Beyond Templates: Revisiting Zero-Shot Remote Sensing through Meta-Prompting

Eirini Baltzi[a], Dionysis Christopoulos[a], Sotiris Spanos[a], Valsamis Ntouskos[a,b], Konstantinos Karantzalos[a]

[a]*Remote Sensing Lab, National Technical University of Athens, Athens, Greece*
[b]*Department of Engineering and Sciences, Universitas Mercatorum, Rome, Italy*

Abstract

Vision-language models (VLMs) have sparked growing interest in zero-shot Earth Observation (EO) downstream tasks, with further gains enabled by remote-sensing-adapted models. We examine this setting across 17 VLM variants and 12 remote sensing (RS) datasets under Meta-Prompting for Visual Recognition (MPVR), and show that zero-shot performance remains highly sensitive to textual design choices, from the meta-prompts used to guide the LLM in generating class descriptions to the descriptions themselves. We explore why semantically rich LLM-generated class descriptions do not translate into consistent gains over simple domain-adapted CLIP-style descriptions. While LLM descriptions are more semantically expressive, they can also introduce noise in the text embedding space, reducing robustness in downstream tasks. We support this observation through a text log-likelihood analysis in the whitened CLIP feature space, comparing LLM-generated and template-based descriptions. Building on this finding, we study query embedding calibration and show that lightweight calibration of the query space consistently yields strong improvements in zero-shot classification and retrieval. Overall, our results provide practical insight into the trade-off between semantic richness and robustness, and identify embedding calibration as a simple and effective tool for improving zero-shot remote sensing performance.

Keywords: vision-language, zero-shot, meta-prompting, earth observation, large language models, embedding calibration, centering, retrieval, scene classification

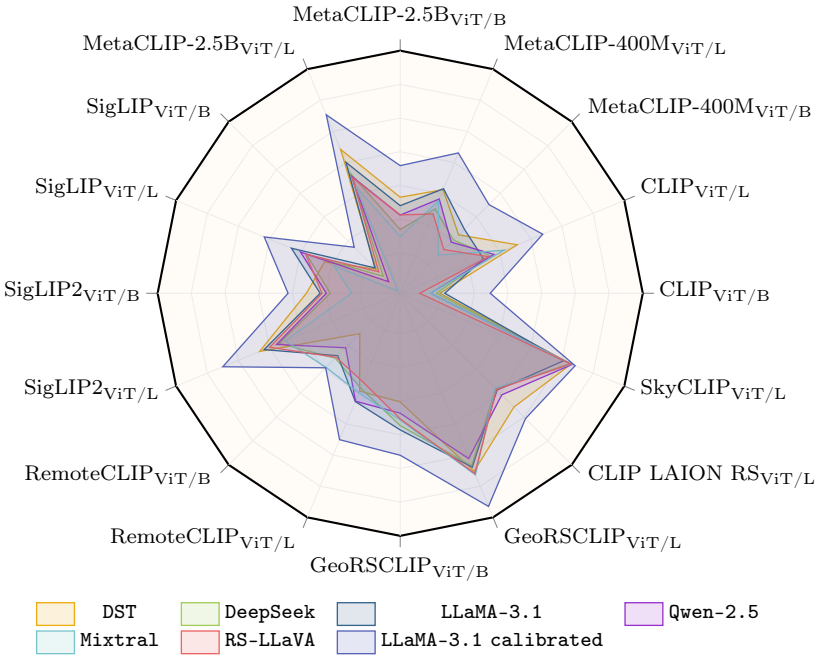

Figure 1: Comparison of template-based `DST` and meta-prompting approaches in terms of classification accuracy across the evaluated VLMs. We also report the `LLaMA-3.1` variant obtained by embedding calibration. `LLaMA-3.1 calibrated` consistently defines the strongest overall performance regime, surpassing all other prompt sources.

1. Introduction

Vision-language models (VLMs), such as CLIP [1], enable zero-shot visual recognition by aligning image and text representations in a shared embedding space. In this setting, classification is performed by comparing image embeddings with textual descriptions of candidate classes. While this paradigm has demonstrated strong generalization on natural-image benchmarks, its behavior in remote sensing (RS) imagery remains underexplored and highly sensitive to prompt formulation. Although RS datasets are often included in broader VLM evaluations, our work provides a targeted, domain-specific analysis of these dynamics.

Remote sensing scenes differ substantially from natural images due to overhead viewing geometry, varying spatial resolution, and scene-level semantics. As a result, the generic templates commonly used in zero-shot evaluation (e.g., "a photo of a class") often fail to capture RS-specific visual cues. Recent work on Meta-Prompting for Visual Recognition (MPVR) [2]

has shown that large language models (LLMs) can automatically generate class- and dataset-aware descriptions, enriching the textual supervision used for zero-shot prediction. Our prior work [3] establishes the feasibility of using MPVR-style prompting with open-source LLMs to generate remote-sensing-aware class descriptions. However, we observed that performance gains vary across datasets and vision-language backbones.

In this work, we revisit meta-prompting for zero-shot remote sensing through a comprehensive empirical evaluation across twelve RS datasets and a broad range of vision-language models. Building on the two-stage prompt generation procedure introduced in MPVR, we adapt prompt construction to remote-sensing scene semantics and examine several components that affect zero-shot performance. Specifically, we investigate the role of prompt aggregation, including a top-k voting variant in addition to prototype averaging, and we study embedding calibration strategies based on image- and text-side centering, analyzing their asymmetric impact on classification and text-to-image retrieval.

Our study provides practical insights into how textual diversity, aggregation, and representation geometry interact in zero-shot remote sensing. The results indicate that prompt design and lightweight calibration substantially influence both performance and robustness across datasets and model families, offering useful guidance for deploying VLMs in RS applications and for future work on prompt-based adaptation under domain shift.

The main contributions of this work are:

- **Large-scale RS study of meta-prompting.** We evaluate MPVR-style prompting across 12 remote-sensing datasets and 17 vision-language model variants.

- **Embedding calibration analysis.** We study simple embedding-centering strategies and show task-dependent effects: image centering consistently benefits classification, while text centering is more effective for retrieval.

- **Dual-task evaluation.** Beyond zero-shot classification, we report text-to-image retrieval results, providing a complementary lens for assessing prompt quality and cross-modal behavior in RS.

- **Analysis of prompt robustness and embedding geometry.** We study why semantically rich LLM-generated descriptions can be less

stable than templates by analyzing their behavior in the whitened CLIP feature space.

2. Related Work

Vision-Language Models for Zero-Shot Recognition. Vision-language models (VLMs), such as CLIP [1], MetaCLIP [4], and SigLIP [5], have established a powerful paradigm for zero-shot visual recognition by learning a shared embedding space between images and natural language. In this setting, classification is performed by matching the image representations with the textual descriptions of candidate classes, enabling strong generalization to unseen categories [1]. Despite their success on natural image benchmarks, VLMs remain highly sensitive to the formulation of textual prompts [6, 7], and their performance can degrade under domain shift, especially when the target visual distribution differs substantially from that observed during pretraining [8]. While much of the literature has focused on improving model architectures or scaling pretraining data, comparatively less attention has been given to understanding how the prompt design affects zero-shot behavior in domain-specific settings.

Vision-Language Models in Remote Sensing. Applying VLMs to remote sensing (RS) imagery introduces additional challenges due to the unique properties of overhead data, including varying spatial resolutions, scale ambiguity, and scene-level semantics [9]. Recent efforts such as RemoteCLIP [10], GeoRSCLIP [11], and SkyCLIP [12] adapt vision-language models to RS data through domain-specific pretraining or large-scale remote sensing vision-language corpora. These models consistently improve zero-shot performance relative to general-purpose backbones, yet the mismatch between generic textual supervision and RS-specific visual content remains an important limitation. This motivates not only the use of domain-aware prompt design, but also a broader analysis of how prompting strategies interact with model domain adaptation in remote sensing.

Prompt Engineering and Meta-Prompting. Prompt engineering has emerged as a central component of zero-shot evaluation, with prior work exploring manually designed templates, prompt ensembling, and automatic prompt generation [1, 6, 7]. Among these directions, meta-prompting approaches such as MPVR [2] are particularly relevant, as they use large language models (LLMs) to generate richer and more diverse class descriptions than fixed

templates. By increasing textual coverage, such methods can improve alignment between text and image embeddings, especially when simple prompts fail to capture the relevant semantics of a class. However, the benefit of richer descriptions is not uniform, and their interaction with the underlying embedding space remains poorly understood, particularly under the domain shift induced by aerial and satellite imagery.

Large Language Models as Prompt Generators in Remote Sensing. Recent advances in large language models have enabled the automatic generation of detailed textual descriptions that can serve to guide frozen VLMs in zero-shot settings. These descriptions may incorporate attributes, contextual cues, and domain-relevant scene characteristics that go beyond class-name-based templates [2]. In remote sensing, this is especially appealing, as overhead scenes often require descriptions that capture spatial layout, land-cover composition, and imaging perspective [9, 13]. In this context, recent open-source and domain-specialized LLMs, including LLaMA-3.1 [14], DeepSeek [15], Qwen-2.5 [16], Mixtral [17], and RS-LLaVA [18], provide a broad space of prompt-generation sources to study. At the same time, the use of LLM-generated prompts in RS remains underexplored from a comparative and analytical perspective. While individual aspects have been explored, a systematic understanding of how different prompt sources interact with both general-purpose and remote-sensing-adapted VLMs remains limited, particularly regarding how the geometry of the resulting text representations affects downstream behavior across tasks. In this work, we address this gap through a large-scale evaluation of template-based and meta-prompt-generated descriptions across multiple RS datasets, VLM families, open-source LLMs, and downstream tasks, while also examining the role of embedding calibration in both classification and retrieval.

3. Evaluation Framework

To systematically study the impact of textual supervision on zero-shot performance in remote sensing, we build upon the unified evaluation framework introduced in MPVR [2]. We formulate our zero-shot remote sensing evaluation pipeline as two explicit stages (Figure 2). In the first stage, we contextualize a selected LLM using dataset-level information to generate dataset-specific queries. In our setup, GPT-5 is used for this meta-prompting step. Subsequently, these queries are used to prompt open-source LLMs,

which produce class-specific textual descriptions for the categories of each downstream target dataset. In the second stage, the generated descriptions are encoded with the frozen VLM text encoder and compared against image embeddings extracted from the frozen VLM visual encoder. Unless stated otherwise, we apply prompt embedding aggregation at the class level and a simple yet highly effective calibration step in the query modality before computing zero-shot predictions. This formulation provides a controlled framework for analyzing how dataset-level contextualization, class-description generation, prompt aggregation, and query-side embedding calibration affect downstream zero-shot performance across different vision-language models (VLMs) and remote sensing benchmarks.

Downstream tasks. We evaluate the generated class representations on two downstream tasks, namely zero-shot image classification and text-to-image retrieval. In the former, image embeddings are compared against the set of aggregated text prototypes, and classification is performed by identifying the class that maximizes similarity. In the latter, each class prototype acts as a text query to retrieve relevant images from the gallery by ranking image embeddings according to cosine similarity. Together, these tasks offer complementary perspectives on the quality of alignment in the shared vision-language embedding space.

3.1. Description Generation Strategies

Template-based descriptions. As a point of comparison, we consider two forms of fixed textual supervision that bypass the dataset-specific query generation stage and construct descriptions directly from class names. In particular, the Simple Template (`ST`) baseline, uses a single generic template ("a satellite image of a {class}"), whereas the Domain-Specific Template (`DST`) baseline, employs multiple remote-sensing aware templates designed to better reflect the overhead imaging perspective (e.g., "an overhead image of a {class}"). These baselines provide simple, controlled, and interpretable textual descriptions, but remain limited to shallow class-name conditioning and therefore have reduced capacity to capture richer class semantics and intra-class visual variability.

