DYNAMIC-ISOCLIP: CROSS-MODAL PRIOR SPECTRAL ROUTING FOR INSTANCE-AWARE CLIP ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose Dynamic-IsoCLIP, a supervised and parameter-efficient framework for adapting frozen CLIP representations to image-to-image retrieval and base-to-new classification. Existing IsoCLIP-style methods improve intra-modal alignment through a discrete topological surgery, namely fixed hard truncation of the CLIP projector spectrum, but apply the same manually selected mask to every input. To address this limitation, Dynamic-IsoCLIP introduces Cross-Modal Prior Spectral Routing (CMPSR), which replaces the binary mask with a continuous, image-conditioned mask over a spectral basis constructed offline from the frozen image and text projectors. The text projector supplies an offline cross-modal geometric prior, while a lightweight MLP router selects spectral directions using only image features at inference, avoiding costly online cross-modal attention. Prototype supervision breaks the routing symmetry, and representation distillation and spectral sparsity act as semantic anchors and training stabilizers for the adaptation. Experiments on eleven benchmarks show that Dynamic-IsoCLIP achieves the highest image-to-image retrieval mAP among the evaluated supervised baselines. Base-to-new evaluation further shows improved performance over the original CLIP reference on five datasets under the specified text-classifier protocol, indicating that adaptive spectral routing can improve downstream discrimination while retaining unseen-class recognition within the evaluated setting.

1 INTRODUCTION

Vision-language models such as CLIP learn transferable representations by aligning images and text in a shared embedding space (Radford et al., 2021). Their image encoders are widely used for image classification and image-to-image retrieval, where the relevant geometry is intra-modal: images should be close when they depict the same category or visual concept. This requirement differs from the original image–text pretraining objective, which does not directly optimize the similarity structure among images.

IsoCLIP addresses this mismatch through the relationship between the frozen image and text projectors (Magistri et al., 2026). It constructs a cross-modal spectral basis and removes selected directions with a fixed binary mask. This operation is efficient, but it assumes that one globally selected subspace is suitable for every image. Fine-grained categories, scene layouts, textures, and object-centric categories may instead depend on different combinations of visual evidence. The central problem is therefore not simply to remove anisotropic directions, but to select useful directions in a sample-dependent manner while retaining the pretrained semantic structure.

We propose Dynamic-IsoCLIP, which preserves the frozen cross-modal spectral basis and learns an image-conditioned soft mask. A small MLP router predicts one weight for each spectral direction; the image feature is projected into the basis, reweighted, reconstructed, and mapped through the original image projector. Text supplies an offline geometric prior through the projector decomposition, but no text feature or online cross-attention is required by the router. This asymmetric design is particularly suitable for image-to-image retrieval, where query and gallery images must be processed independently.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

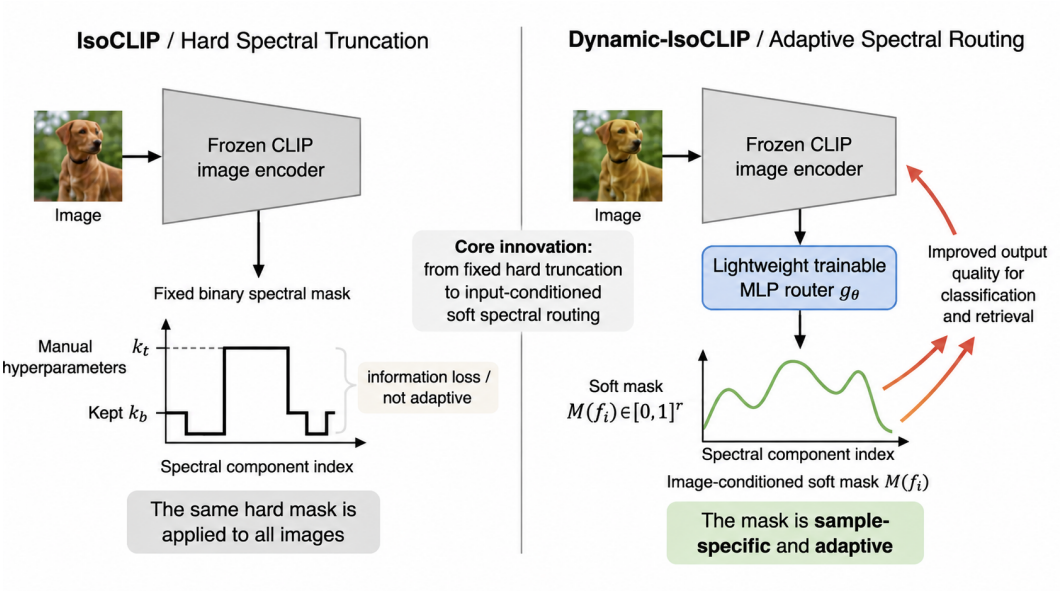

Figure 1: Overview of IsoCLIP and Dynamic-IsoCLIP. IsoCLIP applies a fixed hard spectral truncation to all images, whereas Dynamic-IsoCLIP predicts a sample-specific soft mask with a lightweight router over the cross-modal spectral basis.

The method is a supervised image-side adaptation method rather than a training-free or dynamically text-conditioned model. Class prototypes computed from frozen teacher features provide the task signal. Representation distillation constrains movement away from the teacher representation, while the mean-mask penalty regularizes the routing solution. These auxiliary terms stabilize the optimization but do not by themselves provide class-discriminative supervision. We evaluate the method with image-to-image retrieval mAP, a separate base-to-new classification protocol, loss ablations, and qualitative mask analysis. Figure 1 summarizes the distinction between the global hard selection used by IsoCLIP and the sample-wise soft routing used by Dynamic-IsoCLIP.

Our contributions are threefold:

- We introduce CMPSR, which replaces the fixed hard spectral truncation of IsoCLIP with continuous, input-conditioned weighting of a frozen cross-modal spectral basis.
- We formulate a lightweight supervised router that uses an offline text-guided prior and image-only inference, while representation distillation and spectral sparsity constrain adaptation.
- We evaluate the method on eleven image-to-image retrieval benchmarks, isolate the effect of mask initialization, analyze the auxiliary losses, and test base-to-new classification transfer.

