AVION: Aerial Vision–Language Instruction from Offline Teacher to Prompt-Tuned Network

Supplementary Material

This appendix is organized as follows:

- Section A reports additional quantitative results, including few-shot classification and base-to-novel generalization across six remote sensing benchmarks.
- Section B summarizes the symbols and the default or searched hyperparameter ranges used in AVION.
- Section C presents examples of our *LLM-based domain prompting* module for fine-grained remote sensing classes, comparing simple templates with richer LLM-generated descriptions.
- Section D details the selective prototype aggregation procedure, including the `RS-Flag` rules, MAD-based robust pruning, and pseudo-code for constructing teacher text prototypes.
- Section E describes the masking and renormalization strategy for the logit distillation loss $\mathcal{L}_{\text{logit}}$ in base-to-novel training.
- Section F analyzes the parameter count, FLOPs overhead, runtime, and memory footprint of AVION on top of a frozen GeoRSCLIP backbone.
- Section G presents a hyperparameter sensitivity analysis and our unified selection protocol shared across datasets and tasks.
- Section H provides a qualitative analysis of `RS-Flag`, visualizing how the calibration down-weights RS-agnostic or ground-level descriptions to mitigate hallucinations.
- Section I reports cross-dataset zero-shot transfer results to assess generalization and source-domain overfitting.
- Section J ablates the choice of offline LLM generators for domain prompting, demonstrating that AVION is robust to the specific LLM used.
- Section K extends the evaluation to broader general-domain benchmarks (ImageNet) to study scalability beyond remote sensing.

A. Additional Quantitative Results

We first provide additional few-shot classification results on six remote sensing datasets. Fig. 7 shows the K-shot performance of AVION and competing methods under $K \in \{1, 2, 4, 8, 16\}$ across **AID, EuroSAT, RESISC-45, UCMerced, WHU-RS19, and PatternNet**. As K increases, all methods exhibit smooth performance gains, and AVION remains competitive or superior across datasets, with particularly notable improvements on EuroSAT and UCMerced at higher shot counts. For completeness, Table 7 reports the per-dataset top-1 accuracies for each K. Overall, AVION performs on par with or better than strong prompt-learning and parameter-efficient adaptation baselines across diverse remote sensing scenes.

We further complement these results with base-to-novel generalization experiments on the same six datasets. Table 8 summarizes the detailed Base, Novel, and harmonic-mean (HM) performance for the ViT-B/32 student model, including the difference Δ between **Ours** and the strongest competing baseline for each metric. Fig. 8 provides a radar-chart view of the HM values across the six datasets, offering a compact comparison of overall base-to-novel generalization. AVION achieves the best HM on all six datasets, improving upon the strongest baseline by $+1.08$ to $+3.16$ points while maintaining competitive Base and Novel accuracies. On some datasets (e.g., AID, EuroSAT), AVION trades a small decrease in Novel accuracy for a larger gain on Base and HM, while on the other datasets both Base and Novel improve.

B. Symbols and Default Hyperparameters

To facilitate clarity and reproducibility, Table 9 summarizes the symbols used throughout the paper, together with their definitions and the default values or hyperparameter search ranges adopted in our experiments. The table covers model components (encoders and prompts), loss-function terms (weights and temperatures), and data- or training-related parameters.

