The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4/W14-2025
9th International Workshop on Dynamic and Multi-dimensional GIS (DMGIS 2025), 22–24 August 2025, Beijing, China

Exploring the Potential of VLMs in Remote Sensing through Prompt Optimization

Weibin Ma[1], Ruiqian Zhang[1,*], Xiaogang Ning[1], Hanchao Zhang[1], Yixin Chen[1]

[1]Chinese Academy of Surveying and Mapping, Beijing, People's Republic of China
E-mail address: zhangrq@casm.ac.cn

Keywords: Vision-Language Models(VLMs), Remote Sensing, Prompt Optimization

Abstract

Vision-Language Models(VLMs) have demonstrated impressive capabilities in interpreting natural scene imagery. However, their generalization to domain-specific applications, such as remote sensing, remains underexplored. We address this gap by introducing a refined methodology centered on language-driven prompt optimization, with the aim of enhancing the adaptability of VLMs to remote sensing tasks. Specifically, we adopt a two-stage evaluation framework comprising Zero-Shot Prompting and Prompt-Informed Supervised Fine-Tuning. In the first stage, we assess the influence of prompt formulation on zero-shot performance. In the second stage, we further explore how the incorporation of optimized prompts during supervised fine-tuning can help reveal the model's generalization potential. Within this framework, we introduce two prompting strategies tailored for remote sensing: Cognitively-Guided Prompting (CogPrompt), which employs Chain-of-Thought reasoning to elicit structured and interpretable responses; and Knowledge-Injected Prompting (KnowPrompt), which incorporates domain-specific priors through existence assertions. We conducted a comprehensive evaluation of several open-source VLMs, including Qwen-VL, InternVL, and the LLaVA series, across multiple remote sensing benchmarks, including remote sensing object detection and captioning. To support our analysis, we propose a two-stage evaluation framework, including Zero-Shot Prompting and Prompt-Informed Supervised Fine-Tuning. Extensive experimental results show that prompt optimization consistently enhances overall detection and captioning performance across a range of metrics, and there is still significant room for improvement in the capabilities of VLMs for remote sensing tasks.

1. Introduction

In recent years, the development of Vision Language Models (VLMs) has marked a significant milestone in the field of artificial intelligence. These models, pre-trained on datasets containing billions of image-text pairs (Schuhmann et al., 2022, Jia et al., 2021), have demonstrated unprecedented capabilities in understanding, reasoning and generation, bridging visual perception and natural language. A representative example is CLIP (Radford et al., 2021), which uses contrastive learning to project images and texts in a shared embedding space, enabling robust zero-shot image classification. Building on this paradigm, the combination of Large Language Models (LLMs) with visual encoders has given rise to more advanced VLMs, such as LLaVA (Liu et al., 2023), Qwen2-VL (Wang et al., 2024), and InternVL (Chen et al., 2024). These models support a range of multi-modal tasks, including visual question answering, detailed image captioning, and instruction following, while also demonstrating a degree of commonsense reasoning. With strong generalization across diverse visual tasks, VLMs have become a key component in the development of general-purpose AI systems, offering a scalable and adaptable framework for addressing real-world multimodal challenges.

The notable success of VLMs in the general domain has spurred researchers' interest in transferring their capabilities to more specialized field, such as Remote Sensing. Among various adaptation strategies, Supervised Fine-Tuning has emerged as the predominant strategy. By fine-tuning with remote sensing-specific image-text datasets, researchers aim to extend the knowledge base of VLMs from natural scenes to the remote sensing domain. To improve adaptation efficiency while minimizing computational overhead and mitigating catastrophic forgetting, parameter-efficient fine tuning techniques, particularly Low-Rank Adaptation (LoRA) (Hu et al., 2022), have gained wide adoption. For example, RemoteCLIP (Liu et al., 2024) enhances the cross-modal retrieval capabilities in RS scenarios through continuous contrastive learning in RS-specific image-text pairs. RS-LLaVA (Bazi et al., 2024) represents the first adaptation of the LLaVA architecture to remote sensing, using an instruction-tuning dataset that covers multiple RS tasks to enable basic remote sensing dialogue capabilities. These studies have collectively advanced the application of VLMs in the remote sensing domain, demonstrating the feasibility of adapting general-purpose VLMs to specialized fields through SFT strategies.

Despite recent progress, the direct application of VLMs to the remote sensing domain remains challenging due to the unique characteristics of RS imagery. Unlike natural images (Lin et al., 2014, Deng et al., 2009), RS images are typically acquired from a top-down perspective, exhibit substantial scale variations, and contain numerous domain-specific land cover types (e.g., crop varieties, mining regions) that are rarely present in natural image datasets (Zhang et al., 2023, Zhang et al., 2021). Moreover, spatial relationships and fine-grained texture details in the images often convey more informative cues than those of the discrete objects. Consequently, directly applying prompt templates designed for natural scenes often fails to precisely guide the model to focus on the core elements of remote sensing tasks, leading to comprehension biases and performance degradation.

