IsoCLIP: Decomposing CLIP Projectors for Efficient Intra-modal Alignment Supplementary Material

Simone Magistri[1*] Dipam Goswami[2,3] Marco Mistretta[1] Bartłomiej Twardowski[2,3,4]
Joost van de Weijer[2,3] Andrew D. Bagdanov[1]
[1]University of Florence, Media Integration and Communication Center (MICC), Italy
[2]Department of Computer Science, Universitat Autònoma de Barcelona, Spain
[3]Computer Vision Center, Barcelona, Spain [4]IDEAS Research Institute, Warsaw, Poland
{simone.magistri, marco.mistretta, andrew.bagdanov}@unifi.it
{dgoswami, btwardowski, joost}@cvc.uab.cat

This supplementary material provides additional details and analyses that complement the main paper. We include the full gradient derivation of the CLIP loss (Sec. A), a detailed analysis of how IsoCLIP improves intra-modal retrieval (Sec. B), and an investigation of the effect of adding top and bottom spectral directions to the middle band (Sec. C). We then describe the extension of IsoCLIP to non-linear projection heads (Sec. D), followed by dataset descriptions and implementation details (Sec. E). Additional experimental results, ablations, and comparisons are reported in Sec. F. Finally, we discuss hyperparameter selection (Sec. G), inter-modal degradation (Sec. H), and provide the complete IsoCLIP algorithm pseudocode (Sec. I).

Table of Contents

*Corresponding author.

A. Inter-Modal and Intra-Modal Operators: Gradient Derivation for the CLIP Loss

The symmetric contrastive loss of CLIP is defined as:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}(\mathcal{L}_{i\to t} + \mathcal{L}_{t\to i}), \tag{S1}$$

where $\mathcal{L}_{i\to t}$ moves the embedding of each image i toward the corresponding *positive* paired text t, while pushing it away from all other texts in the mini-batch. In the main paper we present the gradient contribution of the *positive* text g_t; here we provide the complete derivation.

The loss $\mathcal{L}_{i\to t}$ is defined as:

$$\mathcal{L}_{i\to t} = -\log \frac{\exp\big(\text{sim}(f_i, g_t)/\tau\big)}{\sum_{t'} \exp\big(\text{sim}(f_i, g_{t'})/\tau\big)} = -\log \frac{\exp\left(\dfrac{f_i^\top (W_i^\top W_t) g_t/\tau}{\|W_i f_i\|_2 \, \|W_t g_t\|_2}\right)}{\sum_{t'} \exp\left(\dfrac{f_i^\top (W_i^\top W_t) g_{t'}/\tau}{\|W_i f_i\|_2 \, \|W_t g_{t'}\|_2}\right)}, \tag{S2}$$

where f_i and g_t denote the pre-projection image and positive text feature respectively; W_i and W_t denote the image and text projector weights respectively; τ is the temperature; t denotes the *positive text* for image i, and t' ranges over the positive and negative texts in the mini-batch.

Defining the normalization factor

$$\alpha_{t,i} = \frac{1}{\|W_i f_i\|_2 \, \|W_t g_t\|_2}, \tag{S3}$$

and the logit of the positive image–text pair

$$s_t = \alpha_{t,i} \, f_i^\top (W_i^\top W_t) g_t. \tag{S4}$$

the loss becomes:

$$\mathcal{L}_{i\to t} = -\log \frac{\exp(s_t/\tau)}{\sum_{t'} \exp(s_{t'}/\tau)}.$$

By applying the chain rule and isolating the contribution of the *positive* text embedding, we obtain:

$$\frac{\partial \mathcal{L}_{i\to t}}{\partial f_i} = \frac{\partial \mathcal{L}_{i\to t}}{\partial s_t} \frac{\partial s_t}{\partial f_i}. \tag{S5}$$

The first term is the standard derivative of cross-entropy with respect to the logit s_t:

$$\frac{\partial \mathcal{L}_{i\to t}}{\partial s_t} = \frac{1}{\tau}\big(p_t - 1\big), \qquad \text{where} \qquad p_t = \frac{\exp(s_t/\tau)}{\sum_{t'} \exp(s_{t'}/\tau)}$$

is the softmax probability of the positive text t.

For the second term, recalling the definitions in Eq. (S3) and Eq. (S4), we obtain

$$\frac{\partial s_t}{\partial f_i} = \alpha_{t,i}(W_i^\top W_t) g_t + \frac{\partial \alpha_{t,i}}{\partial f_i} \, f_i^\top (W_i^\top W_t) g_t = \alpha_{t,i}(W_i^\top W_t) g_t + \frac{\partial \alpha_{t,i}}{\partial f_i}\left(\frac{s_t}{\alpha_{t,i}}\right) \tag{S6}$$

Exploiting the derivative of the norm, a straightforward calculation yields

$$\frac{\partial \alpha_{t,i}}{\partial f_i} = -\frac{1}{\|W_t g_t\|_2 \|W_i f_i\|_2^2} \frac{\partial \|W_i f_i\|_2}{\partial f_i} = -\alpha_{t,i} W_i^\top \frac{W_i f_i}{||W_i f_i||_2^2}. \tag{S7}$$

Replacing this result in Eq. (S6), gives:

$$\frac{\partial s_t}{\partial f_i} = \alpha_{t,i} \, \overbrace{W_i^\top W_t}^{\Psi} \, g_t - s_t \, \frac{\overbrace{W_i^\top W_i}^{\Psi_i} \, f_i}{\|W_i f_i\|_2^2}, \tag{S8}$$