2 Related Work

2.1 CLIP Adaptation and Parameter-Efficient Transfer

CLIP adaptation methods generally preserve the pretrained encoders while modifying prompts, classifiers, or lightweight trainable components. Early prompt-learning methods such as CoOp and CoCoOp learn task-specific textual contexts while keeping the CLIP backbone frozen (Zhou et al., 2022b;a). These methods demonstrate that downstream supervision can improve CLIP transfer without full model fine-tuning, but their adaptation is primarily performed in the text-prompt space.

Another line of work adapts the classifier or similarity computation on top of frozen image features. CLIP-NCM and IsoCLIP-NCM provide nearest-class-mean references, while Linear Probe learns

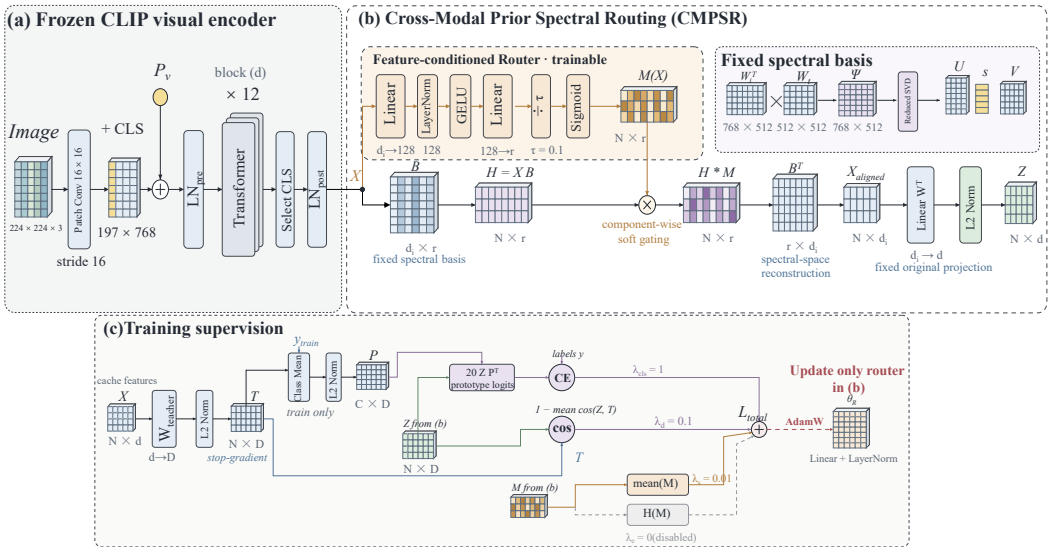

Figure 2: Overview of Dynamic-IsoCLIP. The cross-modal spectral basis is constructed offline from frozen image and text projectors. The image-conditioned router predicts a sample-wise soft mask, while only the router is updated during supervised adaptation.

a supervised linear classifier over the frozen CLIP representation. These approaches are computationally simple and provide useful baselines for separating the effect of classifier learning from the effect of representation adaptation. IsoCLIP-NCM further combines prototype-style classification with the fixed projector-level spectral transformation introduced by IsoCLIP (Magistri et al., 2026).

The supervised baselines considered in Table 1 also include several parameter-efficient adaptation strategies. TaskRes learns task residuals for adapting vision-language representations (Yu et al., 2023). CLIP-AST studies adaptive parameter selection for CLIP tuning (Zhang et al., 2025), while DPC (Li et al., 2025a) and TAC (Hao et al., 2025) explore prompt collaboration and task-aware clustering. LDC addresses few-shot adaptation through logits deconfusion (Li et al., 2025b), and MergeTune continues tuning through model merging or related parameter-efficient updates (Wang et al., 2026). The table further includes DPL, CAPT, and Prompt-OT, which investigate dropout-based prompt learning, confusion-aware prompt tuning, and optimal-transport regularization, respectively (Chen et al., 2026a; Shao et al., 2026; Chen et al., 2026b). Although these methods differ in their optimization targets, most learn a shared adaptation rule that is applied uniformly to all input images.

Dynamic-IsoCLIP is complementary to these approaches. Rather than changing the text prompt, learning only a classifier, or applying one shared residual transformation, it keeps the CLIP encoders, projectors, and cross-modal spectral basis frozen. Its trainable router predicts an image-conditioned soft weighting over projector-induced spectral directions. This design directly addresses input-dependent visual adaptation while retaining the parameter efficiency and image-side inference properties required by the retrieval setting.

2.2 Intra-modal Alignment and Spectral Routing

Several studies analyze the geometric mismatch between CLIP's inter-modal training objective and intra-modal image similarity. Modality inversion and implicit decoding expose this mismatch from a query-side perspective, while topological and covariance-spectrum analyses study the structure of the shared embedding space (Mistretta et al., 2025; D'Orazio et al., 2026; You et al., 2025; Grzywaczewski et al., 2026). IsoCLIP provides a projector-level spectral decomposition for improving intra-modal alignment (Magistri et al., 2026). Other spectral methods target concept discovery, robustness, or test-time adaptation (Caselli et al., 2026; Nie et al., 2026; Baek et al., 2026; Dafnis & Metaxas, 2025). Conditional computation and mixture-of-experts models provide related examples of input-dependent selection (Bengio et al., 2015; Shazeer et al., 2017; Riquelme et al.,

2021). Our method combines these ideas through a frozen cross-modal basis and a supervised, image-conditioned soft mask.

3 METHOD

3.1 FROZEN CROSS-MODAL SPECTRAL BASIS

We use row vectors throughout. Let $f = E_{\text{img}}(x) \in \mathbb{R}^{1\times d_i}$ and let $W_i \in \mathbb{R}^{d\times d_i}$ and $W_t \in \mathbb{R}^{d\times d_t}$ be the frozen image and text projectors. Following IsoCLIP, we define

$$\Psi = W_i^\top W_t = U\Sigma V^\top, \qquad U \in \mathbb{R}^{d_i\times r}, \quad V \in \mathbb{R}^{d_t\times r}. \tag{1}$$

The columns of U provide the image-side spectral basis, and fU contains the coordinates of the image feature in this basis. All projectors and the basis remain fixed during adaptation. A binary mask H gives the fixed IsoCLIP representation

$$z^{\text{Iso}} = \text{Norm}\big(fU\,\text{Diag}(H)U^\top W_i^\top\big), \tag{2}$$

where H removes the selected leading and trailing spectral components. With one-based indices, $H_j = 1$ for $k_t < j \le r - k_b$ and $H_j = 0$ otherwise, where k_t and k_b are the numbers of removed leading and trailing components.