In addition, using large-scale remote sensing data exclusively for full or partial parameter fine-tuning can lead to catastrophic forgetting (Chen et al., 2023), where the model overfits the target domain and loses its previously acquired general knowledge and reasoning capabilities. This undermines both the generalization performance within RS tasks and the model's

broader utility as a versatile AI assistant. Addressing this trade-off—preserving general capabilities while effectively adapting to domain-specific tasks—has become a central challenge. Existing efforts have mainly focused on dataset construction and model architecture adaptation, while comparatively less attention has been paid to another important component: the language interaction interface, i.e., the prompt. Beyond serving as a query mechanism, the prompt can play a critical role in activating domain-relevant knowledge, guiding reasoning processes, and encoding task-specific priors.

To address these challenges, we propose a prompt optimization framework that evaluates the effectiveness of VLMs in remote sensing tasks through language-driven optimization, without the need for architectural modifications. Our overall framework consists of two core components: (1) Zero-Shot Prompting, which employs prompts as inference-time mechanisms to elicit more accurate and interpretable outputs from pretrained models; and (2) Prompt-Informed Training, which integrates prompt design into the fine-tuning process to guide model learning and improve task generalization. Within this dual-stage design, we introduce several advanced prompting strategies that probe a model's reasoning capacity and domain adaptability. Together, these components offer a lightweight and architecture-agnostic approach to enhance the transferability of VLMs in remote sensing scenarios. To validate our framework, we assess the capabilities of VLMs on both object detection and captioning tasks.

In summary, the main contributions of this work are as follows:

(1) We propose a prompt optimization framework for adapting VLMs to remote sensing tasks, which integrates two complementary strategies: Zero-Shot-Prompting and Prompt-Informed Fine-Training to enhance model adaptability during the inference and fine-tuning stages.

(2) We introduce two prompting strategies tailored for the remote sensing domain: (i) Cognitively-Guided Prompting based on Chain-of-Thought reasoning, which improves interpretability and inference quality; and (ii) Knowledge-Injected Prompting, which leverages domain priors through existence assertions and task reformulation to guide robust model behavior.

(3) We conduct comprehensive experiments on multiple VLMs across different remote sensing tasks to validate the effectiveness of the proposed strategies. And the results suggest that prompt design has a notable effect on model performance and transferability.

2. Prompt Optimization Framework

2.1 Framework Overview

Motivated by the limitations of existing VLM adaptation approaches in the remote sensing domain, we propose a prompt-centered optimization framework. The framework is structured in two stages, corresponding to the two core components of our proposed method: Zero-Shot Prompting and Prompt-Informed Supervised Fine-Tuning. This two-stage design allows us to first explore the intrinsic, zero-shot capabilities of existing models through advanced prompting, and then investigate how these optimized prompts can be integrated into the fine-tuning process to cultivate more robust and specialized models.

2.2 Zero-Shot Prompting

To evaluate the ability of VLMs to generalize to remote sensing tasks without any fine-tuning, we investigate a range of Zero-Shot Prompting strategies. In this setting, prompts are used as inference-time control mechanisms to guide model behavior, enabling the evaluation of language interface design without altering model weights. This component aims to isolate and analyze the model's intrinsic reasoning capacity and its adaptability to the unique characteristics of remote sensing imagery, such as complex spatial relations, scale variations, and uncommon object categories. To achieve this, we design two distinct classes of advanced prompting strategies: Cognitive-Guidance Prompting and Knowledge-Injection Prompting. As shown in Table 1 , which includes task description, range and edge handling, instruction and output format of all prompt.

2.2.1 Cognitively-Guided Prompting (CogPrompt) Remote sensing imagery presents unique challenges for VLMs, often characterized by significant scale variation, densely packed scenes, and objects defined more by textural or contextual cues than by distinct visual boundaries. When prompted with a direct query, VLMs may default to superficial pattern matching or memorized associations, leading to inaccurate or uninterpretable results. To address these issues, we propose Cognitively-Guided Prompting, a strategy designed to emulate the step-by-step analytical process employed by human experts. Built upon the CoT prompting paradigm, CogPrompt aims to elicit structured reasoning from pretrained models in a fully zero-shot setting.

Instead of requesting an immediate answer, CogPrompt prompts the model to generate intermediate reasoning steps, enclosed within a predefined "\n<think>...</think>" format. This decomposition of a complex task into interpretable subcomponents facilitates more robust handling of spatial relationships, reduces ambiguity, and activates relevant domain priors. In contrast to basic prompts—which issue direct queries such as "What is in the image?" or "Please locate the airplane"—CogPrompt encourages the model to articulate a coherent reasoning process prior to generating the final output, thereby improving both interpretability and inference quality.

To further enhance the structure and reliability of the reasoning process, we introduce Cognitively Guided Prompting, termed CogPrompt-G. While standard CogPrompt merely suggests that the model "think step-by-step" CogPrompt-G imposes an explicit three-stage reasoning sequence within the prompt. These stages are: (1) describe the context of the image, (2) identify relevant features, (3) determine the object class. An overview of the CogPrompt-G prompting pipeline is presented in Figure 1, which illustrates how each reasoning stage is explicitly encoded into the prompt to simulate a deliberate and expert-like decision-making process. This structured prompting format constrains the model's inference trajectory, promoting consistent and interpretable outputs across varied remote sensing scenes.

