



HOKKAIDO UNIVERSITY

Title	A Study on Visual Recognition with Large Multimodal Models and Its Applications
Author(s)	甘, 耀宗
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16993号
Issue Date	2026-03-25
DOI	https://doi.org/10.14943/doctoral.k16993
Doc URL	https://hdl.handle.net/2115/99837
Type	doctoral thesis
File Information	Yaozong_Gan.pdf, 全文



A thesis for the degree of Doctor of Philosophy

**A Study on Visual Recognition with Large
Multimodal Models and Its Applications**

大規模マルチモーダルモデルを用いた視覚認識
およびその応用に関する研究



Yaozong Gan

Graduate School of Information Science and Technology

Hokkaido University

March, 2026

Contents

1	Introduction	10
1.1	Background	10
1.2	Proposition in this Thesis	11
1.3	Organization of this Thesis	13
2	Related Works	15
2.1	Introduction	15
2.2	General Large Multimodal Model Methods	15
2.3	Traditional Traffic Sign Recognition Methods	18
2.4	Multitask Remote Sensing Recognition Methods	20
2.5	Problems to Be Solved in This Thesis	22
2.6	Conclusion	23
3	Midlevel Feature Matching-Based Zero-Shot Traffic Sign Recognition	24
3.1	Introduction	24
3.2	Method	25
3.2.1	Traffic Sign Detection	25
3.2.2	Midlevel Feature Extraction	27
3.2.3	Zero-Shot Matching	29
3.3	Experiments	30
3.3.1	Experimental Settings	30
3.3.2	Experimental Results	31
3.4	Conclusion	37
4	Cross-domain Traffic Sign Recognition via Few-shot In-context Learning	38
4.1	Introduction	38
4.2	Methodology	38
4.2.1	Traffic Sign Detection	39
4.2.1.1	Traffic Sign-specific Output Module	40
4.2.1.2	Traffic Sign Extraction Module	40
4.2.2	Cross-domain Few-shot In-context Learning for TSR	41

4.3	Experiments	42
4.3.1	Experimental Settings	42
4.3.2	Experimental Results	43
4.3.3	Ablation Studies	46
4.4	Conclusion	49
5	Zero-shot Fine-grained Traffic Sign Recognition via Cross-domain Multi-step Thinking	50
5.1	Introduction	50
5.2	Method	51
5.2.1	Traffic Sign Extraction	51
5.2.1.1	Segmentation	51
5.2.1.2	Extraction	51
5.2.2	Milti-step Thinking	53
5.2.2.1	Prior Knowledge Generation	53
5.2.2.2	Multi-step Reasoning	58
5.3	Experiments	59
5.3.1	Experimental Settings	59
5.3.2	Experimental Results	60
5.3.3	Ablation Studies	62
5.3.3.1	Thinking Strategies	62
5.3.3.2	Hypothesis and Coordinate	69
5.3.3.3	Thinking Orders	71
5.3.3.4	Flexibility and Extensibility	74
5.3.3.5	Inference Speed	74
5.3.3.6	Significant Domain Shift	75
5.3.3.7	Impact of Description Granularity	76
5.4	Discussion	78
5.4.1	Test Set Contamination	78
5.4.2	Significance and Practical Value	78
5.4.3	Limitation	79
5.4.3.1	Automated Identification of Similar Signs	79
5.4.3.2	Robustness under Diverse Weather Conditions	79
5.4.3.3	Computational Complexity and Latency	79
5.4.3.4	Integration of Standardized Sign Taxonomies	80
5.5	Conclusion	81

6	Multitask Remote Sensing Recognition Based on Large Multimodal Models	82
6.1	Introduction	82
6.2	GeoJapan Data Construction	82
6.2.1	Study Region	82
6.2.2	Dataset Curation and Labeling Protocol	84
6.3	Method	86
6.3.1	Visual Feature Extraction and Projection	87
6.3.1.1	Linear Projection and Patch Embedding	87
6.3.1.2	Spatial Attention Calculation	89
6.3.1.3	Visual Feature Projection	90
6.3.2	Multimodal Fusion	90
6.3.2.1	In-Context Learning	91
6.3.2.2	Text Prompt Encoding and Visual Feature Fusion	91
6.3.3	Large Language Model Inference	92
6.3.3.1	Self-Attention Calculation	92
6.3.3.2	Instruction-Tuning Based on LoRA	94
6.4	Experiments	94
6.4.1	Experimental Setup	94
6.4.2	Evaluation Metrics	95
6.4.3	Quantitative Results	96
6.4.3.1	Geospatial Object Classification	96
6.4.3.2	Scene Description	98
6.4.4	Qualitative Results	101
6.4.5	Ablation Studies	106
6.4.5.1	In-Context Learning	106
6.4.5.2	Training Epochs	107
6.4.5.3	Resource Efficiency Analysis	108
6.4.5.4	Attention Map	109
6.5	Discussion	109
6.5.1	Methodological Analysis	109
6.5.2	Key Challenges and Proposed Solutions	110
6.5.3	Implications	110
6.5.4	Limitations	110
6.5.5	Future Directions	111
6.6	Conclusion	111

7	Conclusions	113
7.1	Overview of the Proposition in This Thesis	113
7.2	Future Directions	115
	Acknowledgements	116
	Bibliography	117
	Achievements of the Author	125

List of Figures

2.1	Research map of related research.	23
3.1	Illustration of the proposed TSR pipeline. We first employ the ViT-Adapter to locate target traffic signs within road images. These localized regions are then processed by the TSM module, which utilizes midlevel representations derived from CNNs.	26
3.2	Configuration of the early, midlevel, and last layers across various CNN architectures.	28
3.3	Examples of matching results for different methods on the GTSRB dataset. . . .	34
3.4	Examples of matching results for different methods on the Sapporo urban road dataset.	36
4.1	Overview of the proposed TSR framework. We first apply the proposed TSD network to detect and extract traffic signs from road images. Next, template traffic signs are fed into the LMM to generate description texts. These descriptions encode essential information about the shape, color, and structural components of traffic signs, thereby enhancing the LMM’s reasoning capability for TSR. . . .	39
4.2	Top-1 recognition results of the proposed cross-domain few-shot in-context TSR method. Two representative examples are provided for the benchmark dataset (GTSRB) and the real-world Sapporo urban road dataset. For the Sapporo samples, we illustrate the full procedure, from traffic sign detection in original road images to final recognition by the LMM using the proposed approach.	47
4.3	Top-1 recognition results of the baseline-o, baseline, and our method.	48
4.4	Example of incorrect recognition due to inaccurately generated description texts.	49

