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 ([14, 38, 41]), we align to frozen teacher encoders of publicly available checkpoints and keep the associated text encoders for vision-language models unchanged (for SigLIP2 [41] 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 [20]			
Supervised [14]	ViT-B/16	224×224	`google/vit-base-patch16-224`
Vision–Language Models [17–19]			
SigLIP2 [41]	ViT-B/16	224×224	`google/siglip2-base-patch16-224`
SigLIP2 [41]	ViT-L/16	256×256	`google/siglip2-large-patch16-256`
SigLIP2 [41]	ViT-so400m/16	256×256	`google/siglip2-so400m-patch16-256`
SigLIP2 [41]	ViT-so400m@384/14	384×384	`google/siglip2-so400m-patch16-384`
Self-Supervised Models [29–31]			
DINOv3 [38]	ViT-S/16	224×224	`facebook/dinov3-vits16-pretrain-lvd1689m`
DINOv3 [38]	ViT-B/16	224×224	`facebook/dinov3-vitb16-pretrain-lvd1689m`
DINOv3 [38]	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 [35], CALTECH101 [27], FLOWERS102 [32], FOOD101 [4], FGVC-AIRCRAFT [28], DTD [11], STANFORD CARS [24], SUN397 [42], CIFAR-10 [25], CIFAR-100 [25]. Dense linear probing for depth uses NYUV2 [26, 37].

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 [15] 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 [37] 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 [15] 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 [15] on two different datasets: NAVI [22] and ScanNet [13]), reporting recall at standard error thresholds over view pairs from the same 3D scene.

Zero-shot (SigLIP2 [41]). 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 [33, 41]. We report top-1 for a *single* class-name prompt ("a photo of a {class-name}") and the OpenCLIP 80-prompt template ([10, 21]) on ImageNet [35].

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 [12] Gemma-9B [40] 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 [8], explanations are model-inherent $\mathbf{W}(\mathbf{x})\mathbf{x}$ [3]. For conventional teachers, we visualize AttnLRP [1], LeGrad [5] and CheferCAM [9] (using the authors' original implementation), and Integrated Gradients [39] as well as LIME [34] (from the `captum` library [23]). Where applicable, we use authors' recommended defaults.

Interpretability metrics. We evaluate model attributions with (i) the *Grid Pointing Game (GridPG)* [7, 43] for localization and (ii) *Pixel Deletion* [36] 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. [14]	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. [2]	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 [41]	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. [2]	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 [41]	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 [41]	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 [41]	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 [38]	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 [38]	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. [2]	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 [38]	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 [2] 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. [14]	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. [2]	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 [41]	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. [2]	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 [41]	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 [41]	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 [41]	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 [38]	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 [38]	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. [2]	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 [38]	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 [41]). 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 ([10, 21]), 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 [10, 21]. 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 [41] — ViT-B/16	✗	78.07
B-cosif. [2] — B-ViT-B/16	✓	61.01
ALOE (ours) — B-ViT-B/16	✓	**77.20**
Δ *(ALOE vs. B-cosif.)*		+16.19
Larger architectures		
SigLIP2 [41] — ViT-L/16	✗	82.28
SigLIP2 [41] — ViT-so400m/16	✗	82.56
SigLIP2 [41] — 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 [38]	✗	0.9669	0.2464	0.8706	0.3957
B-cosif. [2]	✓	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 [15], we also evaluated the multiview correspondence on two different datasets (NAVI [22] and ScanNet [13]). The results are averaged over the different viewing angles and are shown for pre-defined 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.

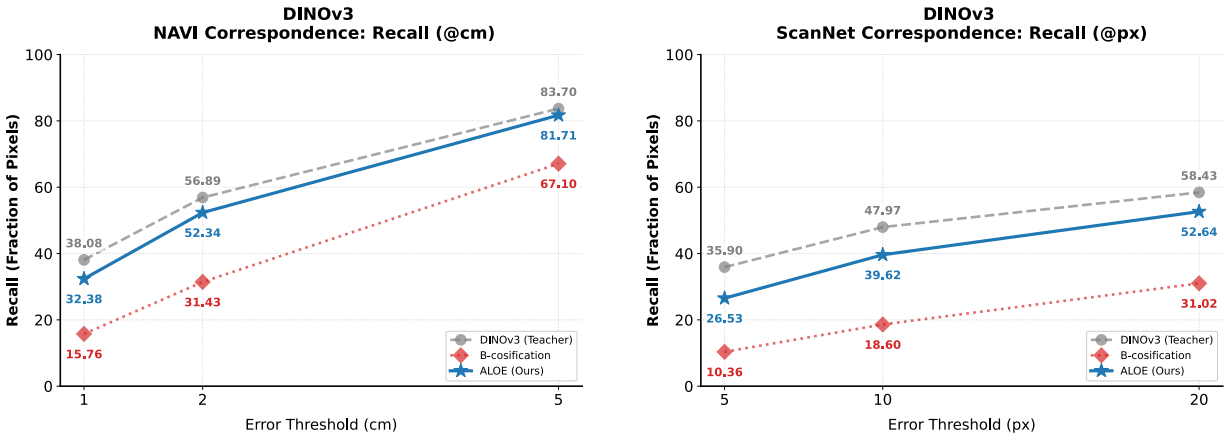

Figure B1. **Multiview correspondence.** Evaluated on two different datasets (left: NAVI [22], right: ScanNet [13]) using Probe3D [15]. ALOE increases performance substantially over B-cosification and closes the gap to the teacher model.

