Supplementary Material for 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`

This document provides supplementary material supporting the camera-ready manuscript "Test-Time Spectral Adaptation for Image-to-Image Multispectral Retrieval" (Paper #71). It includes details on code and data availability, a consolidated mathematical notation summary, graphical visualizations of retrieval performance with detailed analysis, and computational profiling details. Section and equation references match those of the main paper.

S1 Code, Models, and Data Availability

All source code, dataset splitting scripts, evaluation protocols, and pre-extracted feature caches required to reproduce the results in the main paper are publicly available at:

https://github.com/tienesng06/acivs_supplementary.git

Repository Structure and Reproducibility Mapping:

- `src/`: core implementation (retrieval pipeline, manifold-consistency test-time optimization, band encoding) and dataset loaders/5-fold splits for EuroSAT and BigEarthNet-MM.
- `scripts/`: experiment entry points for Tables 1–4 and Figure 2 — all-bands baseline, component ablation, multi-seed BigEarthNet, hyperparameter sensitivity, and runtime profiling.
- `result_for_paper/`: archived CSV outputs and manifests for every table; `checkpoints/`: CLIP ViT-B/16 weights.

S2 Baseline Implementation Details

Full mechanistic descriptions of the six reference points introduced in Sec. 3.3 of the main paper:

- **RGB-CLIP** (main paper ref. [14]): the Sentinel-2 RGB bands (B4, B3, B2) are extracted and encoded with frozen CLIP for query/gallery comparison.

⋆ Corresponding author

Table S1. Consolidated notation used throughout the main paper.

Symbol	Meaning and Operational Scope
$I \in \mathbb{R}^{H\times W\times B}$	Multispectral input patch with $B = 13$ Sentinel-2 bands
$d = 512$	CLIP ViT-B/16 visual/text feature embedding dimension
C	Number of scene class names ($C = 10$ EuroSAT, $C = 10$ BigEarthNet)
$f_v(\cdot),\ f_t(\cdot)$	Frozen CLIP vision and text encoder functions
$I^{(b)} \in \mathbb{R}^{H\times W\times 3}$	Band b replicated across 3 channels as a pseudo-RGB image
$\mathbf{v}^{(b)} \in \mathbb{R}^d$	ℓ_2-normalized CLIP embedding of band b
$V \in \mathbb{R}^{B\times d}$	Stack of B normalized per-band embeddings for an image
$\mathbf{c} \in \mathbb{R}^d$	Universal dataset-level class-name prior embedding
$\ell_b \in \mathbb{R}$	Learnable band-weight logit for band b
$w_b \in (0, 1)$	Softmax band weight, $w = \mathrm{softmax}(\ell)$, $\sum_b w_b = 1$
$\bar{\mathbf{z}}(w) \in \mathbb{R}^d$	Weighted band aggregate vector $\sum_{b=1}^{B} w_b \mathbf{v}^{(b)}$
$\phi(I) \in \mathbb{R}^d$	Final ℓ_2-normalized image descriptor, $\bar{\mathbf{z}}(w_{\mathrm{opt}})/\lVert\bar{\mathbf{z}}(w_{\mathrm{opt}})\rVert_2$
$\mathcal{N}_k(\mathbf{v}^{(b)})$	$k = 5$ nearest within-image neighbour embeddings of band b
$\lambda_m = 0.1,\ \lambda_c = 1.0$	Loss weights for manifold consistency and text prior conditioning
$\alpha_b \in \mathbb{R}$	Per-band attribution score, $w_b \cdot (\mathbf{v}^{(b)})^\top \phi(I)$
$\mathcal{Q},\ \mathcal{G}$	Disjoint query set and gallery set

- **All-bands average**: average of all normalized band embeddings with uniform weights — exactly the descriptor at the uniform initialization (Eq. 3 of the main paper) before any test-time update, isolating the value added by adaptation alone.
- **PCA-to-3-channels** (main paper ref. [15]): each image is reshaped into spectral pixel vectors, the 13 bands are reduced to three principal components, and the resulting pseudo-RGB image is encoded with frozen CLIP.
- **NDVI composite** (main paper ref. [16]): a three-channel composite is built from NDVI, green, and red channels and encoded with frozen CLIP, testing a physics-informed but fixed spectral transformation.
- **Tip-Adapter** (main paper ref. [8]): a key-value cache is built from the fold's training split (RGB CLIP features, one-hot labels); each query/gallery image maps to a normalized class-logit vector, compared via image-to-image similarity.
- **RS-TransCLIP** (main paper ref. [17]): starting from RGB-CLIP image-to-image similarity, scores are refined with gallery-side patch-affinity diffusion.