2.2.2 Knowledge-Injected Prompting (KnowPrompt) To enhance the robustness and domain adaptability of VLMs in remote sensing tasks, we propose KnowPrompt—a prompting strategy that injects external domain priors and reformulates task semantics through prompt engineering. In contrast to cognitively guided methods that simulate step-by-step reasoning, KnowPrompt focuses on reducing ambiguity and reinforcing

Method	Task Description	Range and Edge Handling	Instruction	Output Format
Baseline	Detect all objects belonging to the category 'Plane' in the image.	Provide the bounding boxes (between 0 and 1000, integer) and confidence (between 0 and 1, with two decimal places). If no object belonging to the category 'Plane' in the image, return 'No Objects'.	None	<answer>['Position': [x1, y1, x2, y2]] </answer>
Cog-Prompt	Detect all objects belonging to the category 'Plane' in the image.	Provide the bounding boxes (between 0 and 1000, integer) and confidence (between 0 and 1, with two decimal places). If no object belonging to the category 'Plane' in the image, return 'No Objects'.	Output the thinking process in <think></think>and final answer in <answer></answer>tags.	<think>... <think><answer> ['Position': [x1, y1, x2, y2]] </answer>
Cog-Prompt-G	Detect all objects belonging to the category 'Plane' in the image.	Provide the bounding boxes (between 0 and 1000, integer) and confidence (between 0 and 1, with two decimal places). If no object belonging to the category 'Plane' in the image, return 'No Objects'. Please think as followed: Observe this image carefully, and explain the background context of this picture. Output features unique to a top-down perspective. This includes: geometric shape, texture, color features and spatial relationship. For each candidate 'Plane' identified in step 2, perform a detailed verification. Confirm if it truly is a 'Plane'. Provide a clear reason for your decision.	Output the thinking process in <think></think>and final answer in <answer></answer>tags.	<think>... </think><answer> ['Position': [x1, y1, x2, y2]] </answer>
Know-Prompt	Detect all objects belonging to the category 'Plane' in the image.	There is at least one 'Plane' in this image. Provide the bounding boxes (between 0 and 1000, integer) and confidence (between 0 and 1, with two decimal places). If no object belonging to the category 'Plane' in the image, return 'No Objects'.	None	<answer>['Position': [x1, y1, x2, y2]] </answer>

Table 1. Prompt design comparisons for zero-shot object detection on the DOTA dataset. Each prompt consists of four components: Task Description (i.e., problem statement), Range and Edge Handling, Instruction, and Output Format.

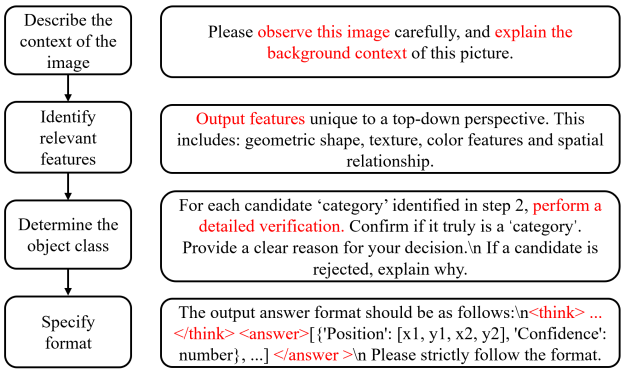

Figure 1. Prompt structure of CogPrompt-G. Each step is encoded in the prompt to support interpretable object analysis in remote sensing imagery.

domain alignment by modifying the informational content of the input query.

The first component of KnowPrompt, introduces explicit assumptions about object presence within the prompt to guide model attention and reduce uncertainty. Rather than requiring the model to jointly infer both existence and location (e.g., "Is there a soccer-ball-field in the image?"), the prompt asserts prior knowledge, such as: "There is at least one soccer-ball-field in the image. Where is it located?". This removes the need for the model to internally resolve object existence, allowing it to concentrate solely on the spatial localization subtask. In remote sensing imagery—where targets may be small, occluded, or visually ambiguous—such declarative priors can help reduce false negatives and improve localization robustness.

In addition to modifying semantic priors, KnowPrompt promotes representational consistency through task reformulation. This approach pairs the same image with varied task instructions to examine the stability of the model's internal representations. For example, a remote sensing image may be evaluated using both an object localization prompt ("Where is the airplane?") and a descriptive captioning prompt ("Describe what you see in the image."). By reformulating the task while keeping the visual input constant, we can assess the model's ability to generalize across analytical paradigms and maintain consistent interpretations of scene content.

2.3 Prompt-Informed Supervised Fine-Tuning

While zero-shot prompting effectively probes the inherent capabilities of pretrained VLMs, it remains limited by the model's fixed parameters and reliance on prompt engineering alone. To further enhance model adaptability and generalization in remote sensing scenarios, we introduce the second component of our framework: Prompt-Informed Supervised Fine-Tuning. This strategy builds on the insights from zero-shot evaluations by embedding optimized prompts into the model's training process. We hypothesize that not only the quantity and quality of training data, but also the structure and logic of the prompts, play a critical role in guiding model behavior.

To enable Prompt-Informed SFT, we first construct a structured instruction dataset based on existing remote sensing benchmarks. These datasets typically contain image–label pairs, like

the caption-level descriptions. Following a few-shot learning paradigm, we randomly sample a small, fixed number of images from each dataset to form the fine-tuning set, while reserving the remainder for inference-based evaluation.

