Meta-Prompting with Open-Source Language Models for Zero-Shot Scene Classification in Remote Sensing

Antonis Promponas[1, *], Eirini Baltzi[1, *], Valsamis Ntouskos[2], Konstantinos Karantzalos[1]

[1] Remote Sensing Lab, NTUA, Athens, Greece - (antonispromponas, eirinibaltzi)@mail.ntua.gr, karank@central.ntua.gr
[2] Department of Engineering and Sciences, Universitas Mercatorum, Rome, Italy - valsamis.ntouskos@unimercatorum.it
[*] Equal contribution

Keywords: Meta-prompting, Remote Sensing, Zero-shot Classification, Vision-language Models (VLMs), Large Language Models (LLMs), Prompt Generation

Abstract

Zero-shot visual recognition with vision-language models (VLMs) has shown strong generalization to unseen categories in natural-image benchmarks, yet its effectiveness in remote-sensing (RS) imagery remains less explored. In this paper, we investigate whether meta-prompting with large language models (LLMs) can improve zero-shot scene classification in RS by automatically generating semantically rich class descriptions. Building on the Meta-Prompting for Visual Recognition (MPVR) framework, we evaluate three open-source LLMs, Mixtral-8×7B, Qwen 2.5 7B, and LLaMA 3.1 8B, as prompt generators across five RS benchmark datasets. The resulting descriptions are encoded with several VLMs, including CLIP, MetaCLIP, RemoteCLIP, and CLIP-LAION-RS, and compared against generic single-template and handcrafted domain-specific prompting baselines. Our results show that LLM-generated prompts are competitive with, and in several cases improve upon, manually designed templates, while revealing that the gains depend on both the dataset and the visual backbone. Overall, the study highlights the potential of open-source LLMs as scalable prompt generators for zero-shot remote-sensing recognition and provides insight into the transferability of meta-prompting beyond natural-image domains.

1. Introduction

Vision–language models (VLMs) such as CLIP have enabled a new paradigm for visual recognition, where models can classify images without task-specific training by matching image embeddings to text descriptions of candidate classes. This zero-shot setting is particularly attractive in remote sensing (RS), where collecting large, well-curated labeled datasets is often expensive, time-consuming, and highly domain-dependent. In practical Earth observation (EO) scenarios, the ability to recognize land-use or land-cover categories from textual descriptions alone could support rapid adaptation to new sensors, regions, and taxonomies without retraining dedicated classifiers.

Despite this promise, transferring zero-shot vision–language recognition to RS imagery remains challenging. Unlike natural-image benchmarks, remote-sensing scenes are captured from aerial or satellite platforms, often exhibiting large intra-class variation. At the same time, the corresponding semantic concepts are typically expressed through spatial layouts, textures, object co-occurrences, and scene composition rather than a single dominant foreground object. In addition, the visual appearance of RS categories is influenced by imaging conditions, scale, seasonal effects, and geographic context (Rolf et al., 2024). As a result, prompt formulations commonly used in natural-image classification, such as “a photo of a {class},” may not adequately describe the semantics of EO imagery. A central factor in zero-shot VLM performance is therefore the quality of the textual prompts used to represent each category.

Prior work has shown that prompt engineering can substantially influence performance, either through handcrafted templates or by using language models to generate richer descriptions (Radford et al., 2021, Mirza et al., 2024). However, manually designing effective prompts for remote-sensing data is difficult to scale, particularly across datasets with different label spaces and scene definitions. This motivates the use of large language models (LLMs) as automatic prompt generators capable of producing dataset-aware and class-specific textual descriptions.

Recent work on Meta-Prompting for Visual Recognition (MPVR) (Mirza et al., 2024) demonstrated that LLMs can be used to generate improved prompts for zero-shot recognition in natural-image settings. Yet, the use of *open-source* LLMs for this task has not been systematically explored. Thus, it remains unclear whether meta-prompting transfers effectively to the remote-sensing domain, where class semantics and viewing conditions differ substantially from those encountered in conventional image–text data. Understanding this transferability is important both from a scientific perspective, to assess the robustness of meta-prompting beyond natural-image domains, and from a practical perspective, since open-source LLMs offer accessible and reproducible alternatives for prompt generation.