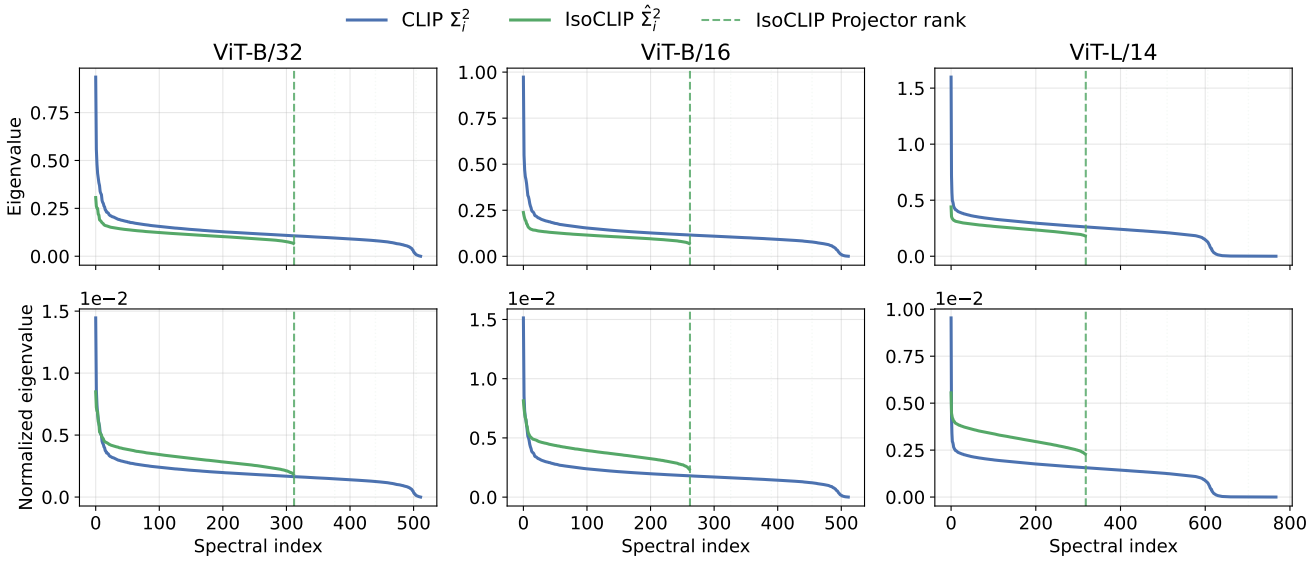

Figure S6. **Spectrum of the intra-modal operator in CLIP (Σ_i^2) and after applying IsoCLIP ($\hat{\Sigma}_i^2$).** (Top) IsoCLIP truncates the spectrum of the intra-modal operator according to the retained subspace defined by the middle-band of the inter-modal operator (Eq. (S9)), resulting in a lower-rank operator. (Bottom) The normalized eigenvalues (obtained by dividing each eigenvalue by the sum of the spectrum) reveal that, after applying IsoCLIP, the intra-modal operator is distributed across more directions than in standard CLIP.

where Ψ and Ψ_i represent respectively the *inter-modal* and *intra-modal* operator, matching the definition in the main paper.

On the role of negative texts. As specified in the main paper and in the previous derivation, we focus on the contribution of the positive text g_t to the gradient (see Eq. (S5)). In general, the full gradient with respect to f_i is obtained by summing over all texts t' in the batch:

$$\frac{\partial \mathcal{L}_{i\to t}}{\partial f_i} = \frac{1}{\tau}\sum_{t'}(p_{t'} - y_{t'})\left[\alpha_{t',i}\Psi g_{t'} - s_{t'}\frac{\Psi_i f_i}{\|W_i f_i\|_2^2}\right],$$

where $y_{t'} = 1$ only for the positive text. Negative texts contribute additional image–text directions $g_{t'}$, but they do not introduce new operators: all interactions between image and text features occur through the *inter-modal* operator Ψ, while Ψ_i acts solely as the *intra-modal* normalization term enforcing unit-length image embeddings. Thus, negatives only reweight the contributions of these two operators: repulsion from negatives and attraction toward the positive act exclusively through the inter-modal operator Ψ, while Ψ_i remains a normalization term.

B. Improving Intra-Modal Retrieval via IsoCLIP

At the end of Section 4 in the main paper we discussed that projecting the image and text projectors W_i and W_t onto the subspaces corresponding to the approximately isotropic region of the inter-modal operator spectrum Ψ improves intra-modal retrieval performance. The projected weights are defined as:

$$\widehat{W}_i = W_i U_{\mathcal{S}_U} U_{\mathcal{S}_U}^\top, \quad \widehat{W}_t = W_t V_{\mathcal{S}_V} V_{\mathcal{S}_V}^\top. \tag{S9}$$

We now provide additional insights into the mechanism responsible for this improvement. Recall that intra-modal similarity between two image features f_i and $f_{\hat{i}}$ is defined as:

$$\text{sim}(f_i, f_{\hat{i}}) = \frac{f_i^\top (W_i^\top W_i) f_{\hat{i}}}{\|W_i f_i\|_2 \, \|W_i f_{\hat{i}}\|_2} \propto f_i^\top (W_i^\top W_i) f_{\hat{i}}. \tag{S10}$$

Let us consider the singular value decomposition of the image projector $W_i = U_i \Sigma_i V_i^\top$. Substituting into the similarity expression, we obtain:

$$\text{sim}(f_i, f_{\hat{i}}) \propto f_i^\top (V_i \Sigma_i U_i^\top U_i \Sigma_i V_i^\top) f_{\hat{i}} = f_i^\top V_i \Sigma_i^2 V_i^\top f_{\hat{i}}. \tag{S11}$$