To establish a baseline for comparison, we implement a standard SFT pipeline using simple, command-style prompts (e.g., "Describe the image" or "Locate the airplane") paired with conventional labels. On this basis, we introduce our Prompt-Informed SFT strategy, denoted as CogPrompt_SFT in which each training sample is reformulated using the CogPrompt strategy introduced in Section 2.2.1.

For each image in the fine-tuning set, we generate a high-quality, reasoning-aware response to match the CogPrompt structure (e.g., in the format `"<think> Reasoning steps... </think> Final answer"`). These structured outputs are synthesized using a strong teacher model (Gemini 2.5 Pro) which enables the generation of fluent and logically coherent responses following the CoT format. This process yields a set of (image, optimized prompt, structured response) triplets that form the training data for CogPrompt_SFT.

We then fine-tune the target VLMs on this prompt-informed dataset using the LoRA technique. By updating only a small number of additional trainable parameters, LoRA provides an efficient and scalable fine-tuning approach that adapts the model to remote sensing tasks while mitigating the risk of catastrophic forgetting of general pre-trained knowledge.

3. Experimental Setup

This section outlines the experimental setup used to evaluate the proposed prompting strategies, covering the datasets and tasks, model backbones, evaluation metrics, and training protocols for both zero-shot inference and supervised fine-tuning.

3.1 Datasets and Tasks

To comprehensively evaluate the performance of various VLMs on remote sensing tasks, we selected four different datasets, covering two basic capabilities: object detection and image captioning.

3.1.1 Object Detection Our object detection task is trained and evaluated on a custom-sampled subset derived from the DOTA(Xia et al., 2018) and DIOR(Li et al., 2020) datasets. Both are prominent large-scale benchmarks renowned for their high-resolution aerial images, which present significant challenges such as complex backgrounds, dense object distributions, and wide scale variations.

Due to the high computational cost of evaluating all 15 annotated categories, we first conducted an initial inference using the Qwen2-7B model to identify object classes with relatively high detection accuracy. Based on this analysis, we selected three representative categories—planes, soccer-ball-fields, and tennis-courts—for focused evaluation. These classes were chosen for their semantic clarity, consistent visual appearance, and relatively low inter-class ambiguity, making them suitable for assessing object-level reasoning in VLMs.

Among them, planes and tennis-courts contain a large number of labeled RGB images. For these, we randomly sampled 500 images per class to ensure sufficient evaluation coverage while maintaining computational efficiency. In contrast, soccer-ball-fields included only 378 images in total, and all were used in the experiment.

3.1.2 Image Captioning For the image captioning task, we evaluate the models' ability to generate descriptive and contextually appropriate captions using three datasets: RSICD (Lu et al., 2017), Sydney Captions (Qu et al., 2016). We also use DIOR-RSVG (Zhan et al., 2023), a captioning-oriented variant derived from the DIOR dataset, which requires describing specific regions and their spatial relationships to assess performance on a more complex visual grounding task.

Both RSICD and Sydney Captions are used in their entirety. Due to computational constraints, however, it was not feasible to process the full DIOR-RSVG dataset. To mitigate this, we randomly sampled approximately 2,300 image-text pairs to construct a representative subset. This subset was subsequently partitioned into training and testing sets using a 3:7 split ratio, ensuring a consistent and reliable evaluation protocol under our experimental settings. We present the outputs of DIOR-RSVG dataset in Figure 2, which provides a qualitative example from the DIOR-RSVG dataset, illustrating the task and a typical model output.

3.2 Model Backbones

To ensure the generalizability of our findings, we evaluate a diverse set of publicly available VLMs that differ in architecture, scale, and pretraining paradigms. To enhance the relevance and reproducibility of our study, we selected widely adopted representatives from three major VLM series: Qwen-VL, LLaVA, and InternVL.

Qwen-VL. We include Qwen2-VL-7B and Qwen2.5-VL-7B (here-after referred to as Qwen2 and Qwen2.5), developed by Alibaba Cloud. These models integrate a Vision Transformer (ViT) with the Qwen large language model, and are optimized for multilingual, multimodal tasks, particularly those requiring fine-grained visual grounding.

LLaVA. We adopt LLaVA-1.5-7B (here-after referred to as LLaVA), an open-source VLM that aligns ViT features with a language model via a lightweight projection matrix. Its architecture emphasizes simplicity and efficiency, and has been widely used as a baseline in multimodal learning research.

InternVL. We also include InternVL3-2B (here-after referred to as InternVL), a large-scale foundation model that scales the vision encoder and progressively aligns it with a language model. It is designed to support a wide range of visual perception and instruction-following tasks across domains.

All selected models are publicly available and used without architectural modifications. Further details on parameter settings and evaluation protocols are provided in subsequent sections.

3.3 Evaluation Metrics

To ensure a rigorous and task-appropriate evaluation of model performance, we adopt standard metrics for both object detection and image captioning tasks.