LLM-based description generation. Under the LLM-based description generation setting, we adopt a meta-prompting strategy ([2]) in which dataset-level information is first used to formulate dataset-specific queries, and these

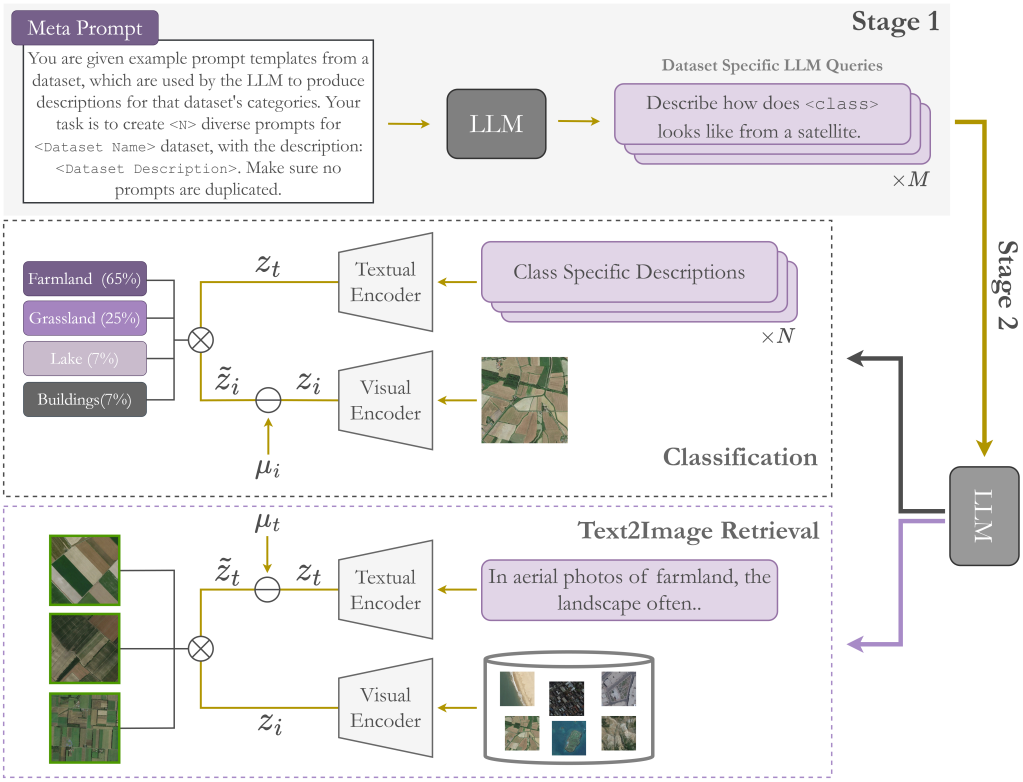

Figure 2: Overview of our evaluation framework. In Stage 1, dataset-level information is used to generate dataset-specific LLM queries. In Stage 2, these queries produce class-specific descriptions, which are encoded by the frozen text encoder, while images are mapped into the same shared embedding space by the frozen visual encoder. For zero-shot classification, image embeddings are centered before matching with aggregated text prototypes. For text-to-image retrieval, text embeddings are centered before matching with image embeddings. Here, μ_i and μ_t denote the mean embedding vectors computed over the query modality of the corresponding downstream task, namely the set of image embeddings for classification and the set of text embeddings for retrieval.

queries are subsequently supplied to open-source LLMs to generate class-specific textual descriptions across the target label space. This way, the language model is not prompted directly with isolated class names, but is first contextualized at the dataset level, allowing the generated descriptions to better reflect the semantics and visual characteristics of the target benchmark. Compared with fixed template baselines, this meta-prompting strategy yields multiple descriptions per class with greater diversity in wording, semantic emphasis, and descriptive detail. In this study, we evaluate several

open-source LLMs, including `LLaMA-3.1` [14], `DeepSeek` [15], `Qwen-2.5` [16], `Mixtral` [17], and the domain-specific `RS-LLaVA` [18]. This design allows us to analyze how the source and characteristics of the meta-prompt-generated descriptions influence the resulting text embedding space and, in turn, zero-shot downstream performance.

3.2. Prototype construction

Prompt Aggregation. Given a set of textual descriptions for each class, obtained using either the template- or LLM-based description generation strategy, we map them into the shared embedding space using the frozen text encoder of the corresponding vision-language model. Formally, let $\mathcal{T}_c = \{t_{c,1}, \ldots, t_{c,N_c}\}$ denote the set of textual descriptions associated with class c, where N_c is the number of descriptions for that class. Each description is encoded as

$$z_{c,n}^{\text{text}} = f_{\text{text}}(t_{c,n}),$$

where $f_{\text{text}}(\cdot)$ is the frozen text encoder. A single class representation is then formed by averaging the corresponding text embeddings,

$$\bar{z}_c^{\text{text}} = \frac{1}{N_c} \sum_{n=1}^{N_c} z_{c,n}^{\text{text}},$$

yielding one class prototype per category.

Embedding calibration. Following standard practice in vision-language models, both visual and textual embeddings are ℓ_2-normalized, and cross-modal matching is performed using cosine similarity. Given an image x with embedding

$$z^{\text{img}} = f_{\text{img}}(x),$$

the similarity between image x and class c is computed as

$$s(x, c) = \cos\big(z^{\text{img}}, \bar{z}_c^{\text{text}}\big),$$

where $f_{\text{img}}(\cdot)$ denotes the frozen visual encoder. Additionally, we apply a lightweight embedding calibration step based on mean-centering. Specifically, prior to similarity computation, we subtract the mean vector of the query modality, where the query side is defined by the downstream task. Let

$\mathcal{Q} = \{q_1, \ldots, q_M\}$ denote the set of query embeddings for the task under consideration. We compute the mean query vector as

$$\mu_{\mathcal{Q}} = \frac{1}{M} \sum_{j=1}^{M} q_j,$$

and center each query embedding according to

$$\tilde{q}_j = q_j - \mu_{\mathcal{Q}}.$$

For zero-shot image classification, this corresponds to centering image embeddings, whereas for text-to-image retrieval, it corresponds to centering text embeddings. In contrast, centering the database-side modality did not provide consistent gains and often reduced performance. We therefore apply calibration only to the query modality. This operation effectively suppresses dominant shared components that reflect generic semantics or modality-level bias rather than class-discriminative content, thereby improving the sensitivity of cosine similarity to meaningful cross-modal differences [19].

4. Experimental Setup

Evaluated datasets. We conduct our evaluation on the full suite of remote sensing scene classification benchmarks adopted in RemoteCLIP [10]. This benchmark collection comprises 12 datasets with substantial variation in dataset scale, number of classes, image resolution, and ground sampling distance (GSD), thereby covering diverse scene categories and levels of semantic granularity. A detailed dataset overview is provided in the appendix (Table A.5). All metrics reported in the following section are obtained by evaluating on the entirety of each dataset, with every available image treated as part of the test set, since the proposed setting does not involve any task-specific training.

Evaluated Vision-Language Models. We evaluate a diverse set of vision-language models (VLMs) spanning both general-purpose and remote sensing-adapted architectures. The evaluated model families include CLIP [1], MetaCLIP [4], SigLIP [5], and SigLIP2 [20], as well as remote sensing-oriented models such as RemoteCLIP [10], GeoRSCLIP [11], CLIP-LAION-RS [12], and SkyCLIP [12]. Whenever available, we consider both ViT-Base and ViT-Large

variants in order to study the consistency of the proposed textual supervision strategies across different backbone scales. For RemoteCLIP, we further evaluate the ResNet-50 variant, thereby extending the comparison to a convolutional visual encoder. Throughout all experiments, both the visual and text encoders are kept frozen, such that performance differences can be attributed solely to the proposed description generation, aggregation, and calibration strategies rather than to any task-specific adaptation of the backbone models.

Evaluated LLMs. For LLM-based prompt generation, we employ a diverse set of open-source large language models (LLMs) with different architectural characteristics, parameter scales, and levels of domain specialization. In particular, we evaluate `LLaMA-3.1` [14] with 8B parameters, `DeepSeek` [15], `Qwen-2.5` [16], and the domain-specific `RS-LLaVA` [18], each in their 7B configuration, together with `Mixtral` [17], which adopts an 8×7B mixture-of-experts design. Each LLM is used within the same meta-prompting pipeline to generate class-specific textual descriptions for every target dataset and for all classes in its label space. More specifically, for each dataset-specific query and each class, we generate four independent textual descriptions, and we keep this repetition factor fixed across all evaluated LLMs and datasets. To ensure a controlled comparison, we use a standard temperature of 0.7 and a maximum generation length of 70 tokens for all models, except for `RS-LLaVA`, where a higher temperature of 1.05 is used to encourage greater descriptive diversity. This fixed generation setup allows us to study how architectural and domain-specific differences among LLMs influence the resulting text descriptions and their effectiveness in downstream zero-shot classification and retrieval.

5. Quantitative Results

5.1. Classification

Vanilla Embeddings. We report results on the zero-shot image classification task, using top-1 accuracy as the evaluation metric. We focus here on the Domain-Specific templates (`DST`) and on LLaMA-generated descriptions (`LLaMA-3.1`) approaches in Table 1 and Table 2 respectively. `LLaMA-3.1` is selected as the representative LLM-based configuration because it proved to be the most consistent choice among the evaluated open-source LLMs throughout our experiments. Tables B.6–B.10 in the Appendix, present the

Table 1: Zero-shot classification accuracy (%) on all evaluated remote sensing datasets under the Domain-Specific Template (`DST`) baseline. Results are reported across the full set of vision-language models and backbone architectures. Best results per dataset are in **bold**, second best are underlined.

Model	Arch	EuroSAT	AID	MLRSNet	Optimal-31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHU-RS19	SIRI-WHU	RSSCN7	Avg
CLIP	ViT-B/16	51.4	67.3	57.2	72.2	64.5	64.4	62.4	30.9	43.0	85.5	55.3	65.4	59.9
	ViT-L/14	63.3	69.4	65.6	80.7	75.9	71.6	66.2	40.2	48.9	85.5	65.2	64.2	66.4
MetaCLIP-400M	ViT-B/16	53.4	71.8	59.8	77.7	69.1	70.0	65.3	33.6	39.4	90.0	60.5	67.1	63.1
	ViT-L/14	50.1	73.9	62.7	80.9	74.4	71.9	64.7	35.5	44.8	89.2	66.5	69.4	65.3
MetaCLIP-2.5B	ViT-B/16	49.1	72.5	64.4	76.0	71.9	70.5	65.0	33.2	41.1	90.7	66.8	68.6	64.1
	ViT-L/14	60.8	78.9	**70.7**	83.2	**82.8**	75.0	63.1	39.5	44.0	90.0	65.8	68.9	68.6
SigLIP	ViT-B/16	43.3	71.0	56.7	75.9	64.2	65.6	62.1	26.2	36.8	85.0	62.5	65.0	59.5
	ViT-L/16	52.8	72.2	60.8	83.1	65.5	69.3	63.4	30.3	41.4	90.7	63.1	64.5	63.1
SigLIP2	ViT-B/16	49.2	72.7	62.5	83.1	71.0	71.2	61.1	34.5	42.7	90.8	60.5	68.1	64.0
	ViT-L/16	54.5	74.7	70.2	85.3	78.9	75.0	70.5	39.8	44.4	93.0	62.9	70.3	68.3
RemoteCLIP	RN-50	23.2	89.4	42.3	63.8	45.9	51.4	67.1	16.3	35.0	92.8	54.0	49.5	52.5
	ViT-B/32	34.5	**91.5**	56.9	77.5	59.3	67.9	62.6	26.2	39.7	**95.1**	63.9	59.8	61.2
	ViT-L/14	42.9	84.2	61.6	82.5	64.1	**75.4**	62.3	33.9	49.3	90.2	**67.0**	64.9	64.9
GeoRSCLIP	ViT-B/32	47.6	72.1	64.1	79.8	76.6	69.2	66.4	29.9	46.4	91.3	60.3	**77.0**	65.1
	ViT-L/14	64.3	76.9	67.8	**86.5**	80.9	73.9	**78.1**	41.4	**55.7**	94.1	63.7	72.4	**71.3**
CLIP LAION-RS	ViT-L/14	68.8	72.5	66.1	82.4	77.8	74.1	68.4	42.3	50.4	89.0	66.6	68.7	68.9
SkyCLIP	ViT-L/14	**69.7**	72.2	68.3	83.5	80.6	73.2	74.9	**48.5**	53.1	89.2	66.5	69.1	70.7

corresponding results for the Simple Template (`ST`) and the remaining LLMs (`DeepSeek`, `Qwen-2.5`, `Mixtral`, `RS-LLaVA`) across all models and datasets.

