Test-Time Spectral Adaptation for Image-to-Image Multispectral Retrieval

Tien-Anh Nguyen [1,2], Thanh-Hai Tran [1,2], and Xuan-Bach Le [1,2⋆]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
{anh.nguyentien1904,tthai.sdh241,lexuanbach}@hcmut.edu.vn

Abstract. Multispectral archives contain evidence that RGB encoders cannot directly ingest. Sentinel-2 images include visible, red-edge, near-infrared, and short-wave infrared bands, whereas models such as CLIP expect three-channel natural images. This research studies image-to-image multispectral retrieval, where query and gallery images are encoded by the same method-specific pipeline and ranked by cosine similarity between descriptors. Starting from a generic RGB-pretrained CLIP backbone, with no Earth-observation pretraining, this study proposes a parameter-frozen test-time adaptation that updates only B band-weight logits per image. The method initializes the logits uniformly, refines them with a lightweight manifold-consistency objective steered by a class-name text prior, and reports band-level attribution scores. On 5-fold EuroSAT retrieval with disjoint query/gallery splits, all-band encoding raises Recall@1 from 80.28% for RGB-CLIP to 86.49% for the proposed adapted descriptor and mAP to 46.50%; the all-bands-average reference is statistically indistinguishable on this benchmark, showing that the main EuroSAT gain comes from exposing the frozen encoder to every band. On a 5-seed BigEarthNet-MM evaluation, the proposed method and the all-bands-average reference perform within one standard deviation across all metrics, while both surpass fixed spectral reductions and cache-based baselines by clear margins. This research therefore frames the contribution as accuracy competitive with full-band averaging, with per-band interpretability and near-zero adaptation overhead in the frozen generic-CLIP regime.

Keywords: Multispectral retrieval · Vision-language models · Test-time adaptation · Spectral band weighting · Remote sensing

1 Introduction

Multispectral remote sensing retrieval asks whether a query patch can retrieve related scenes while using all sensor channels that support the match. The task is increasingly treated as image-embedding search, from large-scale archive retrieval

⋆ Corresponding author

to multispectral foundation-model studies [1,2]. Sentinel-2 imagery makes the descriptor problem richer than ordinary RGB retrieval: visible bands describe color and texture, while near-infrared (NIR), red-edge, and short-wave infrared (SWIR) bands capture vegetation structure, moisture, and material differences that are weak or invisible in RGB [3,4]. A useful descriptor should therefore exploit more than the three channels that happen to match natural-image pretraining.

This sensor mismatch matters for vision-language models (VLMs) such as CLIP. Their visual filters and embedding geometry are learned from RGB image-text pairs, so non-visible bands can produce activations outside the training distribution [5,6]; at the same time, RGB web corpora only partially represent the spectral cues that separate land-cover types [7]. Unlike text-to-image retrieval, the object of study here is the image descriptor itself: query and gallery items are multispectral observations, and both must be encoded by the same rule before cosine similarity is meaningful.

Existing adaptation strategies leave this operating point only partly covered: cache/prompt adapters [8] refine predictions in RGB feature or logit space without modelling spectral-band relationships; remote-sensing VLMs [5,9,10,11] learn stronger representations but change the encoder and require substantial pretraining; and spectral fusion methods [12] combine bands before or after feature extraction, with fixed reductions unable to adapt per image and learned fusion needing task-specific data (Sec. 5 elaborates).

This research targets a practical middle ground: starting from a generic RGB-pretrained CLIP backbone, the encoder stays fixed, paired multispectral captions are avoided, all available bands are used, and the result is an inspectable descriptor that adapts the fusion rule rather than the CLIP parameters. The goal is not to surpass domain-pretrained EO VLMs nor to beat all-bands averaging on saturated ranking metrics, but to define a lightweight adaptation point that provides per-band interpretability while maintaining competitive retrieval accuracy in the multi-label regime at negligible added cost. This research makes three contributions:

- **Query-label-free spectral weighting.** This research initializes per-band logits uniformly and refines them with a manifold-consistency objective guided by a dataset-level class-name prior — using no per-query labels (Secs. 2.2, 2.4); the resulting closed-vocabulary assumption and its open-set relaxation are discussed in Sec. 6.
- **Parameter-frozen test-time adaptation.** This research updates only the B band-weight logits for five gradient steps, leaving every CLIP parameter fixed, and exposes per-band attribution scores for the final descriptor.
- **A controlled image-to-image evaluation.** This research compares fixed spectral reductions, cache/logit adaptation, and graph diffusion under one disjoint query/gallery protocol, with sensitivity, ablation, runtime, and attribution analyses on EuroSAT and BigEarthNet-MM.