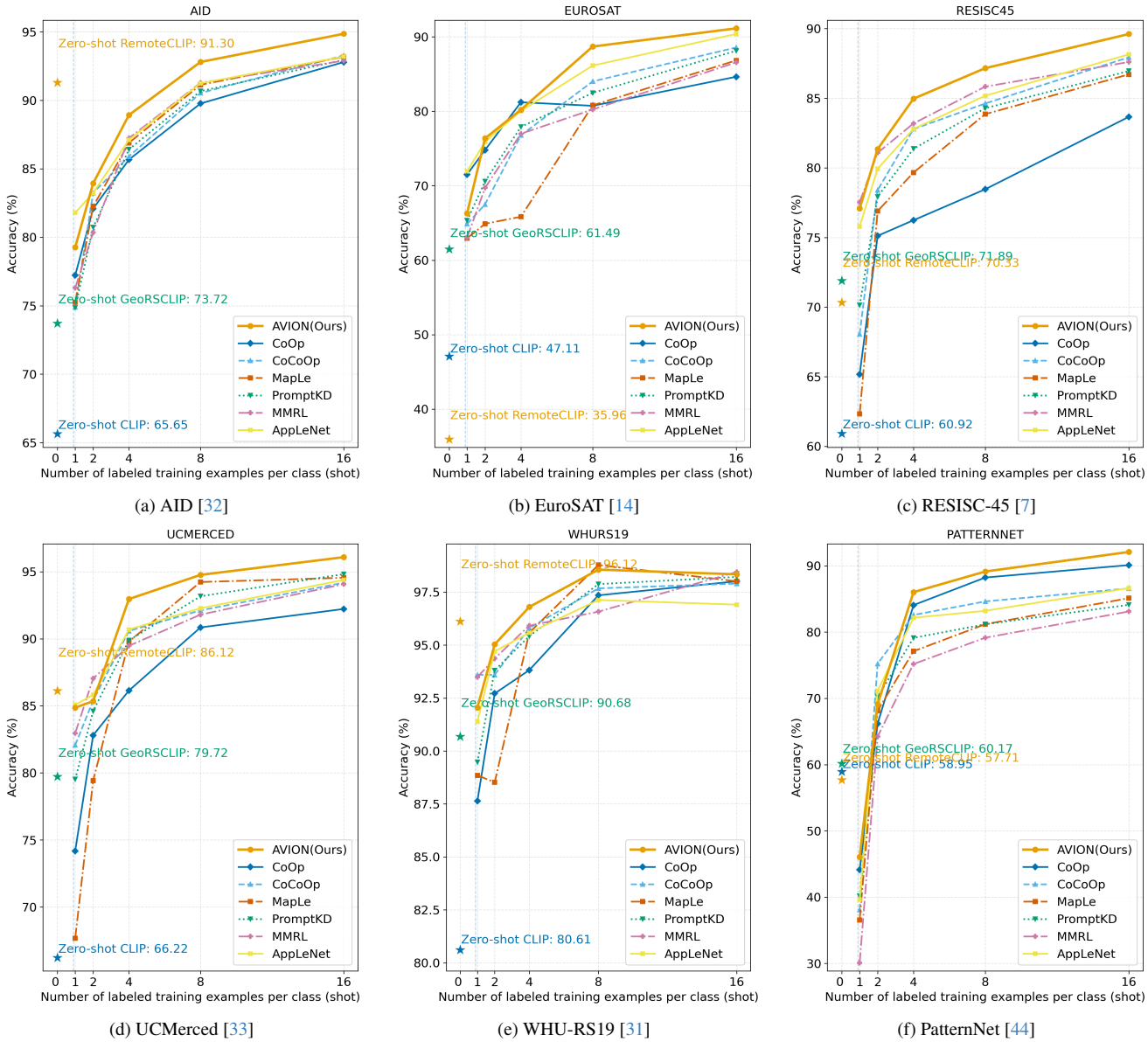
C. Examples of *LLM-based Domain Prompting*

To provide a concrete illustration of our *LLM-based domain prompting* module, this section presents example textual candidates generated for several fine-grained remote sensing classes.

As described in Sec. 3.2, our goal is to mitigate the *semantic poverty* of simple templates (e.g., "a photo of [CLASS]") commonly used in prior work. Instead of relying on a single generic template, we prompt an LLM to generate multiple candidates that are both semantically richer and explicitly aware of the remote sensing imagery domain. Following the query design illustrated in Fig. 3, we use the following generic prompt template for each class `[CLASS]`:

> *"Generate N overhead-view descriptions of `[CLASS]` from satellite imagery, highlighting class-specific scene elements while avoiding any ground-level terms."*

This constraint encourages candidates that are appropriate



(a) AID [32] (b) EuroSAT [14] (c) RESISC-45 [7]

(d) UCMerced [33] (e) WHU-RS19 [31] (f) PatternNet [44]

Figure 7. Few-shot classification performance (top-1 accuracy, %) across six remote sensing datasets (AID [32], EuroSAT [14], RESISC-45 [7], UCMerced [33], WHU-RS19 [31], and PatternNet [44]) for different numbers of shots $K \in \{1, 2, 4, 8, 16\}$.

for overhead perspectives and contain fine-grained, domain-specific details.

Table 10 compares standard simple templates with the candidates produced by our *LLM-based domain prompting* process. The richer descriptions provide a stronger semantic foundation for the teacher model's prototypes, enabling our framework to better distinguish between challenging, fine-grained categories such as "industrial area" and "commercial area".

D. Details of *Selective Prototype Aggregation*

This section provides a detailed definition of our RS-Flag prior and a step-by-step pseudo-code for the *selective prototype aggregation* process introduced in Sec. 3.2.

RS-Flag rules. The RS-Flag indicator $\text{RS-Flag}_{k,j} \in \{0, 1\}$ is set to 1 if a generated caption meets all of the following criteria, and 0 otherwise:

1. **Token constraints:** The caption must contain at least one RS-positive token and no RS-negative tokens. We use the following case-insensitive, word-boundary–matched

Table 7. Few-shot classification (top-1, %) on six remote sensing datasets. Winners are **bold**; runners-up are underlined.