In this work, we investigate *meta-prompting with open-source language models for zero-shot scene classification in remote sensing*. We adapt the two-stage MPVR framework to the RS domain and compare three types of textual supervision: a single generic prompt template, a set of handcrafted remote-sensing-specific templates, and automatically generated class descriptions produced by open-source LLMs. We consider three widely used LLM families, Mixtral 8×7B (Jiang et al., 2024), Qwen 2.5 7B (Qwen et al., 2025), and LLaMA 3.1 8B (Grattafiori et al., 2024) and evaluate them across five representative remote-sensing scene classification benchmarks: EuroSAT (Helber et al., 2019), Optimal31 (Wang et al., 2018), PatternNet (Zhou et al., 2018), RESISC45 (Cheng et al., 2017), and RSSCN7 (Zou et al., 2015). To study whether the effect of prompt generation depends on the underlying visual encoder, we further benchmark multiple VLMs, including CLIP (Rad-

ford et al., 2021), MetaCLIP (Xu et al., 2024), as well as the domain–specific RemoteCLIP (Liu et al., 2024) and CLIP LAION-RS (Schuhmann et al., 2022).

2. Related Work

Recent advances in vision–language pretraining have enabled recognition without task-specific supervision by aligning image and text representations in a shared embedding space. CLIP (Radford et al., 2021) established this paradigm for zero-shot classification and demonstrated that prompt design plays a crucial role in downstream performance. In particular, the choice of textual template used to represent each class can significantly affect alignment quality between image and text embeddings. This observation motivated a large body of literature on prompt engineering, prompt ensembling, and prompt learning for zero-shot recognition (Sahoo et al., 2024, Zhou et al., 2022). Beyond manually designed templates, several methods have explored the use of richer textual descriptions generated by large language models (LLMs), including DCLIP (Menon and Vondrick, 2023), CuPL (Pratt et al., 2023), and Waffle (Roth et al., 2023). These approaches show that augmenting class names with diverse semantic descriptions can improve zero-shot recognition without modifying the visual backbone. However, they often still depend on handcrafted LLM queries, manually curated descriptions, or dataset-specific human intervention.

To reduce this dependency on manual prompt construction, Meta-Prompting for Visual Recognition (MPVR) (Mirza et al., 2024) proposed a two-stage framework in which an LLM first generates task-specific query templates and then produces rich class-level textual descriptions by instantiating these templates. This strategy offers a more automated way to construct prompt sets while preserving the benefits of diverse language supervision. Our work is closely related to this direction but investigates its transferability to remote-sensing scene classification, a domain with substantially different visual and semantic characteristics from natural-image recognition.

Compared with conventional image benchmarks, remote-sensing imagery introduces unique challenges for zero-shot vision–language transfer. Overhead scenes are characterized by nadir views, significant scale variation, spatially distributed semantics, and category definitions that often depend on land-cover configuration or structural arrangement rather than distinct salient objects. These properties make prompt construction particularly important, as generic natural-image templates may fail to capture the semantics of aerial and satellite imagery. While remote-sensing-specific vision–language models such as RemoteCLIP and CLIP-LAION-RS have shown that domain-aware pretraining can improve representation quality, the role of automated prompt generation in this setting remains underexplored. In particular, there has been limited analysis of whether open-source LLMs can generate effective remote-sensing-aware class descriptions across datasets and model families.

Our work addresses this gap by systematically studying meta-prompting with open-source LLMs for zero-shot remote-sensing scene classification. In contrast to prior prompt-generation methods developed primarily for natural-image benchmarks, we evaluate the interaction between prompt source, remote-sensing datasets, and both general-purpose and domain-specialized VLM backbones. This positions our study at the intersection of prompt engineering, automated language supervision, and domain transfer in remote sensing.

3. Methodology

Our framework builds on the two-stage meta-prompting strategy introduced in (Mirza et al., 2024), adapted here for zero-shot remote-sensing scene classification. The central idea is to replace minimal class names or generic templates with richer class-specific descriptions that better reflect the appearance and semantics of aerial and satellite imagery. Figure 1 illustrates the overall pipeline.

3.1 Prompting Configurations

To quantify the effect of textual supervision, we compare three prompting configurations of increasing complexity.

Generic template (*s-temp*). As a minimal zero-shot baseline, we use a single generic template of the form *“a satellite photo of a {class}”*. This setting follows the standard prompt construction commonly used in CLIP-style zero-shot evaluation and serves as a reference point for measuring the benefit of more informative text descriptions.