From the *model perspective*, remote sensing-adapted VLMs consistently achieve stronger performance than general-purpose models, regardless of the description generation strategy. This trend is especially pronounced for recent domain-specialized models such as GeoRSCLIP and SkyCLIP, which deliver the most consistent results. Among the general-purpose VLMs, SigLIP2 and MetaCLIP stand out as the most competitive alternatives.

From the *description-generation perspective*, the template-based baselines provide strong and reliable reference points. In particular, `DST` consistently outperforms `ST`, highlighting the importance of using diverse remote-sensing-aware templates rather than a single generic formulation. On the contrary, the LLM-based strategy does not yield a consistent improvement over template-based descriptions. In fact, it leads to performance degradation for the majority of dataset-model combinations. This behavior can be attributed to the fact that, although LLM-generated descriptions introduce

Table 2: Zero-shot classification accuracy (%) on all evaluated remote sensing datasets under the meta-prompting approach with `LLaMA-3.1`. Results are reported across the full set of vision-language models and backbone architectures. Best results per dataset are in **bold**, second best are underlined.

Model	Arch	EuroSAT	AID	MLRSNet	Optimal-31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHU-RS19	SIRI-WHU	RSSCN7	Avg
CLIP	ViT-B/16	52.5	69.3	56.5	74.5	66.6	65.8	63.6	28.6	37.2	86.3	52.9	70.1	60.3
	ViT-L/14	59.2	70.3	62.7	78.5	74.6	69.5	61.8	32.7	41.2	88.2	57.8	67.8	63.7
MetaCLIP-400M	ViT-B/16	54.0	67.8	59.3	79.3	70.7	68.6	70.1	34.8	41.7	87.2	57.9	73.5	63.7
	ViT-L/14	55.4	71.4	62.6	79.7	77.1	71.5	68.9	30.8	41.0	89.5	62.6	74.2	65.4
MetaCLIP-2.5B	ViT-B/16	44.9	69.2	62.3	78.8	73.4	70.5	66.7	35.4	40.0	89.4	57.3	74.2	63.5
	ViT-L/14	59.4	74.6	66.7	83.8	80.6	73.2	66.4	40.7	39.0	90.9	62.3	72.6	67.5
SigLIP	ViT-B/16	38.5	69.9	58.3	76.5	66.2	65.8	62.9	30.3	41.5	83.3	55.6	67.1	59.7
	ViT-L/16	49.3	76.1	64.7	81.0	71.3	72.0	72.9	33.2	48.7	89.8	58.3	71.7	65.7
SigLIP2	ViT-B/16	41.7	71.8	62.5	80.9	72.1	70.3	64.2	33.2	44.8	87.4	57.2	69.1	62.9
	ViT-L/16	55.0	76.2	68.1	83.1	76.8	74.4	72.8	37.8	47.9	91.7	60.4	71.1	67.9
RemoteCLIP	RN-50	26.1	81.9	45.2	69.5	48.8	56.5	62.3	22.0	31.8	92.9	54.7	65.7	54.8
	ViT-B/32	38.0	**87.3**	58.2	79.9	59.5	72.5	65.5	27.7	39.0	**95.3**	68.3	71.6	63.6
	ViT-L/14	45.8	80.9	60.7	81.2	63.6	74.1	67.9	31.4	51.0	91.2	**72.6**	68.0	65.7
GeoRSCLIP	ViT-B/32	52.8	72.0	65.7	80.6	78.0	71.0	72.7	34.1	48.6	89.9	60.6	**79.7**	67.1
	ViT-L/14	69.2	74.5	**70.2**	**84.9**	80.0	**74.6**	**74.4**	41.3	**53.5**	90.6	61.1	77.5	**71.0**
CLIP LAION-RS	ViT-L/14	**69.9**	70.7	63.5	79.0	74.6	71.3	65.1	39.8	48.9	89.5	60.8	72.8	67.2
SkyCLIP	ViT-L/14	68.5	72.5	68.9	84.1	**82.1**	74.0	72.5	**43.9**	50.8	90.9	61.1	72.2	70.1

semantically richer and more diverse information, they also introduce variability that is not always aligned with the discriminative structure of the downstream visual categories. In practice, these descriptions may contain overly generic attributes, stylistic variation, or semantically adjacent but discriminatively irrelevant details, which can blur the resulting class prototypes in the shared embedding space. As a result, the additional semantic richness of LLM-generated text does not necessarily translate into better alignment with visual representations, and often acts as a source of noise for zero-shot classification.

Figure 3 provides a complementary view to the aforementioned classification results by reporting, for each dataset, the performance of GeoRSCLIP ViT-L/14 model, across all evaluated description-generation strategies. For the LLM-based approaches, each score corresponds to the average classification accuracy obtained from three independently generated description sets per LLM, while the error bars denote the associated standard deviation. This protocol is adopted to account for the intrinsic stochasticity of LLM-based

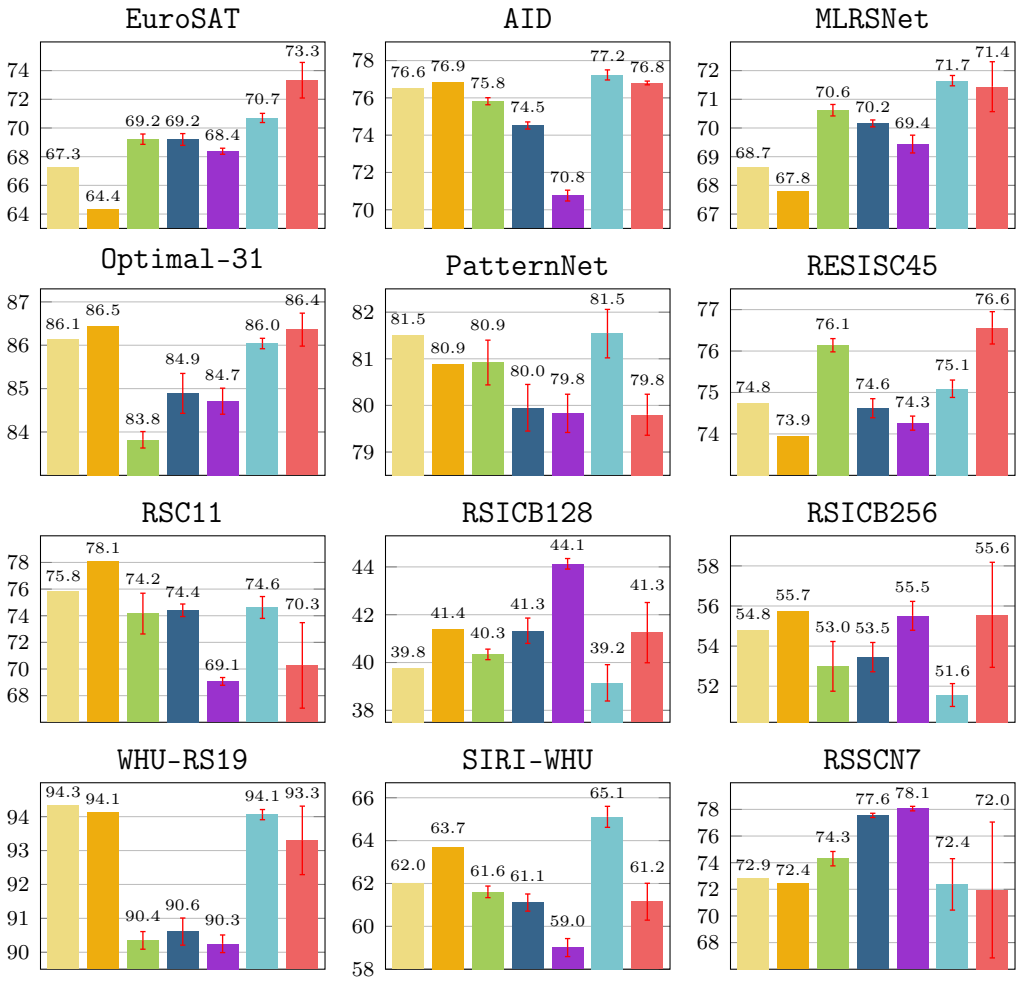

Figure 3: Dataset-wise comparison of template-based (ST, DST) and meta-prompting approaches (DeepSeek, LLaMA-3.1, Qwen-2.5, Mixtral, RS-LLaVA) for zero-shot classification using GeoRSCLIP ViT-L/14. For the meta-prompting approaches the values correspond to the mean classification accuracy over three independently generated description sets, with error bars indicating the associated standard deviation.

generation and to provide a more reliable estimate of the performance variability induced by different sampled description sets.

Although the aggregated results indicate that template-based strategies generally outperform LLM-based ones on average across datasets, Figure 3 shows that this behavior is dataset-dependent. In particular, the number of cases in which DST clearly surpasses all LLM-based approaches, is relatively

limited (e.g., RSC11). For the majority of datasets, the best-performing description source varies across LLMs, indicating that no single LLM consistently dominates across benchmarks. This dataset-specific behavior is illustrated by several contrasting examples. Qwen-2.5 for instance, achieves the highest performance on RSICB128 and RSSCN7, while yielding the weakest results on PatternNet, WHU-RS19, and SIRI-WHU. This instability is further reflected in the relatively large standard deviations observed for several LLMs, most notably RS-LLaVA, suggesting that the effectiveness of meta-prompting is highly contingent on the dataset and prompt formulation, in contrast to the more stable behavior exhibited by template-based formulations.

Table 3: Effect of image-side embedding centering on average zero-shot classification accuracy (%) across all evaluated datasets and description-generation strategies. Reported accuracies correspond to the centered setting, while Δ denotes the relative performance change with respect to the vanilla, non-centered, image embeddings. Positive values indicate gains and negative values indicate degradation.

Model	Arch	ST		DST		LLaMA-3.1		DeepSeek		Qwen-2.5		Mixtral		RS-LLaVA	
		centered	Δ	centered	Δ	centered	Δ	centered	Δ	centered	Δ	centered	Δ	centered	Δ
CLIP	ViT-B/16	62.1	+4.8	62.5	+2.5	63.7	+3.4	63.9	+4.3	63.9	+4.6	62.3	+3.0	62.4	+3.9
	ViT-L/14	66.4	+2.3	66.7	+0.3	68.5	+4.8	68.2	+3.9	69.1	+4.6	67.7	+2.3	68.1	+4.0
MetaCLIP-400M	ViT-B/16	64.5	+2.2	65.7	+2.5	66.3	+2.6	65.8	+3.2	66.5	+4.1	64.2	+3.2	64.3	+2.7
	ViT-L/14	65.6	+2.3	67.4	+2.1	68.3	+2.9	66.9	+3.1	68.2	+3.6	66.2	+1.8	66.2	+2.8
MetaCLIP-2.5B	ViT-B/16	65.7	+3.9	65.6	+1.5	66.5	+3.0	65.9	+4.2	66.5	+3.7	64.9	+3.7	64.8	+2.0
	ViT-L/14	71.2	+3.5	71.1	+2.6	71.3	+3.8	70.5	+3.9	71.4	+5.0	69.1	+3.6	69.9	+3.6
SigLIP	ViT-B/16	58.9	-0.6	59.1	-0.4	61.8	+2.2	61.3	+2.5	61.4	+3.2	60.0	+2.7	60.9	+1.7
	ViT-L/16	65.3	+2.0	65.0	+1.9	67.9	+2.2	67.6	+3.1	67.4	+2.4	66.1	+3.8	67.1	+2.5
SigLIP2	ViT-B/16	65.4	+1.9	65.4	+1.4	65.3	+2.4	65.1	+2.9	65.1	+2.6	63.7	+3.1	64.0	+1.2
	ViT-L/16	70.5	+0.7	70.0	+1.7	71.2	+3.3	70.2	+3.5	70.9	+4.0	69.0	+2.8	69.4	+1.9
RemoteCLIP	RN-50	53.3	+1.9	55.5	+3.0	57.1	+2.3	57.2	+1.0	56.1	+1.6	57.0	+1.0	55.6	+2.0
	ViT-B/32	61.8	+1.2	63.1	+1.9	64.8	+1.3	64.9	+1.1	64.0	+1.4	64.7	-0.1	64.5	+0.8
	ViT-L/14	66.3	+1.9	67.2	+2.3	68.8	+3.1	68.3	+3.7	68.7	+3.0	67.4	+2.4	67.6	+3.4
GeoRSCLIP	ViT-B/32	66.5	+2.2	66.3	+1.3	69.0	+1.9	67.9	+1.1	68.8	+2.9	66.3	-0.1	66.8	+0.5
	ViT-L/14	71.7	+0.5	71.3	+0.0	74.1	+3.1	73.7	+2.8	74.0	+3.7	72.1	+0.5	73.5	+2.0
CLIP LAION-RS	ViT-L/14	68.6	+0.1	68.7	-0.3	70.2	+3.0	69.4	+2.4	70.7	+3.0	68.9	+1.9	69.7	+2.6
SkyCLIP	ViT-L/14	69.4	-1.2	69.5	-1.2	71.0	+0.9	70.5	-0.1	71.5	+0.7	70.2	-0.6	71.1	+0.4

Embedding Calibration. Table 3 reports the effect of image-side embedding centering in the zero-shot classification task, averaged across all evaluated datasets. For each vision-language model and description-generation strategy, we report both the classification accuracy obtained after centering and the corresponding performance change ($\mathbf{\Delta}$) relative to the vanilla setting.