5.1	Schematic of the proposed CdMT framework. The approach initially localizes traffic signs and subsequently employs multi-step reasoning to achieve fine-grained TSR.	52
5.2	Generation of context descriptions. The sign’s coordinates, background context, and prior traffic sign hypothesis are encoded as (orange), (blue), and (yellow), respectively.	54
5.3	Generation of characteristic descriptions. In-context learning is introduced to guide the LMM in capturing the key features of traffic signs.	55
5.4	Generation of differential descriptions. Discriminative features among visually similar signs are emphasized to facilitate the fine-grained reasoning of the LMM.	56
5.5	Qualitative recognition results generated by our method on the Sapporo dataset.	63
5.6	Qualitative recognition results generated by our method on the GTSRB dataset.	64
5.7	Qualitative recognition results generated by our method on the BTSD dataset.	65
5.8	Qualitative recognition results generated by our method on the TT-100K dataset.	66
5.9	Qualitative recognition results generated by our method on the Yokohama dataset.	67
5.10	Recognition examples of baseline and proposed CdMT (Gpt-4o).	70
5.11	Visualization of recognition failures yielded by the baseline and the proposed CdMT (Gpt-4o).	71
5.12	Visual results of traffic sign extraction achieved by the proposed module across various segmentation architectures.	74
5.13	CdMT recognition results on samples with significant domain shifts.	76
5.14	Representative characteristic descriptions across short, medium, and long settings.	78
6.1	Overview of the study area. Kawasaki corresponds to an urban industrial zone, whereas Sapporo includes large natural regions with forests and farmland.	83
6.2	Distribution of geospatial object categories in the GeoJapan dataset.	84
6.3	Examples of geospatial object categories. (a) Comparison between Japanese remote sensing imagery and general remote sensing imagery, showing the same object categories from both the GeoJapan and GeoChat datasets. (b) Illustration of additional geospatial object categories that are newly included in GeoJapan.	85
6.4	Example of instruction-tuning dataset for a GeoJapan image.	86

6.5	Scene description generation.	87
6.6	Overview of the proposed GFF.	88
6.7	Examples of GFF outputs for geospatial object classification. Three correctly recognized cases and one misclassification are shown.	102
6.8	Qualitative results of GFF on the scene description task. Ground truth descrip- tions and outputs from different models are shown for comparison. Text high- lighted in blue indicates portions of the GFF-generated descriptions that match the reference annotations and capture key scene information.	104
6.9	Qualitative results of GFF on the scene description task. Ground truth descrip- tions and outputs from different models are shown for comparison. Text high- lighted in blue indicates portions of the GFF-generated descriptions that match the reference annotations and capture key scene information.	105
6.10	GFF scene description results on the GeoJapan dataset over training epochs. . . .	108
6.11	Visualization of attention maps, with heatmaps showing regions emphasized by GFF.	109

List of Tables

3.1	Dimensions of the different layers used in our method.	29
3.2	Top- k accuracy of different methods on the GTSRB dataset.	32
3.3	Top- k accuracy of different methods on the Sapporo urban road dataset.	33
3.4	Analysis of inference latency per target image across distinct backbones and datasets. “Template Count” denotes the number of reference categories used for matching.	35
4.1	Top- k accuracy of different methods on four datasets.	44
5.1	Quantitative comparison of Top- k zero-shot fine-grained TSR performance across five diverse datasets. We benchmark the proposed approach against state-of-the-art methods. Bold and underlined values denote the best and second-best results, respectively. All metrics represent the mean accuracy calculated over five independent trials.	61
5.2	Evaluation of Top- k zero-shot fine-grained TSR performance across distinct reasoning strategies. The abbreviations “Cont*”, “Char*”, and “Diff*” correspond to context, characteristic, and differential descriptions, respectively. Best results are highlighted in bold. The symbol “-” implies that context descriptions could not be generated due to the unavailability of original road images.	68
5.3	Evaluation of Top- k zero-shot fine-grained TSR performance employing diverse context description generation strategies. The best results are shown in bold. . . .	72
5.4	TSR performance under different reasoning configurations. The labels “w” and “w/o” denote the application or omission of the thinking steps, respectively. . . .	73
5.5	Evaluation of extraction module impact on TSR performance.	75
5.6	Inference speed for each traffic sign. “s” denotes seconds.	75
5.7	TSR results of CdMT-Gpt-4o with different characteristic description lengths. . .	77

5.8	Analysis of computational overhead for CdMT-Gpt-4o. The variables N_C and N_D represent the number of template classes and visually similar signs, respectively. “Phase Type” distinguishes whether a stage is performed online for inference or offline for preprocessing. All time measurements are reported in seconds.	80
6.1	Geospatial object classification performance on the GeoJapan dataset.	97
6.2	Detailed classification performance across geospatial categories on GeoJapan, sorted by F1-Score.	99
6.3	Scene description performance on the GeoJapan dataset.	100
6.4	Impact assessment of in-context learning on GFF classification metrics using the GeoJapan dataset.	106
6.5	Impact assessment of in-context learning on GFF scene description metrics using the GeoJapan dataset.	106
6.6	Temporal evolution of classification performance on the GeoJapan benchmark throughout the training process.	107
6.7	Analysis of resource utilization and latency for GFF across distinct tasks.	108

Chapter 1

Introduction

This chapter outlines the background and objectives of the study, followed by the overall structure of the thesis.

1.1 Background

Deep learning has become a cornerstone of artificial intelligence (AI), driving remarkable progress across different tasks. Since the success of convolutional neural networks (CNNs) in large-scale image recognition, a series of CNN-based architectures, such as ResNet [1], EfficientNet [2], have significantly advanced computer vision by learning hierarchical and highly representative visual features. More recently, Transformer-based models [3] have further revolutionized this field by introducing a new paradigm for global context modeling and attention-based representation learning. Meanwhile, in natural language processing (NLP), models such as BERT [4], the GPT-2 [5] have demonstrated the effectiveness of large-scale pre-training in capturing structured knowledge from data. The convergence of these advances has led AI to evolve from unimodal perception to multimodal understanding, in which information from vision, language, and other modalities is jointly exploited for more comprehensive reasoning and decision-making.

Building on these developments, the emergence of large multimodal models (LMMs) marks a significant paradigm shift in AI research. Unlike traditional unimodal networks, LMMs such as Qwen [6], Gemini [7], and GPT [8] are trained on massive datasets that combine heterogeneous modalities, enabling them to learn shared semantic representations across multimodal inputs,

including texts and images. These models have shown impressive capabilities in vision–language alignment, cross-modal retrieval, multimodal reasoning, and generative tasks.

Despite the remarkable progress achieved by LMMs, several unresolved challenges remain in applying them to visual recognition tasks under real-world conditions. First, the effectiveness of current LMMs still depends on a vast amount of paired multimodal data and extensive computational resources, which limits their scalability to data-scarce or domain-specific applications. Second, the alignment between different modalities is often imperfect, resulting in domain bias, ambiguous reasoning paths, and reduced interpretability when processing complex real-world visual scenes. Third, although pre-trained LMMs exhibit impressive visual recognition ability, their cross-domain generalization and fine-grained discrimination remain limited when facing distribution shifts or unseen visual categories. In summary, although LMMs have substantially advanced the field of visual recognition, key issues concerning data dependency and domain-specific applicability in real-world visual recognition scenes remain a challenge. Addressing these challenges is essential to realize more reliable and scalable multimodal visual understanding systems that can effectively operate in real-world recognition tasks such as intelligent transportation, remote sensing, and other complex visual environments.