Dataset	Method	1	2	4	8	16
AID	CoCoOp	74.91	83.23	85.90	90.53	93.28
	CoOp	77.22	82.09	85.67	89.77	92.79
	MMRL	76.30	80.35	87.23	91.28	92.89
	MaPLe	75.24	82.26	86.89	91.13	93.18
	PromptKD	74.85	80.72	86.41	90.67	92.95
	APPLeNet	**81.79**	83.24	87.11	91.24	93.20
	Ours	79.27	**83.94**	**88.92**	**92.80**	**94.86**
EuroSAT	CoCoOp	64.87	67.49	76.78	84.01	88.57
	CoOp	71.51	74.80	**81.21**	80.74	84.63
	MMRL	63.05	69.72	76.99	80.23	86.56
	MaPLe	62.95	64.89	65.84	80.80	86.83
	PromptKD	65.32	70.58	77.91	82.47	88.12
	APPLeNet	**71.94**	75.89	80.04	86.14	90.38
	Ours	66.29	**76.36**	80.21	**88.69**	**91.14**
RESISC-45	CoCoOp	68.06	78.40	82.78	84.62	87.94
	CoOp	65.18	75.12	76.24	78.46	83.66
	MMRL	**77.53**	81.08	83.19	85.83	87.60
	MaPLe	62.32	76.90	79.66	83.86	86.71
	PromptKD	70.14	77.92	81.37	84.28	86.95
	APPLeNet	75.79	79.94	82.79	85.18	88.15
	Ours	77.08	**81.35**	**84.96**	**87.16**	**89.61**
UCMerced	CoCoOp	82.07	85.40	90.58	92.12	94.18
	CoOp	74.18	82.80	86.14	90.85	92.22
	MMRL	82.96	**87.04**	89.47	91.80	94.07
	MaPLe	67.67	79.42	89.79	94.23	94.55
	PromptKD	79.52	84.61	89.88	93.17	94.82
	APPLeNet	**85.08**	85.82	90.69	92.28	94.39
	Ours	84.87	85.34	**92.96**	**94.76**	**96.09**
WHU–RS19	CoCoOp	**93.60**	93.60	95.81	97.68	97.90
	CoOp	87.64	92.72	93.82	97.35	98.01
	MMRL	93.49	94.37	95.92	96.58	**98.45**
	MaPLe	88.85	88.52	95.59	**98.79**	98.01
	PromptKD	89.47	93.81	95.42	97.88	98.22
	APPLeNet	91.39	94.70	95.59	97.13	96.91
	Ours	92.05	**95.03**	**96.80**	98.56	98.34
PatternNet	CoCoOp	38.09	**75.23**	82.58	84.63	86.59
	CoOp	44.13	66.18	84.07	88.23	90.11
	MMRL	30.09	64.23	75.18	79.15	83.11
	MaPLe	36.57	68.14	77.11	81.19	85.12
	PromptKD	40.16	70.19	79.12	81.18	84.11
	APPLeNet	39.62	71.14	82.16	83.21	86.63
	Ours	**46.07**	69.11	**86.03**	**89.14**	**92.09**

lists:

- **Positive:** {overhead, aerial view, satellite imagery, nadir, orthorectified, multispectral, SAR}.
- **Negative:** {street, indoor, selfie, portrait, close-up, ground level}.

2. **Length constraints:** The caption length must be between 6 and 20 words (whitespace-delimited).
3. **Content cues (optional):** The caption may optionally include class-specific cues (e.g., geometry- or infrastructure-related terms), which are captured implicitly by the LLM and the RS-Flag token lists above.

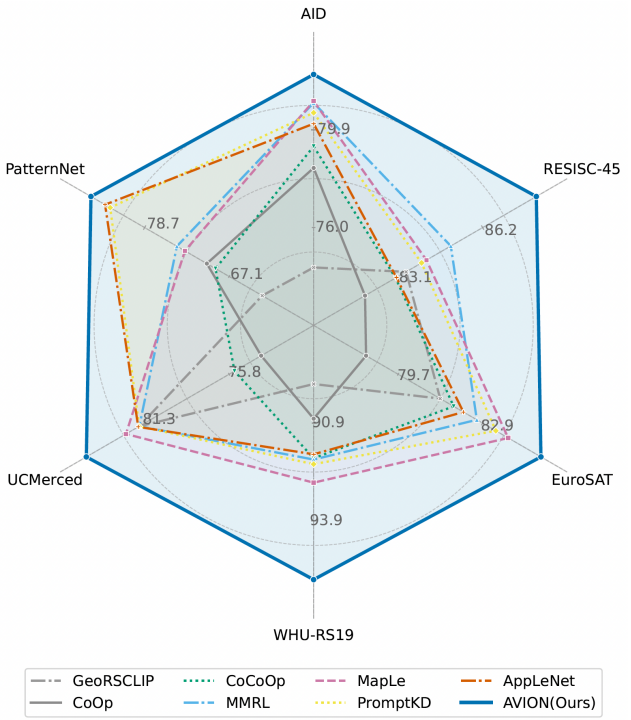


Figure 8. Base-to-novel generalization (HM, %) across six remote sensing datasets (ViT-B/32 student). The radar chart plots the harmonic mean between base and novel accuracies for AVION and competing baselines; a larger enclosed area indicates better overall base-to-novel performance.

