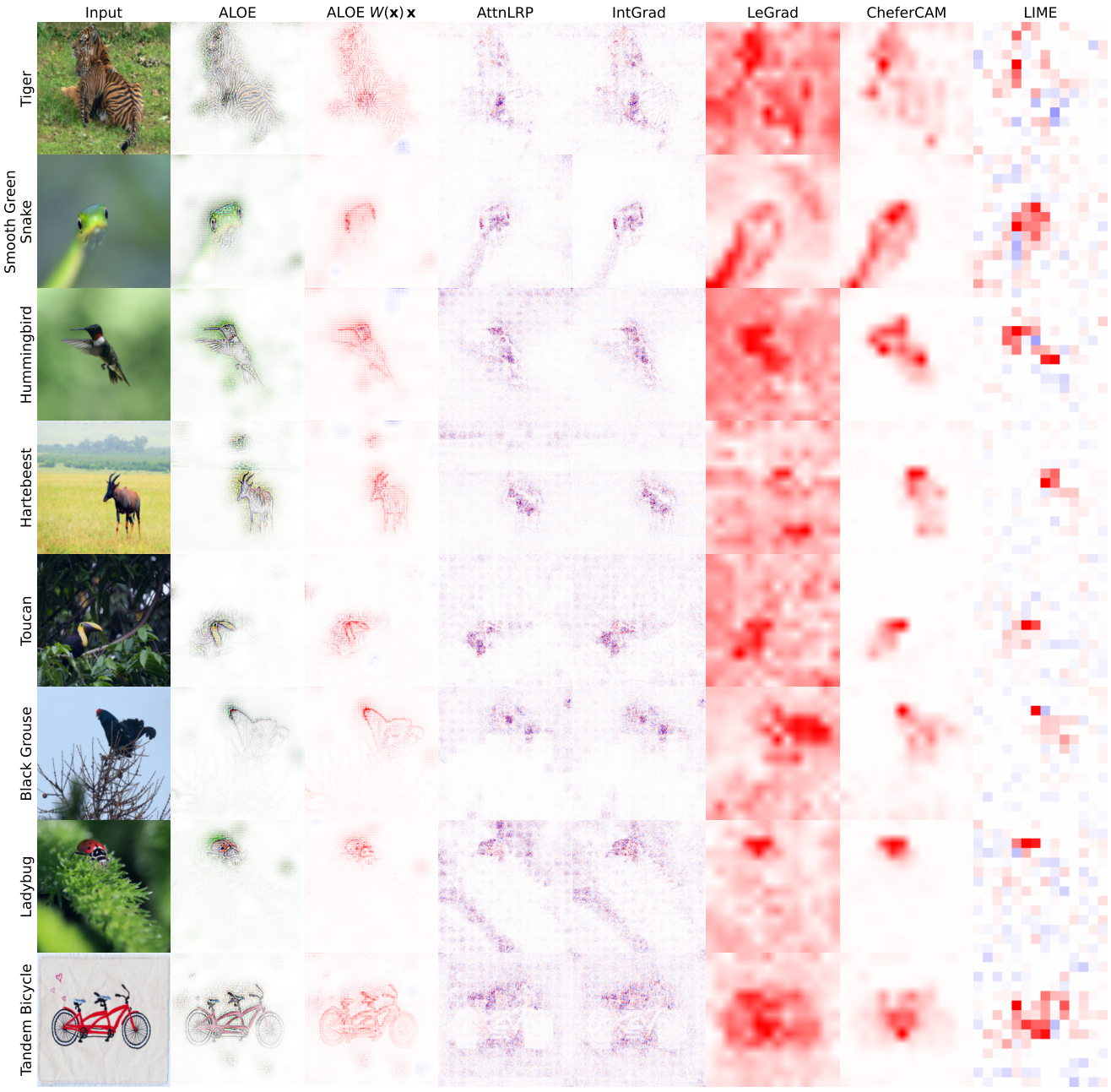


Figure C3. **Qualitative attribution comparisons.** Similar to Figure C2.