The role of this decomposition can be seen from the unnormalized cross-modal similarity

$$\text{sim}(f,t) = \frac{fW_i^\top W_t t^\top}{\|fW_i^\top\|_2\|tW_t^\top\|_2}. \tag{3}$$

The operator $\Psi = W_i^\top W_t$ therefore describes how directions in the text pre-projection space correspond to directions in the image pre-projection space. Its singular values quantify the strength of these directional correspondences, while the columns of U define the image-side coordinates used by the router. In this view, IsoCLIP performs a global spectral filter: with one-based indices, $H_j = 1$ for $k_t < j \le r-k_b$ and $H_j = 0$ otherwise. The same values of k_t and k_b are applied to every image, so the operation removes or retains a direction independently of the visual content of the sample.

This fixed construction is useful as a geometric prior because it is computed once from the pretrained projectors and introduces no trainable backbone parameters. It is nevertheless restrictive for heterogeneous data. The most informative directions for a bird image need not be the same as those for a satellite scene or a food image. CMPSR keeps the basis itself fixed and learns only how strongly each image should use its directions, thereby separating the stable cross-modal coordinate system from the task-dependent routing decision.

3.2 CROSS-MODAL PRIOR SPECTRAL ROUTING

CMPSR predicts a continuous mask for each image using a two-layer MLP:

$$a(f) = \text{GELU}(\text{LN}(fW_1 + b_1))\,W_2 + b_2, \tag{4}$$

$$M(f) = \text{sigmoid}(a(f)/\tau_m) \in (0,1)^r, \tag{5}$$

$$\hat{z} = \text{Norm}\big((fU \odot M(f))U^\top W_i^\top\big). \tag{6}$$

The mask is continuous and sample-specific. The spectral basis carries the offline cross-modal prior, whereas $M(f)$ determines how the current image uses that basis. The router receives only f during inference; text features are not dynamically supplied to it. Setting the mask to a shared binary H recovers the fixed IsoCLIP construction as an external limiting case.

We distinguish two initialization settings. The no-mask-prior configuration does not initialize the mask from IsoCLIP truncation, although it still uses the cross-modal basis in Equation 1. The IsoCLIP-mask-prior configuration initializes low keep probabilities at the spectral extremes and high probabilities in the middle. For a desired initial keep probability p_j^{prior}, the final-layer bias is initialized as

$$b_{2,j} = \tau_m \,\text{logit}\Big(p_j^{\text{prior}}\Big). \tag{7}$$

Both settings use the same router architecture and training objective.

The term "prior" refers to the construction of the routing space and, in the second setting, to the optional starting point of the mask. It does not mean that text features are injected into the router for each query. This distinction gives CMPSR an asymmetric multimodal design: text projectors contribute offline geometric information, whereas the online computation is entirely image-side. For image-to-image retrieval, the same router can independently encode every query and gallery image, avoiding pairwise query–gallery cross-attention and making the representation suitable for large galleries.

The continuous mask should also be distinguished from an exact binary selection. Since the sigmoid output lies in $(0, 1)$ for finite logits, CMPSR softly attenuates spectral components rather than guaranteeing that any component is removed. This formulation allows gradients to pass through all directions during training and lets different inputs use different mixtures of the shared basis. A hard mask is recovered only as an externally specified or limiting configuration, not as the generic output of the learned router.

3.3 Supervised Objective

For the main configuration, the teacher is the original frozen CLIP projector, so

$$z_n^{\mathrm{tea}} = \mathrm{Norm}(f_n W_i^\top). \tag{8}$$

For class c, a fixed prototype is computed from the training examples $\mathcal{D}_c$:

$$p_c = \mathrm{Norm}\left(\frac{1}{|\mathcal{D}_c|}\sum_{n\in\mathcal{D}_c} z_n^{\mathrm{tea}}\right), \qquad s_{n,c} = \alpha \hat{z}_n p_c^\top. \tag{9}$$

The task loss is cross-entropy over prototype similarities:

$$\mathcal{L}_{\mathrm{task}} = -\frac{1}{B}\sum_{n=1}^{B} \log \frac{\exp(s_{n,y_n})}{\sum_{c=1}^{C}\exp(s_{n,c})}. \tag{10}$$

We use cosine representation distillation and a mean-mask penalty:

$$\mathcal{L}_{\mathrm{distill}} = \frac{1}{B}\sum_{n=1}^{B}\left(1 - \hat{z}_n (z_n^{\mathrm{tea}})^\top\right), \tag{11}$$

$$\mathcal{L}_{\mathrm{sparse}} = \frac{1}{Br}\sum_{n=1}^{B}\sum_{j=1}^{r} M_j(f_n), \tag{12}$$

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{task}} + \lambda_d \mathcal{L}_{\mathrm{distill}} + \lambda_s \mathcal{L}_{\mathrm{sparse}}. \tag{13}$$

The task loss breaks the symmetry among otherwise plausible spectral masks. Distillation acts as a semantic anchor, and the mean-mask term limits excessive activation. These terms regularize the learned solution but do not prove semantic preservation or actual sparse computation; a continuous low weight does not automatically remove dense operations. Without task supervision, the corresponding label-free objective is

$$\mathcal{L}_{\text{label-free}} = \lambda_d \mathcal{L}_{\mathrm{distill}} + \lambda_s \mathcal{L}_{\mathrm{sparse}}, \tag{14}$$

which provides no direct criterion for selecting class-discriminative directions.

Breaking the routing symmetry. The cross-modal basis is obtained without knowledge of the downstream class partition. Its components can consequently mix class-relevant visual structure with nuisance variation or domain-specific statistics. If the router is optimized only to remain close to a teacher and to reduce its average mask value, many masks can obtain similar regularization costs without being equally useful for recognition. Prototype supervision supplies the missing discriminative criterion: it increases the relative score of spectral configurations that separate the training classes and decreases the score of configurations that do not. This is why the main method is explicitly supervised, while the label-free objective is retained only as an analysis formulation.