For the object detection task on the DOTA dataset, we report Precision (P), Recall (R), F1 Score (F1), and Average Precision (AP). AP, computed as the area under the precision-recall curve, provides a class-wise summary of detection performance across varying confidence thresholds. Given the high spatial resolution and wide scene coverage of remote sensing imagery, predicted bounding boxes from VLMs often exhibit low spatial overlap with ground-truth annotations, especially compared

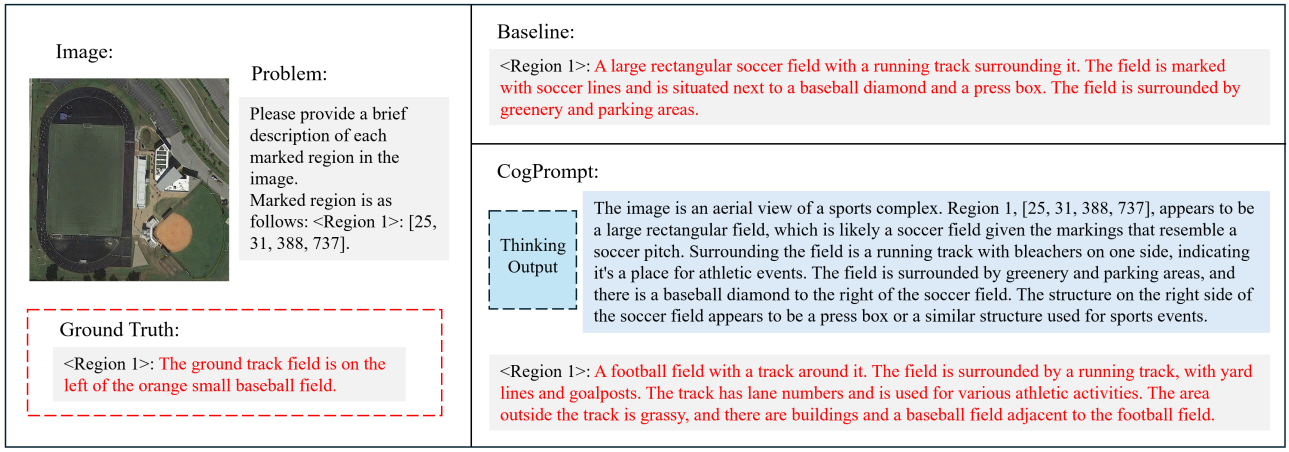

Figure 2. Details of outputs from VLMs when using Baseline and CogPrompt.

to natural image scenarios. To accommodate this, we relax the intersection-over-union (IoU) threshold used to determine true positives from the conventional 0.5 to 0.1, allowing for a more inclusive evaluation of object localization behavior under weakly aligned predictions.

For the image captioning task, evaluated on the RSICD, Sydney Captions, and DIOR-RSVG datasets, we use the BLEU-1 (hereafter referred to as BLEU) metric, which measures single-token precision between generated and reference captions, reflecting basic fluency. When it comes to caculate BLEU, we do not consider the thinking part, but only calculate the similarity between the true value and the predicted result part, that is, calculate the similarity of the part marked in red in the Figure 2.

3.4 Implementation Details

To evaluate the isolated effect of prompt design on model performance, we conduct a controlled zero-shot prompting experiment in which the VLM remains fixed—i.e., without any parameter updates or architectural changes. The prompt itself serves as the sole independent variable. We design a set of prompting strategies that differ in reasoning structure and semantic content, aiming to assess their influence on the model's ability to process and interpret remote sensing imagery.

The prompting strategies compared in this study include:

Baseline: A minimal, direct query such as "What is in the image?" or "Please locate the airplane." This serves as a baseline for evaluating the model's raw zero-shot capability without additional guidance.

CogPrompt: A cognitively guided prompt that activates the model's internal CoT reasoning by instructing it to "think step-by-step" before answering, thereby improving interpretability and reasoning depth.

CogPrompt-G: A guided variant of CogPrompt that enforces a fixed four-stage reasoning process—format specification, key feature identification, candidate scanning, and verification—explicitly encoded into the prompt to structure the model's analytical flow.

KnowPrompt: A knowledge-injected prompt that asserts the existence of the target object. This reduces ambiguity in object presence detection and focuses the model on spatial localization.

In our experiments, a desktop workstation equipped with an Intel(R) Core(TM) i9-10980XE CPU @ 3.00GHz and 4 NVIDIA GeForce A800 GPUs with total 320G of memory was utilized. The operating system employed was Ubuntu 20.04, and the experiments were conducted on the PyTorch platform. During training process, we set the evaluation batch size to 2 per device and an initial learning rate of 2e-5. The model was trained for a total of 20 epochs. And the rank of LoRA matrices is 4. Meanwhile, for Prompt-Informed SFT training, we conducted experiments on the DIOR-RSVG dataset, where we randomly selected 100 images as the training set, and reasoned on the rest of the images.

4. Results and Analysis

This section presents a comprehensive empirical evaluation of the proposed prompt optimization framework. We assess the effectiveness of each core component across multiple remote sensing tasks and VLMs. The analysis is organized in two parts: the first examines the impact of direct prompting strategies under zero-shot settings, and the second investigates the benefits of incorporating prompts into supervised fine-tuning.

4.1 Zero-Shot Prompting Results

To evaluate the effectiveness of our proposed zero-shot prompting strategies—CogPrompt and KnowPrompt—we conduct a series of experiments across two representative remote sensing tasks: object detection and image captioning. These experiments aim to assess the models' inference behavior under different prompt formulations, without any fine-tuning or weight updates. For clarity and structured analysis, we present the results in three subsections: CogPrompt, KnowPrompt in Object Detection and CogPrompt in Captioning. Each subsection includes both quantitative evaluation and qualitative examples to highlight the prompt-induced behavioral differences.