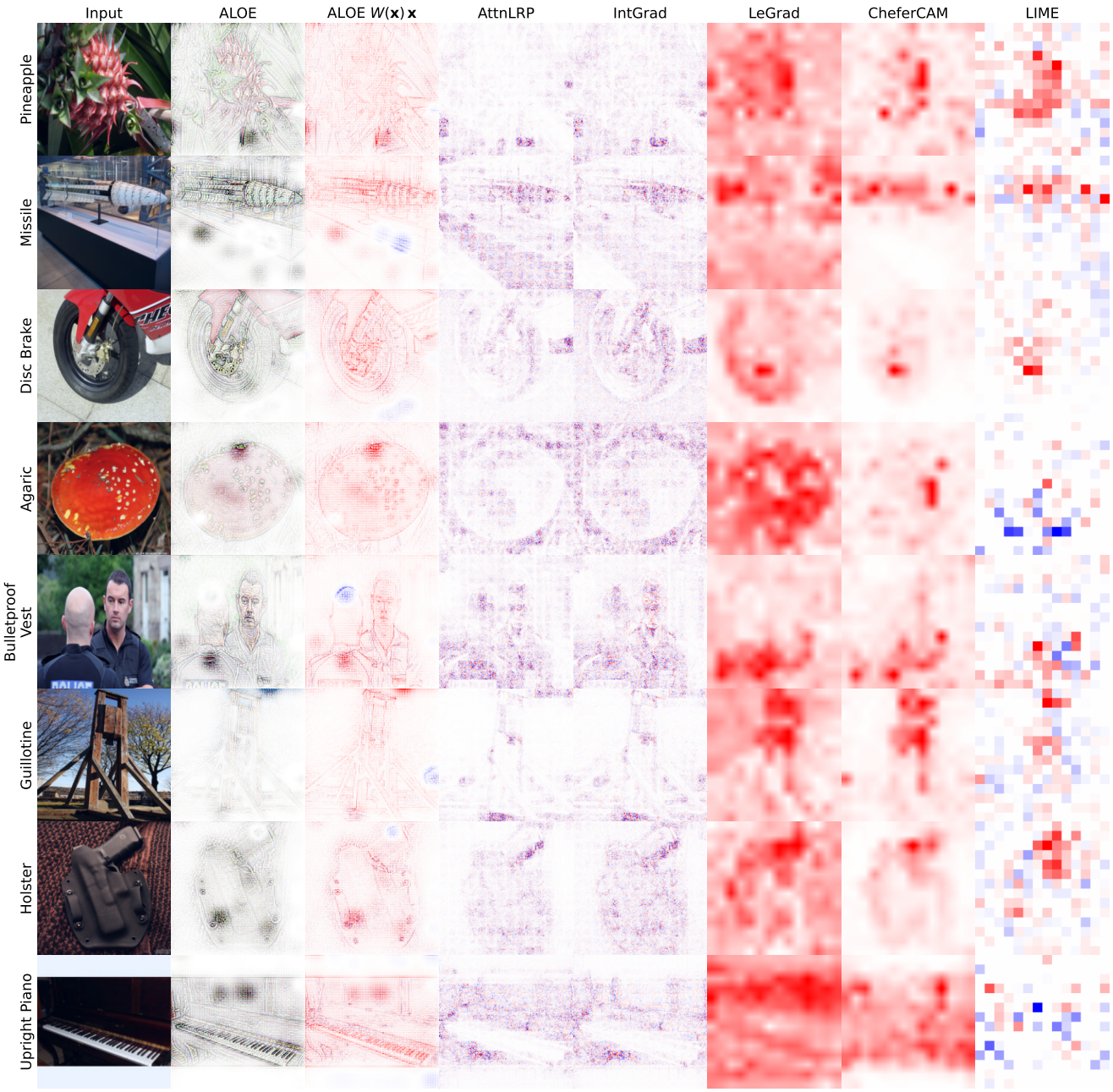


Figure C4. **Qualitative attribution comparisons–not well-localized examples.** Visualizations for **ALOE (ours)**—using model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$—versus popular post-hoc methods (AttnLRP, Integrated Gradients (IntGrad), LeGrad, CheferCAM, and LIME). We highlight failure cases where explanations lack localization, suggesting a reliance on background context. Given that B-cos attributions are mathematically faithful to the model's linear transformation, these diffuse heatmaps could also expose model-level shortcuts rather than "explanation failure."