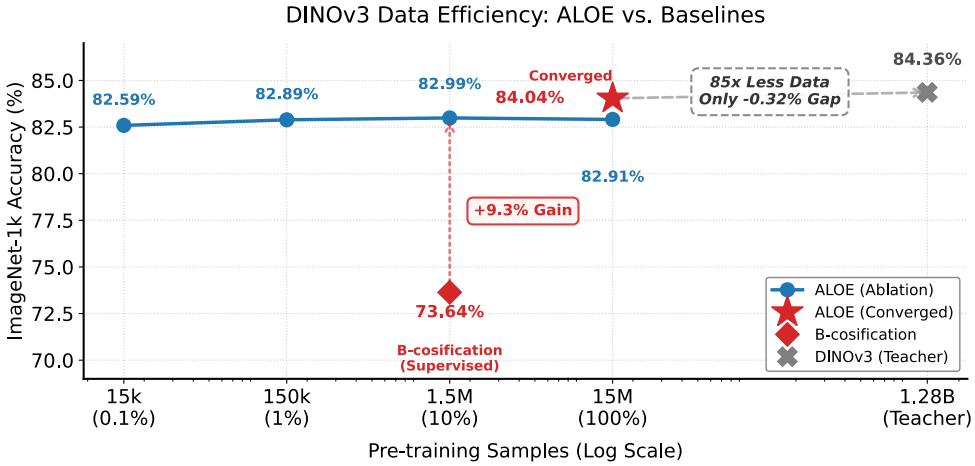

Figure B2. **Data efficiency of ALOE for DINOv3 on YFCC15M (ViT-B/16).** We vary the amount of unlabeled YFCC15M data used for label-free alignment from 0.1% (15k images) to 100% (15M images) and report ImageNet-1k linear-probe accuracy. Across these subsets, ALOE remains nearly flat (82.59%–82.99%), 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 [41]										
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 [38]										
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 [14]										
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) [39], LeGrad (LeG) [5], and CheferCAM (Chf) [9], 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 [39], LeGrad [5], and CheferCAM [9]), 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.

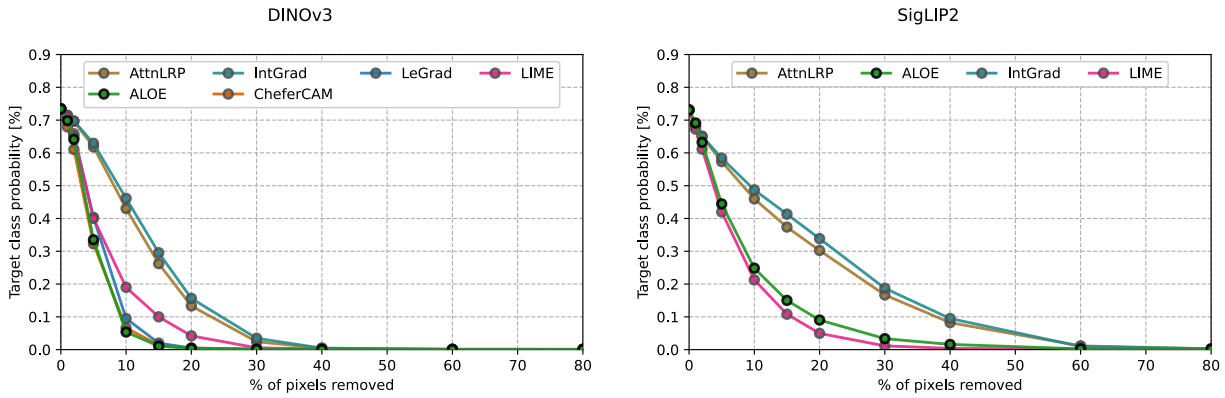

Figure B3. **Perturbation stability of explanations on ViT-B/16.** Target-class probability vs. percentage of top-attributed 16×16 pixel blocks removed (lower curves are better). Across both teachers—**DINOv3** (a) and **SigLIP2** (b)—**ALOE (ours)** using model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$ [8] 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 [34] attributions computed for the SigLIP2 model perform the best, while ALOE still remains competitive.