The results indicate that image centering constitutes a simple yet highly effective calibration step (as shown in Figure 1). Performance gains are widespread, reaching up to $+\mathbf{4.8}\%$. At the same time, performance degradation is relatively rare and is mostly confined to template-based descrip-

tions. For example, in the case of SkyCLIP, both ST and DST decrease by approximately **1.2**%, while DeepSeek and Mixtral exhibit smaller drops of about **0.1**% and **0.6**%, respectively. In contrast, LLaMA-3.1, Qwen-2.5, and RS-LLaVA all benefit from centering. This pattern indicates that image centering improves alignment in a highly consistent manner, with only a few isolated exceptions.

An important observation is that embedding calibration changes the relative ranking between template-based and LLM-based descriptions. Feature centering acts as a significant enabler for high-capacity models like LLaMA-3.1 and Qwen-2.5 (and DeepSeek to a lesser degree), making the LLM-generated descriptions consistently surpass ST and DST. This suggests that a considerable part of the underperformance of meta-prompting in the vanilla setting depends on how their richer and more variable semantics interact with the uncalibrated embedding space. Centering, on the other hand, effectively improves the alignment between image embeddings and the more diverse textual prototypes produced by LLMs. We note that Mixtral and RS-LLaVA lag behind other LLMs even post-calibration and frequently trail the simple ST baseline (e.g., under SigLIP2 ViT-L/16), suggesting an inherent stylistic or structural incompatibility that geometric centering alone cannot resolve.

Reading the table column-wise shifts the focus to architectural capacity and highlights the definitive advantage of remote-sensing-adapted VLMs over general-purpose alternatives. Within any given prompting strategy or calibration state, domain-specific architectures like RemoteCLIP, GeoRSCLIP, and SkyCLIP consistently outperform standard baselines like CLIP or MetaCLIP. This margin is most pronounced for larger baselines. For instance, under the calibrated Qwen-2.5 text generation, GeoRSCLIP ViT-L/14 achieves 74.0%, a 4.9% improvement over the general-purpose CLIP ViT-L/14 counterpart (69.1%). Overall, the results suggest that pre-training on domain-specific imagery establishes a fundamentally more resilient embedding geometry for downstream zero-shot tasks.

5.2. Text-to-Image Retrieval

In this section, we evaluate text-to-image retrieval task, using the mean Average Precision (mAP) metric. Table 4 summarizes the average retrieval performance across the evaluated datasets, for every vision-language model and description-generation strategy. We report results both before and after text-side embedding centering, together with the corresponding performance differences ($\mathbf{\Delta}$). In this task, each aggregated class prototype is used as a text

query to rank the image gallery. Therefore, the query modality corresponds to the text side.

Table 4: Effect of text-side embedding centering on average text-to-image retrieval performance across all evaluated datasets. We report the mAP metric before (mAP) and after (mAP_c) calibration for each description-generation strategy. Δ denotes the relative performance change with respect to the vanilla, non-centered, text embeddings.

Model	Arch	ST			DST			LLaMA-3.1			DeepSeek			Qwen-2.5			Mixtral			RS-LLaVA		
		mAP	mAP_c	Δ	mAP	mAP_c	Δ	mAP	mAP_c	Δ	mAP	mAP_c	Δ	mAP	mAP_c	Δ	mAP	mAP_c	Δ	mAP	mAP_c	Δ
CLIP	ViT-B/16	51.1	62.5	+11.4	52.9	63.5	+10.6	51.8	63.9	+12.1	51.0	64.6	+13.6	50.9	63.7	+12.8	52.7	64.4	+11.7	50.0	64.3	+14.3
	ViT-L/14	55.4	68.0	+12.6	59.2	68.4	+9.2	54.0	69.0	+15.0	53.5	69.5	+16.0	53.6	68.8	+15.2	55.4	69.4	+14.0	52.7	69.7	+17.0
MetaCLIP-400M	ViT-B/16	55.6	64.7	+9.1	59.3	65.2	+5.9	55.6	65.8	+10.2	56.0	66.5	+10.5	54.5	65.5	+11.0	58.6	65.5	+6.9	55.0	66.2	+11.2
	ViT-L/14	56.9	65.9	+9.0	58.5	66.9	+8.4	55.5	68.0	+12.5	55.4	68.3	+12.9	54.6	67.8	+13.2	57.0	67.6	+10.6	53.9	68.0	+14.1
MetaCLIP-2.5B	ViT-B/16	57.6	66.4	+8.8	61.5	66.5	+5.0	57.3	67.0	+9.7	57.3	67.6	+10.3	55.8	66.5	+10.7	60.2	67.2	+7.0	56.7	67.4	+10.7
	ViT-L/14	62.2	70.8	+8.6	63.5	71.0	+7.5	59.7	71.2	+11.5	59.6	71.9	+12.3	58.7	70.6	+11.9	62.5	71.0	+8.5	58.8	71.2	+12.4
SigLIP	ViT-B/16	49.6	61.0	+11.4	54.9	61.2	+6.3	50.8	61.6	+10.8	50.8	62.1	+11.3	49.8	61.6	+11.8	52.9	61.5	+8.6	49.5	62.1	+12.6
	ViT-L/16	54.7	65.7	+11.0	58.6	65.9	+7.3	56.0	67.5	+11.5	55.8	67.9	+12.1	55.1	67.3	+12.2	57.2	67.0	+9.8	54.9	62.6	+7.7
SigLIP2	ViT-B/16	56.9	65.8	+8.9	61.8	66.5	+4.7	54.9	65.7	+10.8	55.3	66.4	+11.1	53.3	65.8	+12.5	59.1	65.8	+6.7	55.4	66.4	+11.0
	ViT-L/16	61.1	70.5	+9.4	64.1	70.4	+6.3	57.8	70.9	+13.1	57.5	70.8	+13.3	56.5	70.7	+14.2	60.2	70.0	+9.8	56.8	70.9	+14.1
RemoteCLIP	RN-50	43.6	56.7	+13.1	52.5	58.8	+6.3	41.5	57.8	+16.3	41.8	60.0	+18.2	38.5	57.7	+19.2	45.0	60.1	+15.1	42.1	58.7	+16.6
	ViT-B/32	52.7	65.1	+12.4	54.3	66.1	+11.8	49.2	66.8	+17.6	49.4	67.4	+18.0	46.6	66.8	+20.2	51.5	67.2	+15.7	49.4	67.3	+17.9
	ViT-L/14	54.3	70.0	+15.7	54.5	70.3	+15.8	50.9	69.8	+18.9	50.9	70.5	+19.6	49.9	69.6	+19.7	51.8	70.3	+18.5	50.5	70.2	+19.7
GeoRSCLIP	ViT-B/32	62.0	67.7	+5.7	64.7	67.8	+3.1	57.9	67.4	+9.5	58.4	68.3	+9.9	55.9	67.3	+11.4	60.7	68.3	+7.6	59.2	68.4	+9.2
	ViT-L/14	68.1	72.4	+4.3	69.8	72.6	+2.8	63.3	72.2	+8.9	64.2	**73.0**	+8.8	61.8	72.3	+10.5	**67.3**	73.1	+5.8	65.8	**73.7**	+7.9
CLIP LAION-RS	ViT-B/32	65.5	71.2	+5.7	67.4	71.3	+3.9	62.6	70.7	+8.1	62.6	71.4	+8.8	61.7	70.8	+9.1	65.0	71.5	+6.5	63.0	71.8	+8.8
SkyCLIP	ViT-L/14	**68.9**	**72.7**	+3.8	**70.7**	**72.9**	+2.2	**66.5**	**72.6**	+6.1	**66.2**	73.0	+6.8	**65.9**	**72.7**	+6.8	66.9	**73.1**	+6.2	**66.4**	73.4	+7.0

Similarly to the classification task, remote sensing-adapted VLMs remain the most robust performers and consistently outperform general-purpose models across all description-generation strategies. This advantage is even more pronounced in retrieval, where models such as RemoteCLIP, GeoRSCLIP, CLIP-LAION-RS, and SkyCLIP define the strongest performance regime. Among the general-purpose VLMs, the MetaCLIP family emerges as the most competitive alternative. However, a substantial performance gap relative to the domain-adapted models still remains.

Additionally, text-to-image retrieval exhibits substantially larger gains from embedding calibration, with improvements reaching more than **+19**%. Crucially, centering does not cause performance degradation for any model or description-generation strategy. This highlights the strong sensitivity of retrieval performance to the geometry of the text embedding space and further confirms the effectiveness and robustness of the proposed calibration step. Both template- and LLM-based descriptions benefit from text-side centering, but the gains are significantly larger for the meta-prompting approaches. This indicates that the richer yet more dispersed semantic structure induced by meta-prompting is especially sensitive to embedding-space biases. By removing dominant shared directions from the text side, centering improves the compatibility between class prototypes and image embeddings, yielding stronger retrieval performance.

5.3. Experimental Analysis

Log-likelihood. To further analyze the geometry of the generated textual descriptions, we compute a text-space log-likelihood score following the formulation of [21]. Specifically, we evaluate the DST baseline and all LLM-based meta-prompting strategies by encoding their text descriptions with CLIP ViT-L/14, then applying the same centering and whitening transformation proposed in [21]. Let $\mathbf{t} \in \mathbb{R}^D$ denote a text embedding, $\boldsymbol{\mu}$ the reference mean feature vector, and $\mathbf{W}$ the whitening matrix. The whitened text representation is defined as

$$\tilde{\mathbf{t}} = (\mathbf{t} - \boldsymbol{\mu})\mathbf{W}. \tag{1}$$

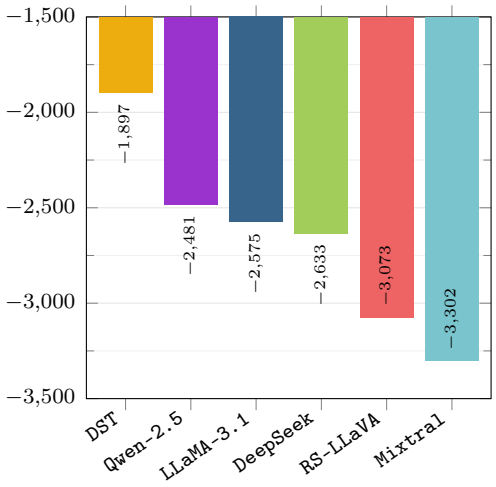

Figure 4: Average text log-likelihood across all evaluated datasets in the CLIP ViT-L/14 text feature space for the different description-generation strategies. Higher values indicate that the corresponding generated descriptions lie in higher-density regions of the whitened text embedding space.

Assuming a standard Gaussian distribution in the whitened space, the log-likelihood of $\mathbf{t}$ is given by:

$$\log p(\mathbf{t}) = -\frac{1}{2}\left(D\log(2\pi) + \|\tilde{\mathbf{t}}\|^2\right). \tag{2}$$

For the LLM-based approaches, the value shown in Figure 4 corresponds to the average log-likelihood across all evaluated datasets. As shown in Figure 4, DST achieves the highest average log-likelihood, whereas all meta-prompting sources produce substantially lower values. This supports the view that, despite their greater semantic richness, LLM-generated descriptions deviate more from the high-density regions of the CLIP text manifold, which is consistent with the noisier and less stable downstream behavior observed in the main evaluation.