Experiments on EuroSAT use 5-fold cross-validation (60% cache training, 20% query, 20% gallery), reporting Recall@K and mAP with cross-fold error

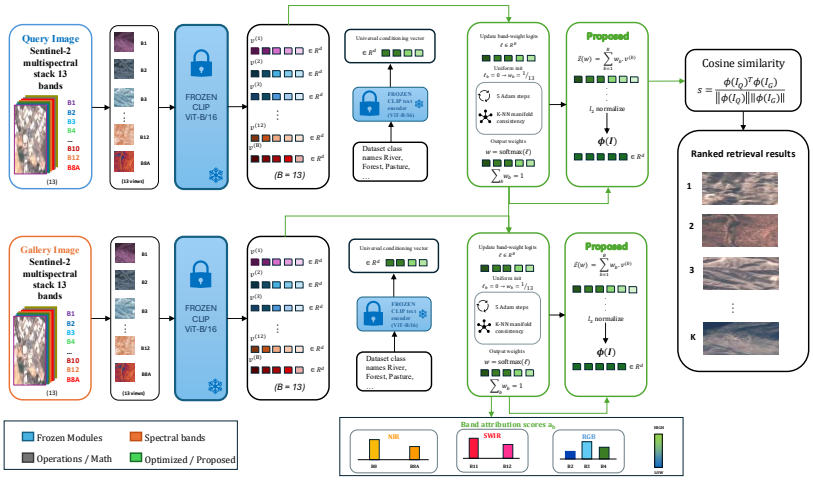

Fig. 1. Image-to-image multispectral retrieval framework. Each band is encoded as a pseudo-RGB view by frozen CLIP; spectral weights start uniform (w_b=$1/B$) and are refined by five manifold-consistency steps that update only the band-weight logits before cosine retrieval.

bars and paired t-tests. A complementary multi-seed BigEarthNet-MM study probes the multi-label setting. This research emphasizes the discriminative R@1 and mAP rather than near-saturated R@10, analyzes computational cost, and outlines future robustness directions regarding atmospheric corruption, mixed land cover, and low spectral contrast.

2 Methodology

This research keeps the retrieval protocol symmetric: every query and gallery image passes through the same method-specific descriptor before ranking. Figure 1 shows the full pipeline; the components below develop it in turn — frozen-encoder setting (Sec. 2.1), band encoding, conditioning, fusion, and manifold adaptation (Secs. 2.2–2.4), then attribution and complexity (Secs. 2.5–2.6); notation is consolidated in the supplementary material.

2.1 Problem Formulation

Let $\mathcal{D} = \{(I_i, y_i)\}_{i=1}^{N}$ be a multispectral archive, where $I_i \in \mathbb{R}^{H\times W\times B}$ contains B bands and y_i is an evaluation annotation, either a single scene class or a multi-label set. This study uses labels only for evaluation and, for supervised baselines, cache construction; the proposed descriptor never sees them at test time. The query set $\mathcal{Q}$ and gallery set $\mathcal{G}$ are disjoint subsets of $\mathcal{D}$. For a query image $I_q \in \mathcal{Q}$, retrieval ranks gallery images $I_g \in \mathcal{G}$ by cosine similarity between method-specific descriptors:

$$s(I_q, I_g) = \phi(I_q)^\top \phi(I_g), \tag{1}$$

where both sides are processed by the same encoder $\phi(\cdot)$. Relevance is defined by the benchmark protocol: exact class equality for single-label EuroSAT, and label-set overlap for multi-label BigEarthNet-MM. This research assumes access to a frozen CLIP vision encoder $f_v : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$ and, only for class-name–guided conditioning, a frozen CLIP text encoder f_t.

Challenge. CLIP expects three-channel inputs, whereas multispectral images have $B > 3$ bands. Selecting RGB bands discards NIR, red-edge, and SWIR evidence; concatenating all bands would require changing the encoder interface. The goal is to construct a discriminative descriptor $\phi(I)$ from all bands while keeping model parameters fixed.

2.2 Spectral Band Encoding and Weighting

This study turns each spectral band into a CLIP-compatible view: each band $b \in \{1, \ldots, B\}$ is replicated into three channels with the usual CLIP preprocessing:

$$I^{(b)} = [I[:,:,b], I[:,:,b], I[:,:,b]]. \tag{2}$$

This yields normalized per-band embeddings $\mathbf{v}^{(b)} = f_v(I^{(b)})/\|f_v(I^{(b)})\|_2 \in \mathbb{R}^d$. These are stacked as $V = [\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(B)}]^\top \in \mathbb{R}^{B\times d}$, so all later operations act on an image-specific set of band descriptors.