Domain-specific templates (*ds-temp*). To better reflect the geometry and semantics of overhead imagery, we define a fixed set of handcrafted remote-sensing-oriented templates:

> “an aerial photo of a {},” “a satellite image of a {},” “aerial view of the {},” “an overhead photo of a {},” “a top-down image showing a {},” “a high-resolution satellite photo of a {},” “an orthophoto depicting a {},” “aerial landscape of the {},” “a vertical aerial view of the {},” “an image captured from above showing a {}.”

These templates incorporate RS-specific phrasing such as *satellite image*, *overhead photo*, and *top-down view*, which are expected to align more closely with the visual statistics of aerial scenes than generic natural-image prompts.

LLM-generated descriptions. Our most expressive setting uses the meta-prompting pipeline to generate richer dataset-aware and class-aware descriptions automatically. Instead of manually specifying all prompt variants, we rely on open-source LLMs to produce textual descriptions that include characteristic attributes, spatial layouts, and contextual cues associated with each scene class.

3.2 Two-Stage Meta-Prompting Pipeline

The LLM-based configuration follows two stages.

Stage 1: Dataset-specific meta-prompt generation. For each benchmark dataset, we first construct a dataset-level instruction that provides the LLM with example prompt templates and a brief dataset description. The purpose of this stage is to adapt the language model to the style and semantics of the target benchmark, encouraging it to produce prompts appropriate for remote-sensing imagery rather than natural photographs. In practice, each dataset is associated with approximately 30–36 meta-prompts, which act as diverse textual instructions for the next generation step.

Stage 2: Class-specific description generation. Given the dataset-level meta-prompts, the LLM generates multiple class-specific descriptions for each semantic category. For every