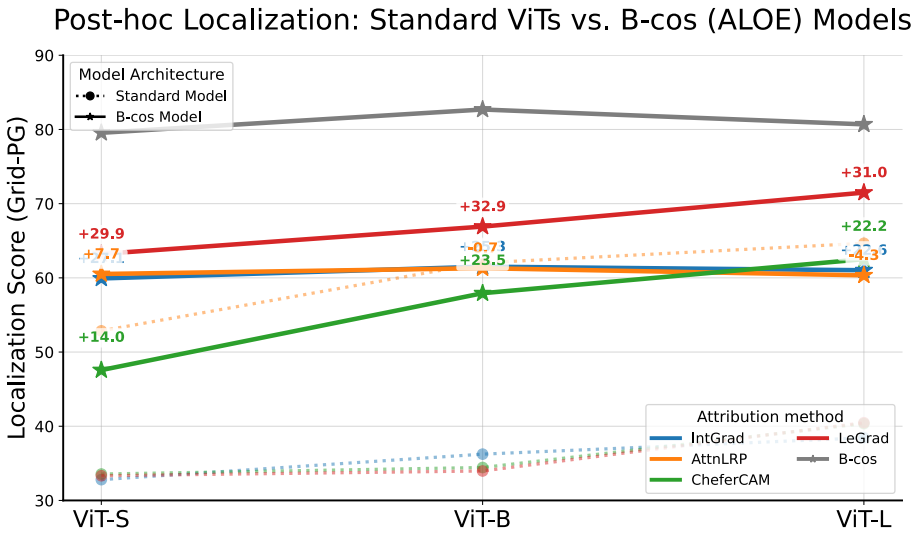

Figure B4. **GridPG on baseline and ALOE.** We computed the GridPG score on an DINOv3 ViT-B/16 model using different post-hoc attribution methods on both the teacher and distilled ALOE model. Applying post-hoc methods which depend on the attention scores and its gradient to ALOE models increase the score 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 [34] 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 [16]). 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 [6] 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.

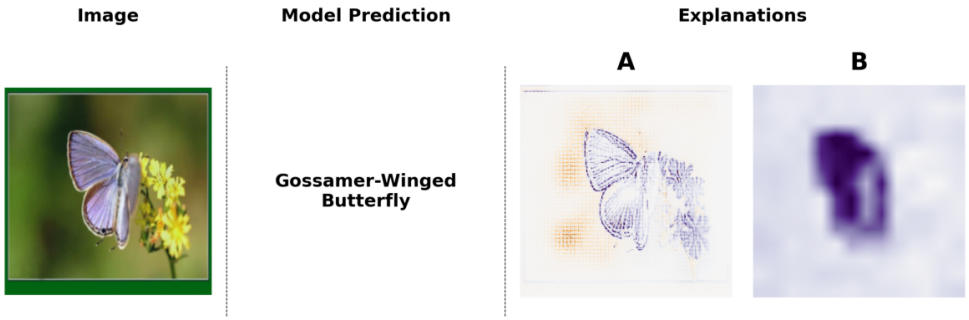

Figure B5. **Example for the User Study.** Users were given two attribution maps from randomly drawn methods and had to decide which attribution map helps them best to understand the model's prediction. Blue shows positive attributions while yellow shows negative ones.

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.