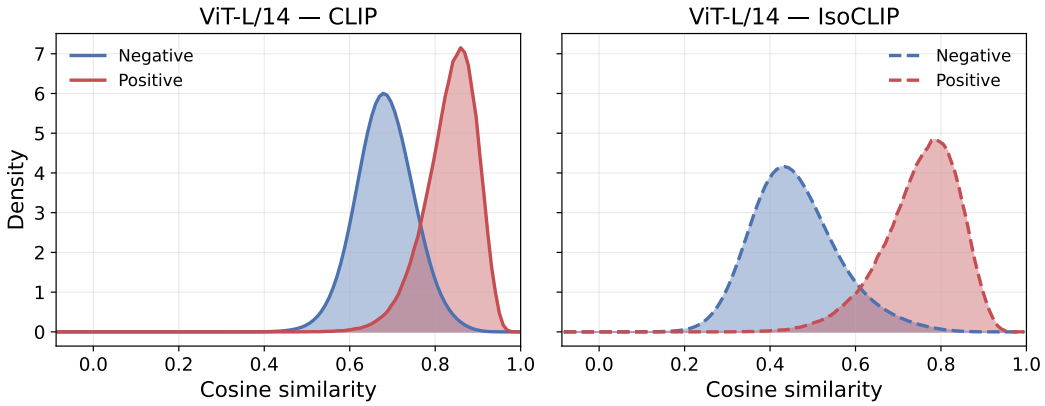

Figure S7. Cosine similarity distribution in image-to-image retrieval for positive and negative pairs on the CUB dataset for CLIP (left) and IsoCLIP (right). Positives are obtained by computing the similarity of each image with all images in the gallery that share the same label, while negatives correspond to the cosine similarity with images having different labels. We observe that the CLIP cosine similarities are concentrated in a narrower range with higher mean values, while IsoCLIP, by weighting more directions in cosine similarity computations, spreads the distribution, shifts the mean similarities to lower values and increases positive and negative separation.

This shows that intra-modal similarity is governed by the intra-modal operator $\Psi_i = W_i^\top W_i = V_i \Sigma_i^2 V_i^\top$. Moreover, this shows that CLIP image-to-image cosine similarity can be interpreted as a weighted summation over the singular directions v_k of W_i, where the weights are given by the squared singular values σ_k^2. In practice, the spectrum Σ_i^2 is highly anisotropic, so that a small number of singular directions receive excessively large weight in similarity computations. Consequently, similarity scores are dominated by these few directions, reducing the separability between positive and negative pairs.

IsoCLIP mitigates this by restricting the projector to the middle band of the inter-modal spectrum (Eq.S9). The resulting filtered projector $\widehat{W}_i$ can be written as $\widehat{W}_i = \hat{U}_i \hat{\Sigma}_i \hat{V}_i$. The resulting similarity becomes:

$$\text{sim}(f_i, f_{\hat{i}}) \propto f_i^\top \hat{V}_i \hat{\Sigma}_i^2 \hat{V}_i^\top f_{\hat{i}}. \tag{S12}$$

Because IsoCLIP removes the highly anisotropic top and bottom spectral directions, the spectrum $\hat{\Sigma}^2$ becomes significantly flatter. As a result, similarity computations distribute weight across a larger number of directions corresponding to the middle band of the inter-modal spectrum, which encode cross-modal semantic alignment.

Fig. S6 visualizes the eigenvalues of the intra-modal operator $W_i^\top W_i$, namely Σ_i^2 for the original CLIP projector and $\hat{\Sigma}_i^2$ for the IsoCLIP projector, for ViT-B/32, ViT-B/16 and ViT-L/14. The top row shows that IsoCLIP truncates the spectrum according to the retained subspace (Eq. (S9)), resulting in a lower-rank operator. The bottom row shows the normalized spectra with the summation of the eigenvalues. Compared to CLIP, the retained IsoCLIP spectrum is less concentrated in few directions, reducing spectral anisotropy and distributing similarity across a larger set of directions.

IsoCLIP similarity distribution and mAP improvement. It is interesting to observe the effect of the spectra Σ_i^2 and $\hat{\Sigma}_i^2$ on the cosine similarity distribution. Figure S7 shows the distribution of cosine similarities for positive and negative image pairs using the original CLIP projector (left) and the IsoCLIP projector (right) on CUB images.

In Fig. S7, we observe that the original CLIP projector (left), due to the strong anisotropy, concentrates both the positive and negative cosine similarity distribution in a narrow range (peaks ≈ 0.6 for negative and ≈ 0.9 for positive). This indicates that projected image features occupy a relatively small region of the hyper-sphere.

In contrast, IsoCLIP projects W_i onto the middle-band of the inter-modal operator spectrum, distributing similarity across more directions (Eq. S12). As a result, cosine similarities become less concentrated and shift toward lower values (around ≈ 0.4 for negatives and ≈ 0.8 for positives), indicating that features occupy a larger region of the hypersphere. The reduced overlap between positive and negative similarities (shaded area) leads to higher mAP.

C. Adding Top and Bottom Directions to the Isotropic Middle Band Directions

We complement the experiments in the main paper by incrementally adding top and bottom directions to the 50 middle-band directions shown in Fig. 3 (main paper) on CUB using ViT-B/16. In Fig. S8 we observe that extending the middle band with either top or bottom directions to define the subspace used to project the projector weights in IsoCLIP (Eq. 10 in the

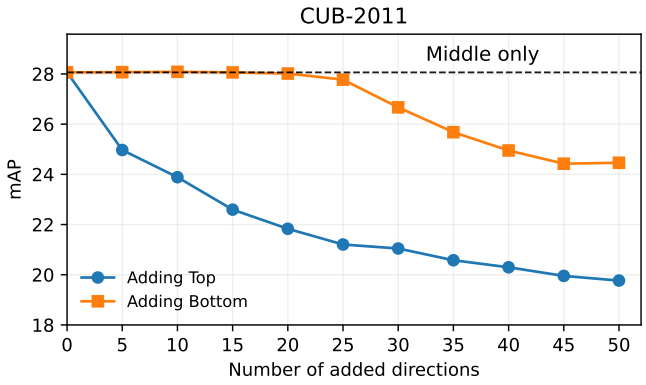

Figure S8. Image-image retrieval mAP performance obtained by iteratively adding top and bottom directions to the 50 middle-band directions when computing IsoCLIP on CUB-2011 dataset using the ViT-B/16.

main paper) degrades mAP retrieval performance, with top directions being more detrimental than bottom ones. This result strengthens the claim that the middle region identified by the inter-modal operator is optimal for improving image-to-image retrieval performance, while the extremes, capturing modality-specific variations, are detrimental.