Regularizers as training stabilizers. The distillation term does not force the routed feature to be identical to the teacher. Instead, it limits excessive displacement from the pretrained image embedding space and acts as a semantic anchor. The mean-mask term discourages indiscriminate activation of every spectral direction, but it should not be interpreted as a proof of computational sparsity because the implementation still evaluates a dense vector unless explicit pruning is added. Their intended role is therefore to shape the optimization landscape and prevent unstable or overly specialized routing, while the task loss remains the primary source of semantic supervision.

3.4 Inference and Computational Cost

At inference, each image independently follows Equations 4–6. The SVD is computed once offline, and normalized routed embeddings are compared by cosine similarity. For hidden width h, the router has

$$N_{\text{router}} = d_i h + hr + 3h + r \tag{15}$$

trainable parameters, including the affine parameters of layer normalization. The additional analytical cost is $O(d_i h + hr + d_i r)$ per image, excluding the frozen encoder and original projection. These are analytical estimates rather than measured latency results.

4 Experiments

4.1 Setup and Evaluation Protocol

We use frozen CLIP ViT-B/16 representations on Caltech101, CUB-200-2011, Stanford Cars, Oxford-IIIT Pets, Flowers102, FGVC Aircraft, DTD, EuroSAT, Food101, SUN397, and UCF101. Only the router is trained. Unless otherwise stated, the reported configuration uses a two-layer MLP with hidden width 128, AdamW for 20 epochs, batch size 512, learning rate 10^{-3}, weight decay 10^{-4}, and no entropy regularizer.

The main task is image-to-image retrieval. Each query and gallery image is independently processed by the same image encoder and router, and gallery images are ranked by cosine similarity. A gallery image is relevant when it has the same class label as the query. Tables 1–3 report mean average precision (mAP), expressed as a percentage. All eleven datasets use test images as queries and training images as the gallery; the CUB evaluation uses the explicit test/train split override. Training labels may be used for prototypes and optimization, while test labels are used only to score retrieval relevance. For query q, if $\mathrm{rel}_q(k)$ indicates same-class relevance at rank k and R_q is the number of relevant gallery images, then

$$\mathrm{AP}(q) = \frac{1}{R_q} \sum_{k=1}^{G} \mathrm{rel}_q(k) \frac{\sum_{j=1}^{k} \mathrm{rel}_q(j)}{k}, \qquad \mathrm{mAP} = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{AP}(q). \tag{16}$$

The benchmark suite is deliberately heterogeneous. Caltech101 and Oxford-IIIT Pets are general object datasets; CUB-200-2011, Stanford Cars, Flowers102, and FGVC Aircraft emphasize fine-grained recognition; DTD evaluates texture; EuroSAT evaluates remote-sensing scenes; Food101 contains object-centric food categories; and SUN397 and UCF101 cover scenes and actions. This diversity is important for testing whether a single global spectral rule is sufficient. Thus, the first three tables evaluate ranking quality rather than nearest-neighbor top-1 accuracy.

Base-to-new evaluation is separate from image-to-image retrieval. The router is trained on base classes, and base and new test images are classified with the corresponding text classifiers. Table 4 reports Base, New, and their harmonic mean H. Thus, the first three result tables measure image-to-image mAP, while Table 4 measures text-classifier classification accuracy.

4.2 Comparison with CLIP Adaptation Baselines

We first compare Dynamic-IsoCLIP with representative CLIP adaptation baselines under the common image-to-image retrieval protocol. The complete numerical comparison is reported in Table 1.

Table 1: Comparison with CLIP adaptation baselines on eleven benchmarks. All entries report image-to-image retrieval mAP (%). "Trainable" indicates whether additional parameters are optimized, and "Label Used" indicates whether training labels are used. The best result in each column is bold.

Method	Trainable	Label Used	Caltech	CUB	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	UCF101
CLIP-NCM	×	✓	80.10	38.53	34.36	53.18	73.62	22.41	33.35	49.05	43.58	31.65	50.99
Linear Probe	✓	✓	86.75	44.44	47.53	66.19	77.49	24.60	42.48	76.43	59.15	43.03	60.83
TaskRes(CVPR 2023)	✓	✓	80.76	33.02	31.71	37.44	71.20	19.60	31.17	52.59	43.57	36.15	50.32
CLIP-AST(CVPR 2025)	✓	✓	81.76	36.10	36.51	41.23	73.11	20.14	32.36	68.23	58.56	39.62	54.19
DPC(CVPR 2025)	✓	✓	77.46	40.46	33.45	60.26	74.87	23.16	33.95	73.94	54.30	31.11	48.76
TAC(CVPR 2025)	✓	✓	83.22	36.20	36.79	43.35	73.96	20.36	33.07	55.63	49.08	38.98	54.11
LDC(CVPR 2025)	✓	✓	81.82	37.31	38.89	49.07	75.67	21.80	33.49	68.88	48.00	37.64	54.32
MergeTune-style(ICLR 2026)	✓	✓	80.59	36.70	34.96	45.56	73.46	19.84	35.59	70.77	48.78	39.38	52.92
DPL(AAAI 2026)	✓	✓	81.50	36.58	37.58	46.47	74.04	21.22	33.62	64.94	47.40	37.59	53.26
CAPT(CVPR 2026)	✓	✓	81.83	37.48	35.42	41.70	73.16	20.14	32.15	59.55	61.07	38.88	51.12
Prompt-OT(WACV 2026)	✓	✓	77.61	40.47	33.46	60.28	74.78	23.16	33.96	73.83	54.09	31.13	48.79
IsoCLIP-NCM(CVPR 2026)	×	✓	81.96	39.03	42.76	62.38	73.62	22.64	35.65	52.37	45.74	30.34	50.36
Dynamic-IsoCLIP	✓	✓	**89.99**	**52.94**	**56.21**	**72.22**	**85.84**	**27.91**	**45.15**	**95.88**	**81.73**	**53.81**	**79.02**

Table 2: Training protocol and spectral initialization analysis. All entries report image-to-image retrieval mAP (%). CLIP and IsoCLIP are training-free references; Dynamic-IsoCLIP variants train the router with the training split.