1.2 Proposition in this Thesis

The purpose of this thesis is to investigate how LMMs can be effectively adapted and extended for real-world visual recognition tasks. Special attention is given to applications such as traffic sign recognition (TSR) and remote sensing, with a particular focus on improving generalization, interpretability, and data efficiency under cross-domain conditions. To address the challenges mentioned earlier, this thesis introduces a progressive series of multimodal visual understanding methods. Each method is designed to mitigate one or more of these issues and collectively contribute to enhancing the adaptability and reasoning capability of LMMs. The studies can be summarized as follows.

In the first part, we focus on feature-level alignment for zero-shot recognition. We investigate how structural relationships among visual categories can be leveraged to perform cross-domain recognition without labels. To this end, we construct a midlevel feature correspondence space that preserves structural similarity in the feature domain, enabling robust recognition of different

categories and improving model generalization under data-scarce conditions.

In the second part, we explore few-shot adaptation through in-context learning. Building upon the feature alignment concept, we propose an approach that utilizes multimodal contextual cues to guide LMMs in learning from a limited number of labeled examples. This method achieves effective knowledge transfer and enhances recognition accuracy across heterogeneous visual environments, demonstrating the feasibility of data-efficient multimodal learning.

In the third part, we extend visual recognition to the reasoning level. We introduce a multi-step reasoning mechanism that combines visual and semantic information to perform fine-grained and interpretable recognition across domains. Inspired by human chain-of-thought reasoning, our approach enhances transparency and semantic understanding in zero-shot recognition tasks, allowing LMMs to perform stepwise inference under domain shifts.

In the final part, we apply the developed methodologies to a large-scale multimodal fusion scenario in remote sensing. We integrate visual imagery with textual and geographic information to construct a multimodal fusion model for regional scene understanding. Through extensive experiments, we verify the scalability and practicality of the proposed multimodal recognition paradigm, demonstrating its potential to extend multimodal reasoning to geospatial and real-world applications.

The contributions of this thesis are summarized as follows:

- This thesis develops a series of multimodal visual understanding methods that progressively enhance multimodal feature-level alignment, data-efficient adaptation, and cross-domain reasoning for visual recognition tasks in real-world applications.
- The proposed approaches systematically demonstrate how LMMs can be adapted as interpretable, generalizable, and data-efficient systems capable of handling zero-shot, few-shot, and cross-domain recognition challenges.
- The studies presented in this thesis provide both theoretical insights into multimodal representation and reasoning, and practical guidance for applying LMMs to domain-specific visual perception in fields such as intelligent transportation and remote sensing.

1.3 Organization of this Thesis

The structure of this thesis is outlined below.

Chapter 2 reviews the related work on deep learning, LMMs, and visual recognition applications in traffic sign and remote sensing. It provides an overview of existing approaches in zero-shot and few-shot learning, visual–language models, and reasoning-based multimodal systems. The chapter also highlights the current challenges in semantic alignment, data efficiency, and cross-domain generalization that motivate the research presented in this thesis.

Chapter 3 investigates visual recognition under zero-shot conditions by analyzing structural relationships among different categories. It focuses on midlevel feature alignment and explores how feature correspondence can improve cross-domain generalization when labeled data are unavailable. The experiments demonstrate that structural similarity in the feature space enables accurate recognition of traffic sign categories and enhances the robustness of vision models.

Chapter 4 expands the investigation to few-shot in-context learning scenarios. It introduces an in-context adaptation approach that allows LMMs to quickly adapt to TSR tasks with template traffic sign samples. The chapter emphasizes efficient knowledge transfer and multimodal contextual reasoning, showing how these mechanisms can improve performance in real-world environments with limited annotated data.

Chapter 5 explores multimodal reasoning for fine-grained and cross-domain visual recognition. It presents a multi-step reasoning strategy that combines visual and semantic representations to achieve more interpretable recognition. By mimicking human-like sequential inference, this approach enhances the transparency and reasoning ability of LMMs while supporting zero-shot generalization across domains.

Chapter 6 extends the proposed methodologies to large-scale multimodal applications in the remote sensing domain. It constructs a multimodal fusion framework that integrates visual, textual, and geographic information for regional scene recognition. This chapter demonstrates the scalability of LMMs and verifies their ability to provide context-aware understanding in complex real-world datasets.

Chapter 7 presents the conclusions of this thesis and the future directions are discussed.

The approaches proposed in each chapter are based on the author’s published papers, which are listed in the final section of this thesis. Chapter 3 presents the method reported in [A-3],

which proposes a zero-shot TSR framework based on midlevel feature matching. Chapter 4 introduces the method described in [B-1], which develops a cross-domain few-shot in-context learning approach for enhancing TSR. Chapter 5 presents the method corresponding to [A-1], which focuses on cross-domain multi-step thinking for zero-shot fine-grained TSR in the wild. Chapter 6 introduces the method proposed in [A-2], which develops the a large-scale multimodal regional remote sensing scene understanding framework.

Chapter 2

Related Works

2.1 Introduction

This chapter reviews the relevant literature. Subsection 2.2 introduces prior work on LMMs, while Subsection 2.3 summarizes existing approaches to TSR. Subsection 2.4 presents related studies on multitask remote sensing recognition. Subsequently, Subsection 2.5 outlines the key research problems addressed in this thesis. Finally, Subsection 2.6 concludes this chapter.

2.2 General Large Multimodal Model Methods

Reference [7]

This paper introduces Gemini, a family of highly capable LMMs that integrates advanced language understanding with multimodal perception and reasoning. Gemini models are trained on large-scale multimodal data consisting of text, images, audio, and code, enabling unified processing and generation across modalities. Compared with previous instruction-tuned LMMs that primarily rely on paired image–text data, Gemini adopts a more comprehensive multimodal training pipeline that combines grounded visual reasoning, conversational understanding, and code-based problem solving. Extensive experiments demonstrate that Gemini models achieve state-of-the-art performance on a wide range of benchmarks, including visual question answering, object localization, and instruction-following tasks, positioning them as one of the most general-purpose LMM frameworks to date.