D. Extension of IsoCLIP to non-linear projection heads

Non-linear projection heads prevent the direct application of IsoCLIP. For instance, models like SigLIP2 [17] employ a Multilayer Perceptron (MLP) head for the image encoder and a linear layer for the text encoder, mapping both modalities into the shared embedding space. We now discuss how to generalize IsoCLIP with non-linear projection heads.

Recent work [20] has shown that linear approximation of Vision Transformer blocks can reveal singular defects in attention feature maps by exploiting data-driven linearization. We propose a simple data-free first-order linearization of the final MLP layer before applying IsoCLIP in models like SigLIP2.

The last MLP visual head of SigLIP2 is defined as:

$$x \leftarrow x + W_2\phi\left(W_1\mathrm{LN}(x) + b_1\right) + b_2, \tag{S13}$$

where $W_1 \in \mathbb{R}^{m\times n}, W_2 \in \mathbb{R}^{n\times m}$ are weight matrices, $b_1 \in \mathbb{R}^m, b_2 \in \mathbb{R}^n$ are bias terms and ϕ denotes the GELU activation.

The Layer Normalization (LN) operator is defined as:

$$\mathrm{LN}(x) = \gamma \odot \frac{x - \mu(x)}{\sqrt{\sigma^2(x) + \varepsilon}} + \beta \tag{S14}$$

where γ and β are affine parameters, ε is a small constant, and $\mu(x), \sigma^2(x)$ denote the mean and the variance of x.

Writing $\mathrm{LN}(x) = \gamma \odot \hat{x} + \beta$, the affine parameters can be absorbed into the first linear layer as $\tilde{W}_1 = W_1\,\mathrm{Diag}(\gamma)$ and $\tilde{b}_1 = W_1\beta + b_1$, yielding

$$x \leftarrow x + W_2\,\phi(\tilde{W}_1\hat{x} + \tilde{b}_1) + b_2.$$

Assuming a normalized regime where LayerNorm acts approximately as identity (i.e. $\hat{x} \approx x$), and approximating GELU by its average slope $\phi(z) \approx \frac{1}{2}z$, we obtain the following linearization

$$x \leftarrow \left(I + \frac{1}{2}W_2\tilde{W}_1\right)x + \frac{1}{2}W_2\tilde{b}_1 + b_2. \tag{S15}$$

Accordingly, the effective image projection matrix can be written as

$$W_i = \left[I + \tfrac{1}{2}W_2\tilde{W}_1 \quad \tfrac{1}{2}W_2\tilde{b}_1 + b_2\right], \tag{S16}$$

which can be used to apply IsoCLIP as in Eq. S9.

Table S5. Ablation study on the number of OVI pseudo-patches P for text-to-text retrieval on Flickr30K validation set. The highest mAP score for each model is highlighted in bold, with the corresponding value of P used in all experiments.

	Number of Pseudo-Patches P				
Backbone	1	2	4	8	16
ViT-B/16	52.8	52.9	**53.1**	51.9	50.8
ViT-B/16-open	**60.2**	59.1	58.0	57.3	57.2

E. Dataset and Implementation Details

In this section, we describe the datasets used and provide implementation details for the modality inversion baselines reported in the tables in the main paper.

E.1. Datasets

For image-to-image retrieval, we consider 13 datasets - ROxford5k [15], RParis6k [15], CUB [19], Stanford Cars [8], Oxford-IIIT Pets [13], Oxford 102 Flowers [12], FGVC Aircraft [10], SUN397 [21], Caltech101 [5], DTD [3], EuroSAT [6], Food101 [2], and UCF101 [16]. We follow the dataset splits proposed in [11] for all datasets. For experiments on ROxford and RParis, we include the R1M distractor set, having about 1 million images, as negative samples for all queries. Here we only report results on the Easy setting from [15]. For the other datasets, we use the training data split as the gallery and the test split as the query set for 12 datasets except CUB where the entire dataset is considered as both the query and gallery.

For text–to-text retrieval, we consider 3 image-captioning datasets - Flickr30k [14], COCO [9], and NoCaps [1]. These datasets contain multiple short captions for each image. Following [11], we consider the first caption of every image as the query and all captions in the dataset as the gallery. The goal is to retrieve the other captions which are associated to the same image. We ignore the images here for the text retrieval task. On an average, COCO and Flickr30K contains 5 captions for each image and the nocaps dataset has 10 captions for each image. We use the captions from the test split for COCO and Flickr30K following the split from [7]. For nocaps, we use the validation set.

We also evaluate image classification on 10 datasets drawn from those used for image retrieval. For this setting, we use the original training splits to compute the class-wise prototypes, and we report performance on the original test splits.

E.2. Implementation Details for Modality Inversion Baselines

We provide implementation details for the Optimization-based Textual Inversion (OTI) and Optimization-based Visual Inversion (OVI) methods used as baselines in our experiments from [11]. These methods map features from one modality to the complementary modality space through iterative optimization.

Optimization-based Textual Inversion (OTI). OTI maps image features to the text embedding space by optimizing a set of pseudo-tokens $v^* = \{v_1^*, \ldots, v_R^*\}$ in the token embedding space. Following the original implementation [11], we use a single pseudo-token ($R = 1$) for all backbones, randomly initialized and concatenated with the template sentence "a photo of". We employ the AdamW optimizer with learning rate 0.02, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.01, performing 150 optimization steps with mixed precision training.

Optimization-based Visual Inversion (OVI). OVI maps text features into the image embedding space by optimizing a set of visual pseudo-patches $w^* = \{w_1^*, \ldots, w_P^*\}$ in the patch embedding space. Following the original implementation [11], we adopt the same optimizer settings as OTI, but we run 1000 optimization steps.