Method	Mask	Training	Init	Caltech	CUB	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	UCF101
CLIP (Image–Image)	None	None	Original CLIP	80.60	31.60	31.00	36.30	70.80	19.00	30.70	51.20	42.80	35.90	49.80
IsoCLIP	Hard	None	k-top/k-bottom	85.00	38.60	41.20	50.70	77.40	20.50	36.00	55.60	53.50	38.50	55.40
Dynamic-IsoCLIP	Soft	Supervised	No prior	89.99	52.94	56.21	72.22	85.84	27.91	45.15	95.88	81.73	53.81	79.02
Dynamic-IsoCLIP	Soft	Supervised	IsoCLIP prior	93.96	55.48	58.71	81.20	89.55	28.23	56.15	96.31	81.95	55.36	81.41

Dynamic-IsoCLIP achieves the highest mAP among the compared methods on all eleven datasets. Relative to Linear Probe, the gains are 3.24, 8.50, 8.68, 6.03, 8.35, 3.31, 2.67, 19.45, 22.58, 10.78, and 18.19 points in the displayed dataset order. The improvements over CLIP-NCM and IsoCLIP-NCM are also consistent, indicating that sample-wise spectral weighting improves the frozen-feature retrieval representation under this protocol. These results establish competitiveness within the listed baselines, not universal superiority over all possible adapters or full fine-tuning methods.

The improvement is visible across different dataset regimes rather than being concentrated in a single category family. On fine-grained datasets such as CUB, Cars, Flowers, and Aircraft, the routed representation can preserve complementary shape, texture, and local-attribute evidence. On EuroSAT, SUN397, and UCF101, the larger gains are consistent with the need to balance global scene or action structure against image-specific details. The table does not by itself prove that the router is the only source of improvement, since the baselines differ in their classifiers and optimization procedures; it does show that CMPSR is effective under the common retrieval protocol and parameter-efficient training setting.

4.3 Training Protocol and Mask Initialization

We next isolate the effects of the training protocol and mask initialization. Table 2 compares the original CLIP representation, training-free IsoCLIP, and supervised Dynamic-IsoCLIP with either no mask prior or an IsoCLIP mask prior.

IsoCLIP improves over original CLIP on all eleven retrieval benchmarks. Supervised soft routing improves over both training-free references in the no-prior setting, and IsoCLIP-prior initialization further improves every column. The prior gains are largest on DTD and Pets and smallest on Food101, Aircraft, and EuroSAT. Since the training protocols differ between the reference and supervised rows, this table is an initialization and protocol analysis rather than a supervision-matched ablation.

The no-prior row is important because it shows that the benefit of CMPSR does not require copying the manually selected IsoCLIP hard mask at initialization. The router can learn a useful soft selection from the task signal while using the same frozen cross-modal basis. The prior-initialized row then measures whether the fixed IsoCLIP preference provides a favorable starting point for further adaptation. Its consistent improvement suggests that the prior is a useful inductive bias, but the dataset-dependent magnitude of the gain indicates that initialization and task optimization should not be conflated.

Table 3: Ablation of the sparsity and distillation losses. All entries report image-to-image retrieval mAP (%).

Dataset	Full	w/o Sparsity	w/o Distill	w/o Both
EuroSAT	**95.88**	94.85	94.66	91.09
Food101	**81.73**	79.41	79.27	73.77
Cars	**56.21**	46.51	47.06	40.31
CUB	**52.94**	44.90	46.43	44.08

Table 4: Base-to-new generalization results. Entries report classification accuracy (%). The router is trained on base classes, and the corresponding text classifiers are used for base and new evaluation.

Dataset	Dynamic-IsoCLIP			CLIP		
	Base	New	H	Base	New	H
Caltech101	**93.89**	**97.12**	**95.48**	92.79	93.78	93.28
Oxford Pets	**97.14**	**93.58**	**95.33**	89.52	88.59	89.05
Food101	**92.77**	**85.42**	**88.94**	86.76	85.05	85.90
EuroSAT	**98.15**	**53.74**	**69.45**	68.44	29.69	41.41
CUB	**78.32**	**61.13**	**68.67**	56.19	54.53	55.35

4.4 Loss Component Ablation

To examine the role of the two auxiliary regularizers, Table 3 reports a controlled ablation of the sparsity and distillation terms on four representative datasets.

The full objective obtains the highest mAP on all four datasets, while removing both regularizers gives the lowest result. The drops from removing both terms are 4.79 points on EuroSAT, 7.96 on Food101, 15.90 on Cars, and 8.86 on CUB. Removing either term also lowers the score in every case. These results support using distillation and the mean-mask penalty as auxiliary stabilizers in the reported retrieval setting, but the score differences alone do not establish a causal mechanism or measured computational sparsity.

The effect is strongest on Cars and CUB, where fine-grained discrimination requires the router to retain useful subtle directions without drifting arbitrarily from the teacher. On EuroSAT and Food101, the same regularizers still improve mAP, but the relative gaps are smaller. This pattern is consistent with the intended role of the auxiliary losses: they are not independent engines of class recognition, but stabilizing constraints whose value depends on the visual complexity of the task. In particular, the task loss is present in every row; removing the two auxiliary terms does not create a label-free model.

4.5 Base-to-New Generalization

Finally, we evaluate whether supervised routing transfers beyond the classes used for router training. Table 4 reports the Base, New, and harmonic-mean accuracies for Dynamic-IsoCLIP and the original CLIP reference.

Dynamic-IsoCLIP improves Base, New, and H over CLIP on all five datasets. New accuracy improves by 3.34, 4.99, 0.37, 24.05, and 6.60 points for Caltech101, Oxford Pets, Food101, EuroSAT, and CUB, respectively; the corresponding harmonic-mean gains are 2.20, 6.28, 3.04, 28.04, and 13.32 points. This is evidence of unseen-class transfer within the evaluated partitions, not a guarantee of open-world or cross-modal retrieval preservation.