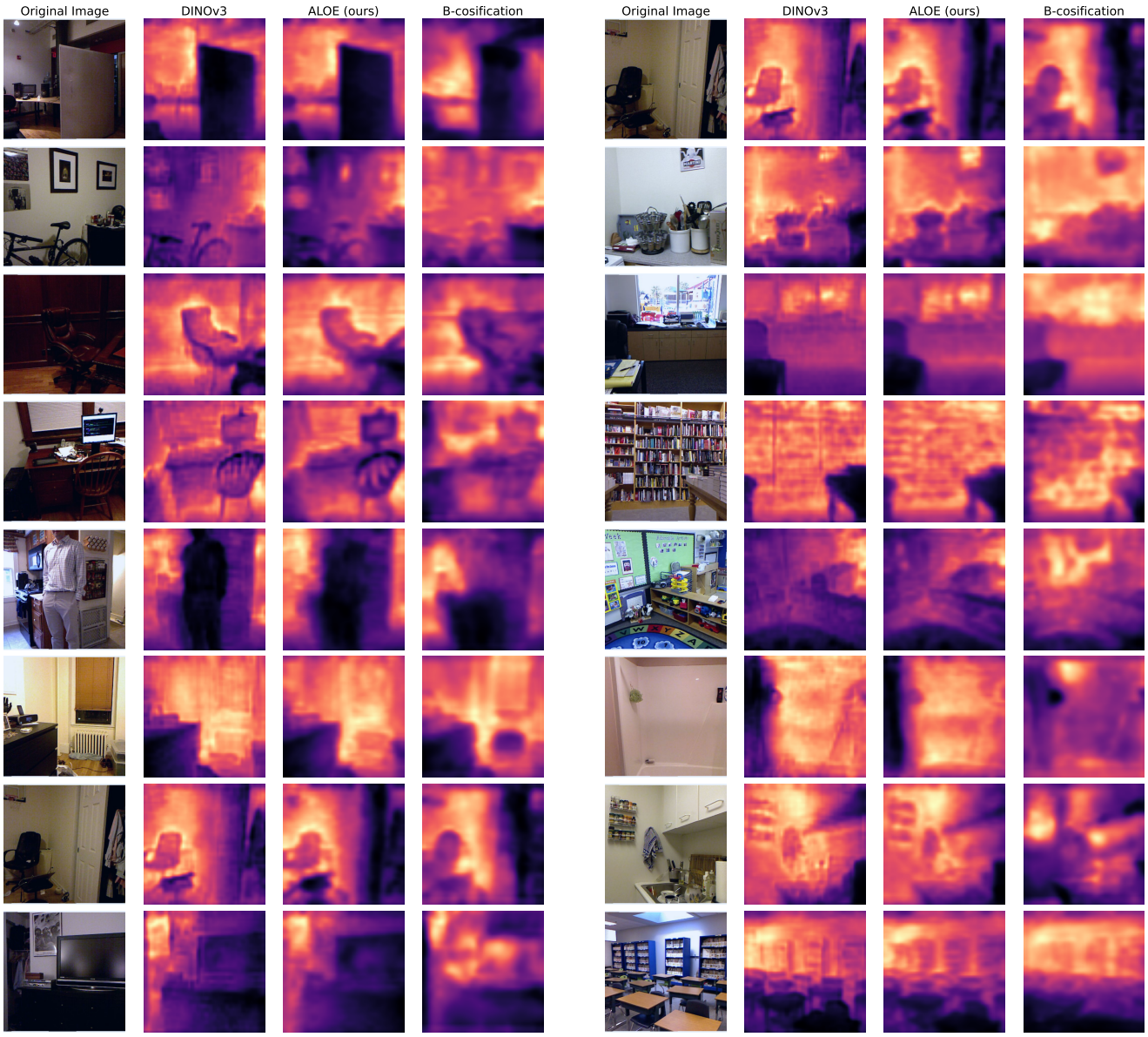


Figure C5. **Depth Estimation.** Visualization of predicted depth maps for **ALOE (ours)**, DINOv3 and B-cosification on the NYUv2 [69] depth-estimation dataset using the Probe3D [26] protocol. ALOE-aligned features yield depth maps with coherent geometry that visually are very similar to the DINOv3 teacher model. The depth maps from the vanilla B-cosification model seem to be more noisy and blurred.

Depth maps (dense linear probing). Figure C5 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 are visually very similar to the DINOv3 teacher. These visuals complement the relative/absolute depth metrics in Tab. B4, 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 noisier and more blurred.

Multimodal large language model explanations. In addition to the explanations in Fig. 8, we provide further examples—including failure cases—in Figs. C6 to C8. Using AttnLRP [1], we propagate relevance through the LLaVA-More [21] Gemma-9B [72] backbone, utilizing B-cos inherent explanations from the ALOE SigLIP2 vision encoder. Notably, Fig. C7 and Fig. C8 demonstrate instances where explanations fail to match predictions. We suspect that this discrepancy likely stems from three factors: the model relying on spurious associations, reliance on previous tokens restricting the vision-to-language

information flow, or AttnLRP fundamentally generating unfaithful explanations. As we discussed earlier, since the current setup still relies on AttnLRP to propagate relevance through the language model, we defer a fuller treatment of end-to-end inherently interpretable MLLMs to future work.