Unlike OTI, the optimal number of pseudo-patches P for OVI depends on the model architecture. Following the procedure described in [11], we validated the pseudo patches $P \in \{1, 2, 4, 8, 16\}$ on text-to-text retrieval on Flickr30k validation set, for both ViT-B/16 and ViT-B/16-open, which we introduce in this paper. Table S5 reports the results of this ablation study. For OpenAI ViT-B/16, the best performance is obtained with $P = 4$ pseudo-patches (53.1% mAP). For OpenCLIP ViT-B/16, a single pseudo-patch ($P = 1$) is sufficient, achieving 60.2% mAP.

F. Additional Results and Ablations

In this section, we provide additional results and ablations that complement those in the main paper.

Table S6. **Image-to-image retrieval** performance using *OpenCLIP* ViT-B/32, pre-trained on DataComp dataset, EVA-02 B/16, pretrained on Merged-2B [4] and SigLIP2 B/16 pre-trained on WebLI [17].

Method	Intra-modal	Backbone	Caltech	CUB	ROxford	RParis	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	UCF101	Avg
Image-Image	✓	ViT-B/32-open	82.3	32.1	50.8	74.7	46.7	44.1	77.0	19.6	36.9	56.4	39.6	36.2	45.7	49.4
OTI (I→T)	✗		83.3	**34.3**	54.4	**75.8**	**50.5**	**50.5**	78.0	**20.1**	**40.9**	54.5	**42.9**	**37.8**	**48.2**	**51.6**
IsoCLIP	✓		**83.4**	34.2	**56.8**	**75.8**	49.9	47.8	**78.2**	19.8	37.6	**57.7**	41.5	36.6	46.3	51.2
Image-Image	✓	EVA-02 B/16	86.9	55.9	52.1	78.3	49.4	55.4	91.5	24.9	35.8	61.2	55.5	41.1	57.0	57.3
IsoCLIP	✓		**89.7**	**58.3**	**53.0**	**80.4**	**55.2**	**62.8**	**92.7**	**25.7**	**40.0**	**62.3**	**57.5**	**42.2**	**57.9**	**59.8**
Image-Image	✓	SigLIP2 B/16	89.2	38.1	53.2	76.5	70.8	56.6	**89.3**	**41.8**	39.0	50.8	59.2	43.6	59.2	59.0
IsoCLIP	✓		**93.1**	**41.4**	**54.2**	**77.9**	**74.5**	**64.0**	87.2	40.9	**43.8**	**53.2**	**63.6**	**46.9**	**61.8**	**61.7**

F.1. Image-to-Image Retrieval on ViT-B/32-open, EVA-02 B/16 and SigLIP2 B/16

In Tab. S6, we compare IsoCLIP against standard image-to-image retrieval (Image-Image) using only the vision encoder ViT-B/32-open pretrained on the DataComp dataset, as well as against the textual inversion–based approach (OTI). Consistent with the results in the main paper, where ViT-B/16-open with the same pre-training was used, IsoCLIP performs significantly better than Image-Image across all datasets, and achieves slightly lower performance than OTI while requiring substantially lower query latency.

We also evaluate IsoCLIP on EVA-02 B/16 pre-trained on the Merged-2B dataset, where it consistently outperforms standard image-to-image retrieval. Finally, we evaluate SigLIP2 B/16 pre-trained on WebLI, and observe that, despite the linearization of the last MLP projection head (Sec. D), IsoCLIP significantly improves performance on most datasets.

F.2. Image Classification on ViT-B/16-open

Table S7. **Image classification** performance using OpenCLIP ViT-B/16, pre-trained on DataComp dataset, on 10 datasets. We compare IsoCLIP with intra-modal NCM classification and zero-shot classification.

Method	Intra-modal	Classifier	Backbone	Caltech	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	UCF101	Average
Image-Text	✗	Zero-Shot	ViT-B/16-open	96.9	89.8	**92.8**	75.3	29.8	58.3	53.4	**87.5**	69.8	67.8	72.1
Image-Image	✓	NCM		96.7	90.6	88.9	98.6	54.0	74.5	84.4	86.7	**74.9**	79.9	82.9
IsoCLIP	✓	NCM		**97.2**	**91.4**	90.6	**98.8**	**54.5**	**75.9**	**85.5**	87.1	74.3	**80.9**	**83.6**

We extend the experiments on image prototype-based classification with the Nearest Class Mean (NCM) classifier, presented in the main paper, by also evaluating IsoCLIP on ViT-B/16-open. For this model, we observe that IsoCLIP slightly outperforms standard NCM on the image encoder on average and provides consistent improvements on most datasets.

F.3. Comparison with Unimodal DINOv2 B/14

We performed a preliminary comparison with DINOv2 B/14 for image-to-image retrieval. The results are highly dataset dependent: DINOv2 achieves 67.0 mAP on CUB (vs. 53.0 for IsoCLIP PE-Core-B/16), but only 22.3 mAP on Cars, far below IsoCLIP PE-Core-B/16 (62.3 mAP), despite ViT-B/14 processing more image patches due to its smaller patch size. These results suggest that CLIP-style models remain competitive for image-to-image retrieval and highlight the utility of IsoCLIP. A more systematic comparison with self-supervised models is left for future work.

F.4. Experiments on Places365 and iNaturalist

We evaluate IsoCLIP on more benchmarks like Places365 [22] and iNaturalist [18] and present the results in Tab. S8 using PE-Core B/16, SigLIP2 B/16 and EVA-02 B/16. We consider the validation set of Places365 having 35k images and iNaturalist 2021 mini-train version having 500k images for both gallery and query images. The values of k_t and k_b are selected using Caltech101. We show that IsoCLIP consistently improves accuracy on both datasets across different models. On average, IsoCLIP improves by 1.89% on PE-Core B/16, 1.24% on SigLIP2 B/16 and 0.83% on EVA02 B/16.

Table S8. Experiments on Places365 & iNaturalist.