Input: teacher features $\{\mathbf{v}_i^T\}$, caption embeddings $\{\mathbf{t}_{k,j}\}$; hyperparameters $(\zeta_s, \beta, \gamma, \varepsilon)$.

For each class k**:**

1) $\mathcal{B}_k \leftarrow \{i \mid y_i{=}k\}$; compute scores $s_{k,j} \leftarrow \frac{1}{|\mathcal{B}_k|}\sum_{i\in\mathcal{B}_k}(\mathbf{v}_i^T)^\top \mathbf{t}_{k,j}$.

2) $m_k \leftarrow \text{median}(\{s_{k,j}\}_j)$; $\Delta_{k,j} \leftarrow |s_{k,j} - m_k|$; $\text{MAD}_k \leftarrow \text{median}(\{\Delta_{k,j}\}_j)$; $z_{k,j} \leftarrow \Delta_{k,j}/(\text{MAD}_k + \varepsilon)$.

3) $\mathcal{J}_k \leftarrow \{j \mid z_{k,j} \le \zeta_s\}$ *// robust pruning of outlier captions*

4) Compute weights $w_{k,j}$ for $j \in \mathcal{J}_k$ using Eq. 3, where the RS-Flag prior is added as a calibration term to the similarity scores before softmax normalization.

5) $\mathbf{t}_k^{T*} \leftarrow \dfrac{\sum_{j\in\mathcal{J}_k} w_{k,j}\mathbf{t}_{k,j}}{\left\|\sum_{j\in\mathcal{J}_k} w_{k,j}\mathbf{t}_{k,j}\right\|_2}$ *// ℓ_2-normalization*

Output: $\{\mathbf{t}_k^{T*}\}_{k=1}^C$.

Selective prototype aggregation and robust pruning. The *selective prototype aggregation* algorithm (shown in the box above) is performed offline during training. Its core steps involve scoring caption candidates (Step 1), performing robust pruning (Steps 2-3), and then aggregating the remaining

Table 8. Detailed base-to-novel results (%) on six remote sensing datasets (ViT-B/32 student). Baselines: GeoRSCLIP [41], CoOp [43], CoCoOp [42], MMRL [13], MaPLe [20], PromptKD [24], and APPLeNet [28]. We report Base, Novel, and HM; Δ denotes the difference between **Ours** and the best-performing baseline for each metric.

AID [32]	Base	Novel	HM
GeoRSCLIP [41]	71.96	**76.90**	74.35
CoOp [43]	95.53	66.41	78.35
CoCoOp [42]	94.83	68.03	79.22
MMRL [13]	95.27	70.40	80.97
MaPLe [20]	95.10	70.60	81.04
PromptKD [24]	94.90	70.00	80.57
APPLeNet [28]	94.60	69.50	80.13
Ours	**95.80**	71.85	**82.11**
Δ	+0.27	-5.05	+1.08

(a) AID.

RESISC-45 [7]	Base	Novel	HM
GeoRSCLIP [41]	85.76	81.17	83.40
CoOp [43]	90.19	74.97	81.88
CoCoOp [42]	92.97	74.97	83.01
MMRL [13]	**94.10**	77.58	85.05
MaPLe [20]	93.50	76.50	84.15
PromptKD [24]	93.80	76.00	83.97
APPLeNet [28]	93.00	75.00	83.04
Ours	93.90	**83.17**	**88.21**
Δ	-0.20	+2.00	+3.16

(b) RESISC-45.

EuroSAT [14]	Base	Novel	HM
GeoRSCLIP [41]	80.60	**82.00**	81.29
CoOp [43]	91.09	68.94	78.48
CoCoOp [42]	92.61	73.25	81.80
MMRL [13]	93.07	74.37	82.68
MaPLe [20]	92.95	76.40	83.87
PromptKD [24]	92.70	75.80	83.40
APPLeNet [28]	92.20	74.10	82.17
Ours	**94.90**	77.15	**85.11**
Δ	+1.83	-4.85	+1.24

(c) EuroSAT.

WHU-RS19 [31]	Base	Novel	HM
GeoRSCLIP [41]	85.62	94.18	89.70
CoOp [43]	93.08	88.54	90.75
CoCoOp [42]	93.08	90.84	91.95
MMRL [13]	93.92	90.14	91.99
MaPLe [20]	93.60	91.80	92.69
PromptKD [24]	93.40	90.90	92.13
APPLeNet [28]	93.20	90.50	91.83
Ours	**96.85**	**94.45**	**95.63**
Δ	+2.93	+0.27	+2.94

(d) WHU-RS19.

UCMerced [33]	Base	Novel	HM
GeoRSCLIP [41]	87.27	**76.67**	81.63
CoOp [43]	94.88	60.22	73.68
CoCoOp [42]	93.69	63.11	75.42
MMRL [13]	96.58	70.67	81.62
MaPLe [20]	96.20	72.20	82.49
PromptKD [24]	96.00	71.00	81.63
APPLeNet [28]	95.80	71.20	81.69
Ours	**97.90**	75.21	**85.07**
Δ	+1.32	-1.46	+2.58

(e) UCMerced.