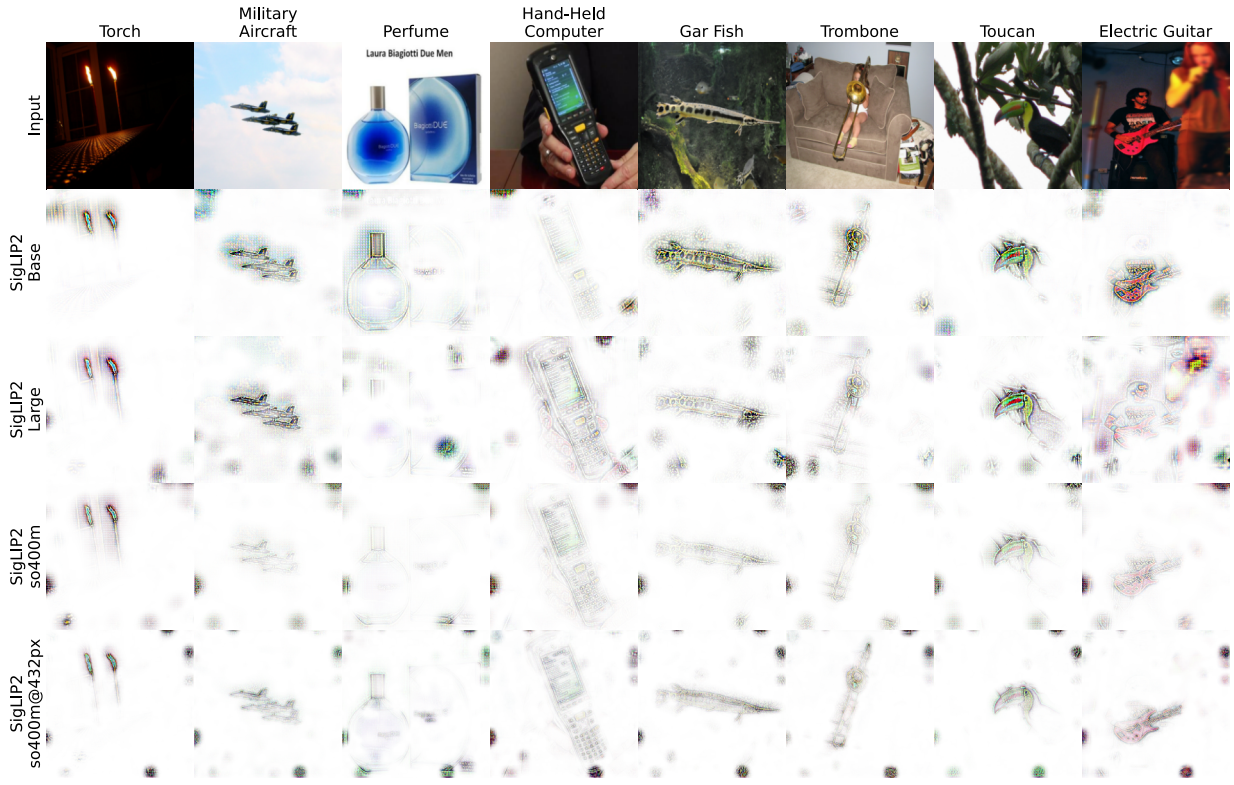

Figure C1. **Zero-shot, model-inherent VLM explanations.** Qualitative comparison of **ALOE (ours)** using $\mathbf{W}(\mathbf{x})\mathbf{x}$ attributions vs. AttnLRP, given the fixed text prompt *"A photo of a {class-name}"*. Our explanations are sharply localized on class-relevant regions, whereas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

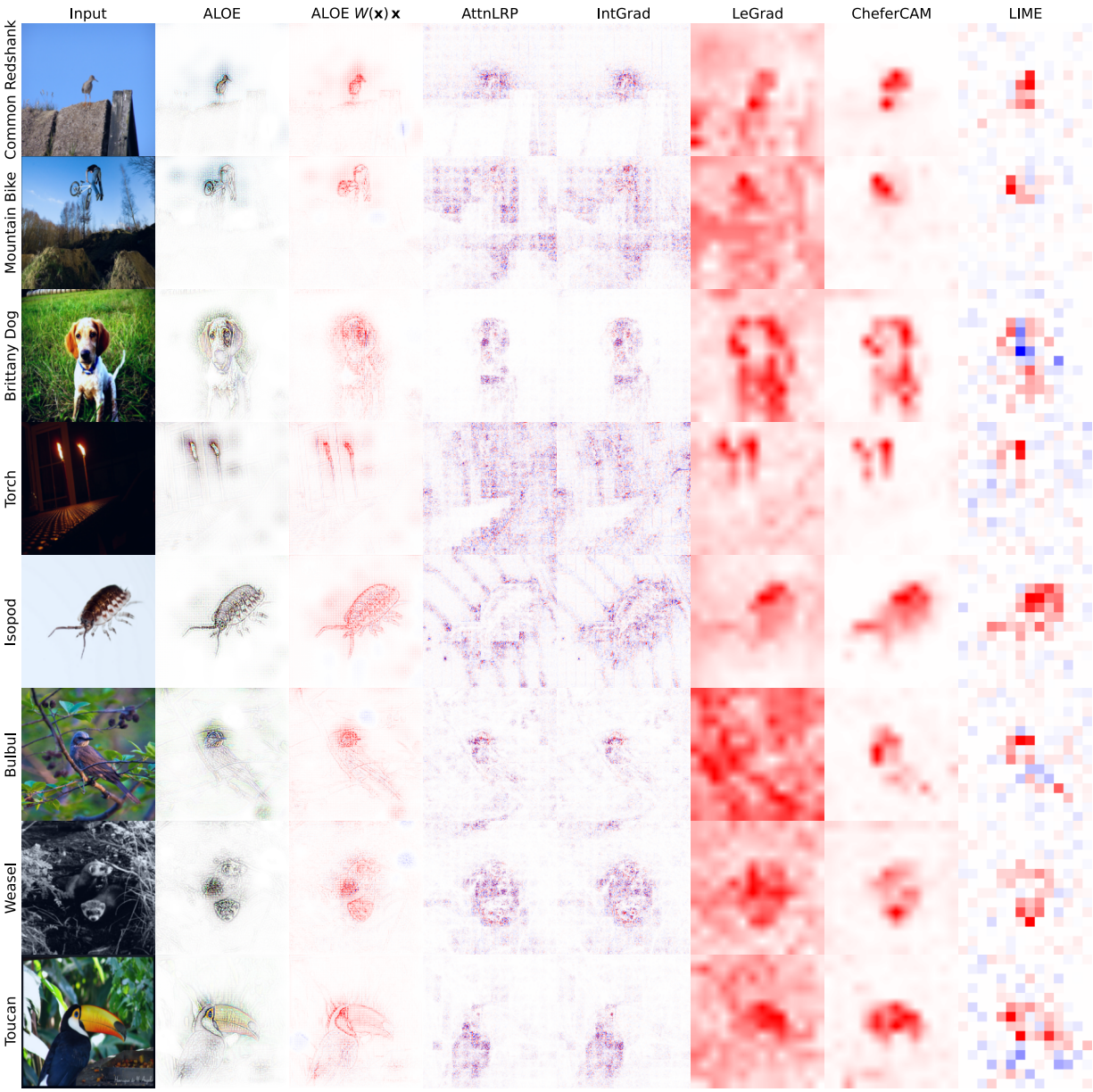

Figure C2. **Qualitative attribution comparisons.** Example 1: Visualizations for **ALOE (ours)**—using model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$—versus popular post-hoc methods (AttnLRP, Integrated Gradients (IntGrad), LeGrad, CheferCAM, and LIME). ALOE produces sharper, better-localized,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