4.1.1 Results of CogPrompt in Object Detection To validate the effectiveness of the proposed CogPrompt, we conduct a series of comparative experiments on a standard remote sensing object detection task using multiple representative VLMs. Detailed model configurations and implementation settings are provided in Sections 3.4 and 3.2.

The experimental results for all prompting strategies across different object categories are summarized in Table 2. These results demonstrate that overall, the performance of current VLMs

VLM	Method	Soccer-ball-field				Tennis-court				Plane			
		P (%)	R (%)	F1 (%)	AP (%)	P (%)	R (%)	F1 (%)	AP (%)	P (%)	R (%)	F1 (%)	AP (%)
LLaVA	Baseline	16.06	4.87	7.48	0.97	19.44	2.14	3.86	0.64	34.66	4.47	7.92	1.7
	CogPrompt	20.29	17.45	7.92	1.7	25	6.94	18.77	4.24	28.38	6.58	10.86	1.97
	CogPrompt-G	26.23	7.35	11.48	2.48	22.86	4.04	6.87	1.11	29.93	3.1	5.62	1.19
	KnowPrompt	17.96	7.81	10.89	1.66	27.48	3.64	6.43	1.21	31.42	4.29	7.55	1.53
Qwen2	Baseline	55.36	19.5	28.84	12.65	62.34	15.20	24.4	9.95	68.09	15.28	24.94	10.81
	CogPrompt	57.36	30.03	39.42	20.56	47.92	17.65	25.8	8.92	47.2	15.31	23.13	8.13
	CogPrompt-G	31.4	24.69	27.64	10.25	26.84	6.27	10.16	1.87	34.79	7.05	11.72	2.68
	KnowPrompt	55.68	23.11	32.67	14.03	62.85	16.08	25.6	11.03	67.5	15.7	25.48	11.05
Qwen2.5	Baseline	38.21	23.43	29.04	10.18	35.09	9.49	14.94	3.68	15.07	5.7	8.27	0.98
	CogPrompt	49.26	21.07	29.52	11.22	37.44	8.06	13.27	3.32	17.34	5.35	8.19	1.07
	CogPrompt-G	39.68	22.97	29.1	10.57	38.57	8.19	13.51	3.39	19.42	4.12	6.8	0.91
	KnowPrompt	32.66	22.8	26.85	10.44	30.68	7.79	12.42	2.75	16.18	6.03	8.78	1.21
InternVL	Baseline	13.29	9.95	11.38	1.97	18.79	5.97	9.06	1.48	24.64	3.51	6.14	0.98
	CogPrompt	21.23	15.31	17.79	4.58	33.73	11.32	16.69	6.01	25.05	4.05	6.98	1.3
	CogPrompt-G	12.57	12.86	12.72	2.78	25.28	6.88	10.81	2.24	18.24	3.17	5.4	0.75
	KnowPrompt	15.27	14.24	14.74	2.72	24.26	6.67	10.47	1.71	23.09	4.33	7.29	1.21

Table 2. Performance comparison of VLMs in remote sensing object detection on DOTA dataset.

on remote sensing object detection remains modest, with considerable variance across object categories. This highlights the substantial challenge of transferring pretrained VLM capabilities to the remote sensing domain. Within this context, our proposed CogPrompt strategy consistently improves model performance relative to the baseline prompts.

Notably, the impact of CogPrompt is especially pronounced for models with limited initial performance, such as LLaVA and InternVL. Across most object categories, CogPrompt delivers substantial gains. For example, InternVL's AP on the "Tennis court" category increases from 1.48% to 6.01%, representing a fourfold improvement. Similarly, LLaVA's AP on the same category rises from 0.64% to 4.24%, a more than sixfold increase. These gains are not confined to AP alone; recall and F1 score also improve significantly. For instance, LLaVA's recall on the "Soccer-ball field" category improves from 4.87% to 17.45%. These results indicate that for models unable to directly establish strong visual-to-semantic mappings in complex remote sensing scenes, CogPrompt serves as an effective cognitive scaffold, guiding the model through structured reasoning and enhancing its localization ability. In contrast, the effect of CogPrompt on stronger base models, such as Qwen2, is more nuanced. In some categories—such as "Soccer-ball field"—CogPrompt still yields notable improvements. Qwen2's AP improves from 12.65% to 20.56%, and Qwen2.5's Precision increases from 38.21% to 49.26%.

As shown in Table 3, we also utilize Qwen2.5 to conduct a detailed comparative analysis on DIOR dataset to evaluate the object detection performance of CogPrompt. The experiment is performed on three distinct and challenging categories from remote sensing imagery: Dam, Golf Field, and Expressway Toll Station. We report AP and F1 as the primary evaluation metrics. The results clearly indicate that CogPrompt consistently and significantly outperforms the Baseline across all tested categories. For the 'Dam' and 'Golf Field' categories, our method demonstrates solid improvements. For instance, in the 'Dam' category, the AP increases from 0.6764 to 0.7368, and the F1 rises from 0.8000 to 0.8571. This points to a more balanced and accurate detection capability. The most remarkable advantage of our method is observed in the 'Expressway Toll Station' category, which is notoriously difficult due to small, densely packed objects and complex backgrounds. CogPrompt achieves an AP of 0.2777, a more than five-fold improvement over the Baseline's 0.0501. This substantial gain underscores CogPrompt's enhanced robustness and its superior capability in handling complex scenes where the baseline system evidently struggles.