PatternNet [44]	Base	Novel	HM
GeoRSCLIP [41]	57.38	67.59	62.48
CoOp [43]	82.22	58.06	70.14
CoCoOp [42]	83.94	53.95	68.94
MMRL [13]	89.56	59.00	74.28
MaPLe [20]	85.89	60.45	73.17
PromptKD [24]	91.21	75.80	83.50
APPLeNet [28]	94.22	74.20	84.21
Ours	**94.50**	**77.83**	**86.16**
Δ	+0.28	+2.03	+1.95

(f) PatternNet.

Table 9. Symbols and default/search ranges used in AVION. In all experiments we fix K_p=30, β=10, γ=2, ζ_s=3.0.

Symbol	Meaning	Default / Search
N	# training images	–
C	# classes	–
D	image/text embedding dimension	inherited from the VLM
f_T, g_T	teacher image/text encoders	frozen
f_S, g_S	student image/text encoders	backbones frozen
$\mathbf{v}_i^T$	ℓ_2-normalized teacher image feature for x_i	unit ℓ_2-norm
$\mathbf{t}_k^{T*}$	teacher text prototype for class k	unit ℓ_2-norm
$\mathbf{v}_i^S$	ℓ_2-normalized student image feature for x_i	unit ℓ_2-norm
$\mathbf{t}_k^S$	student text feature for class k	unit ℓ_2-norm
K_p	# captions per class from *LLM-based domain prompting*	10–50
β, γ	weighting in *selective prototype aggregation* (Eq. 3)	$\beta \in [5, 20]$, $\gamma \in \{0, 2\}$
ζ_s	MAD-based pruning threshold (Eq. 9)	2.5–3.5
ε	numerical stability constant (Eq. 9)	10^{-8}
τ_s, τ_t	student/teacher logit scales	τ_t via grid search; τ_s learnable
τ	temperature of $\mathcal{L}_{\text{logit}}$ (Eq. 7)	τ=2
$\lambda_{\text{img}}, \lambda_{\text{text}}, \lambda_{\text{logit}}$	loss weights	0.5, 0.5, 1.0 (fixed; see Appx. G)
$P^{(v)}$	# visual prompt tokens per layer	4–16
$P^{(t)}$	# text prompt tokens per layer	2–8
L_v, L_t	# prompt-injected layers (vision/text)	match ViT/Transformer depth

Table 10. Examples of LLM-generated candidate descriptions from our *LLM-based domain prompting* module, compared to standard simple templates. The generated candidates provide richer semantic and visual cues specific to remote sensing imagery.

Class Name	Simple Template	LLM-Generated Candidates
`airport`	a photo of an airport	- Overhead airport imagery reveals aprons (with parked aircraft at stands) adjacent to terminals, connected to the airfield via taxilanes and the taxiway system. - An aerial view of an airport showing multiple long, paved runways and taxiways. - Satellite imagery capturing a large airfield complex with terminal buildings and hangars.
`industrial area`	a photo of an industrial area	- A top-down view of an industrial zone characterized by large factory buildings and smokestacks. - Aerial imagery of a manufacturing plant complex with large warehouses and shipping depots. - A satellite photo showing a dense cluster of large-scale industrial structures and storage tanks.
`commercial area`	a photo of a commercial area	- An overhead view of a commercial district featuring multi-story office buildings and retail centers. - Nadir-view imagery of a dense downtown business area with high-rise buildings and streets. - An aerial photo of a shopping center or mall surrounded by a large, organized parking lot.
`forest`	a photo of a forest	- A vast expanse of dense, green forest canopy as seen from a satellite. - An aerial view looking straight down at a dense woodland of coniferous or deciduous trees. - A top-down remote sensing image of a large, contiguous area covered by trees.

candidates using the RS-Flag prior (Steps 4-5) to obtain the final teacher text prototype for each class.

The robust z-score used in Step 2 is computed using the median absolute deviation (MAD). Let $m_k = \mathrm{median}(\{s_{k,j}\}_j)$, $\Delta_{k,j} = |s_{k,j} - m_k|$, and $\mathrm{MAD}_k = \mathrm{median}(\{\Delta_{k,j}\}_j)$. The z-score is defined as

$$z_{k,j} = \frac{\Delta_{k,j}}{\mathrm{MAD}_k + \varepsilon}, \tag{9}$$

which downweights or removes outlier captions whose scores deviate strongly from the median.