Class-Name–Guided Conditioning The image-to-image protocol does not use the query image's ground-truth class as text. Instead, this study forms one universal conditioning vector $\mathbf{c} = \frac{1}{C}\sum_{j=1}^{C} \widehat{f_t(\text{name}_j)}$ by averaging the CLIP text embeddings of the C dataset class names and then ℓ_2-renormalising. The same $\mathbf{c}$ is used for every image, so it supplies dataset-level ontology knowledge but no per-query label (a closed-vocabulary assumption, discussed in Sec. 6). It enters the adaptation objective as a weak scene-vocabulary alignment term (Eq. 5); setting its weight to zero gives the "w/o class-name conditioning" ablation (Sec. 4.2). In deployments without a known class vocabulary, coarse land-cover prompts can replace dataset names (an open-set relaxation discussed in Sec. 6), while fully prompt-free conditioning remains future work.

Uniform Weight Initialization This research initialises all band-weight logits to zero, $\ell_b^{(0)} = 0$, so that softmax yields uniform weights:

$$w_b^{(0)} = \text{softmax}(\ell^{(0)})_b = \frac{1}{B}. \tag{3}$$

This avoids data-dependent initialisation bias; CLIP's near-isotropic band geometry makes uniform weights a strong starting point.

2.3 Image-Conditioned Spectral Fusion

Given weights $\{w_b\}_{b=1}^B$, the fused descriptor is projected onto the unit hypersphere:

$$\bar{\mathbf{z}}(w) = \sum_{b=1}^{B} w_b \mathbf{v}^{(b)}, \quad \phi(I; w) = \frac{\bar{\mathbf{z}}(w)}{\|\bar{\mathbf{z}}(w)\|_2}. \tag{4}$$

2.4 Test-Time Manifold Consistency Adaptation

Starting from the uniform initialisation (Eq. 3), this study refines the weight logits with a small test-time optimisation while keeping all CLIP parameters frozen. Weights are obtained by $w = \mathrm{softmax}(\ell)$. Let $\mathcal{N}_k(\mathbf{v}^{(b)})$ denote the k nearest neighbouring band embeddings within the image. The objective has two terms — a manifold-consistency term and a class-name–guided conditioning term:

$$\mathcal{L}_{\mathrm{TTA}}(\ell) = \underbrace{\frac{\lambda_m}{Bk} \sum_{b=1}^{B} \sum_{\mathbf{v}^{(j)} \in \mathcal{N}_k(\mathbf{v}^{(b)})} \left\| \bar{\mathbf{z}}(w) - \mathbf{v}^{(j)} \right\|_2}_{\text{manifold consistency}} - \underbrace{\lambda_c \, \mathbf{c}^\top \frac{\bar{\mathbf{z}}(w)}{\|\bar{\mathbf{z}}(w)\|_2}}_{\text{class-name conditioning}} \tag{5}$$

where λ_m weights neighbourhood preservation and λ_c weights alignment to the universal prior $\mathbf{c}$ (Sec. 2.2). The first term averages L2 distances between the fused descriptor and all within-image band neighbours, keeping the descriptor near the local band manifold. The second steers spectral emphasis toward channels whose aggregate is consistent with the dataset-level vocabulary, without using per-image labels ($\lambda_c = 0$ gives the "w/o class-name conditioning" row of Table 4). During adaptation $\bar{\mathbf{z}}(w)$ is not L2-normalised in the manifold term; final normalisation is applied only after obtaining w_{opt}.

Because $\bar{\mathbf{z}}(w)$ is always a convex combination of per-band embeddings and only logits are updated, the descriptor cannot drift from the observed image bands before final normalisation; the loss changes spectral emphasis while keeping the descriptor tied to the frozen embedding geometry.

At inference, this study optimises the weight logits ℓ for 5 Adam steps while keeping f_v and f_t frozen. The final descriptor is $\phi(I) = \bar{\mathbf{z}}(w_{\mathrm{opt}})/\|\bar{\mathbf{z}}(w_{\mathrm{opt}})\|_2$. Both query and gallery images follow this same pipeline before cosine similarity is computed.

2.5 Interpretable Band Attribution

For each image, per-band contribution scores $\alpha_b = w_b \cdot (\mathbf{v}^{(b)})^\top \phi(I)$ indicate which spectral channels shape the final descriptor. For example, high α_{NIR} for agricultural query images is consistent with vegetation-sensitive NIR evidence, while high α_{SWIR} for water images is consistent with water's near-total SWIR absorption. This research uses the scores as inspection signals, not as causal explanations. Mean α_b scores for all EuroSAT classes across all 13 Sentinel-2 bands are tabulated in the supplementary material (Table S3).