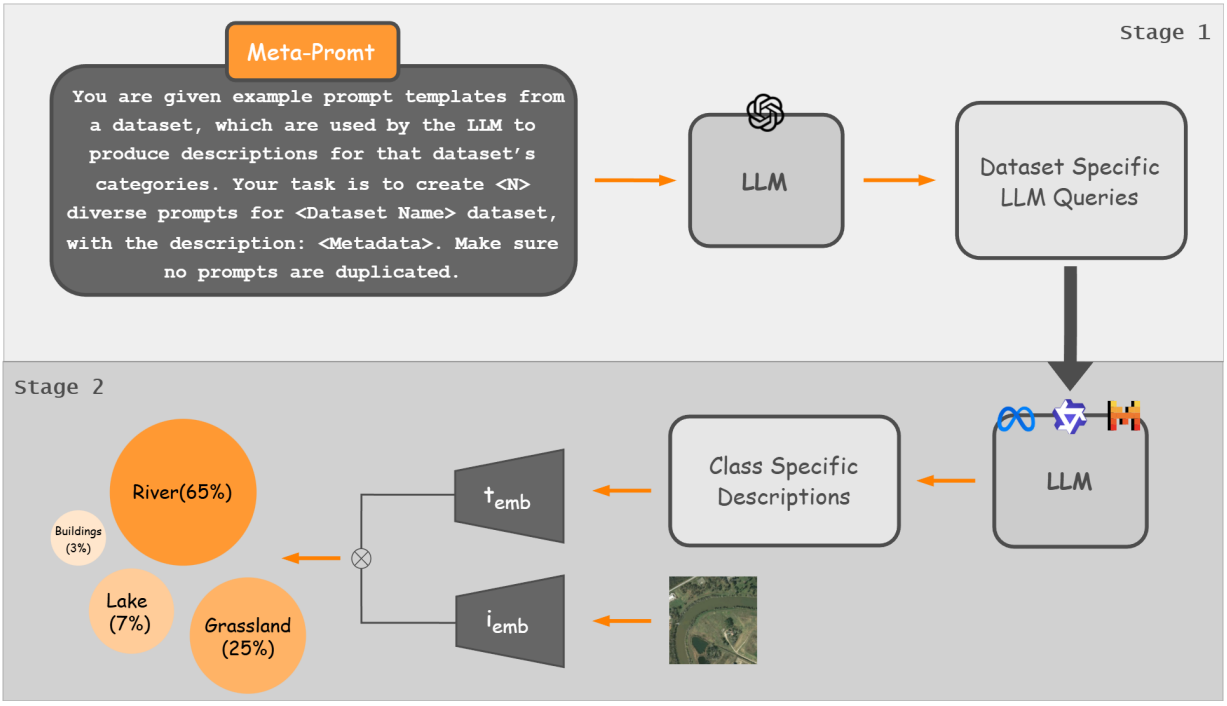

Figure 1. **Overview of the proposed meta-prompting pipeline for zero-shot remote-sensing scene classification.** In Stage 1, a dataset-level meta-prompt is provided to an open-source large language model (LLM), which generates dataset-specific query templates tailored to the semantics of the target remote-sensing benchmark. In Stage 2, these queries are used to produce class-specific textual descriptions for each scene category. The generated descriptions are encoded by the text branch of a frozen vision–language model (VLM) and aggregated into class prototypes, while the input remote-sensing image is encoded by the corresponding image branch. Zero-shot prediction is then obtained by comparing image and text embeddings through similarity matching.

class, we obtain four textual descriptions per meta-prompt, resulting in a diverse pool of candidate descriptions that vary in wording, level of abstraction, and emphasis on visual attributes. This process is intended to capture complementary aspects of a remote-sensing class, such as land-cover composition, spatial arrangement, man-made structures, vegetation patterns, or water-body appearance.

We evaluate three open-source LLMs in this generation stage: Mixtral 8×7B, Qwen 2.5 7B, and LLaMA 3.1 8B. These models differ in architecture and language generation style, which makes them suitable for studying whether prompt quality depends on the underlying LLM. Empirically, they provide different forms of textual variation, ranging from concise attribute-focused descriptions to more context-rich and spatially grounded formulations.

3.3 Text Prototype Construction and Zero-Shot Inference

All generated prompts are encoded using the text encoder of a frozen vision–language model. Letting $\mathcal{C}$ denote the set of classes considered in the dataset, and $\mathcal{T}_c = \{t_{c,1}, t_{c,2}, \ldots, t_{c,N_c}\}$ the set of textual descriptions associated with class $c \in \mathcal{C}$, each description is mapped to an embedding vector through the VLM text encoder:

$$\mathbf{z}_{c,n}^{text} = f_{\text{text}}(t_{c,n}).$$

We then form a single class prototype by averaging all text embeddings for that class:

$$\bar{\mathbf{z}}_c^{text} = \frac{1}{N_c} \sum_{n=1}^{N_c} \mathbf{z}_{c,n}^{(t)}.$$

All prompts contribute equally to the prototype; no ranking, filtering, or weighting scheme is applied.

At inference time, an input image x is encoded by the corresponding VLM image encoder:

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

The predicted label is obtained by computing cosine similarity between the image embedding and each class text prototype:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \cos\left(\mathbf{z}^{img}, \bar{\mathbf{z}}_c^{text}\right).$$

This preserves the standard zero-shot formulation: no fine-tuning is performed on the target remote-sensing datasets, and all predictions rely exclusively on the alignment between frozen image and text representations.

4. Experimental Setup

4.1 Vision–Language Models

To evaluate whether prompt generation interacts differently with general-purpose and remote-sensing-aware VLMs, we benchmark four model families: CLIP, MetaCLIP, RemoteCLIP, and CLIP-LAION-RS. For CLIP, MetaCLIP, and RemoteCLIP, we consider both ViT-B/32 and ViT-L/14 variants when available, while CLIP-LAION-RS is evaluated with a ViT-L/14 backbone. This selection enables us to study the effect of both architecture scale and domain specialization on zero-shot scene classification performance.

Table 1. Zero-shot classification performance (Top-1 Accuracy, %) across different prompt configurations. s-temp denotes the simple baseline template, ds-temp the handcrafted domain-specific template, and the remaining settings use LLM-generated descriptions. The highest value in each setting is shown in **bold**.