E. Masking and Renormalization for $\mathcal{L}_{\text{logit}}$ in Base to Novel Training

Let $\mathcal{Y}_{\text{base}} \subset \mathcal{Y}$ denote the set of base classes and $\mathcal{Y}_{\text{novel}} = \mathcal{Y} \setminus \mathcal{Y}_{\text{base}}$ the set of novel classes. During base-to-novel training, we restrict the distillation distributions to base classes by masking scores for $\mathcal{Y}_{\text{novel}}$ and renormalizing over $\mathcal{Y}_{\text{base}}$:

$$\tilde{q}_{i,k}^{(T)} = \begin{cases} \dfrac{\exp\big(s_{i,k}^{(T)}/\tau\big)}{\sum_{k' \in \mathcal{Y}_{\text{base}}} \exp\big(s_{i,k'}^{(T)}/\tau\big)} & k \in \mathcal{Y}_{\text{base}}, \\ 0 & k \in \mathcal{Y}_{\text{novel}}, \end{cases} \tag{10}$$

$$\tilde{p}_{i,k}^{(S)} = \begin{cases} \dfrac{\exp\big(s_{i,k}^{(S)}/\tau\big)}{\sum_{k' \in \mathcal{Y}_{\text{base}}} \exp\big(s_{i,k'}^{(S)}/\tau\big)} & k \in \mathcal{Y}_{\text{base}}, \\ 0 & k \in \mathcal{Y}_{\text{novel}}. \end{cases} \tag{11}$$

Here $s_{i,k}^{(T)}$ and $s_{i,k}^{(S)}$ are the teacher and student logits for image x_i and class k, and τ is the distillation temperature used in $\mathcal{L}_{\text{logit}}$ (Eq. 7). The $\mathcal{L}_{\text{logit}}$ term then becomes

$$\mathcal{L}_{\text{logit}}(x_i) = \tau^2 \,\mathrm{KL}\Big(\tilde{q}_{i,\cdot}^{(T)} \,\Big\|\, \tilde{p}_{i,\cdot}^{(S)}\Big),$$

which ensures that novel classes neither appear in the denominator nor contribute to the divergence during training.

F. Complexity and Inference Overhead

We quantify the additional parameters and computational overhead introduced by AVION on top of a frozen GeoRSCLIP backbone. Unless otherwise stated, all numbers below correspond to our default ViT-B/32 student with deep prompts in both encoders.

Parameter count. For vision and text branches with widths D_v and D_t and L_v, L_t transformer layers that receive prompts, the total number of prompt parameters is

$$\#\theta_{\text{prompts}} \;=\; L_v\, P^{(v)}\, D_v \;+\; L_t\, P^{(t)}\, D_t,$$

where $P^{(v)}$ and $P^{(t)}$ denote the number of learnable prompt tokens per layer on the vision and text sides, respectively (see Tab. 9). In our main configuration (GeoRSCLIP ViT-B/32 student), the vision encoder has D_v=768, L_v=12, $P^{(v)}$=8, and the text encoder has D_t=512, L_t=12, $P^{(t)}$=4. Then

$$12{\cdot}8{\cdot}768 \;+\; 12{\cdot}4{\cdot}512 \;=\; 98{,}304$$

trainable prompt parameters, which is well below 1% of a ViT-B/32 backbone. All other backbone weights remain frozen; beyond these prompts, only a handful of scalar hyperparameters (e.g., the learnable logit scale τ_s) are updated.

FLOPs overhead (self-attention and MLP). Let N be the per-layer token count and P the number of prompt tokens added to that layer. The self-attention cost scales from N^2 to $(N{+}P)^2$, so the *relative* attention overhead is

$$r_{\text{attn}} \;\approx\; \frac{(N+P)^2 - N^2}{N^2} \;=\; 2\frac{P}{N} + \left(\frac{P}{N}\right)^2.$$

The MLP cost scales approximately linearly with sequence length and therefore grows proportionally to P/N.

Examples (our setup). For the vision branch, ViT-B/32 at input resolution 224^2 has N_v=49 patch tokens; with $P^{(v)}$=8,

$$r_{\text{attn}}^{(v)} \approx 2 \cdot \tfrac{8}{49} + \left(\tfrac{8}{49}\right)^2 \approx 35.3\%,$$

and the per-layer MLP overhead is $\approx P^{(v)}/N_v \approx 16.3\%$. (We ignore the class token here for a rough estimate.) For the text branch, the CLIP text encoder uses N_t=77 tokens; with $P^{(t)}$=4,

$$r_{\text{attn}}^{(t)} \approx 2 \cdot \tfrac{4}{77} + \left(\tfrac{4}{77}\right)^2 \approx 10.7\%,$$

and the MLP overhead is $\approx P^{(t)}/N_t \approx 5.2\%$. Because ViT-B/32 has shorter vision sequences than ViT-B/16, the same P yields a larger *relative* attention overhead; reducing P or using shallower prompting (smaller L_v, L_t) lowers this cost. In practice, these per-layer increases translate into a moderate end-to-end FLOPs and latency overhead compared to the frozen backbone.

Runtime and memory. Table 11 reports the measured runtime of our ViT-B/32 student (frozen backbone + deep prompts). We report latency, throughput, and memory consumption over different batch sizes at 224^2 resolution. Throughput is computed as *batch size / latency*; exact numbers may vary across hardware and software stacks. The teacher is used only during training, so it incurs zero cost at inference.