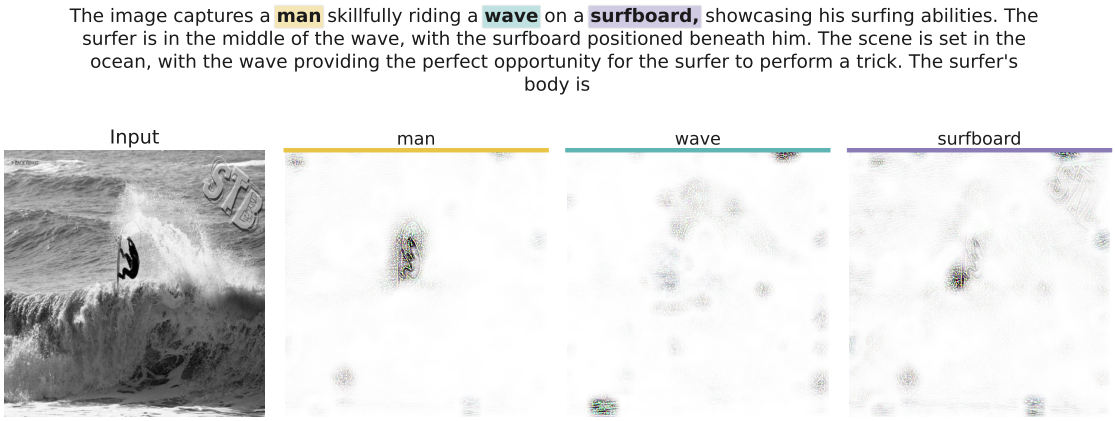


Figure C6. **Token-level Visual Grounding via AttnLRP and ALOE.** We generate visual explanations by leveraging the inherent B-cos interpretability of the ALOE SigLIP2 vision backbone, using AttnLRP to propagate relevance from LLaVA-MORE's GEMMA-9B language model.

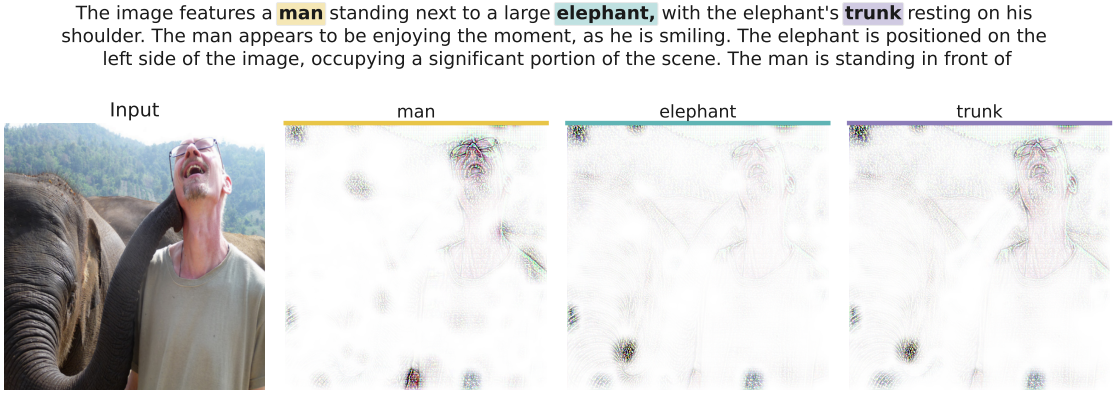


Figure C7. **Token-level Visual Grounding via AttnLRP and ALOE.** Similar to Fig. C6

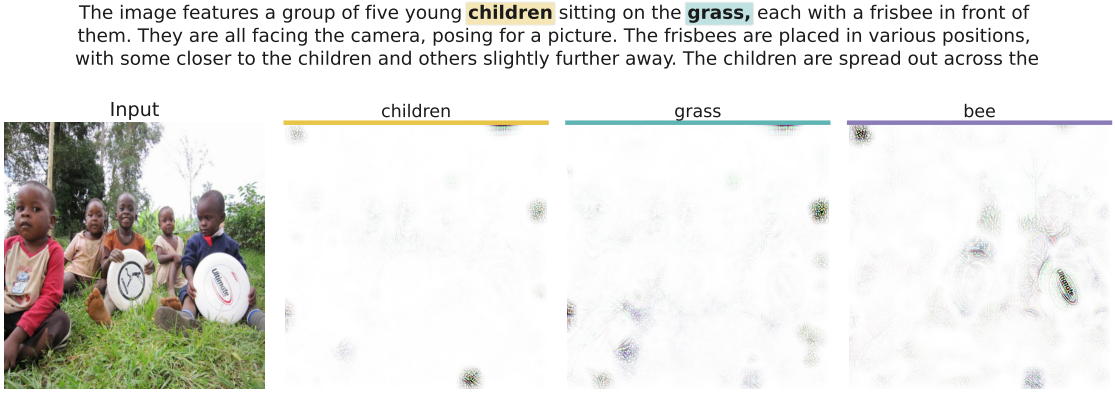


Figure C8. **Token-level Visual Grounding via AttnLRP and ALOE.** Similar to Figs. C6 and C7