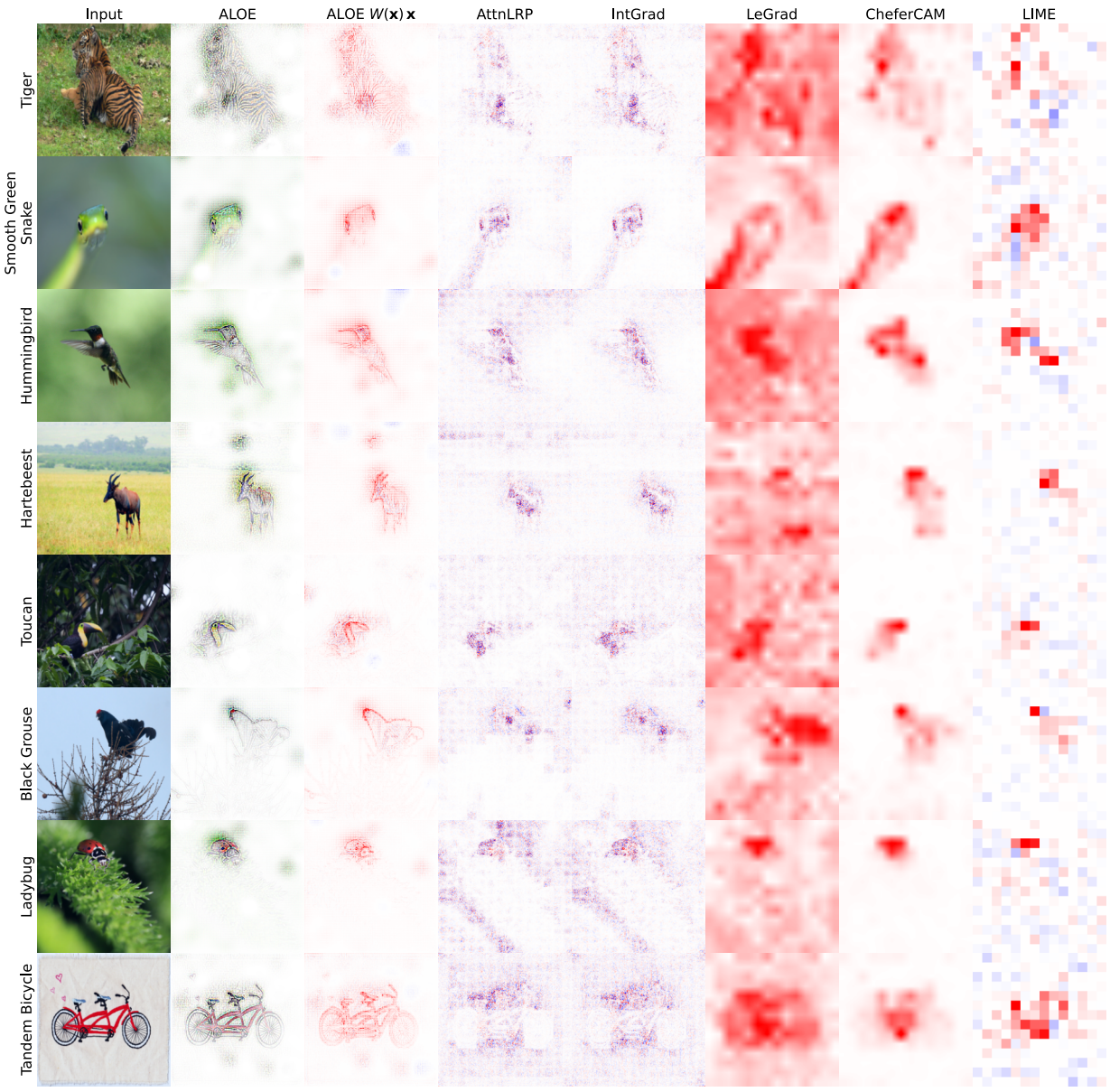

Figure C3. **Qualitative attribution comparisons.** Similar to Figure C2.