Table 11. Runtime with ViT-B/32 student (frozen backbone + deep prompts). Teacher is train-time only and has zero inference cost. Throughput is computed as batch/latency; exact numbers may vary by hardware and software stack.

Setting	Batch	Res.	Lat. (ms)	img/s	Mem (MB)
Student	64	224^2	27.4	2334	1500
Student	256	224^2	109.7	2334	5800
Teacher (train-time only)	64	224^2	n/a	n/a	n/a

G. Hyperparameter Sensitivity Analysis

In this section, we provide a detailed justification of the hyperparameter choices used for our AVION framework, focusing on (i) our parameter selection protocol, (ii) the effect of loss weights (λ) and scheduling, and (iii) our choice of distillation temperature (τ).

Parameter Selection Protocol and Fixed Settings. To ensure robustness and avoid over-tuning, all hyperparameters were selected *once* on a validation split derived from the **AID** dataset's base classes. These exact settings (e.g., learning rates, loss weights, warm-up schedule) were then frozen and applied to all six classification datasets and both retrieval datasets across all three experimental protocols (few-shot, base-to-novel, and retrieval). The strong performance across all tasks suggests the robustness of this single set of hyperparameters. Furthermore, following prior work on logit-based distillation [24], we use a default distillation temperature of τ=2 and do not tune it as part of our selection process.

Analysis of Loss Weights and Scheduling. We conducted a detailed ablation study, summarized in Table 12, to validate our final loss weight configuration ($\lambda_{\text{img/text}}$=0.5, λ_{logit}=1.0) and the 30% linear warm-up strategy on the $\mathcal{L}_{\text{logit}}$ term. The results lead to three key conclusions:

Table 12. Ablation on loss weights (λ) and scheduling strategies. We report the average harmonic mean (HM, %) for base-to-novel generalization.

Setting	λ_{img}	λ_{text}	λ_{logit}	Avg. HM
Only $\mathcal{L}_{\text{logit}}$	0	0	1.0	83.14
All 1.0 (2-stage)	1.0	1.0	1.0	85.47
All 1.0 (Warm-up)	1.0	1.0	1.0	86.91
Ours (Warm-up)	**0.5**	**0.5**	**1.0**	**87.05**

1. **The tri-aspect alignment losses are critical and complementary.** Relying only on $\mathcal{L}_{\text{logit}}$ results in a significantly lower HM (83.14%), indicating that the representation alignment $\mathcal{L}_{\text{img}}$ and semantic alignment $\mathcal{L}_{\text{text}}$ play a critical role in achieving top performance.
2. **Our chosen weights ($\lambda_{\text{img/text}}$=0.5) provide a better balance.** Setting all three weights to 1.0 yields a strong 86.91% HM, but it is slightly outperformed by our final configuration at 87.05% HM. Comparing "All 1.0 (Warm-up)" and "Ours (Warm-up)" isolates the effect of λ_{img} and λ_{text}, suggesting that giving these embedding-level alignment terms a higher weight (1.0 vs. 0.5) can slightly over-regularize the student and hinder its optimal adaptation to the task.
3. **The 30% warm-up is the superior scheduling strategy.** Comparing "All 1.0 (2-stage)" (85.47%) and "All 1.0 (Warm-up)" (86.91%) isolates the impact of the scheduling policy. The smoother 30% warm-up schedule yields a clear HM improvement over the hard 2-stage scheme, and our final setting ("Ours (Warm-up)") further reaches 87.05%. This supports our choice of a warm-up strategy for $\mathcal{L}_{\text{logit}}$ as a more stable and robust schedule.

H. Qualitative Analysis of RS-Flag

To empirically justify the reliability of the RS-Flag mechanism and demonstrate how it mitigates hallucinations, we use the "stadium" class as an example to visualize the normalized weights w obtained from Eq. (3), as shown in Figure 9. The candidates along the x-axis are ranked by their initial visual-textual similarity scores (blue bars).

As highlighted by the dashed circles, several candidates—most notably the top-ranked ones (indices 1, 2, and 3)—are significantly suppressed after applying the RS-Flag calibration (orange bars). This occurs because these candidates, despite having high initial visual alignment with the image, lack explicit remote sensing contextual features or contain ground-level negative words (i.e., RS-Flag=0). Taking index 3 as an example, a description like "A massive concrete ring structure surrounding a rectangular green field." is not

inherently incorrect visually, but it lacks an explicit aerial perspective. Therefore, our mechanism does not directly eliminate it, but instead reduces its relative weight. This calibration allows candidates with stronger RS-awareness (e.g., indices 4, 5, 6, and 7), which receive a positive logit shift (RS-Flag=1), to correctly dominate the final aggregated prototype.