Setting	Model	Arch	EuroSAT	Optimal31	PatternNet	RESISC45	RSSCN7
s-temp	CLIP	ViT-B/32	39.25	66.18	54.74	55.43	49.86
	CLIP	ViT-L/14	62.97	76.08	71.85	66.99	61.50
	MetaCLIP-400M	ViT-B/32	50.39	71.77	58.97	61.55	58.64
	MetaCLIP-400M	ViT-L/14	52.61	79.95	74.49	69.98	64.46
	RemoteCLIP	ViT-B/32	38.35	76.24	58.51	66.29	58.82
	RemoteCLIP	ViT-L/14	43.14	**82.31**	64.06	**75.07**	63.21
	CLIP LAION-RS	ViT-L/14	**70.43**	80.48	**76.68**	72.01	**66.46**
ds-temp	CLIP	ViT-B/32	48.03	70.48	58.82	60.54	59.54
	CLIP	ViT-L/14	63.34	80.70	75.93	71.60	64.25
	MetaCLIP-400M	ViT-B/32	48.70	71.29	63.65	64.10	63.50
	MetaCLIP-400M	ViT-L/14	50.09	80.86	74.38	71.87	**69.36**
	RemoteCLIP	ViT-B/32	34.47	77.53	59.32	67.90	59.79
	RemoteCLIP	ViT-L/14	42.90	**82.53**	64.12	**75.40**	64.89
	CLIP LAION-RS	ViT-L/14	**68.84**	82.42	**77.81**	74.08	68.68
Mixtral 8×7B	CLIP	ViT-B/32	52.65	70.75	61.82	61.32	59.54
	CLIP	ViT-L/14	63.74	80.54	75.07	69.46	65.89
	MetaCLIP-400M	ViT-B/32	57.50	70.59	62.89	59.82	58.82
	MetaCLIP-400M	ViT-L/14	58.16	80.59	**75.82**	71.29	68.82
	RemoteCLIP	ViT-B/32	41.35	80.43	61.87	72.45	66.68
	RemoteCLIP	ViT-L/14	44.79	**81.51**	63.99	**74.05**	63.71
	CLIP LAION-RS	ViT-L/14	**70.43**	80.32	75.25	70.41	**72.00**
Qwen 2.5 7B	CLIP	ViT-B/32	49.62	71.61	61.56	61.83	68.86
	CLIP	ViT-L/14	64.00	75.22	72.89	70.12	71.32
	MetaCLIP-400M	ViT-B/32	55.86	69.25	59.97	59.10	64.96
	MetaCLIP-400M	ViT-L/14	54.71	76.18	73.25	70.37	75.68
	RemoteCLIP	ViT-B/32	37.68	78.82	58.62	71.88	72.61
	RemoteCLIP	ViT-L/14	47.71	**82.26**	63.53	**74.32**	69.43
	CLIP LAION-RS	ViT-L/14	**73.35**	78.23	**74.56**	71.46	**75.86**
LLaMA 3.1 8B	CLIP	ViT-B/32	46.76	74.62	61.62	63.35	70.39
	CLIP	ViT-L/14	60.49	79.46	74.65	69.62	68.43
	MetaCLIP-400M	ViT-B/32	47.45	71.77	62.73	61.08	66.86
	MetaCLIP-400M	ViT-L/14	55.06	80.22	**76.93**	71.50	**74.79**
	RemoteCLIP	ViT-B/32	36.85	79.57	59.59	72.57	71.64
	RemoteCLIP	ViT-L/14	45.67	**81.08**	64.18	**74.10**	68.79
	CLIP LAION-RS	ViT-L/14	**70.10**	79.52	74.47	70.97	72.79

4.2 Datasets

We evaluate on five widely used remote-sensing scene classification benchmarks: EuroSAT, Optimal31, PatternNet, RESISC45, and RSSCN7. Together, these datasets cover different sensing platforms, scene definitions, spatial layouts, and levels of semantic complexity, making them suitable for assessing the robustness of zero-shot recognition across diverse remote-sensing conditions.

EuroSAT is based on Sentinel-2 satellite imagery and contains land-use and land-cover categories observed at multispectral satellite scale. Compared with aerial benchmarks, it reflects a more satellite-oriented distribution, where classes are often characterized by broader spatial texture and land-cover composition rather than highly localized objects.

Optimal31 consists of aerial scene categories spanning both natural and man-made environments. Its classes include structurally distinct land-use types, making it useful for evaluating whether textual prompts can capture scene-level semantics expressed through layout, materials, and spatial organization.