Performance distributions. Figures 5 and 6 visualize the distributions of downstream performance before and after query-side embedding centering for zero-shot image classification and text-to-image retrieval, respectively. Across both figures, the post-centering distributions are shifted to the right

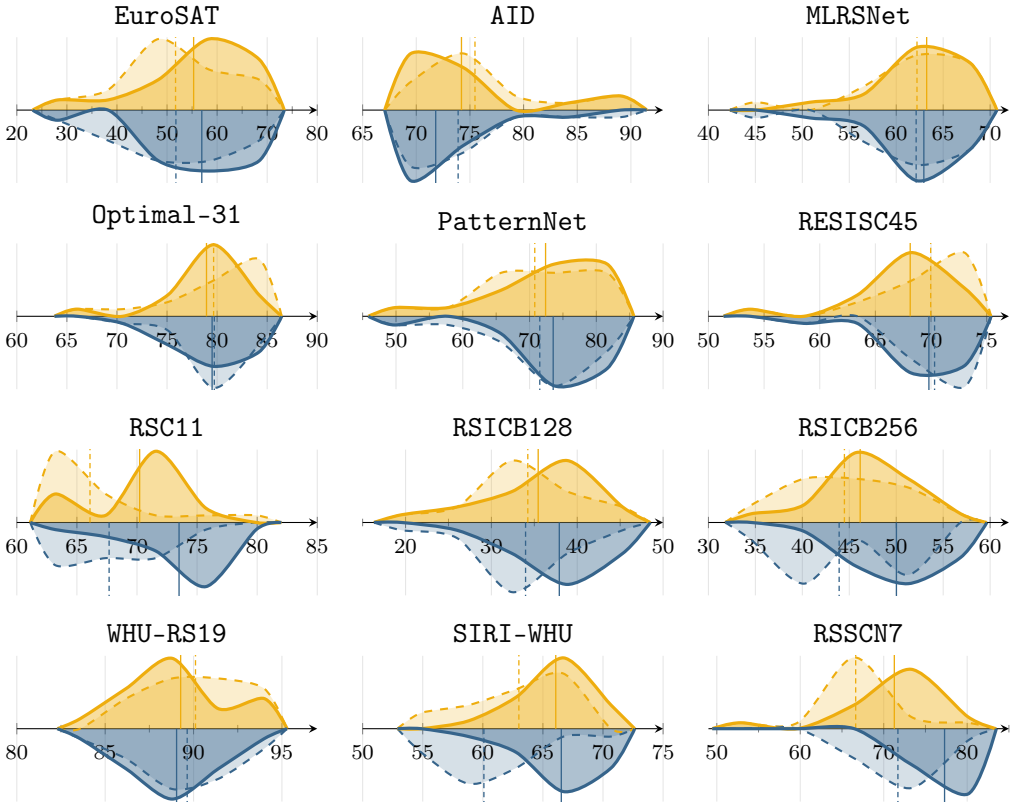

Figure 5: Dataset-wise fixed-bin histograms of zero-shot classification accuracy, across all evaluated models for DST and LLaMA-3.1 generated descriptions. Dashed: results under the non-centered setting, solid: results under the image-side embedding centering.

for the vast majority of datasets, indicating systematic improvements in both classification accuracy and retrieval mAP. Beyond the improvement in average performance, centering often shifts the entire distribution toward a higher-performing regime, with the post-centering curves in many cases becoming not only right-shifted but also more concentrated. This indicates that the effect of centering is broad and stabilizing, improving performance consistently across model and backbone instantiations rather than merely benefiting a small number of favorable cases. The improvement is generally more pronounced for LLaMA-3.1 than for DST, particularly in the retrieval setting, where the blue solid distributions frequently exhibit larger displacements than the yellow ones. At the same time, the magnitude of the shift remains dataset-dependent, with some benchmarks showing moderate gains and others exhibiting a much stronger redistribution toward higher scores.

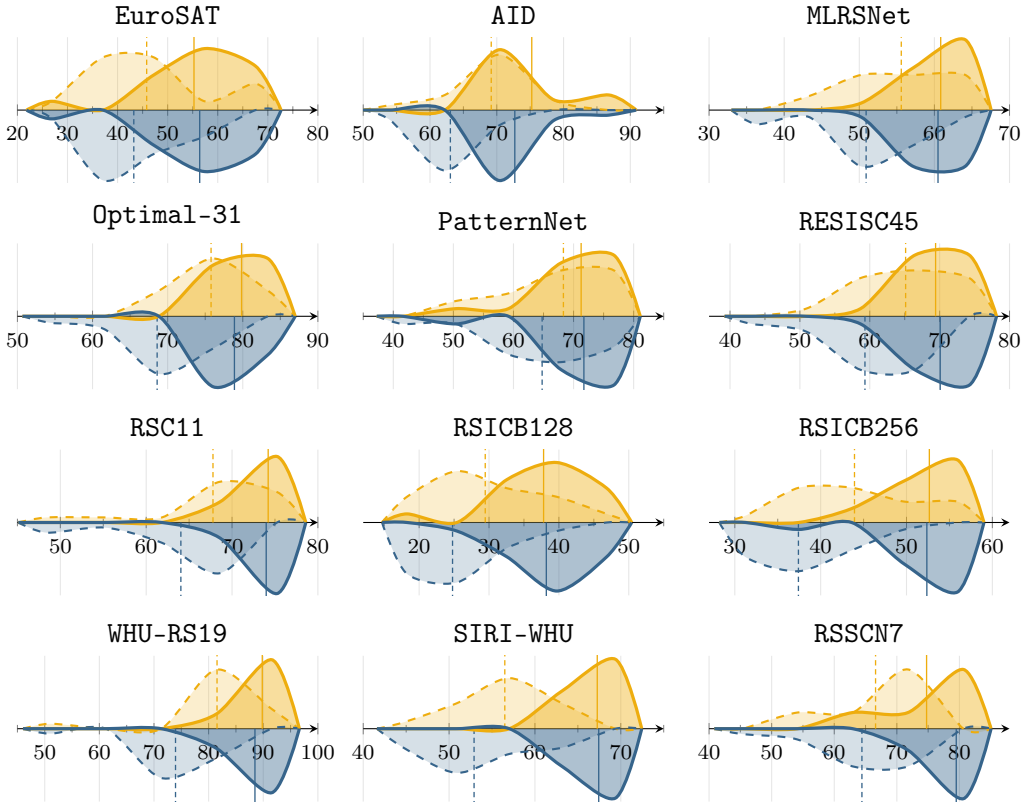

Figure 6: Dataset-wise fixed-bin histograms of text-to-image retrieval mAP, across all evaluated models for DST and LLaMA-3.1 generated descriptions. Dashed: results under the non-centered setting, solid: results under the text-side embedding centering.

5.4. Ablation Study

Top-k voting. We further investigate an alternative strategy for zero-shot image classification based on top-k voting. Instead of averaging all description embeddings of a class into a single prototype, we retain the individual text embeddings and, for each image, rank all candidate descriptions based on cosine similarity. The final class prediction is then obtained by majority voting among the classes associated with the top-k most similar text embeddings. Figure 7 reports the average classification accuracy across datasets for different values of k, considering both DST and LLaMA-3.1 descriptions across the evaluated VLMs. The optimal value of k is model-dependent, indicating that the contribution of this approach varies with the underlying geometry of each vision-language model. In the majority of cases, simple embedding averaging remains a strong upper reference, with top-k voting

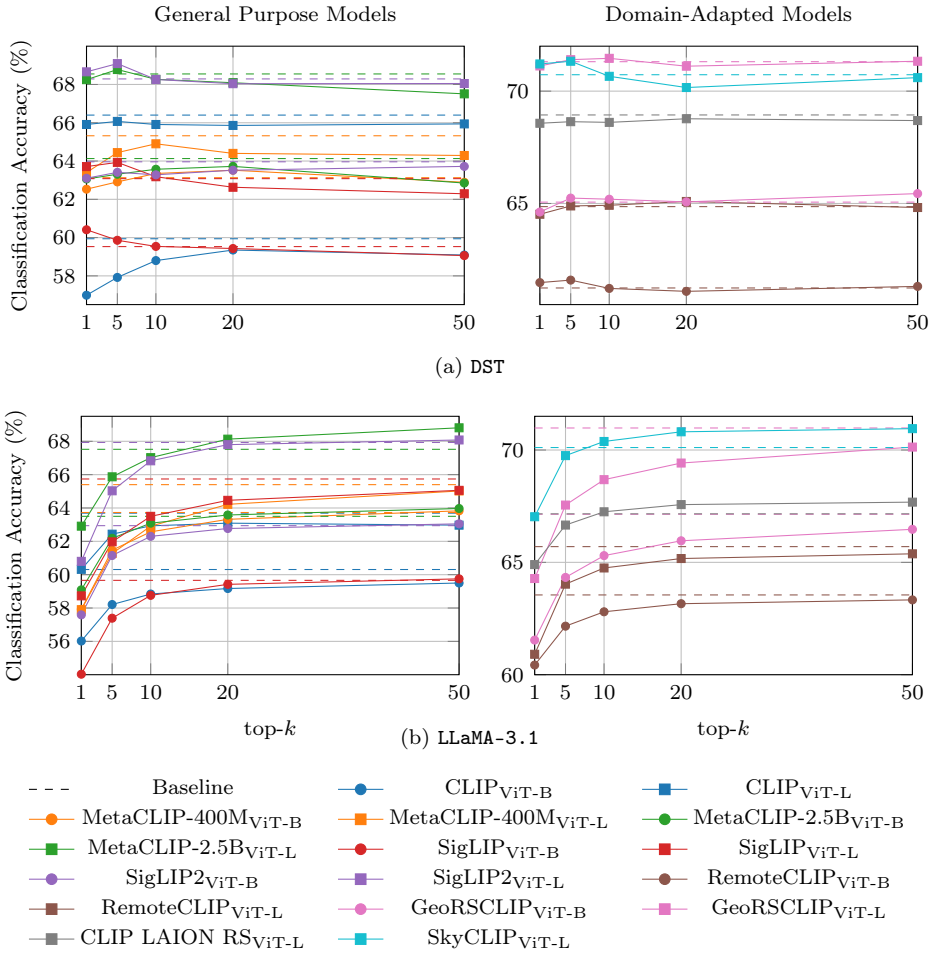

Figure 7: Zero-shot classification accuracy under the top-k voting scheme for the evaluated vision-language models. Solid curves report performance for different values of k, while the dashed horizontal lines indicate the corresponding average accuracy achieved by the default prompt aggregation approach.

yielding at most modest gains for the best selected k. Overall, the results indicate that although top-k voting can occasionally yield marginal gains, it does not consistently outperform simple prototype averaging, whose stability and effectiveness make it the preferred default choice.

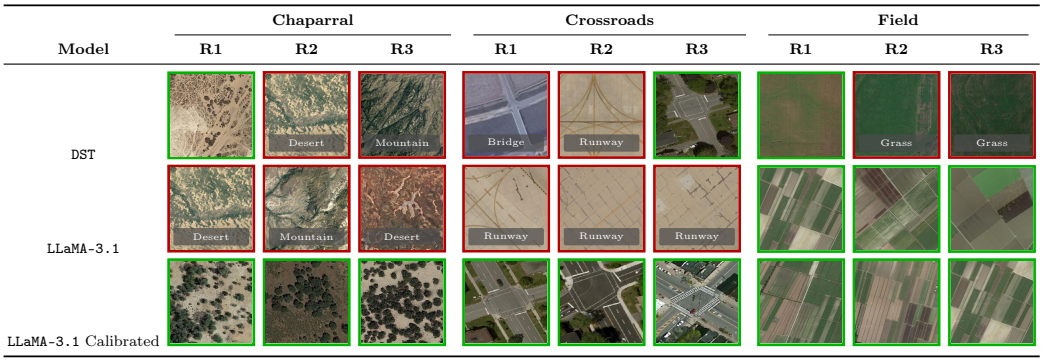

Figure 8: Qualitative text-to-image retrieval examples for the GeoRSCLIP ViT-L/14 vision-language model. For each class query, we show the top-3 retrieved images under three prompting variants: `DST`, `LLaMA-3.1`, and `LLaMA-3.1` with calibration. The selected classes (*Chaparral*, *Crossroads*, and *Field*) originate from different datasets. Green borders denote correct retrievals, whereas red borders indicate false positives.