The gains on the New split are particularly relevant because the router is trained only on base classes. Nevertheless, this protocol remains a within-dataset evaluation: the text classifiers, class partitions, and visual domains are fixed for each benchmark. It should therefore be interpreted as evidence that the constrained image representation can retain useful generic structure under the tested partitions, rather than as a universal guarantee of zero-shot preservation. The image-to-image mAP results and the base-to-new classification results answer different questions and are not directly interchangeable.

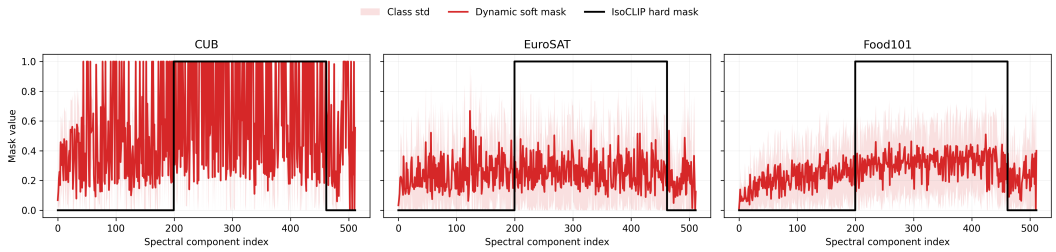

Figure 3: Comparison of the fixed IsoCLIP mask and sample-wise Dynamic-IsoCLIP masks on CUB, EuroSAT, and Food101.

4.6 Qualitative Analysis

Figure 3 compares the fixed hard mask with sample-wise soft masks on CUB, EuroSAT, and Food101. The continuous patterns vary across samples and datasets, supporting the intended sample-wise routing behavior. This qualitative evidence is illustrative and should not be interpreted as an aggregate metric.

The hard-versus-soft comparison shows the qualitative distinction between a global binary rule and a continuous sample-wise rule. Dynamic-IsoCLIP does not simply reproduce one fixed central interval: the relative weights vary with the input and can preserve non-zero contributions across a broader range of directions. On CUB, the variation is consistent with the need to preserve subtle bird attributes; on EuroSAT and Food101, the routing patterns are more concentrated, reflecting more stable scene- or category-level cues. These examples illustrate the intended input-conditioned behavior without claiming that the visualization alone proves causality.

5 Limitations and Scope

Dynamic-IsoCLIP is a supervised image-side adaptation method. Its router uses training labels, and the current study does not establish preservation of image-to-text or text-to-image retrieval. The base-to-new experiment evaluates unseen classes within the same dataset rather than cross-domain open-world transfer. In addition, the continuous mask is a soft weighting mechanism: it does not automatically produce sparse hardware execution, and the reported computational analysis is analytical rather than a measured latency benchmark. Finally, the experiments use one frozen CLIP backbone and do not include a matched global soft-mask control. These limitations define the scope of the reported claims and motivate future evaluation of cross-modal retention, measured efficiency, and label-free routing.

6 Conclusion

We presented Dynamic-IsoCLIP, a parameter-efficient Cross-Modal Prior Spectral Routing method for frozen CLIP image representations. CMPSR retains a cross-modal spectral basis constructed offline from frozen projectors and predicts a continuous image-conditioned mask at inference. Task supervision provides the discriminative signal, while distillation and the mean-mask penalty constrain the learned solution.

Across eleven image-to-image retrieval benchmarks, the main configuration obtains the highest mAP among the listed supervised baselines. The initialization analysis shows a consistent benefit from an IsoCLIP mask prior, and the ablation study supports the use of the two auxiliary terms in the evaluated setting. Base-to-new evaluation further shows improved classification accuracy over CLIP on the five reported datasets under the specified text-classifier protocol.

The conclusions are intentionally limited to the evaluated protocols. The current study does not measure image-to-text or text-to-image retrieval preservation, does not include a matched global soft-mask control, and does not report measured latency or repeated-run uncertainty. Extending evaluation to these settings and improving label-free routing are important directions.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and LaTeX consistency checks. They were not used to generate or alter experimental data, fabricate results, conduct experiments, or make final scientific decisions. All AI-assisted suggestions were reviewed by the authors, who take responsibility for the final content.

ETHICS STATEMENT

This work uses standard visual recognition benchmarks and does not involve human subjects or personal data collection. The datasets are used for research evaluation in accordance with their respective terms of use.

REPRODUCIBILITY STATEMENT

This manuscript specifies the model, objective, evaluation metrics, protocols, and reported configurations. Anonymous source code and configuration files can be provided as supplementary material when available.

REFERENCES

Sunghyun Baek, Jaemyung Yu, Seunghee Koh, Minsu Kim, Hyeonseong Jeon, and Junmo Kim. IMSE: Intrinsic mixture of spectral experts fine-tuning for test-time adaptation. In *International Conference on Learning Representations (ICLR)*, 2026.

Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, and Doina Precup. Conditional computation in neural networks for faster models. *arXiv preprint arXiv:1511.06297*, 2015.

Lorenzo Caselli, Marco Mistretta, Simone Magistri, and Andrew D Bagdanov. SpectralGCD: Spectral concept selection and cross-modal representation learning for generalized category discovery. In *International Conference on Learning Representations (ICLR)*, 2026.

Biao Chen, Lin Zuo, Mengmeng Jing, Kunbin He, and Yuchen Wang. Dropout prompt learning: Towards robust and adaptive vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19987–19995, mar 2026a. doi: 10.1609/aaai.v40i24.39083. AAAI 2026 Technical Track on Machine Learning I.

Xiwen Chen, Wenhui Zhu, Peijie Qiu, Hao Wang, Huayu Li, Haiyu Wu, Xuanzhao Dong, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. Prompt-OT: An optimal transport regularization paradigm for knowledge preservation in vision-language model adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 667–676, 2026b.

Konstantinos M Dafnis and Dimitris N Metaxas. Test-time spectrum-aware latent steering for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Antonio D'Orazio, Maria Rosaria Briglia, Donato Crisostomi, Dario Loi, Emanuele Rodolà, and Iacopo Masi. Implicit inversion turns CLIP into a decoder. In *International Conference on Learning Representations (ICLR)*, 2026.

Jakub Grzywaczewski, Dawid Płudowski, and Przemysław Biecek. Your CLIP has 164 dimensions of noise: Exploring the embeddings covariance eigenspectrum of contrastively pretrained vision-language transformers. *arXiv preprint arXiv:2605.14893*, 2026.