Reference [9]

This paper introduces a method called LLaVA, which extends the instruction-tuning paradigm

from NLP to the vision–language domain. The study aims to align visual representations with language instructions by leveraging large-scale multimodal datasets composed of image–text pairs and task-specific instructions. Unlike traditional multimodal pre-training that focuses solely on captioning or visual question answering, Visual Instruction Tuning trains models to follow high-level human instructions across diverse vision–language tasks, thereby enabling more general and flexible visual reasoning. Experiments on multiple datasets demonstrate that the proposed tuning strategy significantly enhances model generalization and instruction-following capabilities across zero-shot and few-shot visual understanding benchmarks.

Reference [10]

This paper introduces Qwen-VL, a versatile LMM designed for comprehensive multimodal understanding tasks such as image captioning, visual question answering, object localization, and text recognition. The model is built on a large-scale transformer-based architecture that integrates visual and textual encoders with an instruction-tuned alignment module. Unlike conventional LMMs that are constrained to a single capability or dataset, Qwen-VL unifies diverse perception and reasoning tasks within a single framework and supports high-resolution image inputs for fine-grained visual understanding. Experimental results demonstrate that Qwen-VL achieves state-of-the-art performance across multiple benchmarks, showing strong generalization in both perception-oriented and instruction-following multimodal tasks.

Reference [11]

This paper introduces InstructBLIP, an LMM that extends the BLIP-2 architecture through instruction-tuning with diverse multimodal directives. The model retains the frozen pre-trained vision and language encoders of BLIP-2 while incorporating an instruction-aware training objective to align multimodal reasoning with human-given prompts. Unlike standard LMMs that rely solely on image–caption alignment, InstructBLIP explicitly teaches the model to follow natural language instructions for a wide range of vision–language tasks. Experiments demonstrate that the instruction-tuned model achieves stronger generalization and more consistent reasoning compared with BLIP-2 and other contemporary

LMMs.

Reference [12]

This paper presents MiniGPT-4, an LMM designed to connect a pre-trained vision encoder with a language model using a simple linear projection trained on paired image–text conversations. The study shows that large-scale instruction data enable the model to perform tasks such as visual description, caption-based reasoning, and user-guided dialogue generation. In contrast to heavily engineered training pipelines, MiniGPT-4 adopts a lightweight alignment stage while still achieving impressive performance on open-ended multimodal understanding tasks. This simplicity highlights the effectiveness of instruction-based alignment for rapid adaptation of LMMs.

Reference [13]

This paper introduces Kosmos-2, an LMM developed for grounded multimodal understanding that supports referring expression comprehension, object grounding, and free-form description. The model expands upon the Kosmos-1 framework by combining high-resolution vision encoders with language models trained on visually grounded text. Unlike pure instruction-following LMMs, Kosmos-2 focuses on localization and grounding, bridging perception and reasoning through spatially aligned multimodal tokens. It achieves competitive results across visual question answering, grounding, and captioning benchmarks, demonstrating the potential of grounding-based alignment in LMMs.

Reference [14]

This paper proposes Multimodal GPT, an extension of the GPT framework that unifies vision and language within a single generative architecture. By introducing visual tokens into the transformer layers, MMGPT processes and generates multimodal content without relying on separate visual encoders. This design allows end-to-end training across image understanding, question answering, and multimodal reasoning tasks using unified token representations. The experiments confirm that integrating visual and textual modalities at the token level enables smoother modality fusion and improved visual reasoning consistency.

2.3 Traditional Traffic Sign Recognition Methods

Reference [15]

This paper introduces an efficient CNN framework for real-time TSR in complex outdoor environments. The authors design a lightweight CNN architecture that balances recognition accuracy and computational cost, enabling deployment on embedded and vehicle-mounted systems. Unlike traditional cascade or template-matching approaches, this study demonstrates that compact CNNs can achieve robust performance under varying illumination, scale, and background conditions. Experiments conducted on real-world traffic sign datasets show that the proposed model achieves real-time inference and competitive recognition accuracy, paving the way for practical intelligent transportation applications.

Reference [16]

This study presents a unified framework for real-time traffic sign detection and classification based on color–shape features and CNNs. The authors combine region proposals with CNN-based classification to improve detection speed without sacrificing accuracy. Compared with traditional handcrafted feature methods, the proposed system significantly enhances robustness to environmental variations such as lighting, rotation, and occlusion. Comprehensive experiments demonstrate that the framework achieves near real-time performance on both public and self-collected road datasets, making it suitable for driver-assistance systems.

Reference [17]

This paper proposes a practical CNN design for real-time TSR with a focus on efficient feature extraction and computational optimization. The authors evaluate various CNN configurations and training strategies to identify architectures that provide the best trade-off between recognition accuracy and processing speed. Unlike heavy deep networks requiring high-end GPUs, the proposed design maintains high accuracy while operating efficiently on low-power hardware platforms. The experimental results demonstrate that an appropriately structured CNN can deliver reliable recognition performance for real-time intelligent transportation systems.

Reference [18]

This paper proposes a multi-task CNN architecture for TSR that jointly performs detection, classification, and structural analysis within a single framework. The multi-task setting allows the network to share low-level visual features among related subtasks, improving generalization and reducing computational redundancy. Compared with single-task CNNs, the proposed approach significantly enhances accuracy and robustness under challenging conditions such as occlusion and variable illumination. Extensive experiments on benchmark traffic sign datasets verify that the multi-task learning framework achieves faster convergence and superior performance for real-world intelligent transportation applications.

Reference [19]

This study introduces a few-shot learning method for TSR that incorporates clustering inductive bias into a random neural network. The key idea is to enhance representation learning for scarce categories by grouping semantically similar signs and leveraging random weight initialization to improve diversity in feature extraction. Unlike conventional deep CNNs requiring large-scale annotated data, the proposed approach achieves competitive performance with only a few samples per class. Experimental results demonstrate that the method effectively mitigates data scarcity and improves generalization across unseen traffic sign categories, offering a promising direction for data-efficient recognition systems.

Reference [20]

This work presents an efficient traffic sign recognition framework built upon the Extreme Learning Machine (ELM). The approach leverages ELM’s fast training capability and strong generalization to construct a lightweight classifier that operates on features extracted from traffic sign images. By combining robust feature descriptors with the rapid single-hidden-layer feedforward architecture of ELM, the method achieves competitive recognition accuracy while significantly reducing computational cost. Experimental evaluations on benchmark datasets demonstrate that the ELM-based system provides a practical balance between efficiency and performance, making it well suited for real-time TSR applications in resource-constrained environments.

Reference [21]

This paper presents SignParser, an end-to-end framework designed to advance TSR to-

ward holistic understanding. By integrating visual recognition, text reading, and semantic parsing components within a unified transformer-based architecture, SignParser enables simultaneous recognition and text interpretation of road signs. Compared with prior modular systems, the end-to-end framework improves consistency between visual and linguistic elements and reduces error propagation across stages. Extensive evaluations on diverse real-world benchmarks validate the method’s state-of-the-art capabilities in sign detection and text parsing, highlighting the potential of multimodal frameworks for comprehensive traffic sign understanding.