S3 Mathematical Notation Summary

Table S1 consolidates the mathematical symbols, dimensions, and operational definitions used throughout the formulation (Section 2) and evaluation (Section 3) of the main paper.

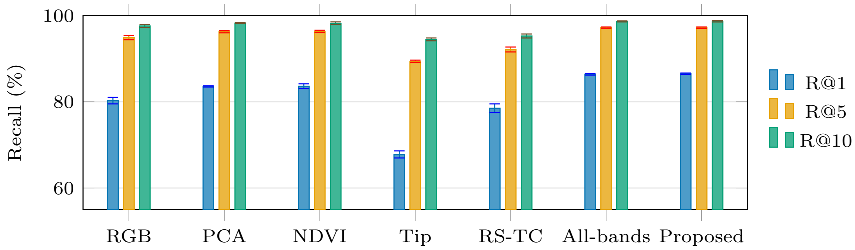

Fig. S1. EuroSAT image-to-image retrieval performance across methods (5-fold mean $\pm$ std). R@1 shows the strongest discriminative separation (+6.21 pp over RGB-CLIP), while R@5 and R@10 saturate near 94–98% across methods.

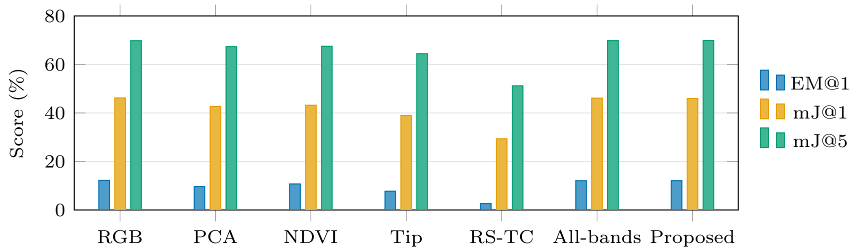

Fig. S2. BigEarthNet-MM strict set-overlap quality metrics across methods (5-seed mean). ExactMatch@1 (EM@1) measures exact label-set identity; MeanJaccard@1/5 (mJ@1, mJ@5) measures Jaccard similarity $|L_q \cap L_g|/|L_q \cup L_g|$.

Relation to Main Paper Formulation: As established in Section 2.1, retrieval ranks gallery items by $s(I_q, I_g) = \phi(I_q)^\top \phi(I_g)$. The descriptor $\phi(I)$ is constructed from V via test-time adaptation of ℓ without changing f_v or f_t.

S4 Retrieval Performance Visualizations

Figures S1 and S2 provide graphical visualizations of the multi-baseline retrieval benchmarks presented in Tables 1 and 2 of Section 3 in the main paper.

EuroSAT Analysis (Figure S1): Visualizing Recall@K highlights two structural phenomena in single-label Sentinel-2 retrieval:

1. **Discriminative Signal Concentrates at Top-1:** Both Recall@5 and Recall@10 saturate above 94% for nearly all methods, because EuroSAT has clear land-cover class boundaries. Recall@1 is therefore the primary discriminative benchmark.
2. **Full-Band Advantage:** Exposing the frozen CLIP encoder to all 13 spectral channels (all-bands average 86.40%, proposed 86.49%) lifts R@1 by +6.12 and +6.21 pp over 3-channel RGB-CLIP at 80.28%, and by +18.61 and +18.70 pp over the cache-based Tip-Adapter at 67.79%.

Table S2. Hardware environment and profiling breakdown per query image.

Parameter / Stage	**Specification / Measured Value**
System Processor	Apple M4 SoC (10 CPU cores: 4 perf. + 6 efficiency)
Graphics Hardware	Integrated 8-core GPU, 16 GB unified memory
OS & Execution Environment	macOS 26.2, Darwin 25.2.0, PyTorch MPS backend
Backbone Architecture	Frozen CLIP ViT-B/16 ($d = 512$)
Per-Band CLIP Forward Passes ($B = 13$)	272.0 ms (99.12% of total query time)
Manifold Adaptation (5 Adam steps)	2.42 ms (0.88% of total query time)
Total (sum of stage medians)	**274.42 ms**; end-to-end median 274.73 ms (272.06 ms for All-bands avg.)