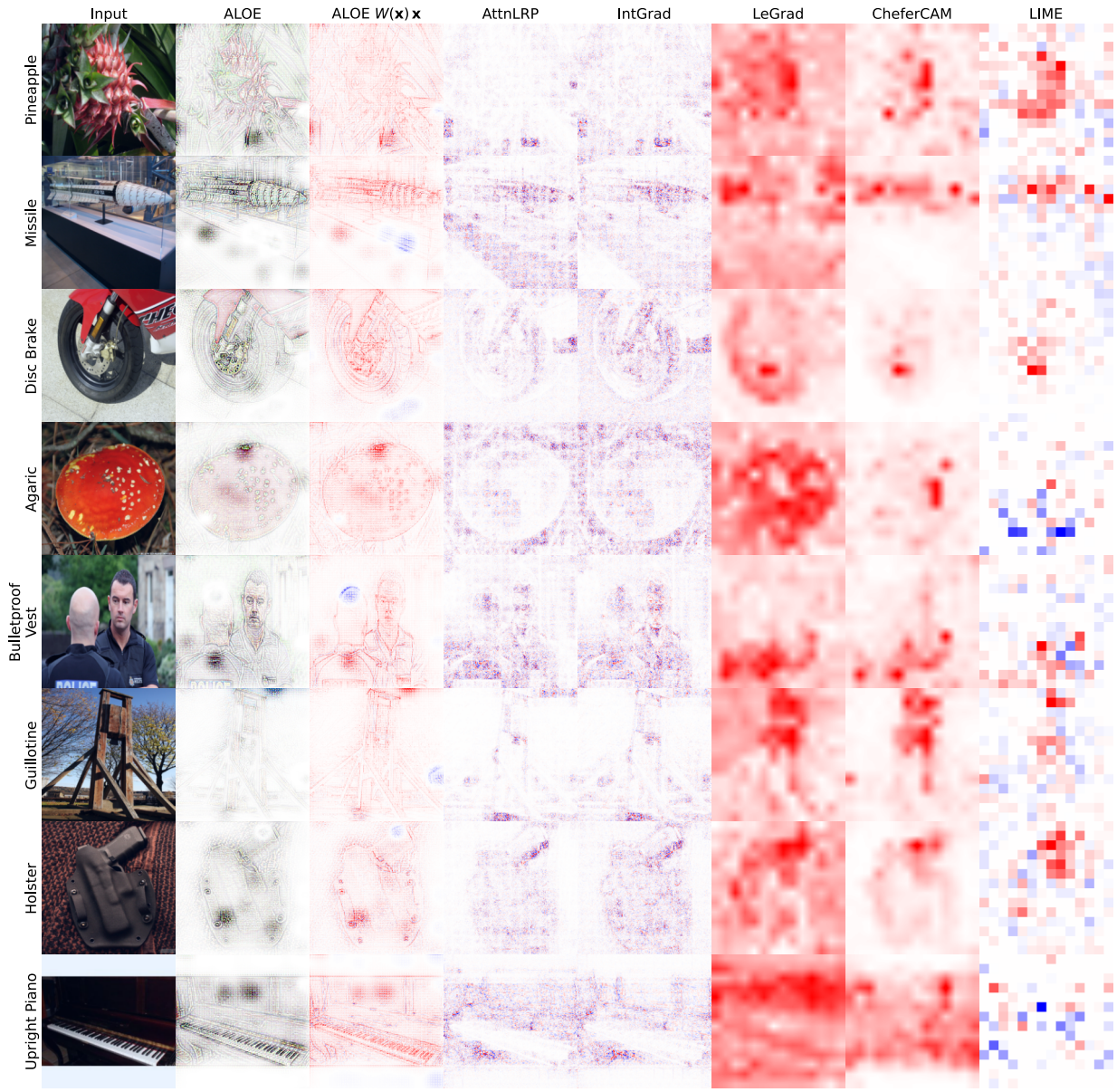

Figure C4. **Qualitative attribution comparisons–not well-localized examples.** Visualizations for **ALOE (ours)**—using model-inherent B-cos attributions $\mathbf{W}(\mathbf{x})\mathbf{x}$—versus popular post-hoc methods (AttnLRP, Integrated Gradients (IntGrad), LeGrad, CheferCAM, and LIME). We highlight 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."