2.4 Multitask Remote Sensing Recognition Methods

Reference [22]

This paper introduces GeoChat, an LMM designed to enable grounded multimodal understanding and dialogue within remote sensing applications. GeoChat integrates a vision encoder trained on high-resolution satellite imagery with a large language model to support tasks such as scene interpretation, object localization, and spatial reasoning through conversational interaction. Unlike conventional remote sensing models that rely on visual classification, GeoChat allows human-like communication grounded in geographic context, bridging perception and natural language reasoning. Comprehensive experiments show that the model achieves strong spatial grounding and descriptive capabilities across diverse remote sensing datasets.

Reference [23]

This paper introduces GeoGround, a unified LMM framework for visual grounding in remote sensing imagery. The model integrates fine-grained geographic features with natural language supervision to align spatial regions in satellite images with textual descriptions. Compared with standard LMMs that focus on general vision–language understanding, GeoGround emphasizes spatial correspondence and grounding precision, enabling accurate localization of referenced objects or patterns in complex aerial scenes. Experimental evaluation demonstrates promising performance on multiple visual grounding benchmarks, highlighting the potential of unified multimodal learning in remote sensing.

Reference [24]

This study introduces SkySenseGPT, an instruction-tuning dataset and vision-language model tailored for remote sensing image understanding. The dataset is constructed to capture diverse geographical scenes, object categories, and hierarchical semantic attributes, enabling models to learn precise and context-aware reasoning over high-resolution aerial imagery. Building on this dataset, the proposed SkySenseGPT model leverages multi-modal instruction-tuning to enhance capabilities such as detailed scene description, object localization, and geospatial reasoning. Experiments demonstrate that the model achieves superior performance across various remote sensing benchmarks, highlighting the effectiveness of fine-grained instruction design in improving multimodal comprehension within the remote sensing domain.

Reference [25]

This paper presents CDChat, an LMM specifically developed for change detection and description in remote sensing imagery. The model combines multi-temporal visual inputs with a language generation module to produce natural language descriptions of land-cover changes. Unlike traditional change detection methods that rely on pixel-level comparison, CDChat formulates the problem as multimodal reasoning, where the model interprets spatial and temporal differences through dialogue-like responses. Experimental results confirm that CDChat significantly improves both the accuracy and explainability of remote sensing change analysis.

Reference [26]

This study presents RS-LLaVA, an LMM developed to jointly perform image captioning and visual question answering for remote sensing imagery. The model extends the LLaVA architecture to the geospatial domain by incorporating high-resolution satellite data and remote-sensing-specific vocabulary. Compared with generic LMMs trained on natural images, RS-LLaVA is capable of producing domain-aware captions and context-consistent answers that reflect geographical semantics. Experimental results on multiple benchmark datasets show that RS-LLaVA significantly improves both descriptive and reasoning accuracy, demonstrating its potential for human-AI interaction in earth observation analysis.

Reference [27]

This paper introduces EarthMarker, an LMM that employs visual prompting to enhance multimodal perception and reasoning for remote sensing tasks. The model combines a large language model with visual encoders guided by prompt-based cues to adaptively focus on relevant spatial details in satellite imagery. Unlike previous LMMs that rely solely on global image–text alignment, EarthMarker incorporates structured visual prompts that allow the model to interpret complex earth observation scenes more effectively. Experiments conducted on several public remote sensing datasets demonstrate superior performance in visual question answering, captioning, and object localization, confirming the advantages of prompt-aware multimodal learning.

Reference [28]

This paper proposes VHM, a versatile and honest LMM tailored for remote sensing image understanding and grounded reasoning. The model integrates multi-level visual features with language instructions, enabling it to handle diverse tasks, including captioning, question answering, and change detection, within a unified architecture. VHM emphasizes response reliability by introducing an honesty evaluation mechanism that estimates confidence in generated text outputs, addressing the hallucination problem common in modern LMMs. Extensive experiments confirm that VHM achieves competitive results across multiple remote sensing benchmarks and provides more trustworthy multimodal outputs.

2.5 Problems to Be Solved in This Thesis

Following the analysis of existing methods, we now outline the primary research problems tackled by this study. Existing LMMs have demonstrated strong visual understanding capabilities, but their performance still degrades under data-scarce and cross-domain conditions due to insufficient modality alignment, limited generalization, and a lack of interpretability. The purpose of this thesis is therefore to enhance the adaptability and reasoning ability of LMMs for visual recognition tasks by improving data efficiency, domain generalization, and semantic consistency. To achieve this goal, this study proposes several multimodal learning methods, including feature-level alignment for zero-shot recognition, in-context adaptation for few-shot

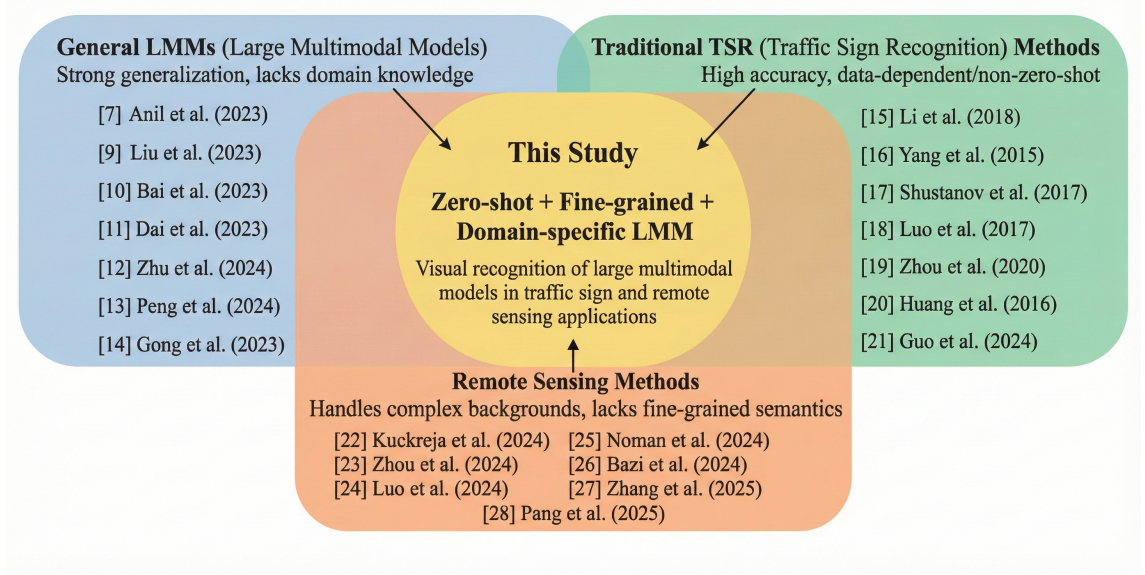


Figure 2.1: Research map of related research.

learning, cross-domain semantic reasoning through multi-step thinking, and fusion modeling for multimodal remote sensing analysis. These approaches are designed to address the aforementioned challenges in a progressive manner, combining representation learning, reasoning, and application-level validation. The effectiveness of the proposed methodologies is evaluated in two real-world domains: TSR, which represents structured visual understanding under diverse conditions, and remote sensing, which requires large-scale multimodal perception and spatial reasoning. To summarize the discussed approaches, we present a research map in Figure 2.1.