PatternNet is a large-scale remote-sensing scene classification dataset designed to reduce ambiguity and improve visual consistency across classes. It contains diverse scene categories with relatively clear semantic definitions, which makes it an informative benchmark for studying how different prompting strategies align textual descriptions with remote-sensing imagery.

RESISC45 is one of the most widely used large-scale benchmarks in remote-sensing scene classification. It includes 45 scene classes with substantial intra-class diversity and inter-class similarity, making it particularly challenging for zero-shot recognition. Its scale and variability make it a strong testbed for evaluating the robustness of prompt-based class representations.

RSSCN7 contains scene categories observed at multiple scales and under varying imaging conditions. Although smaller in label space than RESISC45, it remains challenging due to viewpoint and scale variations, which can affect the consistency between textual descriptions and visual appearance.

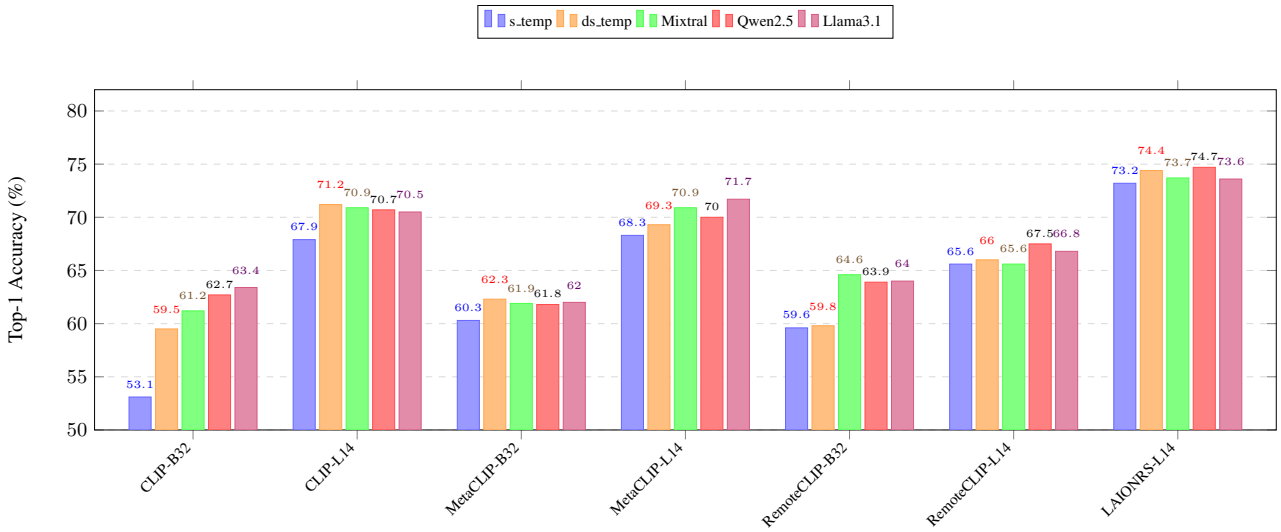

Figure 2. Average Top-1 accuracy (%) of seven vision–language model (VLM) variants based on CLIP, MetaCLIP, RemoteCLIP, and CLIP-LAION-RS, evaluated under five prompting configurations (s_temp, ds_temp, Mixtral-8×7B, Qwen 2.5 7B, and LLaMA 3.1 8B). Each color represents a prompting approach, while each bar group corresponds to a specific VLM. The comparison highlights how different text-generation sources influence visual–language alignment and zero-shot scene-classification accuracy.