6. Qualitative Results

Qualitative analysis of embedding calibration. Figures 8 and 9 present qualitative text-to-image retrieval examples for the GeoRSCLIP ViT-L/14 vision-language model by showing the top-3 retrieved images for three prompting variants, namely `DST`, `LLaMA-3.1`, and `LLaMA-3.1` with calibration. A clear visual trend emerges across all illustrated classes: the calibrated `LLaMA-3.1` setup yields the most consistent retrieval behavior, returning correct top-3 matches throughout the presented examples. In contrast, both `DST` and the uncalibrated `LLaMA-3.1` setup produce multiple false positives, as indicated by the red borders. Importantly, the classes shown in these figures are drawn from different datasets, suggesting that the improvement is not specific to a single benchmark but instead reflects a more general enhancement in the compatibility between text queries and image embeddings. Overall, these qualitative examples are in strong agreement with the quantitative retrieval results, showing that meta-prompt-generated descriptions become substantially more reliable and precise once calibration is applied.

Qualitative comparison across vision-language models. Figures 10 and 11 further examine qualitative text-to-image retrieval under the strongest prompting configuration, namely `LLaMA-3.1` with calibration, by comparing the top-3 retrieved images across different vision-language models. Specifically, the figures report results for CLIP ViT-L, MetaCLIP-2.5B ViT-L, SigLIP2 ViT-

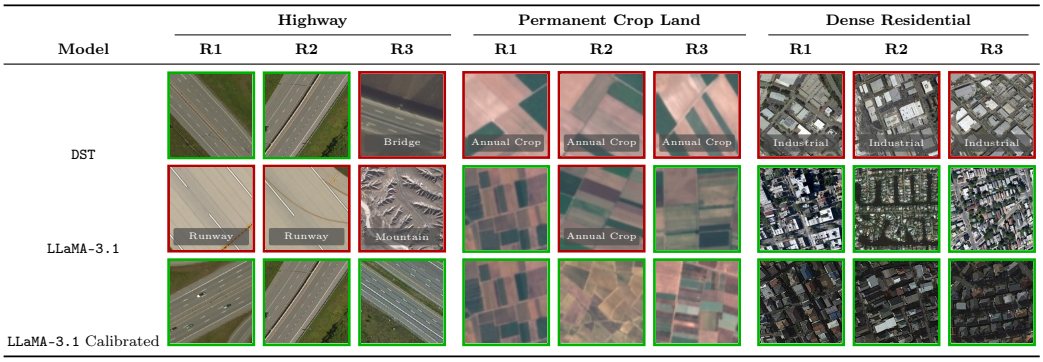

Figure 9: Additional qualitative text-to-image retrieval examples for the GeoRSCLIP ViT-L/14 vision-language model, showing the top-3 retrieved images for `DST`, `LLaMA-3.1`, and `LLaMA-3.1` with calibration. The displayed classes (*Highway*, *Permanent Crop Land*, and *Dense Residential*) are drawn from different datasets. Green borders indicate correct retrievals and red borders indicate incorrect ones.

L, RemoteCLIP ViT-L, GeoRSCLIP ViT-L, and SkyCLIP ViT-L on classes drawn from different datasets. Even under this shared and highly consistent retrieval setup, clear differences remain across models. In particular, the remote-sensing-adapted backbones exhibit more stable and semantically precise retrieval behavior overall, with GeoRSCLIP ViT-L and SkyCLIP ViT-L emerging as the most consistent performers across the illustrated examples. By contrast, the general-purpose models show a larger number of residual false positives or less reliable ranking of the top retrieved images. These examples therefore complement the quantitative results by showing that, once calibration reduces prompt-side ambiguity, the quality of retrieval is increasingly governed by the underlying visual-language representation itself.

7. Discussion

Our study shows that semantic richness, by itself, is not sufficient to guarantee stronger zero-shot performance. In the remote sensing setting, richer LLM-generated descriptions can expand the semantic coverage of a class, but they also introduce variability that is not always well aligned with the frozen vision-language shared embedding space. This suggests that the benefit of richer language supervision is closely tied to the underlying learning regime. An important direction for future work is therefore to examine how semantic richness behaves beyond the strictly zero-shot setting, for example in few-shot scenarios or under trainable adaptation strategies that explicitly

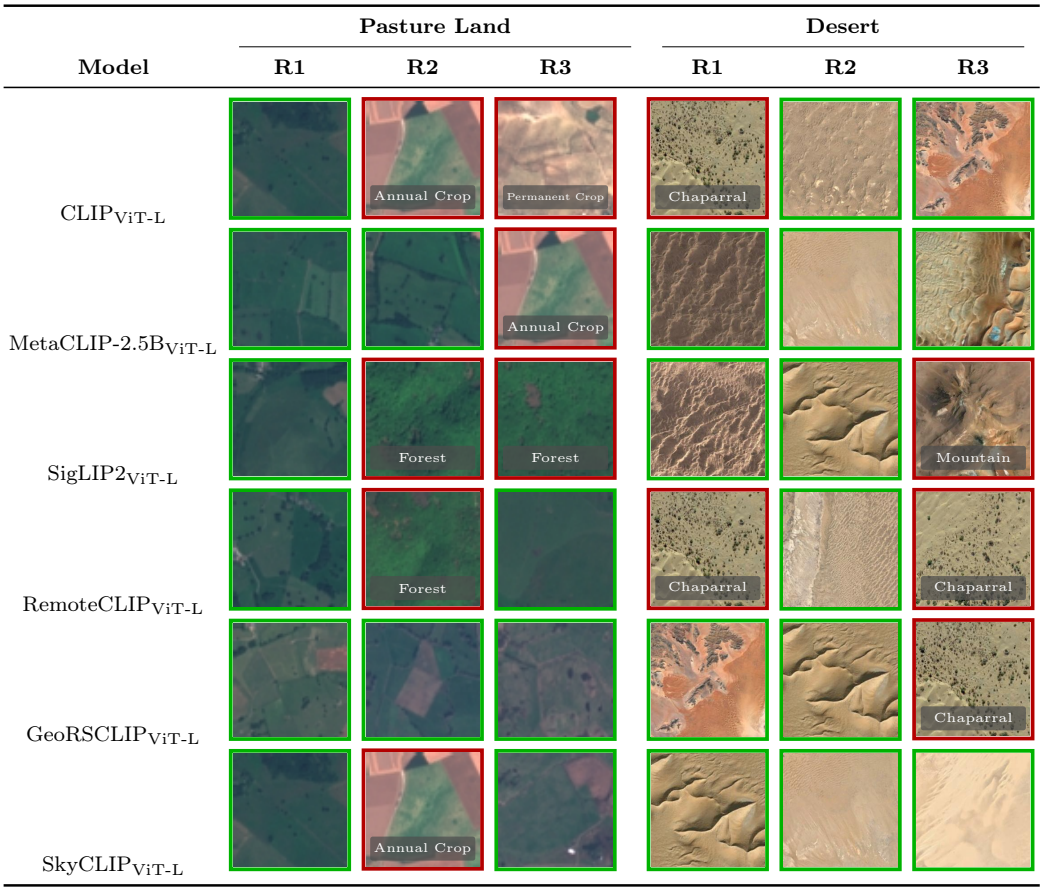

Figure 10: Qualitative comparison of text-to-image retrieval across different vision-language models under the `LLaMA-3.1` with calibration setup. For each class query, we show the top-3 retrieved images for CLIP ViT-L, MetaCLIP-2.5B ViT-L, SigLIP2 ViT-L, RemoteCLIP ViT-L, GeoRSCLIP ViT-L, and SkyCLIP ViT-L. The selected classes (*Pasture Land* and *Desert*) come from different datasets. Green borders denote correct retrievals, while red borders indicate incorrect ones.

align longer and more expressive textual descriptions with image embeddings in the shared representation space.

8. Conclusion

In this work, we revisit zero-shot remote sensing through a systematic study of meta-prompting, prototype construction, and embedding calibration across 12 remote sensing benchmarks, multiple vision-language model

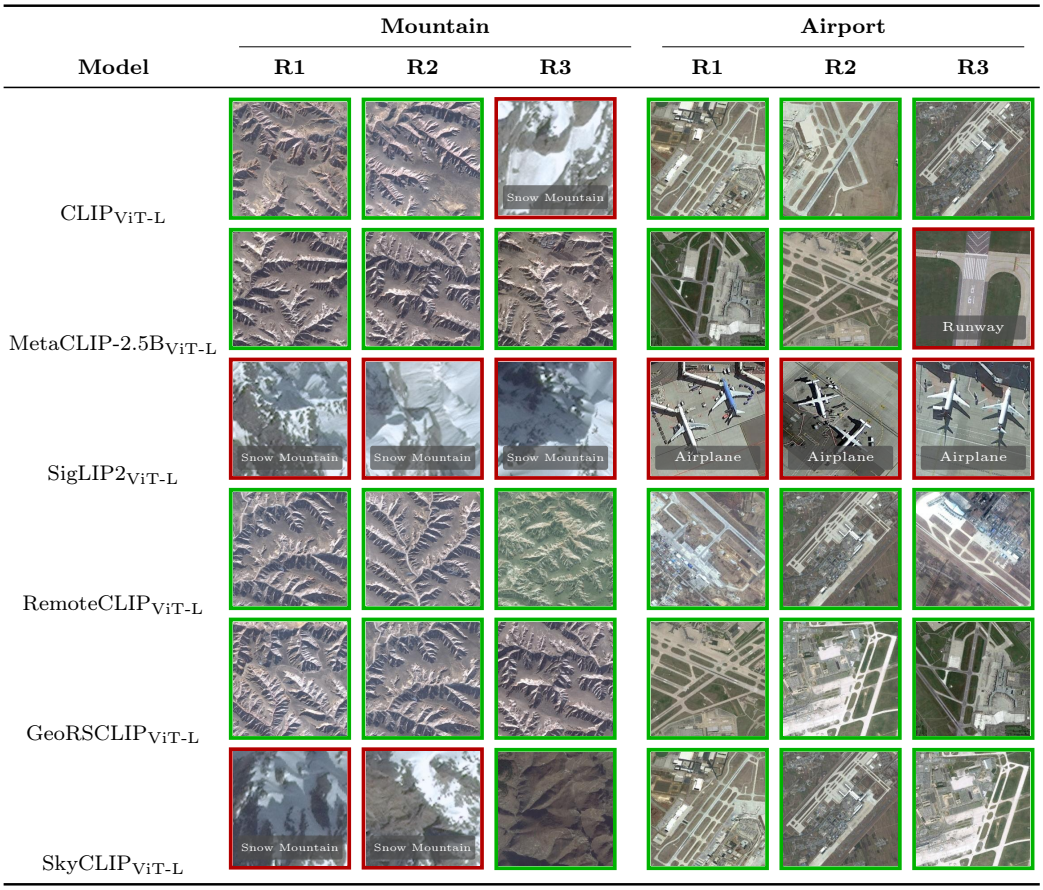

Figure 11: Additional qualitative comparison of text-to-image retrieval across different vision-language models using `LLaMA-3.1` with calibration. We report the top-3 retrieved images for CLIP ViT-L, MetaCLIP-2.5B ViT-L, SigLIP2 ViT-L, RemoteCLIP ViT-L, GeoRSCLIP ViT-L, and SkyCLIP ViT-L on the classes *Mountain* and *Airport*, which are drawn from different datasets. Green borders indicate correct retrievals and red borders indicate false positives.

families, and 5 open-source LLMs. Our results show that, although LLM-generated descriptions provide richer semantic content, they do not consistently improve downstream performance and often introduce additional variability in the text embedding space, whereas simple domain-specific templates remain strong and robust baselines. At the same time, we demonstrate that a significant part of this limitation is not only due to the generated descriptions themselves, but also to how they interact with the shared embedding space: query-side centering yields broad and consistent gains,

with image-side centering improving zero-shot classification and text-side centering producing even stronger benefits for text-to-image retrieval. Across both tasks, remote-sensing-adapted VLMs consistently outperform general-purpose alternatives, confirming the value of domain-specialized representations for satellite imagery. In summary, our findings highlight a trade-off between semantic richness and robustness and show that simple embedding centering constitutes a highly effective mechanism for improving zero-shot remote sensing performance.

9. Acknowledgments

This work was supported by the Edge SpAIce Horizon Europe project under Grant 101135358. We gratefully acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through an EuroHPC Development Access call.

10. Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT specifically to aid in the polishing and refinement of the text. LLMs were *not* used for ideation, technical contributions, experimental design, data analysis, or related work retrieval. All research ideas, methodology, experiments, and conclusions presented are solely the work of the authors. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A. Dataset Details

Table A.5 provides an overview of the 12 remote sensing scene classification datasets considered in this work, including the number of classes, number of images, image resolution, and when available, the corresponding ground sampling distance (GSD). Together, these benchmarks cover a wide spectrum of remote-sensing scenarios, with substantial variation in dataset size, spatial scale, and semantic complexity. This heterogeneity is particularly relevant for our evaluation setting, as it allows us to examine the robustness of zero-shot prompting and retrieval behavior across benchmarks with different visual characteristics and label-space structures.

Table A.5: Overview of remote sensing scene classification datasets used in our experiments.

Dataset	**# Classes**	**# Images**	**Resolution**	**GSD (m)**
RSSCN7 [22]	7	2,800	400×400	–
EuroSAT [23]	10	27,000	64×64	10
RSC11 [24]	11	1,232	512×512	0.2
SIRI-WHU [25]	12	2,400	200×200	2
WHU-RS19 [26]	19	1,005	600×600	≤ 0.5
AID [27]	30	10,000	600×600	0.5–8
Optimal-31 [28]	31	1,860	256×256	–
RSICB256 [29]	35	24,747	256×256	0.3–3
PatternNet [30]	38	30,400	256×256	0.06–4.7
RESISC45 [31]	45	31,500	256×256	0.2–30
RSICB128 [29]	45	36,707	128×128	0.3–3
MLRSNet [32]	46	109,161	256×256	0.1–10

Appendix B. Extended Results

Table B.6: Zero-shot classification accuracy (%) on all evaluated remote sensing datasets under the Simple Template (ST) baseline. Results are reported across the full set of vision-language models and backbone architectures.

Model	Arch	EuroSAT	AID	MLRSNet	Optimal-31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHU-RS19	SIRI-WHU	RSSCN7	Avg
CLIP[1]	ViT-B/16	44.1	66.6	54.4	70.1	62.9	62.1	62.9	26.9	38.4	82.8	52.2	64.3	57.3
	ViT-L/14	63.0	69.5	63.9	76.1	71.8	67.0	67.1	38.0	47.6	85.1	58.8	61.5	64.1
MetaCLIP-400M[4]	ViT-B/16	55.2	71.9	57.0	76.5	68.2	66.2	65.9	31.0	41.0	89.0	59.7	66.6	62.3
	ViT-L/14	52.6	69.3	60.1	80.0	74.5	70.0	55.2	33.4	45.3	88.8	65.9	64.5	63.3
MetaCLIP-2.5B[4]	ViT-B/16	46.6	71.0	62.5	76.0	70.0	68.5	61.7	31.1	40.5	88.2	62.9	63.1	61.9
	ViT-L/14	55.4	78.3	70.1	83.1	**83.5**	74.1	67.0	39.1	41.9	91.0	64.3	65.7	67.8
SigLIP[5]	ViT-B/16	45.0	70.9	57.3	74.5	63.8	64.4	64.5	24.2	38.1	84.4	62.2	65.6	59.6
	ViT-L/16	53.3	73.5	60.2	81.3	67.6	68.5	69.1	29.4	42.5	88.9	62.3	62.7	63.3
SigLIP2 [20]	ViT-B/16	46.4	75.3	63.5	82.6	71.6	71.1	61.9	30.3	43.2	88.8	61.8	65.6	63.5
	ViT-L/16	54.1	75.3	**73.3**	**87.1**	78.8	**77.8**	**77.8**	37.3	47.4	92.0	66.1	70.7	69.8
RemoteCLIP [10]	RN-50	20.5	87.2	41.7	63.0	44.1	49.7	63.8	16.3	33.2	93.6	55.7	48.3	51.4
	ViT-B/32	38.4	**89.3**	55.3	76.2	58.5	66.3	60.3	27.1	42.0	93.7	60.8	58.8	60.6
	ViT-L/14	43.1	82.3	62.0	82.3	64.1	75.1	62.6	34.4	48.9	86.7	**67.8**	63.2	64.4
GeoRSCLIP[11]	ViT-B/32	48.6	70.9	65.3	79.6	75.9	70.1	65.9	26.9	45.9	90.0	56.2	**76.1**	64.3
	ViT-L/14	67.3	76.6	68.7	86.1	81.5	74.8	75.8	39.8	**54.8**	**94.3**	62.0	72.9	**71.2**
CLIP LAION-RS [12]	ViT-L/14	**70.4**	72.8	66.5	80.5	76.7	72.0	70.9	41.7	50.6	89.7	64.2	66.5	68.5
SkyCLIP [12]	ViT-L/14	70.0	74.0	69.1	82.3	79.8	72.0	73.1	**49.2**	53.4	90.0	66.2	68.8	70.7

Table B.7: Zero-shot classification accuracy (%) on all evaluated remote sensing datasets under the meta-prompting approach with `DeepSeek`. Results are reported across the full set of vision-language models and backbone architectures.

Model	Arch	EuroSAT	AID	MLRSNet	Optimal-31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHU-RS19	SIRI-WHU	RSSCN7	Avg
CLIP	ViT-B/16	52.8	67.1	56.6	72.9	65.1	64.9	62.5	28.6	37.7	85.2	56.6	66.2	59.7
	ViT-L/14	64.5	70.9	64.0	78.0	73.3	70.1	62.0	32.3	43.7	87.4	61.8	63.0	64.3
MetaCLIP-400M	ViT-B/16	55.5	67.0	60.3	75.8	69.7	67.6	67.9	31.4	40.5	87.5	57.1	70.9	62.6
	ViT-L/14	53.0	70.8	63.2	76.7	75.7	71.4	61.1	30.5	40.4	88.7	64.2	69.9	63.8
MetaCLIP-2.5B	ViT-B/16	49.3	67.9	63.9	75.2	71.8	68.4	59.5	33.0	38.1	84.5	60.7	68.6	61.7
	ViT-L/14	63.0	73.0	67.1	80.8	79.5	72.8	63.2	39.5	41.9	87.5	62.6	68.3	66.6
SigLIP	ViT-B/16	44.6	67.9	58.7	73.6	63.7	66.4	62.1	25.9	37.8	80.9	58.1	65.9	58.8
	ViT-L/16	53.6	73.7	64.5	79.3	68.8	73.1	70.6	30.3	42.8	88.0	60.2	69.5	64.5
SigLIP2	ViT-B/16	46.8	72.0	63.6	78.4	67.8	70.6	64.9	30.2	42.1	85.9	58.5	65.6	62.2
	ViT-L/16	58.4	74.4	68.4	79.7	74.3	76.0	68.1	34.4	43.4	90.4	61.8	71.3	66.7
RemoteCLIP	RN-50	25.2	83.8	46.7	70.2	46.5	59.0	63.3	23.2	32.7	93.6	64.3	65.9	56.2
	ViT-B/32	38.7	**87.8**	58.5	79.3	60.9	71.6	67.5	26.2	38.8	**96.0**	**71.5**	68.9	63.8
	ViT-L/14	43.1	81.2	61.4	81.5	63.3	73.7	64.7	30.3	47.2	91.5	70.6	67.1	64.6
GeoRSCLIP	ViT-B/32	55.1	72.7	66.1	79.3	76.0	70.3	72.6	32.3	49.7	88.9	60.3	**78.6**	66.8
	ViT-L/14	69.2	75.8	**70.6**	**83.8**	80.9	**76.1**	**74.2**	40.3	53.0	90.4	61.6	74.3	**70.9**
CLIP LAION-RS	ViT-L/14	68.6	72.4	65.0	78.9	75.5	71.3	66.0	37.4	48.4	89.7	62.5	68.6	67.0
SkyCLIP	ViT-L/14	**72.2**	74.6	69.2	81.9	**80.9**	74.1	70.5	**45.3**	**55.1**	90.8	63.3	69.7	70.6

Tables B.6–B.10 present the full zero-shot classification results for the additional description-generation strategies, including the Simple Template (`ST`) baseline and the remaining LLM-based prompt sources DeepSeek, Qwen-2.5, Mixtral, and RS-LLaVA. These tables extend the analysis of Section 5.1, where the main paper emphasizes DST and LLaMA-3.1 as the most representative template-based and LLM-based configurations, respectively. Consistent with the main results, the extended tables show that remote-sensing-adapted vision-language models remain the most robust backbones across datasets, while general-purpose models benefit less consistently from prompt variation. The ST baseline is generally weaker than DST, which further supports the importance of domain-aware prompt design even in the absence of meta-prompting. In addition, the remaining LLM-based strategies exhibit the same non-uniform behavior discussed in the main paper: although they can outperform template-based prompts on specific dataset-model pairs, their gains are not systematic and their performance varies substantially across benchmarks. This again highlights that, in the vanilla embedding space, richer LLM-generated descriptions may introduce useful semantic detail but also additional variability that is not always aligned

Table B.8: Zero-shot classification accuracy (%) on all evaluated remote sensing datasets under the meta-prompting approach with `Qwen-2.5`. Results are reported across the full set of vision-language models and backbone architectures.

Model	Arch	EuroSAT	AID	MLRSNet	Optimal-31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHU-RS19	SIRI-WHU	RSSCN7	Avg
CLIP	ViT-B/16	53.0	64.7	55.9	72.7	63.9	63.9	58.6	31.0	39.5	84.3	53.0	70.7	59.3
	ViT-L/14	62.5	67.8	63.8	76.3	72.9	69.5	60.7	37.4	49.0	86.6	58.0	70.1	64.5
MetaCLIP-400M	ViT-B/16	58.5	64.0	58.2	76.9	67.8	66.8	63.6	34.0	43.0	89.2	54.1	72.3	62.4
	ViT-L/14	56.4	69.4	62.5	76.3	73.5	70.3	63.2	33.8	44.1	88.1	62.1	75.2	64.6
MetaCLIP-2.5B	ViT-B/16	50.6	66.7	63.2	77.4	70.0	69.2	54.8	37.2	44.6	89.6	57.7	72.8	62.8
	ViT-L/14	65.0	70.4	67.2	80.4	78.7	71.2	55.8	41.7	44.9	88.3	60.8	72.2	66.4
SigLIP	ViT-B/16	44.3	65.2	57.9	74.1	65.0	65.3	62.3	26.2	39.4	79.5	51.1	68.4	58.2
	ViT-L/16	54.6	70.9	64.5	78.8	70.9	71.9	69.0	31.6	47.5	87.7	59.4	73.3	65.0
SigLIP2	ViT-B/16	49.3	69.9	62.6	78.2	69.3	70.2	61.3	32.8	46.2	86.5	56.1	67.3	62.5
	ViT-L/16	59.7	71.9	67.4	81.2	76.8	**74.7**	67.9	35.7	47.8	90.1	59.2	70.9	66.9
RemoteCLIP	RN-50	26.2	73.9	45.8	69.7	46.4	56.8	62.3	22.8	33.6	89.8	59.6	66.8	54.5
	ViT-B/32	38.3	**79.9**	58.5	79.0	59.1	71.6	67.7	28.1	38.3	**94.6**	65.5	72.0	62.7
	ViT-L/14	47.8	76.5	61.5	81.9	63.2	74.1	**70.2**	30.5	49.2	91.2	**72.5**	69.4	65.7
GeoRSCLIP	ViT-B/32	51.2	68.8	65.3	79.9	77.2	69.8	66.8	35.1	47.6	89.8	60.2	**79.0**	65.9
	ViT-L/14	68.4	70.8	**69.4**	**84.7**	79.8	74.3	69.1	44.1	55.5	90.3	59.0	78.1	70.3
CLIP LAION-RS	ViT-L/14	73.0	68.6	64.6	78.5	74.9	71.1	63.9	38.3	52.5	89.7	61.6	75.3	67.7
SkyCLIP	ViT-L/14	**75.7**	70.4	69.0	82.0	**80.5**	73.4	68.3	**47.2**	**56.0**	91.1	61.3	74.7	**70.8**

with the discriminative structure of the downstream classes. Overall, the extended results reinforce the main conclusion that prompt effectiveness in zero-shot remote sensing depends on a non-trivial interaction between description source, dataset characteristics, and the underlying VLM.