Method	Backbone	Places365	iNat	Avg
Image-Image	PE-Core B/16	16.72	9.61	13.17
IsoCLIP		**17.04**	**13.07**	**15.06**
Image-Image	SigLIP2 B/16	16.17	7.56	11.87
IsoCLIP		**17.78**	**8.43**	**13.11**
Image-Image	EVA-02 B/16	14.54	10.03	12.29
IsoCLIP		**15.01**	**11.23**	**13.12**

F.5. Additional Ablations

Table S9. We compare IsoCLIP against using pre-projection image features for retrieval (Image–Image [Pre]) and against whitening the CLIP image projection weights (W_i^{white}), using ViT-B32 OpenAI and Open ViT-B32 and ViT-B16 pre-trained on Datacomp dataset.

Method	Backbone	Caltech	CUB	ROxford	RParis	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	UCF101	Average
Image-Image	ViT-B/32	77.1	22.9	42.6	67.9	24.6	30.5	62.0	14.5	28.1	47.9	32.3	34.3	47.1	40.9
Image-Image [Pre]		78.3	23.3	44.9	70.5	24.5	32.7	63.4	14.4	29.5	50.8	34.2	36.0	**49.7**	42.5
W_i^{white}		78.4	24.5	45.0	70.7	25.9	33.2	63.7	14.3	29.4	50.3	34.2	35.1	48.9	42.6
IsoCLIP		**80.8**	**27.0**	**47.2**	**73.8**	**30.0**	**40.8**	**66.5**	**14.9**	**30.9**	**51.5**	**38.0**	**36.4**	48.4	**45.1**
Image-Image	ViT-B/32-open	82.3	32.1	50.8	74.7	46.7	44.1	77.0	19.6	36.9	56.4	39.6	36.2	45.7	49.4
Image-Image [Pre]		**83.6**	31.6	52.8	73.9	46.1	44.6	76.5	19.5	37.3	57.5	39.9	**36.9**	**46.9**	49.8
W_i^{white}		83.2	32.9	52.2	74.1	45.6	45.7	77.0	18.7	37.3	**57.7**	39.8	35.4	46.0	49.7
IsoCLIP		83.4	**34.2**	**56.8**	**75.8**	**49.9**	**47.8**	**78.2**	**19.8**	**37.6**	**57.7**	**41.5**	36.6	46.3	**51.2**
Image-Image	ViT-B/16-open	85.7	42.8	65.3	83.2	55.8	50.4	84.6	23.1	39.9	57.8	51.1	39.5	52.9	56.3
Image-Image [Pre]		86.3	42.2	66.9	83.1	56.1	52.4	84.0	23.2	39.9	58.0	52.4	**40.9**	**54.1**	56.9
W_i^{white}		86.4	43.8	64.5	83.5	56.2	54.0	84.4	22.4	40.2	58.1	52.8	39.8	53.3	56.9
IsoCLIP		**87.6**	**45.9**	**67.3**	**85.0**	**60.7**	**57.8**	**85.8**	**23.5**	**42.5**	**58.6**	**54.7**	39.3	53.4	**58.6**

We provide additional ablations complementing those in the main paper, using ViT-B/32 OpenAI, Open ViT-B/32, and Open ViT-B/16 models pre-trained on the DataComp dataset. As in the main paper, we compare IsoCLIP against using the *raw pre-projection features* f_i (Image-Image [Pre]) and against applying whitening to the image projector weights (W_i^{white}).

In Tab. S9, we observe that, in general, IsoCLIP outperforms all other approaches across most datasets. Consistent with the results reported in the main paper, IsoCLIP achieves larger gains than both alternatives on the majority of benchmarks. However, when using the models pre-trained on DataComp (ViT-B/32-open and ViT-B/16-open), IsoCLIP attains performance comparable to Image-Image and other baselines on SUN397 and UCF101.

G. Analysis of Hyperparameter Selection

All results reported in Tab. 1 and Tab. 2 of the main paper use the same procedure for selecting the top k_t and bottom k_b singular directions that define the middle-spectrum band used to align the projectors.

Selection of k_t and k_b. As described in the main paper, we select (k_t, k_b) using *Caltech101*, a generic object recognition dataset, for image-to-image retrieval and *COCO* for text-to-text retrieval, and apply these values to *all* the backbones. Table S10 reports the selected values for each model. These hyperparameters differ across backbones because the shape and dimensionality of the singular spectrum of the inter-modal operator depend on the embedding dimensionality and the pre-training dataset, consistent with our analysis of the isotropic middle region in Fig. 2 of the main paper.

For ViT-B/16, the selected values for image-to-image retrieval are $k_t = 200$ and $k_b = 50$, validated on Caltech101 (see Figure 5 left in the main paper).

Table S10. Selection of k_t and k_b for image-to-image and text-to-text retrieval using IsoCLIP. For image classification tasks, we use the same values selected for image-to-image retrieval.

Task	Backbone	k_t	k_b
Image-Image	ViT-B/32	150	50
	ViT-B/32-open	50	50
	ViT-B/16	200	50
	ViT-L/14	250	200
	ViT-B/16-open	100	100
	EVA-02 B/16	150	0
	PE-Core-B-16	300	50
	SigLIP2 B/16	350	50
Text-Text	ViT-B/32	20	100
	ViT-B/16	10	50
	ViT-L/14	10	300
	ViT-B/16-open	2	150

Figure S10 shows that, although these values are not necessarily optimal for every dataset, they generalize well and consistently yield strong performance across all datasets. These plots also indicate that dataset-specific tuning of (k_t, k_b) could further improve performance for some datasets if desired. A similar behavior is observed for text-to-text retrieval in Fig. S9, where varying (k_t, k_b) produces similar trends across datasets.