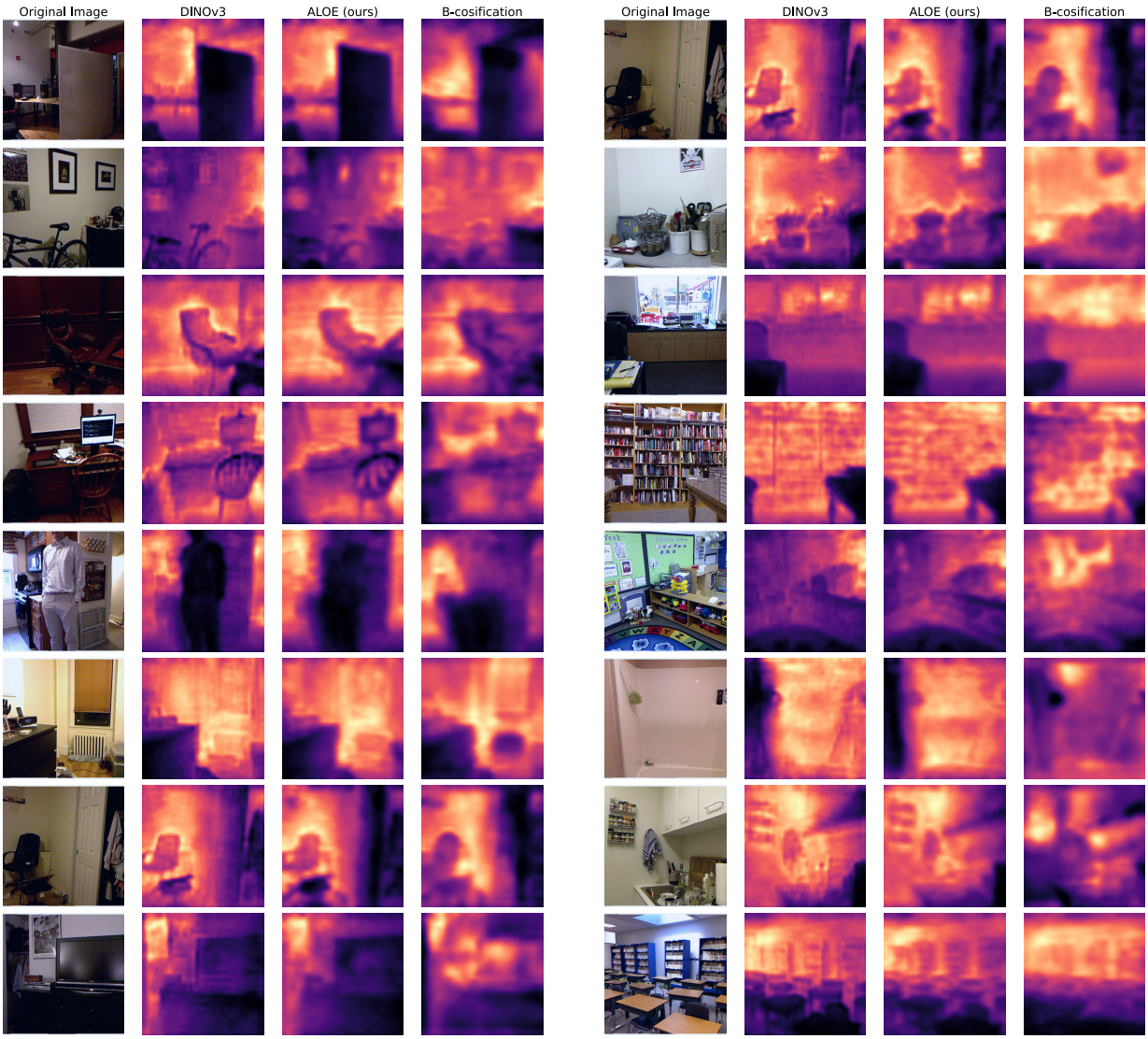

Figure C5. **Depth Estimation.** Visualization of predicted depth maps for **ALOE (ours)**, DINOv3 and B-cosification on the NYUv2 [37] depth-estimation dataset using the Probe3D [15] protocol. ALOE-aligned features yield depth maps with coherent geometry that visually are very similar to the DINOv3 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.

VLM 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 [12] Gemma-9B [40] 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.

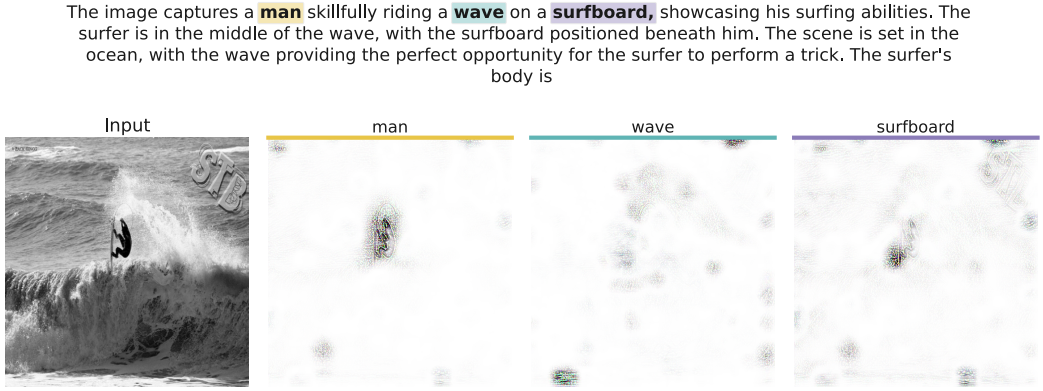

Figure C6. **Token-level Visual Grounding via AttnLRP and ALOE.** We generate visual explanations by leveraging the inherent B-cos interpretability of the ALOE SigLIP2 vision backbone, using AttnLRP to propagate relevance from LLaVA-MORE's GEMMA-9B language model.