2.6 Computational Complexity

The dominant cost is per-band encoding, $O(B \cdot T_{\mathrm{CLIP}})$ per image, where T_{CLIP} is one CLIP forward pass. The five manifold-consistency steps cost $O(5kBd)$, small beside the B=13 forward passes, and add 2.42 ms in the runtime profile once per-band encodings are available. In contrast, single-view baselines (RGB-CLIP, PCA-to-3, NDVI) use one CLIP pass and no optimisation; all-band fusion and the proposed method use B=13 passes, with the proposed method adding five

logit-update steps. Tip-Adapter and RS-TransCLIP use one pass plus a cache lookup or gallery-side diffusion. Since band encodings can be cached once, this is mainly an offline indexing overhead rather than a per-query cost.

3 Experiments

The evaluation asks two questions: how much all-band encoding changes frozen generic CLIP on a controlled single-label benchmark, and whether test-time spectral weighting helps in a multi-label archive where loose Recall@K can saturate. This section defines the datasets, metrics, baselines, and implementation, then reports results.

3.1 Datasets and Image-to-Image Protocol

EuroSAT [3]: 27,000 Sentinel-2 patches, 10 land-use classes, and 13 spectral bands (B1–B12 plus B8A). This study uses the multispectral version and evaluates image-to-image retrieval with 5-fold cross-validation. In each fold, 60% of images are reserved for cache-based baselines such as Tip-Adapter, 20% form the query set, and 20% form the gallery set. Query and gallery sets are disjoint. A retrieved gallery image is relevant when its class label matches the query label.

BigEarthNet-MM [4]: A large-scale multi-label Sentinel-2 archive containing over 590,000 patches across 43 scene classes; this study uses a filtered subset following the standard protocol of Neumann et al. [13]. Patches with seasonal snow or cloud/shadow cover are discarded; from the remaining pool, up to 100,000 patches are sub-sampled before drawing 1,000 query images and 10,000 gallery images across five random seeds (42, 123, 456, 789, 2024). Relevance is judged over the ten most frequent classes of the 19-class nomenclature [4]: patches with no label in this set are dropped, as are labels outside it. A gallery image is relevant to a query if their label sets intersect. All methods encode the same images through their respective pipelines; no text queries are used at retrieval time.

3.2 Evaluation Metrics

EuroSAT metrics use mean ± std over five folds. **Recall@K** ($K \in \{1, 5, 10\}$) is the fraction of queries with at least one same-class gallery image in the top-K; **mAP** averages, over queries, the average precision computed over the full gallery ranking (no top-K truncation).
BigEarthNet-MM uses multi-label metrics (mean ± std over 5 seeds). Here a gallery image is positive if its label set overlaps the query's; **Recall@K** ($K \in \{1, 5\}$) counts queries with such an image in the top-K, **ExactMatch@1** counts top-1 retrievals with an identical label set, and **MeanJaccard@K** averages $|L_q \cap L_g|/|L_q \cup L_g|$ over the top-K.

For EuroSAT ablations, this research uses two-sided paired t-tests over the five fold-level measurements and report p-values, read alongside effect sizes and cross-fold standard deviations.

Table 1. EuroSAT I2I retrieval (5-fold mean ± std). Bold: best per column.

Method	R@1 ↑	R@5 ↑	R@10 ↑	mAP ↑
RGB-CLIP	$80.28_{\pm 0.77}$	$94.90_{\pm 0.52}$	$97.59_{\pm 0.37}$	$40.42_{\pm 0.30}$
PCA-to-3-channels	$83.56_{\pm 0.14}$	$96.23_{\pm 0.24}$	$98.24_{\pm 0.10}$	$42.77_{\pm 0.15}$
NDVI composite	$83.61_{\pm 0.55}$	$96.33_{\pm 0.27}$	$98.21_{\pm 0.32}$	$45.58_{\pm 0.22}$
Tip-Adapter logits	$67.79_{\pm 0.82}$	$89.37_{\pm 0.29}$	$94.47_{\pm 0.33}$	$36.63_{\pm 0.28}$
RS-TransCLIP	$78.51_{\pm 1.00}$	$92.13_{\pm 0.57}$	$95.26_{\pm 0.46}$	$43.26_{\pm 0.31}$
All-bands avg. (uniform init)	$86.40_{\pm 0.21}$	$\mathbf{97.21}_{\pm 0.14}$	$98.64_{\pm 0.12}$	$46.44_{\pm 0.28}$
Proposed	$\mathbf{86.49}_{\pm 0.18}$	$97.19_{\pm 0.16}$	$\mathbf{98.66}_{\pm 0.15}$	$\mathbf{46.50}_{\pm 0.27}$