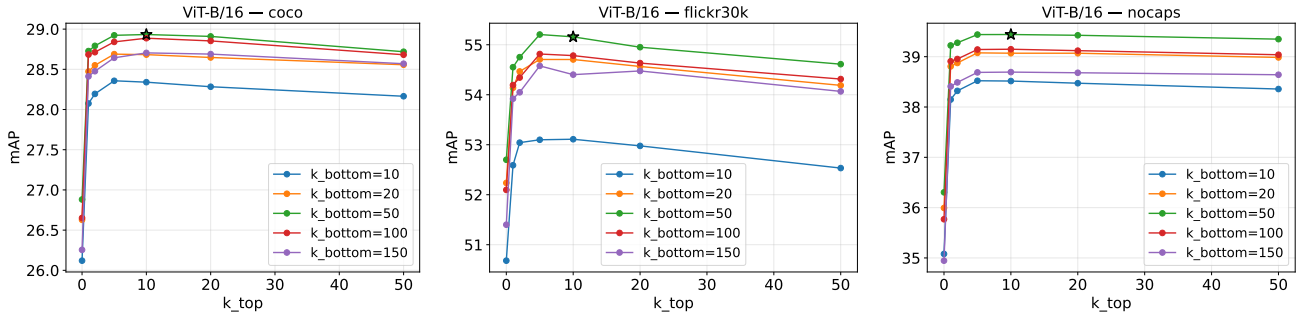

Figure S9. Analysis showing the impact of varying k_t and k_b for the middle band selection across datasets for text-to-text retrieval. The values used in our reported results based on selection from COCO are denoted with stars.

H. Inter-modal Degradation after Applying IsoCLIP

We empirically observe that replacing CLIP projectors with IsoCLIP ones (Eq. S9) reduces CLIP performance on inter-modal tasks such as text-to-image retrieval. This is expected, since the inter-modal operator used to compute inter-modal similarities is explicitly optimized during CLIP training for this purpose, making the original projectors optimal for inter-modal tasks. However, because IsoCLIP modifies only the projector weights, it remains computationally efficient.

In practical settings where a single image gallery is used for both text–image and image–image retrieval, one can store the pre-projection embeddings in the gallery and use the original CLIP projectors for inter-modal similarity while applying IsoCLIP projectors for intra-modal similarity. This introduces only minimal overhead compared to standard zero-shot CLIP inference, since it requires only an additional matrix multiplication with the projection weights.

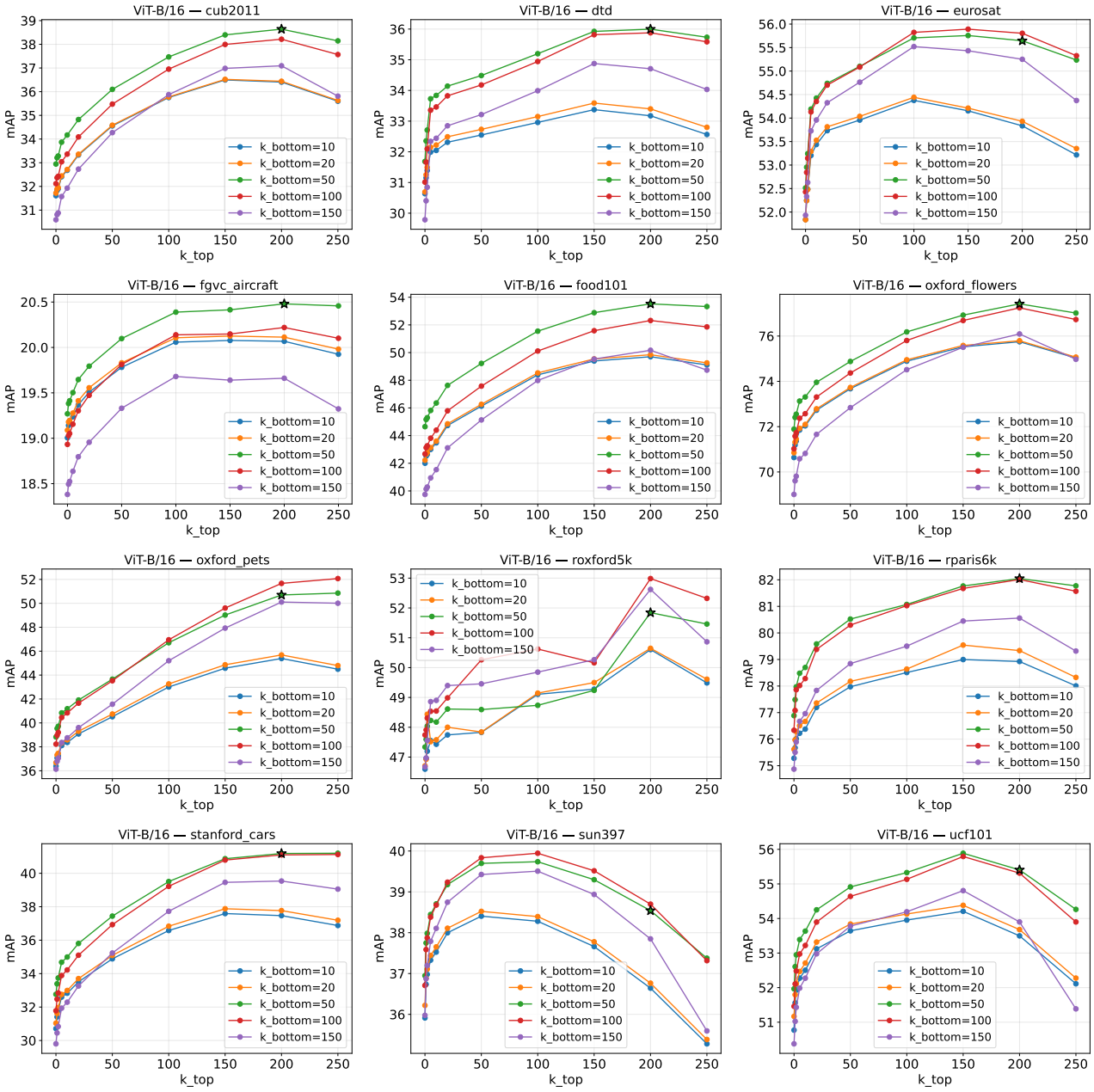

Figure S10. Analysis showing the impact of varying k_t and k_b for the middle band selection across datasets for image-to-image retrieval. The values used in our reported results based on selection from Caltech101 are denoted with stars.