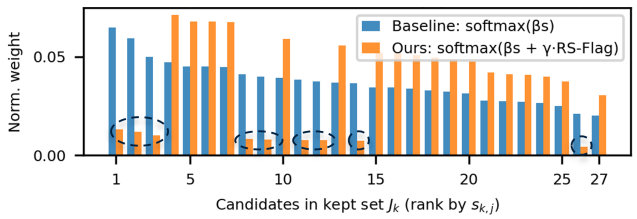


Figure 9. Qualitative analysis of the RS-Flag mechanism using the "stadium" class. Candidates are ranked by their initial similarity ($s_{k,j}$). The visualization compares the baseline softmax weights with our calibrated weights. Dashed circles highlight candidates (e.g., indices 1–3, 8–9, 11–12, 14, 26) that lack explicit remote sensing context and are consequently down-weighted to emphasize true RS-aware descriptions.

I. Cross-dataset Evaluation

To further validate the generalization capabilities of AVION and verify that it reduces source-domain overfitting, we conduct a cross-dataset evaluation. We train AVION under a 16-shot setting (using all classes) on the PatternNet dataset and report the zero-shot transfer performance on three target datasets: RESISC-45, UCMerced, and AID (all classes). For fairness, we re-implement all evaluated methods on the identical GeoRSCLIP (ViT-B/32) backbone.

As demonstrated in Table 13, AVION significantly outperforms APPLeNet on the source domain (+5.46%) and consistently transfers better to unseen target domains: +2.88% on RESISC-45, +1.87% on UCMerced, and +1.59% on AID. Overall, AVION generalizes better across diverse remote sensing datasets, suggesting that our framework effectively mitigates the common issue of source-domain overfitting.

Table 13. Cross-dataset evaluation (%). Models are trained on PatternNet (16-shot) and evaluated zero-shot on targets. All methods utilize the GeoRSCLIP (ViT-B/32) backbone. ZS: Zero-shot.

Method	Source	Target		
	PatternNet	RESISC-45	UCMerced	AID
GeoRSCLIP (ZS)	60.17	71.89	79.72	73.72
APPLeNet	86.63	73.25	80.18	76.40
AVION (Ours)	**92.09**	**76.13**	**82.05**	**77.99**

J. Ablation on LLM Domain Prompting

To ensure that the AVION framework is not overly sensitive to the specific large language model employed for domain prompting, we conduct an ablation study evaluating different offline LLM generators. Table 14 reports the average few-shot and base-to-novel generalization results across all six remote sensing datasets using text candidates generated by GPT-5, Llama-3.1-70B-Instruct, and our default Gemini 2.5 Flash.

The performance differences are marginal across all metrics. This stability confirms that AVION's performance gains stem fundamentally from the architectural design of the verification and distillation mechanisms (e.g., RS-Flag and tri-aspect alignment), rather than relying on the idiosyncrasies of a specific language model.

Table 14. Ablation on offline LLM choices. We report average few-shot accuracy and base-to-novel generalization (Base, Novel, HM) in % across six RS datasets. The framework remains stable regardless of the LLM generator used.

Offline LLM	Few-Shot		Base-to-Novel		
	4-shot	16-shot	Base	Novel	HM
GPT-5	**88.45**	92.85	95.15	78.45	86.00
Llama-3.1-70B-Inst.	87.12	93.65	**95.72**	79.55	86.89
Gemini 2.5 Flash (Ours)	88.31	**93.69**	95.64	**79.94**	**87.05**

K. Comparisons on Broader Benchmarks

While AVION is explicitly tailored to address the unique challenges of aerial and satellite imagery, we also explore its scalability and general applicability on the broader, general-domain ImageNet benchmark.

For this experiment, we evaluate the base-to-novel generalization setting using a standard CLIP (ViT-B/16) student model. To adapt our pipeline to general-domain images, we modify the offline LLM generator query to a general template: "*In one sentence, describe the distinctive appearance of [CLASS].*" We generate 30 candidate descriptions per class using Gemini 2.5 Flash. Additionally, because ImageNet consists primarily of ground-level, human-centric photographs, we disable the RS-Flag filtering mechanism. Instead, the framework relies solely on the visual-textual similarity (Eq. (2)) and the MAD-based robust pruning for selective prototype aggregation.

As shown in Table 15, our offline-teacher-to-student distillation paradigm remains highly effective outside the remote sensing domain. When guided by a ViT-H/14 CLIP teacher, AVION outperforms the recent state-of-the-art general prompt tuning method MMRL in terms of the Harmonic Mean (HM). This suggests that the core mechanism of distilling semantically rich, visually verified prototypes is a

versatile approach that generalizes well beyond aerial contexts.

Table 15. Base-to-novel generalization (%) on the ImageNet benchmark. The student model is CLIP (ViT-B/16). AVION demonstrates competitive transferability on general-domain images without the RS-Flag constraint.

Method	Base	Novel	HM
MMRL	77.90	71.30	74.45
AVION (Teacher: ViT-L/14)	77.27	74.02	75.63
AVION (Teacher: ViT-H/14)	**79.15**	**74.79**	**76.94**