3.3 Baselines

Six reference points are compared against the proposed method: **RGB-CLIP** [14] (3-channel CLIP embedding), **All-bands average** (uniform-weight fusion, exactly the descriptor at initialization, Eq. 3, before any test-time update — reported as an explicit reference row in Tables 1 and 2), **PCA-to-3-channels** [15] and **NDVI composite** [16] (fixed spectral reductions), **Tip-Adapter** [8] (RGB-space cache adapter), and **RS-TransCLIP** [17] (RGB-space graph diffusion). Full per-baseline implementation details are given in the supplementary material (Sec. S2). Domain-pretrained EO VLMs are positioned in Sec. 5, since comparable result rows require same-protocol reruns.

3.4 Implementation Details

This study uses CLIP ViT-B/16 as the frozen backbone for all methods (C=10 class names for EuroSAT/BigEarthNet-MM, Sec. 2.2). Band weight logits are initialised uniformly ($\ell_b^{(0)} = 0$, giving $w_b^{(0)} = 1/B$ via softmax) and refined with manifold consistency using $k = 5$ within-image neighbours, 5 Adam steps, learning rate $\eta = 0.01$, manifold weight $\lambda_m = 0.1$, and conditioning weight $\lambda_c = 1.0$ (Eq. 5). These hyperparameters were selected on EuroSAT held-out-fold cache splits (60% per fold, disjoint from the query/gallery sets), avoiding leakage, and transferred without modification to BigEarthNet. Tip-Adapter uses $\alpha = 20.0$ and $\beta = 5.5$ (tuned on the EuroSAT training split). No method updates any CLIP parameters, and all per-fold configurations are fixed for reproducibility.

3.5 EuroSAT Evaluation

Table 1 reports EuroSAT image-to-image retrieval performance across five folds. Figure S1 (supplementary) visualizes the Recall@K columns from the same data.

Key observations. R@5 and R@10 are near-saturated for every method ($\geq$ 94%), so R@1 and mAP carry the clearest signal. The proposed method reaches R@1 86.49% ± 0.18% (+6.21 pp over RGB-CLIP, +18.70 pp over Tip-Adapter logits, +7.98 pp over RS-TransCLIP) and mAP 46.50% (vs. 40.42% for RGB-CLIP). These are the largest deltas among the adapted descriptor and single-view/cache baselines in Table 1, indicating that non-visible channels help when included in the descriptor. On the saturated R@10, the proposed method leads the all-bands-average reference by 0.02 pp (98.66% ± 0.15% vs. 98.64%) and

Table 2. BigEarthNet I2I retrieval (5-seed mean ± std). Bold: best mean per column.

Method	R@1 ↑	R@5 ↑	mAP ↑	ExactMatch@1 ↑	MeanJaccard@1 ↑	MeanJaccard@5 ↑
RGB-CLIP	$91.02_{\pm 0.96}$	$99.36_{\pm 0.09}$	$70.62_{\pm 0.64}$	$\mathbf{12.20}_{\pm 0.69}$	$\mathbf{46.20}_{\pm 0.74}$	$69.81_{\pm 0.75}$
PCA-to-3	$89.02_{\pm 1.14}$	$99.22_{\pm 0.16}$	$69.07_{\pm 0.77}$	$9.64_{\pm 1.54}$	$42.68_{\pm 1.06}$	$67.35_{\pm 0.53}$
NDVI	$88.92_{\pm 1.11}$	$99.16_{\pm 0.25}$	$69.71_{\pm 0.61}$	$10.74_{\pm 0.90}$	$43.18_{\pm 0.81}$	$67.49_{\pm 0.85}$
Tip-Adapter	$86.56_{\pm 1.04}$	$99.16_{\pm 0.24}$	$70.60_{\pm 0.67}$	$7.74_{\pm 1.11}$	$38.97_{\pm 1.23}$	$64.44_{\pm 0.83}$
RS-TransCLIP	$76.70_{\pm 1.91}$	$97.54_{\pm 1.68}$	$67.97_{\pm 0.90}$	$2.64_{\pm 1.61}$	$29.37_{\pm 1.38}$	$51.20_{\pm 2.52}$
All-bands avg.	$\mathbf{91.12}_{\pm 0.96}$	$\mathbf{99.42}_{\pm 0.22}$	$71.49_{\pm 0.61}$	$12.08_{\pm 0.85}$	$46.11_{\pm 0.92}$	$69.83_{\pm 0.86}$
Proposed	$91.06_{\pm 1.02}$	$99.38_{\pm 0.23}$	$\mathbf{71.51}_{\pm 0.60}$	$12.10_{\pm 1.05}$	$45.95_{\pm 0.99}$	$\mathbf{69.86}_{\pm 0.60}$

leads single-view baselines by 1.07 pp over RGB-CLIP; all R@10 differences are small effects at this saturation level.