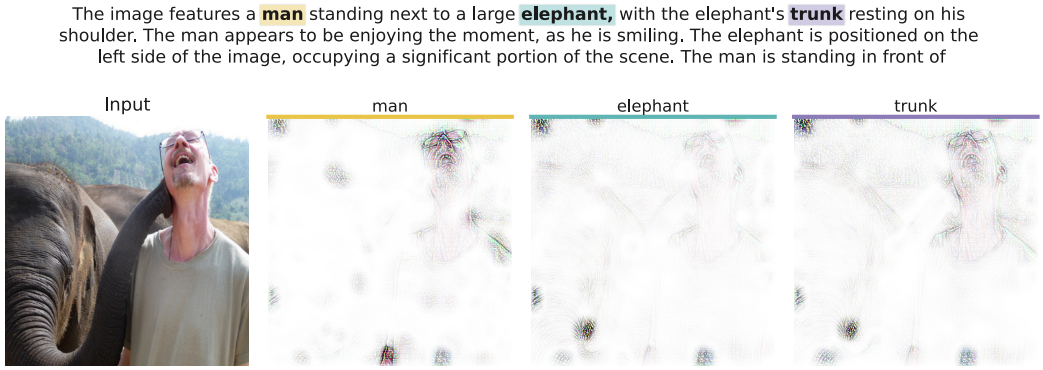

Figure C7. **Token-level Visual Grounding via AttnLRP and ALOE.** Similar to Fig. C6

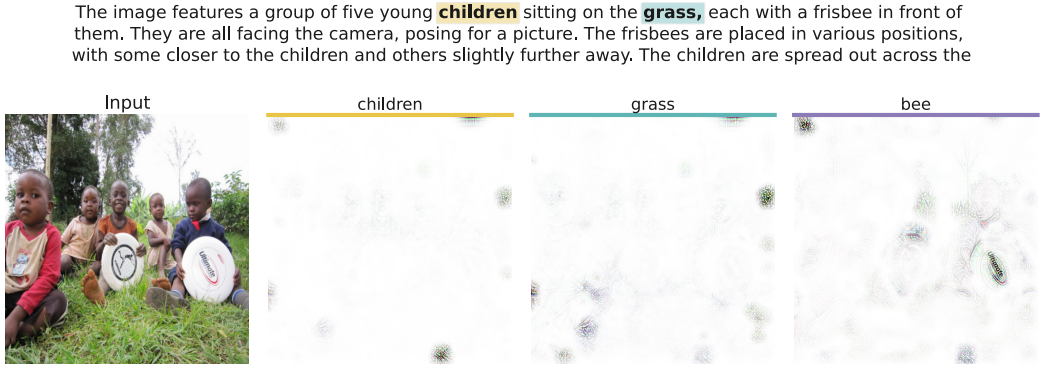

Figure C8. **Token-level Visual Grounding via AttnLRP and ALOE.** Similar to Figs. C6 and C7

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. 2, 3, 7, 14

[2] Shreyash Arya, Sukrut Rao, Moritz Böhle, and Bernt Schiele. B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable. In *NeurIPS*, 2024. 7, 8, 3, 4, 5

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

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*. Springer, 2014. 2

[5] 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. 3, 7

[6] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4), 1952. 8

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

[8] Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers. *IEEE TPAMI*, 2024. 3, 7

[9] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 3, 7

[10] 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. 2, 4, 5

[11] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 2

[12] 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. 8, 2, 14

[13] 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. 2, 5, 6

[14] 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. 2, 3, 4, 7

[15] 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. 2, 5, 6, 14

[16] Siddhartha Gairola, Moritz Böhle, Francesco Locatello, and Bernt Schiele. How to Probe: Simple Yet Effective Techniques for Improving Post-hoc Explanations. In *ICLR*, 2025. 8

[17] Google Research. Siglip2 vision encoder checkpoint (vit-b/16, 224), 2025. Identifier: google/siglip2-base-patch16-224. 2

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

[19] Google Research. Siglip2 vision encoder checkpoint (so400m/16, 256), 2025. Identifier: google/siglip2-so400m-patch16-256. 2

[20] Google Research. Vit-b/16 supervised checkpoint (224), 2025. Identifier: google/vit-base-patch16-224. 2

[21] 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. 2, 4, 5

[22] 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. 2, 5, 6

[23] 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. 3

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 2

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. *Technical Report, Computer Science Department, University of Toronto*, 2009. 2

[26] L'ubor Ladickỳ, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*. Springer, 2014. 2

[27] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022. 2

[28] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[29] Meta AI. Dinov3 vit-b/16 pretrain checkpoint (224), 2025. Identifier: facebook/dinov3-vitb16-pretrain-lvd1689m. 2

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

[31] Meta AI. Dinov3 vit-s/16 pretrain checkpoint (224), 2025. Identifier: facebook/dinov3-vits16-pretrain-lvd1689m. 2

[32] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 2

[33] 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. 2

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of any Classifier. In *KDD*, 2016. 3, 7, 8

[35] 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

[36] 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. 3

[37] 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. 2, 14

[38] 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. 2, 3, 4, 5, 7

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *ICML*, 2017. 3, 7

[40] 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, 2, 14

[41] 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. 2, 3, 4, 5, 7

[42] 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. 2

[43] 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. 3