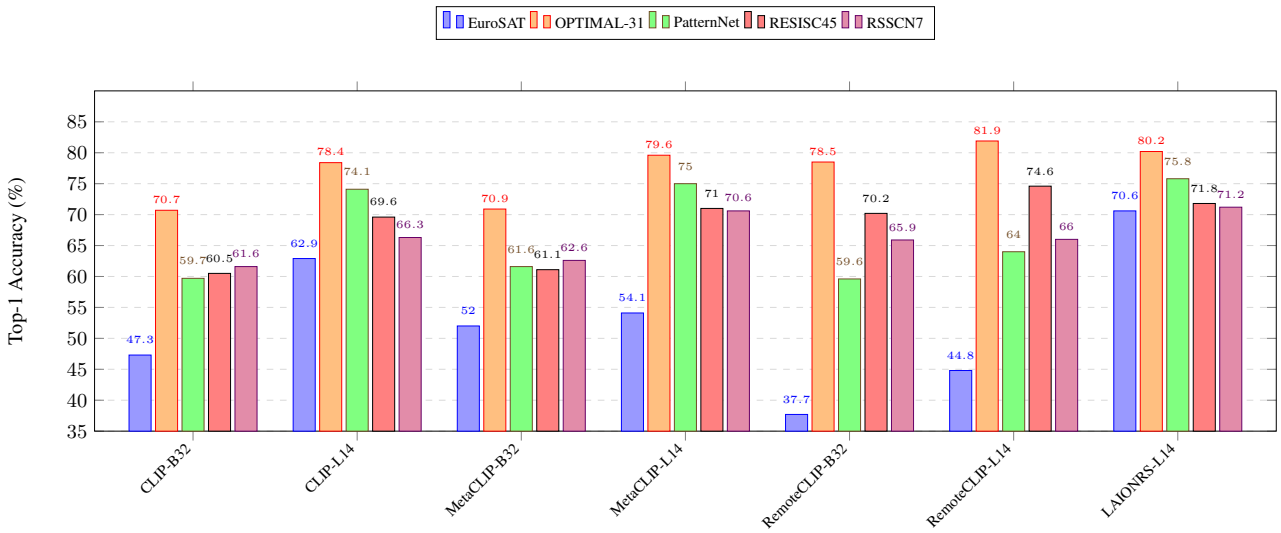

Figure 3. Top-1 accuracy (%) of seven vision–language model (VLM) variants based on CLIP, CLIP, MetaCLIP, RemoteCLIP, and CLIP-LAION-RS, averaged over five prompting configurations (s_temp, ds_temp, Mixtral-8×7B, Qwen 2.5 7B, and LLaMA 3.1 8B). Each color denotes a dataset (EuroSAT, OPTIMAL-31, RESISC45, PatternNet, RSSCN7), and each group of bars corresponds to a specific VLM. The figure illustrates how model performance varies across datasets when predictions are averaged over multiple prompting strategies.

4.3 Evaluation Protocol

We follow a standard zero-shot evaluation protocol with no fine-tuning on the target datasets. For each prompting configuration, class-level text representations are constructed as described in Section 3. For LLM-based prompt generation, the open-source LLMs Mixtral 8×7B, Qwen 2.5 7B, and LLaMA 3.1 8B are used. The prompts are encoded and matched against the corresponding image embeddings from frozen VLM encoders. Performance is assessed in terms of Top-1 classification accuracy.

5. Results

Table 1 reports zero-shot Top-1 accuracy across all datasets, prompt configurations, and VLM backbones. Overall, the results indicate that prompt design has a substantial effect on remote-sensing zero-shot classification, but its impact is not uniform across models or datasets.

A first observation is that the generic single-template baseline (*s-temp*) is often outperformed by richer alternatives, confirming that textual formulation is a critical factor in RS zero-shot recognition. This is especially expected in overhead imagery, where class semantics are less object-centric and require descriptions that better capture scene layout, surface composition, and viewpoint.

The handcrafted domain-specific setting (*ds-temp*) already improves over the generic baseline in many cases, suggesting that explicitly encoding the aerial or satellite perspective is beneficial. However, the magnitude of this improvement varies, indic-

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-233-2026 | © Author(s) 2026. CC BY 4.0 License.

ating that domain-aware wording alone is not always sufficient to fully align text embeddings with the visual structure of RS scenes.

The LLM-generated settings further demonstrate that automatically produced descriptions can be competitive with, and in several cases superior to, manually designed templates. Importantly, their effect depends on both the chosen prompt generator and the underlying VLM. Some backbones appear to benefit more from richer and more varied textual descriptions, while others show smaller gains or dataset-specific fluctuations. This suggests that prompt quality should be analyzed jointly with model specialization rather than in isolation.

Across backbones, CLIP-LAION-RS frequently achieves the strongest performance, highlighting the value of remote-sensing-aware pretraining. At the same time, general-purpose VLMs such as CLIP and MetaCLIP also benefit from improved prompts, indicating that textual supervision can partially compensate for domain mismatch. RemoteCLIP exhibits strong results on several datasets as well, particularly when paired with more informative prompt sets.