Method	Category	F1(%)	AP(%)
Baseline	Baseball Field	74.11	55.68
	Golf Field	52.38	53.69
	Basketball Field	59.15	40.86
CogPrompt	Baseball Filed	74.74	58.69
	Golf Field	77.42	65.63
	Basketball Field	71.79	54.28

Table 3. Performance comparison of Baseline and CogPrompt on the DIOR dataset.

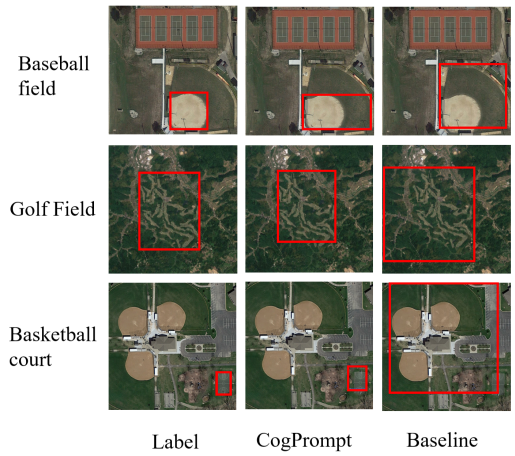

Figure 3. Qualitative comparison of Baseline and CogPrompt using Qwen2.5 on the DIOR dataset.

Figure 3 illustrates that after using CoT, compared with the second and third col, it is obvious that after using CogPrompt, the model's output can better grasp the key point of the question. After output thinking process, model successfully recog-

nizes basketball court. However, the model without CogPrompt incorrectly identified the baseball field as a basketball court. Moreover, it exhibits higher accuracy in localizing objects.

However, in simpler tasks where the model already performs well, the added reasoning steps introduced by CogPrompt may impose unnecessary inference overhead, potentially reducing localization precision. These observations suggest that the effectiveness of CogPrompt is task-dependent. Its benefits are most evident in complex scenes requiring multi-step contextual reasoning, while in simpler scenarios, its application must be more selective to avoid diminishing returns.

4.1.2 Results of CogPrompt-G in Object Detection As shown in Table 2, we further evaluate the performance of CogPrompt-G, the guided variant of our reasoning-based prompting strategy. While CogPrompt-G introduces a more structured multi-step reasoning process during inference, its effectiveness is observed to be highly inconsistent across models and categories. While notable improvements are observed in certain settings—such as the LLaVA model on the "Soccer-ball-field" category, where both precision and recall increase substantially—the strategy often yields inconsistent or even negative effects in other scenarios. For example, performance degradation is observed on several categories under Qwen2 and InternVL, suggesting limited robustness. These results suggest that although guided reasoning may benefit models struggling with contextual organization, it can become counterproductive for models already equipped with strong intrinsic reasoning capabilities. Overall, CogPrompt-G should be applied selectively, depending on both the model architecture and the task complexity. Given these mixed results, we did not extend CogPrompt-G to the captioning task.

VLM	Method	Sydney	RSICD	DIOR-RSVG
LLaVA	Baseline	5.92	7.23	3.88
	CogPrompt	17.38	7.57	14.29
Qwen2	Baseline	17.68	6.1	16.38
	CogPrompt	27.83	10.9	21.61
Qwen2.5	Baseline	13.06	6.7	9.05
	CogPrompt	27.59	9.06	27.17
InternVL	Baseline	15.36	7.17	9.41
	CogPrompt	18.68	7.32	16.02

Table 4. Performance comparison of VLMs in remote sensing captioning.

4.1.3 Results of KnowPrompt in Object Detection To evaluate the effect of KnowPrompt, we compare it with the baseline across all VLMs in the object detection task (see Table 2). KnowPrompt introduces a declarative prior by stating that a target object exists in the image, aiming to encourage detection in lower-confidence scenarios. Quantitative results show that this strategy has the potential to improve detection performance across multiple evaluation metrics, including Precision, Recall, F1, and AP, although the extent and consistency of improvements vary across models and categories. For example, LLaVA's Recall on the "Soccer-ball-field" category increases from 4.87% to 7.81%, while InternVL improves from 9.95% to 14.24%. Qwen2 also exhibits upward trends across all three categories. Similarly, albeit smaller, gains can be observed in other metrics such as F1 and AP in several cases.

However, these improvements are not consistent across all models and categories. In several cases, performance on Precision, Recall, F1, or AP declined slightly. For example, among the 12 model–category combinations evaluated, 10 showed improvements in AP, suggesting a generally positive but non-uniform impact. This suggests that while KnowPrompt can guide the model to recover low-confidence targets, the assertive nature of the prompt may also introduce misdetections when the asserted object presence does not align with the model's internal visual evidence. As a result, the overall effect is sensitive to both model characteristics and category-specific complexity.