2.6 Conclusion

This chapter provided an overview of the related works and clarified the technical challenges addressed in this thesis. Specifically, it summarized prior studies on LMMs, TSR, and multitask remote sensing analysis, highlighting their limitations in data efficiency, domain generalization, and multimodal semantic reasoning. Based on these insights, the key research problems of this thesis were formulated, motivating the development of more adaptable, interpretable, and robust multimodal learning methods. The next chapters describe the proposed approaches in detail and demonstrate how they effectively tackle the identified challenges.

Chapter 3

Midlevel Feature Matching-Based Zero-Shot Traffic Sign Recognition

3.1 Introduction

TSR is a difficult and complex task, yet it remains a widely studied topic because it can help assist drivers during driving and contribute to reducing traffic accidents. Most current TSR approaches are based on CNNs and can reach high recognition accuracy. However, they typically depend on large-scale, carefully prepared traffic sign datasets for training. In addition, traffic signs vary across countries and exist in many different categories, so these models usually must be fine-tuned when they are applied to previously unseen types of traffic signs. To address these challenges, we introduce a matching-based framework designed for zero-shot recognition. Unlike traditional pipelines, this method operates in a training-free manner, identifying categories by evaluating the similarity between target inputs and reference templates. By utilizing robust mid-level features derived from standard CNN architectures, we eliminate the need for additional training or fine-tuning. Our analysis reveals that leveraging these intermediate representations significantly enhances zero-shot discrimination. Experimental validation on both the German TSR Benchmark and a real-world dataset collected in Sapporo, Japan, confirms the effectiveness and generalization potential of this approach.

3.2 Method

Figure 3.1 depicts the architectural overview of the proposed TSR method. The primary workflow involves isolating traffic signs from input road imagery and subsequently deriving feature representations via mid-level CNN layers, which form the foundation for TSM. The system is organized into three sequential phases: TSD, mid-level feature extraction, and zero-shot matching. These modules are systematically detailed in the following subsections.

3.2.1 Traffic Sign Detection

The pipeline is initiated by the TSD phase, which plays a pivotal role in the precise localization of traffic signs within cluttered urban environments. In this subsection, we elaborate on the deployed methodology. We leverage the Vision Transformer Adapter (ViT-Adapter) [29], an architecture built upon recent advancements in vision transformers [3]. While initially engineered for generic object parsing, we tailor this model to address the domain-specific constraints of TSD. This modification introduces inductive biases effectively, facilitating robust identification and categorization of traffic signs across diverse road scenarios.

The process begins by feeding the original road images into the ViT-Adapter. ViT-Adapter generates segmentation maps that delineate multiple object categories with high precision. In our setting, these categories correspond to different types of traffic signs. In the resulting segmentation outputs, traffic signs in urban scenes are represented using distinct colors, with each color denoting a particular object category and thereby aiding their separation and identification. To make the ViT-Adapter outputs more suitable for traffic sign discrimination, we then transform the color-encoded segmentation maps into binary masks. This step simplifies later processing and provides a clear outline of traffic sign regions. Through this binary representation, traffic signs are effectively isolated from the background and other objects, which substantially improves subsequent detection and recognition.

To isolate the traffic sign boundaries, we apply the topological structural analysis method described in [30] to the binary masks. For a given pixel grid N , the algorithm adopts a border-following approach to distinguish foreground pixels (the traffic sign) from the background. By locating the connected components within the foreground, the system traces their perimeters to

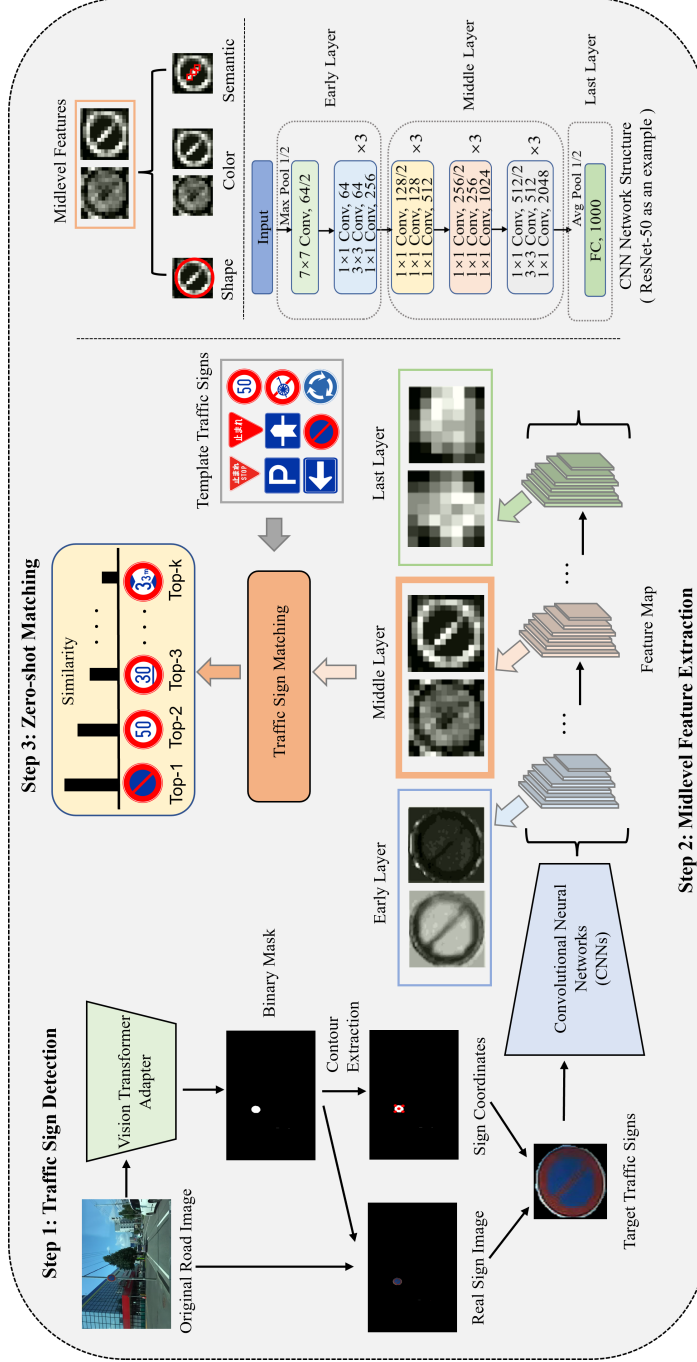


Figure 3.1: Illustration of the proposed TSR pipeline. We first employ the ViT-Adapter to locate target traffic signs within road images. These localized regions are then processed by the TSM module, which utilizes midlevel representations derived from CNNs.

determine the shape of the detected sign I_d . The resulting set of points is calculated as follows:

$$I_d = \{(i, j) \mid N_{ij} = 1\}, \quad (3.1)$$

where N_{ij} corresponds to the foreground pixels of the target sign.