On EuroSAT the dominant factor is per-band encoding, not the adaptation step: the all-bands-average reference already reaches R@1 86.40%±0.21%, statistically indistinguishable from the proposed method's 86.49% ± 0.18% (paired t-test, R@1/R@5/R@10 n.s.; mAP +0.06 pp, p=0.0007), so no substantial EuroSAT ranking gain is claimed beyond this small mAP edge. The relevant gap is the +2.79 to +18.61 pp R@1 both share over the single-view and cache baselines, confirming that encoding all bands is the key driver; the adaptation instead pays off more in the harder multi-label regime below.

3.6 BigEarthNet-MM Evaluation

Table 2 reports the BigEarthNet image-to-image retrieval performance (mean ± std over 5 seeds). Figure S2 (supplementary) visualizes the ExactMatch and MeanJaccard columns from the same data.

Key observations. Over five random seeds, all full-band methods perform consistently. The proposed adaptation edges the all-bands-average reference on mAP (71.51% vs. 71.49%, p=0.004), while R@1 (91.06% vs. 91.12%) remains within one standard deviation. RGB-CLIP alone attains the highest ExactMatch@1 (12.20%, n.s. vs. both), so no strict-match advantage is claimed here; the edge is confined to MeanJaccard@5 (69.86% vs. 69.83%/69.81%). Both full-band methods surpass fixed reductions (PCA-to-3: 89.02% R@1, NDVI: 88.92%) and cache/diffusion baselines (Tip-Adapter: 86.56%, RS-TransCLIP: 76.70%) by clear margins, at negligible cost (+0.88% wall-clock).

3.7 Runtime and Cost

Table 3 reports per-query inference time profiled over 20 EuroSAT samples (5 warm-up runs) with CLIP ViT-B/16 on Apple M4 MPS (hardware details and warm-up protocol in supplementary). The gallery is encoded offline and cached; figures cover only the query image.

Encoding dominates; adaptation is nearly free. The all-bands-average reference and the proposed method run the same 13 CLIP forward passes; the reference incurs a ×8.9 cost and the proposed method ×9.0 over single-view baselines—this overhead is band encoding, not the test-time update. Per-band encoding is 272.0 ms (99.1%) of the 274.42 ms total, while the manifold optimization adds only 2.42 ms (0.88%), negligible beside the $O(B \cdot T_{\text{CLIP}})$ encoding

Table 3. Per-query inference time (CLIP ViT-B/16, Apple M4 MPS; gallery pre-cached).

Method	**CLIP fwds** ↓	**MPS median** (ms) ↓	**Overhead** vs RGB-CLIP ↓
RGB-CLIP	1	30.6	×1.0 (ref.)
PCA-to-3-channels	1	33.1	×1.1
NDVI composite	1	31.9	×1.0
Tip-Adapter logits	1	26.7	×0.9
RS-TransCLIP	1	27.2	×0.9
All-bands avg.	13	272.1	×8.9
Proposed	13	274.7	×9.0

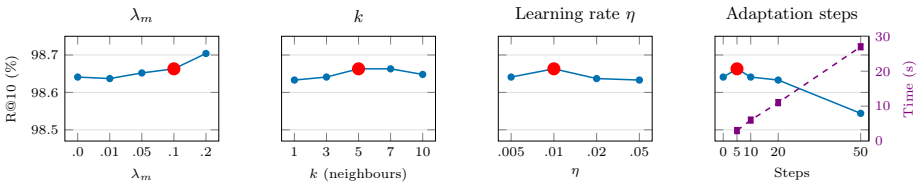

Fig. 2. Hyperparameter sensitivity on EuroSAT R@10. Red markers: default setting. Dashed violet (right axis): total adaptation time per fold (5,400 queries).

(Sec. 2.6). Thus on BigEarthNet the test-time adaptation performs comparably to the all-bands-average reference (R@1 91.06% vs 91.12%; mAP 71.51% vs 71.49%) for only +0.88% additional wall-clock time, while the shared band-encoding cost is the main practical trade-off.

4 Sensitivity and Ablation

This research checks hyperparameter sensitivity and component removal on the same EuroSAT folds as Sec. 3.