BigEarthNet-MM Analysis (Figure S2): In multi-label retrieval, loose Recall@K (which marks any single-label overlap as positive) saturates near 99%. The strict metrics in Figure S2 reveal key differences:

1. **Strict Top-1 Quality:** RGB-CLIP attains the highest ExactMatch@1 at 12.20%, ahead of the proposed adaptation (12.10%) and the all-bands baseline (12.08%). None of these gaps is significant (two-sided paired t-test, $p > 0.8$). The proposed method retains the best MeanJaccard@5 at 69.86%, against 69.83% for the all-bands average and 69.81% for RGB-CLIP, but this margin is also small and not significant ($p = 0.83$).
2. **Failure of Fixed Reductions and Graph Diffusion:** Fixed spectral reductions lose critical non-visible cues (PCA-to-3 reaches 9.64% ExactMatch@1 and NDVI 10.74%), while post-processing graph diffusion degrades the feature geometry badly under multi-label ambiguity (RS-TransCLIP, 2.64%).

S5 Hardware Specification and Profiling Details

To complement Section 3.7 of the main paper, Table S2 provides an expanded breakdown of per-query execution latency and hardware parameters.
Profiling Protocol: Timings are medians over 20 random EuroSAT query images following 5 warm-up executions to ensure GPU frequency stabilization. The gallery embedding index is pre-computed offline and cached; timings strictly measure single-query inference overhead.

S6 Band Attribution Analysis

For each query image the method produces per-band contribution scores $\alpha_b = w_b \cdot (\mathbf{v}^{(b)})^\top \phi(I)$ (Sec. 2.5 of the main paper), where w_b is the learned softmax weight and $\phi(I)$ is the final L2-normalised descriptor. Table S3 reports the mean α_b over 200 randomly sampled EuroSAT query images per class (fold 0); bold entries mark the highest-scoring band in each row.

Key patterns. B08 (NIR) is the dominant band for 8 of 10 classes, consistent with near-infrared reflectance being the primary discriminator of vegetation structure (Forest, Pasture), built-up surfaces (Industrial, Residential, Highway), and mixed land cover (PermanentCrop, River). The two exceptions are AnnualCrop and SeaLake, both top-scored by B09 (Water Vapour): for AnnualCrop this reflects moisture-sensitive crop signatures; for SeaLake it co-occurs with elevated

Table S3. Mean per-band attribution score α_b per EuroSAT class (200 samples, fold 0). Bold: highest scoring band per class. Band abbreviations follow Sentinel-2 convention: VIS = B02–B04 (Blue/Green/Red), RE = B05–B07 (Red-edge), NIR = B08/B8A, WV = B09 (Water vapour), SWIR-c = B10, SWIR = B11–B12.

Class	**B01**	**B02**	**B03**	**B04**	**B05**	**B06**	**B07**	**B08**	**B8A**	**B09**	**B10**	**B11**	**B12**
AnnualCrop	0.34	0.13	0.13	0.19	0.15	0.21	0.26	0.34	0.41	**0.88**	0.20	0.15	0.28
Forest	0.34	0.24	0.12	0.11	0.26	0.14	0.24	**0.92**	0.28	0.45	0.14	0.18	0.30
HerbaceousVegetation	0.25	0.23	0.31	0.46	0.26	0.27	0.25	**0.72**	0.25	0.69	0.19	0.22	0.24
Highway	0.39	0.25	0.29	0.43	0.22	0.12	0.14	**0.91**	0.35	0.31	0.15	0.18	0.15
Industrial	0.29	0.63	0.70	0.75	0.15	0.12	0.07	**0.94**	0.29	0.22	0.16	0.16	0.04
Pasture	0.38	0.22	0.13	0.10	0.19	0.16	0.29	**0.95**	0.27	0.23	0.10	0.15	0.33
PermanentCrop	0.37	0.32	0.46	0.70	0.21	0.27	0.25	**0.75**	0.38	0.38	0.21	0.19	0.24
Residential	0.42	0.45	0.53	0.75	0.34	0.20	0.14	**0.92**	0.42	0.15	0.28	0.24	0.13
River	0.44	0.21	0.20	0.25	0.20	0.14	0.21	**0.86**	0.29	0.41	0.12	0.17	0.24
SeaLake	0.48	0.61	0.56	0.34	0.29	0.18	0.19	0.20	0.65	**0.85**	0.60	0.62	0.20

SWIR-cirrus (B10 = 0.60) and SWIR-1 (B11 = 0.62) absorption features characteristic of open water bodies. These patterns are consistent with established Sentinel-2 spectral indices (NDVI, MNDWI) and confirm that the learned weights recover physically meaningful spectral priorities without any domain pretraining.