4.1.4 Results of CogPrompt in Captioning To further evaluate the generalizability of the Chain-of-Thought strategy, we extended its application from the detection task to image captioning. Experiments were conducted on three public remote sensing datasets: DIOR-RSVG, Sydney, and RSICD, to analyze the comprehensive impact of Chain-of-Thought prompting on the model's scene understanding and caption capabilities. The experimental results are shown in Table 4, respectively. For tasks such as remote sensing image captioning and visual question answering, which require reasoning beyond basic object recognition, the CoT-based prompting strategy demonstrates clear advantages.

Compared to the Baseline prompt, the proposed CogPrompt strategy consistently improved performance across all tested models and datasets. Notably, the Qwen2.5 model achieved a significant increase in BLEU score on the DIOR dataset, from 9.05% to 27.17%. Similarly, the LLaVA model on the Sydney dataset saw its BLEU score rise from 5.92% to 17.38%, representing a nearly threefold improvement. These results indicate that CogPrompt enhances the model's capacity for semantic abstraction and logical structuring, leading to more coherent and contextually relevant caption outputs.

4.2 Results of Prompt-Informed SFT in Captioning

We investigate whether fine-tuning VLMs on a dataset enriched with advanced prompting strategies can produce more accurate, robust, and generalizable models compared to standard SFT methods. Table 4 and Table 5 show the performance of each model on the DIOR-RSVG dataset before and after SFT, respectively. All SFT for these models was performed on an instruction dataset that included these advanced prompts.

VLM	Method	BLEU (%)
LLaVA	Baseline_SFT	18.41
	CogPrompt_SFT	22.25
Qwen2	Baseline_SFT	22.94
	CogPrompt_SFT	28.07
Qwen2.5	Baseline_SFT	17.41
	CogPrompt_SFT	31.97
InternVL	Baseline_SFT	17.35
	CogPrompt_SFT	21.79

Table 5. Performance comparison of VLMs in remote sensing captioning via SFT on the DIOR-RSVG dataset.

The result using CogPrompt is significantly and comprehensively superior to the Baseline method, regardless of whether it is in a zero-shot or post-fine-tuning setting, and regardless of the VLM in question. In all 8 comparative tests shown in the charts, the BLEU score of the CoT group was higher than that of the Base group. This provides initial proof that CoT, as a universal prompting strategy, has a generally positive effect on improving

a model's scene understanding and text generation capabilities for the remote sensing image captioning task. This indicates that when a model relies entirely on its pre-trained knowledge to understand unfamiliar remote sensing images, the guidance from CoT is crucial. By introducing CoT in the prompt to SFT, all VLMs' BLEU have a significant and consistent improvement over their Bases and CoTs.

5. Conclusion

While VLMs have achieved remarkable success in general-domain tasks, their application to remote sensing remains limited due to domain-specific challenges such as scale variability, top-down viewpoints, and sparse semantic alignment. These issues hinder the direct transfer of pre-trained models to geospatial contexts. Therefore, in this work, we investigate the adaptation of VLMs to remote sensing tasks through a language-driven prompt optimization framework. Firstly, we systematically investigate the effects of this approach under both zero-shot inference and supervised fine-tuning scenarios. Furthermore, we propose zero-shot Prompting Results and Prompt-Informed SFT Results exploring CogPrompt, CogPrompt-G and Know-Prompt as a simple and broadly applicable method for solving remote sensing tasks. Through experiments on both object detection tasks and captioning tasks processed on remote sensing, we find that with the use of chain-of-thought prompt, VLMs would have a better understanding of the relationship between different regions and more precise localization. Moreover, our results also show that the effectiveness of guided prompts is highly dependent on the model's intrinsic capabilities , while knowledge-injected prompts boost overall detection performance by improving recall.

Acknowledgment

This work was supported by the National Natural Science Foundation of China [grants number 42201440 and 42401500] and the Fundamental Research Funds for Chinese Academy of Surveying and Mapping [grant number AR2410].

References

Bazi, Y., Bashmal, L., Al Rahhal, M. M., Ricci, R., Melgani, F., 2024. RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9), 1477.

Chen, Y., Sikka, K., Cogswell, M., Ji, H., Divakaran, A., 2023. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L. et al., 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 24185–24198.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 248–255.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al., 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2), 3.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *International conference on machine learning*, PmLR, 4904–4916.

Li, K., Wan, G., Cheng, G., Meng, L., Han, J., 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159, 296–307.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Zitnick, C. L., 2014. *Microsoft COCO: Common Objects in Context*. Springer International Publishing.

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.

Liu, H., Li, C., Wu, Q., Lee, Y. J., 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36, 34892–34916.

Lu, X., Wang, B., Zheng, X., Li, X., 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4), 2183–2195.

Qu, B., Li, X., Tao, D., Lu, X., 2016. Deep semantic understanding of high resolution remote sensing image. *international conference on computer, information and telecommunication systems*, IEEE, 1–5.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *International conference on machine learning*, PmLR, 8748–8763.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35, 25278–25294.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W. et al., 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. Dota: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974–3983.

Zhan, Y., Xiong, Z., Yuan, Y., 2023. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–13.

Zhang, R., Newsam, S., Shao, Z., Huang, X., Wang, J., Li, D., 2021. Multi-scale adversarial network for vehicle detection in UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180, 283–295.

Zhang, R., Zhang, H., Ning, X., Huang, X., Wang, J., Cui, W., 2023. Global-aware siamese network for change detection on remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 199, 61–72.