4.1 Hyperparameter Sensitivity

Figure 2 summarises one-factor sweeps over λ_m, neighbourhood size k, learning rate η, and adaptation steps, averaged across the same five EuroSAT folds as Sec. 3 with λ_c=1.0 held fixed. Across all tested ranges, R@10 varies by less than 0.2 pp and R@1 by at most 0.59 pp, confirming that the reported defaults are not a narrow optimum. The adaptation-step curve is well-behaved within 0–20 steps; the effect of λ_c is isolated in the ablation below.

4.2 Ablation Study

Table 4 isolates the contribution of each component on EuroSAT.

Key findings. Class-name conditioning is the dominant contributor: removing it lowers R@10 by 0.06 pp (p<0.05) and mAP by 0.24 pp (p<0.001), with a

Table 4. Component ablation, EuroSAT (5-fold mean ± std; p-values: paired t-test vs. full model).

Configuration	R@1 ↑	R@10 ↑	mAP ↑
Proposed	**86.49**$_{\pm 0.18}$	**98.66**$_{\pm 0.15}$	**46.50**$_{\pm 0.27}$
w/o Test-time adaptation[a]	86.40$_{\pm 0.25}$ (n.s.)	98.64$_{\pm 0.13}$ (n.s.)	46.44$_{\pm 0.28}$ (p<0.001)
w/o Class-name conditioning	86.19$_{\pm 0.30}$ (n.s.)	98.60$_{\pm 0.13}$ (p<0.05)	46.26$_{\pm 0.29}$ (p<0.001)

[a] `num_steps` = 0, i.e. the all-bands-average configuration of Table 1, re-run inside the ablation pipeline (means agree to two decimals; the small std difference reflects the separate run).

consistent 0.30 pp R@1 trend. The "w/o test-time adaptation" row (no optimisation; identical to the all-bands-average reference in Table 1) shows the combined cost of removing both terms: mAP drops 0.06 pp (p<0.001). The isolated effect of the manifold term is small: within the sensitivity sweep (Fig. 2, whose default point is R@1 86.56%/mAP 46.52%), setting λ_m=0 while retaining conditioning and 5 steps gives R@1 86.56% and mAP 46.54% — indistinguishable from that sweep's default — indicating that the mAP gap vs. the no-optimisation baseline is driven primarily by class-name conditioning. The manifold term acts as a regulariser that preserves inter-band local structure without adding a measurable ranking gain on its own.

5 Related Work

Prior work differs mainly in where adaptation occurs: the encoder, an inference cache, the spectral input representation, or, in this work, only the test-time fusion weights.

CLIP adaptation and EO VLMs. Tip-Adapter [8] and related adapters change inference around frozen RGB-compatible features or logits [18,19]. Remote-sensing VLMs such as RS-CLIP, RemoteCLIP, GeoRSCLIP, DOFA-CLIP, and Beyond the Visible adapt or pretrain the representation itself on large Earth-observation corpora [5,7,9,10,11]. They are complementary to this setting: they change the encoder, whereas the scope of this research is the *generic, frozen RGB-pretrained CLIP* regime, without domain-pretrained weights, paired captions, or dataset-scale training. The reported numbers are therefore not positioned as state-of-the-art over domain-pretrained EO VLMs; a same-protocol comparison remains open.

Spectral fusion. Subspace, attentive, PCA, and index-based methods fuse channels before or after feature extraction [12,15,16,20], while multispectral retrieval studies show that compression affects ranking quality [2]. The proposed method differs by learning per-image band weights at test time via manifold consistency, without modifying the encoder or requiring task-specific training data.

6 Conclusion

This research presented a parameter-frozen test-time adaptation for image-to-image multispectral retrieval that refines only B band-weight logits of a generic RGB-pretrained CLIP and reports band attribution without touching the encoder. Across both EuroSAT (5-fold cross-validation) and BigEarthNet-MM (5-seed evaluation), encoding all bands achieves clear performance gains over single-view/cache baselines (reaching R@1 $\sim$ 86.5% on EuroSAT and $\sim$ 91.1% on BigEarthNet). Test-time adaptation matches the retrieval quality of the strong all-bands-average reference within one standard deviation on both benchmarks. The value of the proposed approach is thus a favourable trade-off rather than an unconstrained ranking win: accuracy competitive with strong all-band baselines, per-band interpretability/attribution, and a near-zero test-time adaptation cost (+0.88% wall-clock overhead over band encoding).