By identifying connected components within N and tracing their external boundaries, the algorithm generates a sequence of coordinates that defines the contour I_d for each traffic sign. This precise localization is a critical component of the TSD phase. Once the spatial extent I_d is determined, it is employed to crop the specific traffic sign region I from the original road imagery. The following subsection provides a detailed discussion on how midlevel features are subsequently extracted from these cropped samples.

3.2.2 Midlevel Feature Extraction

The extraction of midlevel features constitutes a central component of our method. It enables zero-shot recognition by exploiting the pretrained CNNs. Within this framework, traffic sign images are propagated through different layers of the network, ranging from shallow layers that encode low-level cues such as shape, color, and edges to deeper layers that contain richer semantic information [31, 32]. Our approach specifically focuses on exploiting the midlevel features of CNNs. These intermediate layers simultaneously encode abstract semantic information and fundamental visual attributes like color and contour, thereby providing a comprehensive and expressive description of traffic signs. To offer an intuitive illustration of this process, Figure 3.2 presents the structure of the proposed CNN-based scheme, highlighting the early, midlevel, and final layers. The early layers, usually consisting of a series of convolution and pooling operations, are primarily tasked with learning the most basic visual patterns of traffic signs [32, 33]. At this stage, the network predominantly captures elementary features such as edges, simple shapes, and color distributions.

In contrast, the midlevel layers are crucial for encoding the rich semantic characteristics of traffic signs. At this stage, the network forms a sophisticated integration of shape, color, and semantic cues, enabling it to capture fine-grained details that are critical for TSR. It should be noted that the depth and dimensionality of these midlevel layers vary across different CNN architectures, such as ResNet-50, DenseNet-121, and EfficientNet-B0. This variability provides the

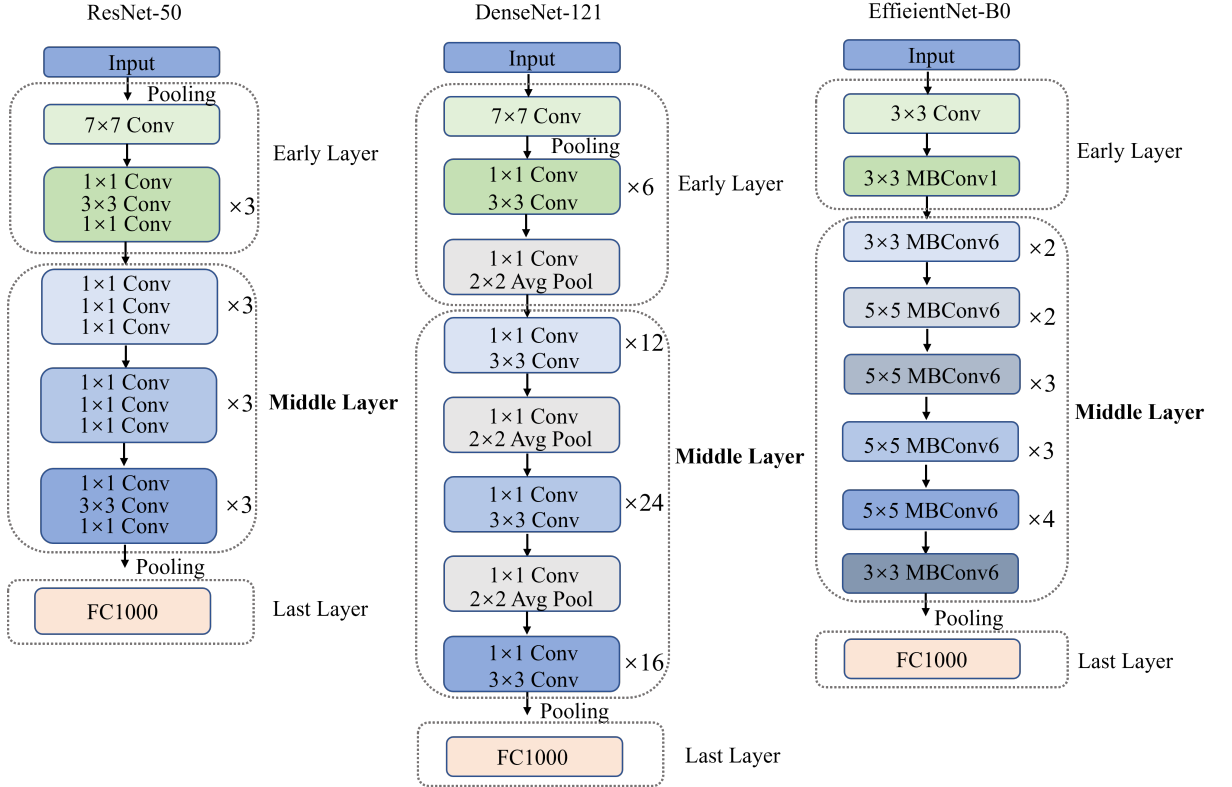


Figure 3.2: Configuration of the early, midlevel, and last layers across various CNN architectures.

flexibility to select the most suitable backbone for TSM.

The final layer is generally implemented as a fully connected layer, which maps the extracted midlevel features to concrete traffic sign categories. In this way, it serves as the bridge between the internal feature representations of the network and the ultimate recognition of traffic signs.

Table 3.1 summarizes the dimensionality of the midlevel features employed in the proposed TSR framework. It presents the optimal intermediate layers selected from various CNNs for TSR, where the corresponding feature sizes indicate the amount and diversity of information captured by each network. These dimensions illustrate how original road images are gradually encoded into structured representations that support reliable recognition. The extraction of midlevel features serves as a key component of the method, allowing the model to retain both fine-grained visual cues and higher-level semantic attributes of traffic signs. This integrated representation motivates the subsequent zero-shot matching stage, which constitutes the core of the TSM procedure.

The extraction of midlevel features F from a traffic sign image I through a CNN is formulated

Table 3.1: Dimensions of the different layers used in our method.