Figures 2 and 3 provide complementary views of these trends. The first averages performance across datasets for each prompting strategy and shows how different VLMs respond to changes in textual supervision. The second averages over prompting settings and instead emphasizes dataset-specific behavior.

Regarding the impact of LLM-generated prompts compared to hand-crafted ones, we observe that the gain is more pronounced in the case of ViT-B/32 based models. There is also no consistent gain when generic and domain-specific VLMs are compared. Overall, we observe that CLIP-L14, MetaCLIP-L14 and LAIONRS-L14 models significantly outperform the other models, with the latter achieving the highest average performance.

As far as the performance on individual datasets is considered, we observe significant variations of the per-dataset results with the average performance of each model. For instance, while RemoteCLIP-L14 trails the leading models in average accuracy, it emerges as a top-performing model for the OPTIMAL-31 and the challenging RESISC45 datasets. Conversely, it shows subpar performance for EuroSAT and PatternNet datasets, justifying the gap observed in average performance.

Overall, these results combined show that zero-shot RS classification is governed by a non-trivial interaction between dataset semantics, visual backbone, and prompt source.

6. Conclusion

We present a study of meta-prompting with open-source language models for zero-shot scene classification in remote sensing. By adapting the MPVR framework to aerial and satellite imagery, we compared generic prompting, handcrafted domain-specific templates, and LLM-generated class descriptions across five benchmarks and multiple VLM families.

Our results suggest that richer textual supervision can improve zero-shot performance, but the gains are not always consistent across visual-language models and datasets, as one might expect. Instead, they depend on the interaction between dataset characteristics, the language model used for prompt generation, and the visual backbone used for inference. These findings highlight both the promise and the limitations of prompt-based transfer in remote sensing.

Overall, this study provides one of the first systematic analyses of open-source LLM-based meta-prompting for remote-sensing scene classification and offers practical insight into when richer textual supervision is most beneficial. The results suggest that future work should investigate adaptive prompt selection, prompt filtering, and tighter coupling between RS-specialized language generation and vision–language pretraining.

References

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10), 1865–1883.

Grattafiori, A., Dubey, A., et al., A. J., 2024. The Llama 3 Herd of Models. https://arxiv.org/abs/2407.21783.

Helber, P., Bischke, B., Dengel, A., Borth, D., 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2217–2226.

Jiang, A. Q., Sablayrolles, A., et al., A. R., 2024. Mixtral of Experts. https://arxiv.org/abs/2401.04088.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J., 2024. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–16.

Menon, S., Vondrick, C., 2023. Visual classification via description from large language models. *International Conference on Learning Representations (ICLR)*.

Mirza, M. J., Karlinsky, L., Lin, W., Doveh, S., , Micorek, J., Kozinski, M., Kuhene, H., Possegger, H., 2024. Meta-Prompting for Automating Zero-shot Visual Recognition with LLMs. *European Conference on Computer Vision (ECCV)*.

Pratt, S. M., Covert, I., Liu, R., Farhadi, A., 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 15645–15655.

Qwen, Yang, A., Yang, B., et al., B. Z., 2025. Qwen2.5 Technical Report. https://arxiv.org/abs/2412.15115.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. M. Meila, T. Zhang (eds), *International Conference on Machine Learning (ICML)*, 139, 8748–8763.

Rolf, E., Klemmer, K., Robinson, C., Kerner, H., 2024. Position: Mission critical – satellite data is a distinct modality in machine learning. *International Conference on Machine Learning (ICML)*.

Roth, K., Kim, J., Koepke, A. S., Vinyals, O., Schmid, C., Akata, Z., 2023. Waffling around for performance: Visual classification with random words and broad concepts. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 15700–15711.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., Chadha, A., 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 1.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J., 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. *Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Wang, Q., Liu, S., Chanussot, J., Li, X., 2018. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2), 1155–1167.

Xu, H., Xie, S., Tan, X., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C., 2024. Demystifying CLIP data. *International Conference on Learning Representations (ICLR)*.

Zhou, K., Yang, J., Loy, C. C., Liu, Z., 2022. Learning to prompt for vision-language models. *International journal of computer vision (IJCV)*, 130(9), 2337–2348.

Zhou, W., Newsam, S., Li, C., Shao, Z., 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145, 197–209.

Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and remote sensing letters*, 12(11), 2321–2325.