Fusheng Hao, Fengxiang He, Fuxiang Wu, Tichao Wang, Chengqun Song, and Jun Cheng. Task-aware clustering for prompting vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14745–14755, 2025.

Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. DPC: Dual-prompt collaboration for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25623–25632, 2025a.

Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with CLIP for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25411–25421, June 2025b.

Simone Magistri, Dipam Goswami, Marco Mistretta, Bartłomiej Twardowski, Joost van de Weijer, and Andrew D Bagdanov. IsoCLIP: Decomposing CLIP Projectors for efficient intra-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29315–29324, 2026.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Sen Nie, Jie Zhang, Zhuo Wang, Shiguang Shan, and Xilin Chen. Contrastive spectral rectification: Test-time defense towards zero-shot adversarial robustness of CLIP. In *International Conference on Machine Learning (ICML)*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8583–8595, 2021.

Maoyuan Shao, Yutong Gao, Xinyang Huang, Lijuan Sun, Guoshun Nan, and Chuang Zhu. CAPT: Confusion-aware prompt tuning for reducing vision-language misalignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3154–3164, 2026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Wenqing Wang, Da Li, Xiatian Zhu, and Josef Kittler. MergeTune: Continued fine-tuning of vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Junwon You, Dasol Kang, and Jae-Hun Jung. Topological alignment of shared vision-language embedding space. *arXiv preprint arXiv:2510.10889*, 2025.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10899–10909, 2023.

Yi Zhang, Yi-Xuan Deng, Meng-Hao Guo, and Shi-Min Hu. Adaptive parameter selection for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4280–4290, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022b.

A Additional Method Details

A.1 Cross-Modal Operator and Fixed Mask

Let $f \in \mathbb{R}^{1\times d_i}$ and $t \in \mathbb{R}^{1\times d_t}$ denote pre-projection image and text features. With $z = fW_i^\top$ and $z^t = tW_t^\top$, their cosine similarity is

$$\mathrm{sim}(f,t) = \frac{fW_i^\top W_t t^\top}{\|fW_i^\top\|_2 \|tW_t^\top\|_2}. \tag{17}$$

This expression motivates the cross-modal operator

$$\Psi = W_i^\top W_t \in \mathbb{R}^{d_i\times d_t}, \qquad \Psi = U\Sigma V^\top, \tag{18}$$

where U and V contain the image-side and text-side singular directions, respectively, and Σ contains their singular values. The spectral coordinates are $\tilde{f} = fU$. When all directions are retained, reconstruction from the reduced basis is $\tilde{f}U^\top = fUU^\top$; the mask modifies this reconstruction by weighting each coordinate before projection.

Using one-based indices, let k_t and k_b be the numbers of leading and trailing components removed. The fixed mask is

$$H_j = \begin{cases} 1, & k_t < j \le r - k_b, \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1,\dots,r. \tag{19}$$

It removes exactly k_t and k_b components. In projector form, the corresponding fixed image projector is

$$W_i^{\mathrm{Iso}} = W_i U \,\mathrm{Diag}(H) U^\top. \tag{20}$$

Thus, the hard mask is equivalent to restricting the original image projector to one globally selected spectral subspace. Replacing H with an input-dependent soft mask $M(f)$ gives the routed embedding

$$\mathrm{Norm}\big(fU\,\mathrm{Diag}(M(f))U^\top W_i^\top\big).$$

Since the mask is diagonal and symmetric,

$$fU\,\mathrm{Diag}(M(f))U^\top W_i^\top = ((fU)\odot M(f))U^\top W_i^\top.$$

A.2 Expanded CMPSR Derivation

For completeness, the image-side routing operation can be written in three equivalent forms. First, the feature is represented in the fixed spectral coordinate system and reweighted component-wise:

$$\tilde{f}^{\mathrm{CMPSR}} = (fU)\,\mathrm{Diag}(M(f)) = (fU)\odot M(f). \tag{21}$$

Second, it is reconstructed in the original image feature space:

$$f^{\mathrm{CMPSR}} = \tilde{f}^{\mathrm{CMPSR}} U^\top = fU\,\mathrm{Diag}(M(f))U^\top. \tag{22}$$

Finally, the frozen image projector maps the reconstructed feature to the normalized CLIP embedding:

$$\hat{z} = \mathrm{Norm}\big(fU\,\mathrm{Diag}(M(f))U^\top W_i^\top\big). \tag{23}$$

Equivalently, CMPSR induces the input-dependent projector

$$\widehat{W}_i(f) = W_i U\,\mathrm{Diag}(M(f))U^\top, \qquad \hat{z} = \mathrm{Norm}\Big(f\widehat{W}_i(f)^\top\Big). \tag{24}$$

When $M(f) = H$ for every input, this expression reduces exactly to Equation 20. The distinction is therefore entirely in the mask-generation rule: IsoCLIP uses a manually specified global binary vector, whereas CMPSR predicts a continuous vector separately for each image.

A.3 Initialization, Objective, and Cost Details

For the IsoCLIP-prior initialization, let p_j^{prior} denote the desired initial keep probability of spectral component j. The final router bias can be initialized as

$$b_{2,j} = \tau_m \operatorname{logit}\left(p_j^{\text{prior}}\right), \tag{25}$$

so that the initial sigmoid output approximately follows the selected prior. The no-prior setting does not impose the top/bottom truncation pattern at initialization; both settings use the same basis, router, and training objective.

Because both routed and teacher embeddings are normalized, the cosine distillation term has the equivalent form

$$1 - \cos(\hat{z}, z^{\text{tea}}) = \frac{1}{2}\|\hat{z} - z^{\text{tea}}\|_2^2. \tag{26}$$

This identity makes explicit why the distillation term acts as a semantic anchor: it penalizes excessive displacement while leaving the task loss free to improve class discrimination. The sparsity term operates on the mean soft-mask activation and does not by itself imply hardware-level conditional execution.