Two limitations scope the contribution. Runtime is dominated by band encoding ($\times$9.0 per query, Table 3), shared with the all-bands-average reference; the optimization itself adds only +0.88% and encodings are cached offline. The class-name–guided prior is label-free per query yet assumes the dataset's class vocabulary (a closed-set assumption); the λ_c=0 ablation (Table 4) shows graceful degradation without it (R@1 86.19%, n.s.), with conditioning's benefit concentrated in mAP (p<0.001); coarse land-cover prompts offer an open-set relaxation. Open questions include same-protocol comparisons against domain-pretrained EO VLMs, reducing the $O(B \cdot T_{\text{CLIP}})$ cost, prompt-free conditioning, and robustness under atmospheric corruption and mixed land cover.

Acknowledgments. The authors acknowledge Ho Chi Minh City University of Technology (HCMUT), VNUHCM for supporting this study.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Song, W., Gao, Z., Dian, R., Ghamisi, P., Zhang, Y., Benediktsson, J.A.: Asymmetric hash code learning for remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. **60**, 1–14 (2022).
2. Blumenstiel, B., Moor, V., Kienzler, R., Brunschwiler, T.: Multi-spectral remote sensing image retrieval using geospatial foundation models. In: Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS). pp. 7286–7291 (2024).
3. Helber, P., Bischke, B., Dengel, A., Borth, D.: EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **12**(7), 2217–2226 (2019).
4. Sumbul, G., de Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., Markl, V.: BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval. IEEE Geosci. Remote Sens. Mag. **9**(3), 174–180 (2021).
5. Xiong, Z., Wang, Y., Yu, W., Stewart, A.J., Zhao, J., Lehmann, N., Dujardin, T., Yuan, Z., Ghamisi, P., Zhu, X.X.: DOFA-CLIP: Multimodal vision-language foundation models for earth observation. arXiv:2503.06312 (2025).

6. Ulku, I., Tanriover, O.O., Akagunduz, E.: LoRA-NIR: Low-rank adaptation of vision transformers for remote sensing with near-infrared imagery. IEEE Geosci. Remote Sens. Lett. **21**, 1–5 (2024).
7. Marimo, C.T., Blumenstiel, B., Nitsche, M., Jakubik, J., Brunschwiler, T.: Beyond the visible: Multispectral vision-language learning for earth observation. In: Proc. ECML PKDD. LNCS, vol. 16018, pp. 359–375. Springer (2025).
8. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In: Proc. Eur. Conf. Comput. Vis. (ECCV). LNCS, vol. 13695, pp. 493–510 (2022).
9. Li, X., Wen, C., Hu, Y., Zhou, N.: RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision. Int. J. Appl. Earth Obs. Geoinf. **124**, 103497 (2023).
10. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: RemoteCLIP: A vision language foundation model for remote sensing. IEEE Trans. Geosci. Remote Sens. **62**, 1–16 (2024).
11. Zhang, Z., Zhao, T., Guo, Y., Yin, J.: RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. IEEE Trans. Geosci. Remote Sens. **62**, 1–23 (2024).
12. Fang, Q., Wang, Z.: Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. Pattern Recognit. **130**, 108786 (2022).
13. Neumann, M., Pinto, A.S., Zhai, X., Houlsby, N.: In-domain representation learning for remote sensing. arXiv preprint arXiv:1911.06721 (2019).
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Machine Learning (ICML). PMLR, vol. 139, pp. 8748–8763 (2021).
15. Karpowicz, F., Kepinski, W., Staszynski, B., Sarwas, G.: Examination of PCA utilisation for multilabel classifier of multispectral images. Computer Science Research Notes **3501**(1), 247–255 (2025).
16. Stratoulias, D., Nuthammachot, N., Suepa, T., Phoungthong, K.: Assessing the spectral information of sentinel-1 and sentinel-2 satellites for above-ground biomass retrieval of a tropical forest. ISPRS Int. J. Geo-Inf. **11**(3), 199 (2022).
17. El Khoury, K., Zanella, M., Gerin, B., Godelaine, T., Macq, B., Mahmoudi, S., De Vleeschouwer, C., Ben Ayed, I.: Enhancing remote sensing vision-language models for zero-shot scene classification. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP). pp. 1–5 (2025).
18. Bendou, Y., Ouasfi, A., Gripon, V., Boukhayma, A.: ProKeR: A kernel perspective on few-shot adaptation of large vision-language models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 25092–25102 (2025).
19. Barbier, C., Abeloos, B., Herbin, S.: Bridging the modality gap: Training-free adaptation of vision-language models for remote sensing via visual prototypes. In: Proc. IEEE/CVF CVPR Workshops. pp. 3082–3091 (2025).
20. Ramirez, J., Vargas, H., Martinez, J.I., Arguello, H.: Subspace-based feature fusion from hyperspectral and multispectral image for land cover classification. In: Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS). pp. 3003–3006 (2021).