Table B.9: Zero-shot classification accuracy (%) on all evaluated remote sensing datasets under the meta-prompting approach with `Mixtral`. Results are reported across the full set of vision-language models and backbone architectures.

Model	Arch	EuroSAT	AID	MLRSNet	Optimal-31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHU-RS19	SIRI-WHU	RSSCN7	Avg
CLIP	ViT-B/16	53.9	68.5	56.7	73.9	66.8	63.7	62.1	26.3	35.8	84.7	54.5	64.4	59.3
	ViT-L/14	65.5	72.2	65.0	80.6	75.0	68.8	65.6	34.1	46.2	87.5	59.7	64.0	65.4
MetaCLIP-400M	ViT-B/16	55.9	68.6	58.0	76.8	69.7	64.0	68.3	28.3	36.0	85.0	57.0	64.3	61.0
	ViT-L/14	59.2	70.9	63.8	79.2	75.6	70.3	64.3	30.4	39.7	88.4	62.6	68.8	64.4
MetaCLIP-2.5B	ViT-B/16	52.1	70.3	61.2	75.3	71.8	66.4	62.8	28.6	39.0	84.0	61.3	61.9	61.2
	ViT-L/14	68.8	74.5	64.1	81.0	79.0	70.9	59.8	34.2	41.0	85.2	62.5	65.6	65.5
SigLIP	ViT-B/16	44.0	67.7	58.0	70.4	64.7	64.0	63.6	25.9	37.3	79.8	56.4	55.6	57.3
	ViT-L/16	55.0	71.9	63.3	76.8	71.2	68.9	73.0	26.0	40.5	84.4	58.8	58.2	62.3
SigLIP2	ViT-B/16	46.2	72.2	62.6	78.0	69.1	68.4	66.0	28.3	39.1	83.6	56.7	57.1	60.6
	ViT-L/16	59.5	74.9	67.3	81.4	77.0	73.0	71.9	32.3	46.7	89.6	58.6	62.4	66.2
RemoteCLIP	RN-50	21.9	83.3	47.0	70.3	49.0	60.1	65.7	22.2	33.5	93.7	59.5	66.4	56.0
	ViT-B/32	42.3	**88.9**	59.1	80.8	62.0	72.3	67.0	27.1	43.0	**96.5**	71.1	67.3	64.8
	ViT-L/14	44.7	81.9	62.1	81.6	64.1	74.0	67.1	32.5	46.3	91.8	**71.2**	63.0	65.0
GeoRSCLIP	ViT-B/32	51.8	72.6	66.7	80.5	77.3	70.2	72.0	29.3	46.4	91.9	62.9	**75.2**	66.4
	ViT-L/14	70.7	77.2	**71.7**	**86.0**	**81.5**	**75.1**	**74.6**	39.2	51.6	94.1	65.1	72.4	**71.6**
CLIP LAION-RS	ViT-L/14	70.6	73.6	65.0	79.9	75.2	69.8	65.7	35.0	46.0	90.9	63.3	68.4	67.0
SkyCLIP	ViT-L/14	**72.2**	74.3	69.1	84.2	81.2	70.6	72.9	**44.9**	**52.8**	90.0	64.4	72.6	70.8

Table B.10: Zero-shot classification accuracy (%) on all evaluated remote sensing datasets under the meta-prompting approach with `RS-LLaVA`. Results are reported across the full set of vision-language models and backbone architectures.

Model	Arch	EuroSAT	AID	MLRSNet	Optimal-31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHU-RS19	SIRI-WHU	RSSCN7	Avg
CLIP	ViT-B/16	50.4	68.3	57.0	73.6	64.4	63.8	58.2	25.5	36.4	85.1	52.9	65.9	58.5
	ViT-L/14	66.7	71.9	64.8	78.7	73.8	68.7	59.8	31.9	40.7	88.5	57.9	65.8	64.1
MetaCLIP-400M	ViT-B/16	55.1	68.4	58.6	77.1	68.7	66.3	67.3	27.6	37.3	85.6	57.0	70.0	61.6
	ViT-L/14	49.9	70.6	62.7	76.8	75.3	71.4	62.3	30.0	39.1	88.4	62.6	71.4	63.4
MetaCLIP-2.5B	ViT-B/16	50.1	71.0	63.5	78.9	72.6	68.9	62.2	31.5	38.6	85.3	60.5	70.8	62.8
	ViT-L/14	61.8	74.4	68.0	82.2	77.9	72.8	60.5	37.0	39.3	88.8	62.5	70.5	66.3
SigLIP	ViT-B/16	44.4	70.7	59.1	76.5	65.1	66.0	60.9	25.8	39.1	83.3	57.9	62.3	59.3
	ViT-L/16	53.9	74.7	64.3	80.9	70.9	72.4	68.4	29.3	46.1	88.1	58.7	67.4	64.6
SigLIP2	ViT-B/16	46.8	73.7	63.4	82.0	69.4	71.1	62.0	29.9	42.6	87.8	59.9	65.0	62.8
	ViT-L/16	60.5	76.0	69.9	83.0	75.8	75.7	66.8	33.4	46.4	91.8	60.8	69.9	67.5
RemoteCLIP	RN-50	21.7	80.8	44.2	68.5	42.0	58.6	60.4	20.1	33.4	92.8	59.4	61.4	53.6
	ViT-B/32	41.1	**88.7**	58.2	79.6	58.8	71.5	65.7	27.6	42.9	**95.8**	71.0	64.0	63.7
	ViT-L/14	45.0	81.5	60.0	80.9	62.3	73.4	65.4	29.7	47.3	90.8	**71.3**	62.2	64.2
GeoRSCLIP	ViT-B/32	52.9	73.0	65.9	81.7	76.9	71.0	66.2	32.3	46.1	91.2	61.1	**77.5**	66.3
	ViT-L/14	73.3	76.8	**71.4**	**86.4**	79.8	**76.6**	**70.3**	41.3	**55.6**	93.3	61.2	72.0	**71.5**
CLIP LAION-RS	ViT-L/14	74.3	73.9	65.0	78.6	73.8	70.2	62.5	34.8	47.3	91.7	60.6	72.6	67.1
SkyCLIP	ViT-L/14	**75.9**	75.8	70.3	83.8	**81.4**	73.0	68.0	**44.7**	50.1	92.5	61.6	71.4	70.7

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Int. Conf. Mach. Learn., Vol. 139, 2021, pp. 8748–8763. URL https://proceedings.mlr.press/v139/radford21a.html

[2] M. J. Mirza, L. Karlinsky, W. Lin, S. Doveh, , J. Micorek, M. Kozinski, H. Kuhene, H. Possegger, Meta-Prompting for Automating Zero-shot Visual Recognition with LLMs, in: Eur. Conf. Comput. Vis., 2024, pp. 370–387. doi:10.1007/978-3-031-72627-9_21.

[3] A. Promponas, E. Baltzi, V. Ntouskos, K. Karantzalos, Meta-prompting with open-source language models for zero-shot scene classification in remote sensing, in: Int. Arch. ISPRS, 2026, p. to appear.

[4] H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, C. Feichtenhofer, Demystifying clip data, in: Int. Conf. Learn. Represent., 2024, pp. 47812–47831.

[5] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Int. Conf. Comput. Vis., 2023, pp. 11975–11986.

[6] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348. doi:10.1007/s11263-022-01653-1.

[7] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Conditional prompt learning for vision-language models, in: IEEE Conf. Comput. Vis. Pattern Recog., 2022, pp. 16816–16825.

[8] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, et al., Robust fine-tuning of zero-shot models, in: IEEE Conf. Comput. Vis. Pattern Recog., 2022, pp. 7959–7971.

[9] Y. Li, H. Zhang, X. Xue, Y. Jiang, Q. Shen, Deep learning for remote sensing image classification: A survey, WIRES Data Min. Knowl. 8 (6) (2018) e1264. doi:10.1002/widm.1264.

[10] C. Liu, et al., RemoteCLIP: A vision language foundation model for remote sensing, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–16. doi:10.1109/TGRS.2024.3390838.

[11] Z. Zhang, T. Zhao, Y. Guo, J. Yin, RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–23. doi:10.1109/TGRS.2024.3449154.

[12] Z. Wang, R. Prabha, T. Huang, J. Wu, R. Rajagopal, Skyscript: A large and semantically diverse vision-language dataset for remote sensing, AAAI 38 (6) (2024) 5805–5813. doi:10.1609/aaai.v38i6.28393.

[13] X. Lu, B. Wang, X. Zheng, X. Li, Exploring models and data for remote sensing image caption generation, IEEE Trans. Geosci. Remote Sens. 56 (4) (2017) 2183–2195. doi:10.1109/TGRS.2017.2776321.

[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The Llama 3 Herd of Models (2024). arXiv:2407.21783.

[15] DeepSeek-AI, :, X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al., DeepSeek LLM: Scaling Open-Source Language Models with Longtermism (2024). arXiv:2401.02954.

[16] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, K. Dang, Y. Fan, Y. Zhang, A. Yang, R. Men, F. Huang, B. Zheng, Y. Miao, S. Quan, Y. Feng, X. Ren, X. Ren, J. Zhou, J. Lin, Qwen2.5-coder technical report (2024). arXiv:2409.12186.

[17] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, et al., Mixtral of experts (2024). arXiv:2401.04088.

[18] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, F. Melgani, Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery, Remote Sensing 16 (9) (2024) 1477. doi:10.3390/rs16091477.

[19] M. Y. Levi, G. Gilboa, The double-ellipsoid geometry of CLIP, in: Int. Conf. Mach. Learn., Vol. 267, 2025, pp. 33999–34019.

[20] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, X. Zhai, Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025). arXiv:2502.14786.

[21] R. Betser, M. Y. Levi, G. Gilboa, Whitened CLIP as a likelihood surrogate of images and captions, in: Int. Conf. Mach. Learn., Vol. 267, 2025, pp. 4069–4095.

[22] Q. Zou, L. Ni, T. Zhang, Q. Wang, Deep learning based feature selection for remote sensing scene classification, IEEE Geoscience and Remote Sensing Letters 12 (11) (2015) 2321–2325. doi:10.1109/LGRS.2015.2475299.

[23] P. Helber, B. Bischke, A. Dengel, D. Borth, 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) (2019) 2217–2226. doi:10.1109/JSTARS.2019.2918242.

[24] L. Zhao, P. Tang, L. Huo, Feature significance-based multibag-of-visual-words model for remote sensing image scene classification, J. Appl. Remote Sens. 10 (3) (2016) 035004–035004. doi:10.1117/1.JRS.10.035004.

[25] B. Zhao, Y. Zhong, G.-S. Xia, L. Zhang, Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 54 (4) (2015) 2108–2123. doi:10.1109/TGRS.2015.2496185.

[26] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, H. Maître, Structural high-resolution satellite image indexing, in: ISPRS TC VII Symposium, Vol. 38, 2010, pp. 298–303.

[27] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, X. Lu, Aid: A benchmark data set for performance evaluation of aerial scene classification, IEEE Trans. Geosci. Remote Sens. 55 (7) (2017) 3965–3981. doi:10.1109/TGRS.2017.2685945.

[28] Q. Wang, S. Liu, J. Chanussot, X. Li, Scene classification with recurrent attention of vhr remote sensing images, IEEE Trans. Geosci. Remote Sens. 57 (2) (2018) 1155–1167. doi:10.1109/TGRS.2018.2864987.

[29] H. Li, X. Dou, C. Tao, Z. Wu, J. Chen, J. Peng, M. Deng, L. Zhao, RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data, Sensors 20 (6) (2020) 1594. doi:10.3390/s20061594.

[30] W. Zhou, S. Newsam, C. Li, Z. Shao, Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval, ISPRS J. Photogramm. Remote Sens. 145 (2018) 197–209. doi:10.1016/j.isprsjprs.2018.01.004.

[31] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art, Proceedings of the IEEE 105 (10) (2017) 1865–1883. doi:10.1109/JPROC.2017.2675998.

[32] X. Qi, P. Zhu, Y. Wang, L. Zhang, J. Peng, M. Wu, J. Chen, X. Zhao, N. Zang, P. T. Mathiopoulos, Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding, ISPRS J. Photogramm. Remote Sens. 169 (2020) 337–350. doi:https://doi.org/10.1016/j.isprsjprs.2020.09.020.