For a router hidden width h, the trainable parameter count is

$$N_{\text{router}} = d_i h + hr + 3h + r. \tag{27}$$

The additional per-image operations are approximately

$$\mathcal{O}(d_i h + hr + d_i r), \tag{28}$$

excluding the frozen image encoder and original projection. If $A = U^\top W_i^\top$ is cached, the final embedding can be evaluated as

$$\hat{z} = \operatorname{Norm}(((fU) \odot M(f))A), \tag{29}$$

which is algebraically identical to Equation 23.

A.4 Symmetry and Label-Free Objective

The spectral components can contain both semantic signal and domain noise. Distillation and mean-mask regularization alone do not specify which directions are discriminative, so many routing configurations can have similar regularization cost. The task loss supplies the signal that breaks this symmetry in the supervised formulation. Removing the task loss gives

$$\mathcal{L}_{\text{label-free}} = \lambda_d \mathcal{L}_{\text{distill}} + \lambda_s \mathcal{L}_{\text{sparse}}. \tag{30}$$

This objective is retained as a conceptual analysis variant rather than a reported training protocol. A low mean mask also does not automatically imply fewer executed operations, because the representation remains dense unless explicit pruning or conditional execution is added.

B Experimental Protocol Details

B.1 Retrieval Metric

For query q, let $\text{rel}_q(k)$ indicate whether the gallery image at rank k has the same class label. If R_q relevant gallery images exist, then

$$\text{AP}(q) = \frac{1}{R_q}\sum_{k=1}^{G} \text{rel}_q(k)\frac{\sum_{j=1}^{k}\text{rel}_q(j)}{k}, \qquad \text{mAP} = \frac{1}{|Q|}\sum_{q\in Q}\text{AP}(q). \tag{31}$$

The query and gallery images are processed independently for image-to-image retrieval. Base-to-new evaluation instead uses text classifiers and reports classification accuracy.

B.2 Recorded Configuration

Setting	Reported value
Backbone	CLIP ViT-B/16
Trainable module	Two-layer MLP router with layer normalization
Hidden width	128
Optimizer	AdamW
Epochs	20
Batch size	512
Learning rate	10^{-3}
Weight decay	10^{-4}
Loss terms	Task, distillation, mean-mask penalty
Entropy term	Disabled

All eleven datasets use test images as queries and training images as the gallery. CUB uses the explicit test/train override. No test labels are used for parameter optimization or routing; they are used only to score retrieval relevance.

C Additional Qualitative Results

The qualitative analysis examines sample-wise routing through three mask heatmaps and representative top-10 retrieval results. The figures are illustrative rather than aggregate performance estimates, and they are interpreted together with the retrieval protocol and the meaning of the mask weights.

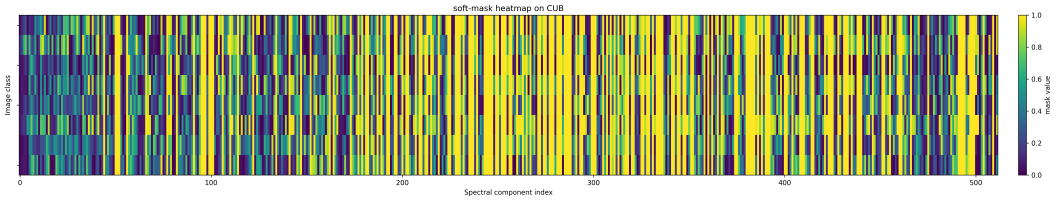

Figure 4: Representative sample-wise soft-mask visualization on CUB. Each row corresponds to an input sample, each column to a spectral component, and the color indicates the predicted soft-mask weight.

For CUB, the rows exhibit noticeable variation across spectral components. This pattern is consistent with fine-grained bird recognition, where small differences in plumage, shape, and local attributes can make different directions useful for different samples. The heatmap also shows that the router does not reduce every input to one identical central interval, although the visualization alone cannot determine which individual component corresponds to a particular semantic attribute.

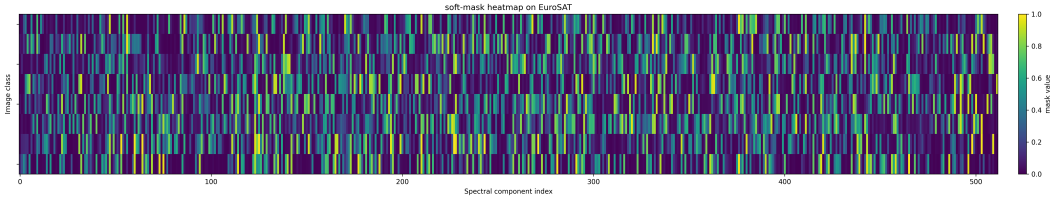

Figure 5: Representative sample-wise soft-mask visualization on EuroSAT. Each row corresponds to an input sample, each column to a spectral component, and the color indicates the predicted soft-mask weight.

For EuroSAT, the mask patterns are comparatively smoother and more concentrated. This behavior is compatible with remote-sensing categories that often depend on recurring scene-level structures, such as land-cover layout and large-scale texture. Nevertheless, different rows still retain differ-

ent weight profiles, indicating that the routing is conditional on the individual image rather than determined only by the dataset.



Figure 6: Representative sample-wise soft-mask visualization on Food101. Each row corresponds to an input sample, each column to a spectral component, and the color indicates the predicted soft-mask weight.

For Food101, the learned weights remain continuous but display category- and sample-dependent changes. Food images can share global composition while differing in local appearance, texture, or ingredient cues, so retaining a weighted mixture of directions is preferable to imposing one binary filter. Across all three datasets, the heatmaps support the existence of continuous sample-wise routing; they do not by themselves establish that routing causes every quantitative improvement.



Figure 7: Representative top-10 image-to-image retrieval results on CUB, EuroSAT, and Food101. Green borders indicate same-label gallery images and red borders indicate different-label images.

The retrieval examples provide a complementary view of the routed embeddings. On CUB, the Dynamic-IsoCLIP row contains more label-consistent neighbors in the displayed ranking, which is important for fine-grained categories with visually similar birds. On EuroSAT, the retrieved images show more consistent scene-level structure under the routed representation. On Food101, the ranking is generally more category-consistent, although visually similar food classes remain difficult to separate. These are selected examples rather than aggregate estimates, so their role is to illus-

trate how sample-wise routing can alter neighborhood ordering, not to replace the reported mAP measurements.