I. The IsoCLIP Algorithm

In this section we provide the complete pseudocode for applying IsoCLIP to pre-trained CLIP models for intra-modal retrieval tasks. The method consists of a one-time preprocessing step that decomposes and aligns the projector weights with the isotropic subspace, followed by the actual retrieval procedure.

Algorithm 1 IsoCLIP Projector Alignment (Training-Free)

Require: Pre-trained CLIP projectors $W_i \in \mathbb{R}^{d \times d_i}$, $W_t \in \mathbb{R}^{d \times d_t}$
Require: Hyperparameters: k_t (top directions to remove), k_b (bottom directions to remove)
Ensure: Aligned projectors $\widehat{W}_i$, $\widehat{W}_t$

// Step 1: Construct inter-modal operator
$\Psi \leftarrow W_i^\top W_t \in \mathbb{R}^{d_i \times d_t}$

// Step 2: Singular Value Decomposition
$U, \Sigma, V^\top \leftarrow \text{SVD}(\Psi)$ ▷ $U \in \mathbb{R}^{d_i \times r}$ (image-side), $V \in \mathbb{R}^{d_t \times r}$ (text-side), $\Sigma \in \mathbb{R}^{r \times r}$

// Step 3: Select middle-band isotropic subspace
Identify the subspace $\mathcal{S}_U = \text{span}\{u_j \mid j \in [k_t, r - k_b]\}$ ▷ Image subspace
$U_{\mathcal{S}_U} \leftarrow [u_{k_t}, u_{k_t+1}, \ldots, u_{r-k_b}]$ ▷ Extract columns from U

Identify the subspace $\mathcal{S}_V = \text{span}\{v_j \mid j \in [k_t, r - k_b]\}$ ▷ Text subspace
$V_{\mathcal{S}_V} \leftarrow [v_{k_t}, v_{k_t+1}, \ldots, v_{r-k_b}]$ ▷ Extract columns from V

// Step 4: Align projectors to the isotropic subspace
$\widehat{W}_i \leftarrow W_i U_{\mathcal{S}_U} U_{\mathcal{S}_U}^\top$ ▷ Project image projector
$\widehat{W}_t \leftarrow W_t V_{\mathcal{S}_V} V_{\mathcal{S}_V}^\top$ ▷ Project text projector

return $\widehat{W}_i$, $\widehat{W}_t$

Algorithm 2 IsoCLIP for Image-to-Image Retrieval

Require: Aligned image projector $\widehat{W}_i$ (from Algorithm 1)
Require: Image encoder $f_\theta(\cdot)$
Require: Query image i_q
Require: Projected gallery features $\{\widehat{F}_{i_1}, \widehat{F}_{i_2}, \ldots, \widehat{F}_{i_N}\}$ where $\widehat{F}_{i_n} = \widehat{W}_i f_\theta(i_n)$
Ensure: Ranked list of gallery images

// Step 1: Extract and project query feature
$f_{i_q} \leftarrow f_\theta(i_q)$ ▷ Extract pre-projection query feature $\in \mathbb{R}^{d_i}$
$\widehat{F}_{i_q} \leftarrow \widehat{W}_i f_{i_q}$ ▷ Project to aligned subspace $\in \mathbb{R}^d$

// Step 2: Compute cosine similarities with pre-computed gallery
for $n = 1$ to N **do**
$s_n \leftarrow \dfrac{\widehat{F}_{i_q}^\top \widehat{F}_{i_n}}{\|\widehat{F}_{i_q}\|_2 \, \|\widehat{F}_{i_n}\|_2}$ ▷ IsoCLIP similarity
end for

// Step 3: Rank by similarity
return Gallery images ranked by $\{s_1, s_2, \ldots, s_N\}$ in descending order

Algorithm 3 IsoCLIP for Text-to-Text Retrieval

Require: Aligned text projector $\widehat{W}_t$ (from Algorithm 1)
Require: Text encoder $g_\phi(\cdot)$
Require: Query text t_q
Require: Projected gallery features $\{\widehat{G}_{t_1}, \widehat{G}_{t_2}, \ldots, \widehat{G}_{t_M}\}$ where $\widehat{G}_{t_m} = \widehat{W}_t g_\phi(t_m)$
Ensure: Ranked list of gallery texts

// Step 1: Extract and project query feature
$g_{t_q} \leftarrow g_\phi(t_q)$ ▷ Extract pre-projection query feature $\in \mathbb{R}^{d_t}$
$\widehat{G}_{t_q} \leftarrow \widehat{W}_t g_{t_q}$ ▷ Project to aligned subspace $\in \mathbb{R}^d$

// Step 2: Compute cosine similarities with pre-computed gallery
for $m = 1$ to M **do**
    $s_m \leftarrow \dfrac{\widehat{G}_{t_q}^\top \widehat{G}_{t_m}}{\|\widehat{G}_{t_q}\|_2 \, \|\widehat{G}_{t_m}\|_2}$ ▷ IsoCLIP similarity
end for

// Step 3: Rank by similarity
return Gallery texts ranked by $\{s_1, s_2, \ldots, s_M\}$ in descending order

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019. 6

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 6

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[4] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 7

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[6] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226, 2019. 6

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 6

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[10] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[11] Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6

[13] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[14] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 6

[15] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 6

[16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[17] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 5, 7

[18] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 7

[19] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[20] Haoqi Wang, Tong Zhang, and Mathieu Salzmann. SINDER: Repairing the singular defects of dinov2. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 5

[21] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[22] Bolei Zhou, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *Journal of Vision*, 17(10):296–296, 2017. 7