Network	Layer	Dimension
ResNet-50	Early layer	$256 \times 56 \times 56$
	Middle layer	$1024 \times 14 \times 14$
	Last layer	$2048 \times 1 \times 1$
DenseNet-121	Early layer	$256 \times 56 \times 56$
	Middle layer	$1024 \times 14 \times 14$
	Last layer	$1024 \times 7 \times 7$
EfficientNet-B0	Early layer	$16 \times 112 \times 112$
	Middle layer	$320 \times 7 \times 7$
	Last layer	$1280 \times 7 \times 7$

as

$$F_{\text{mid}}(I) = \text{CNN}_{\text{mid}}(I), \quad (3.2)$$

where CNN_{mid} denotes the intermediate module responsible for producing midlevel representations. The resulting $F_{\text{mid}}(I)$ serves as the midlevel descriptor of the traffic sign and is subsequently utilized for zero-shot matching.

3.2.3 Zero-Shot Matching

In this section, we focus on the zero-shot matching stage, a crucial component of our TSR framework built upon midlevel feature representations. This recognition procedure is organized into two principal elements: target and template traffic signs. The target set consists of traffic sign images captured from various geographic regions, whereas the template set corresponds to a comprehensive collection of officially approved national traffic sign templates.

To initiate the matching process, we compute the feature divergence between the target traffic sign I_{target} and every template T_i (indexed by i). This comparison is formulated using the Euclidean distance metric:

$$\text{Dissimilarity}(I_{\text{target}}, T_i) = \sqrt{\sum_{j=1}^n (F_{\text{mid}}(I_{\text{target}}^j) - F_{\text{mid}}(T_i^j))^2}. \quad (3.3)$$

In this equation, $F_{\text{mid}}(I_{\text{target}})$ and $F_{\text{mid}}(T_i)$ correspond to the extracted midlevel features of the target and template samples, respectively. This metric quantifies the alignment between the target and template representations in the feature space. Based on these values, we select the top-(k) templates exhibiting the minimum distances, which serve as the primary candidates for determining the identity of the input sign.

The proposed TSR strategy removes the conventional requirement for large-scale training datasets. Instead, it performs recognition by directly exploiting midlevel features, enabling adaptation to diverse traffic sign designs and regions. By emphasizing feature-space similarity rather than explicit training for every sign category, the method offers a flexible and effective way to address the challenges inherent in TSR.

3.3 Experiments

This section presents the experimental results and evaluation of the proposed TSM method. We demonstrate the effectiveness and robustness of our approach in real-world environments.

3.3.1 Experimental Settings

This subsection outlines the experimental setup employed in our study. We evaluate the proposed method on two distinct datasets: the German Traffic Sign Recognition Benchmark (GTSRB) [34] and a real-world urban dataset collected in Sapporo City, Japan. These datasets were chosen to assess the effectiveness of the TSM framework. The GTSRB dataset comprises 1,213 images spanning 43 categories. For our experiments, we manually selected representative samples from all 43 classes to serve as template signs, while the remaining images were utilized as target queries. Notably, GTSRB images are provided as cropped patches, aligning directly with the input requirements of our pipeline. Regarding the Sapporo urban dataset, we employed the designed TSD module to extract traffic signs from original road scenes. This procedure resulted in 71 target images covering 18 distinct sign types. For this scenario, the template set was constructed based on the 111 categories defined in the official Japanese traffic sign specifications.

To evaluate the effectiveness of our method, we performed comparative experiments with conventional approaches based on HOG [35] and SIFT [36]. This comparison enables a direct

performance benchmark against well-established techniques and highlights the benefits of our approach. For midlevel feature extraction, we utilized CNNs with ResNet-50, DenseNet-121, and EfficientNet-B0 as backbone architectures. All these networks were pretrained on ImageNet, and their original architectures were kept unchanged to preserve generality and ensure applicability across different TSR settings. To standardize the input and enable fair comparison, both target and template traffic sign images were resized to 224×224 pixels. To evaluate recognition performance, we utilize the Top- k accuracy metric. This measure is formally defined as:

$$\text{Acc}_k = \frac{H_k}{Q}. \quad (3.4)$$

In this formulation, Q denotes the total number of target query images, while H_k represents the count of instances where the ground-truth template is successfully retrieved within the top- k candidates. Considering the specific constraints of the zero-shot setting, namely the lack of training data and the high similarity among templates, this metric serves as a robust criterion for assessment.

3.3.2 Experimental Results

In this section, we present the experimental outcomes and provide a detailed evaluation of the proposed TSM method. Specifically, we assess the effectiveness and robustness of our approach across multiple datasets and different environmental conditions.

The results are summarized in Tables 3.2 and 3.3. Table 3.2 reports the Top- k TSM performance of various methods on the GTSRB dataset. Our approach, which relies on CNN midlevel features, attains superior performance compared with traditional methods based on handcrafted features. Across the three CNN architectures, ResNet-50, DenseNet-121, and EfficientNet-B0, the proposed method consistently achieves higher Top- k accuracy, confirming its effectiveness. In addition, we evaluated recognition performance using features from the early, middle, and final layers listed in Table 3.1. The results indicate that the midlevel-feature-based scheme yields the highest TSM accuracy for all three CNNs. This finding supports our hypothesis that midlevel features effectively integrate the detailed information encoded in low-level features with the semantic representations of high-level features, thereby enabling more accurate zero-shot TSR.

To further verify the generality of the proposed TSR method, we present in Table 3.3 the Top-

k TSM results of different approaches on the Sapporo urban road dataset. Unlike the setup for GTSRB, the template traffic signs used here are the standard traffic sign templates commonly adopted in Japan. The proposed method again achieves higher Top- k accuracy than the conventional baselines on this dataset. Moreover, among the features from the early, middle, and final layers, the midlevel-feature-based variant attains the best TSM performance for all three CNNs, providing additional evidence of the effectiveness of our design. It is also noteworthy that the entire recognition pipeline operates without any extra traffic sign images for training. The method is able to derive strong feature representations for traffic signs from different countries and delivers high zero-shot TSR accuracy.

Table 3.2: Top- k accuracy of different methods on the GTSRB dataset.

Method		Top1	Top5	Top10
HOG [35]		0.089	0.196	0.329
SIFT [36]		0.238	0.551	0.709
ResNet-50	Early Layer	0.141	0.352	0.569
	Middle Layer (PM)	0.521	0.781	0.930
	Last Layer	0.148	0.359	0.559
DenseNet-121	Early Layer	0.081	0.227	0.374
	Middle Layer (PM)	0.468	0.769	0.910
	Last Layer	0.394	0.680	0.864
EfficientNet-B0	Early Layer	0.245	0.520	0.687
	Middle Layer (PM)	0.444	0.767	0.921
	Last Layer	0.333	0.678	0.848

Table 3.3: Top- k accuracy of different methods on the Sapporo urban road dataset.

Method		Top1	Top5	Top10
HOG [35]		0.014	0.296	0.465
SIFT [36]		0.127	0.310	0.338
ResNet-50	Early Layer	0.014	0.211	0.338
	Middle Layer (PM)	0.338	0.761	0.873
	Last Layer	0.028	0.070	0.141
DenseNet-121	Early Layer	0.014	0.141	0.268
	Middle Layer (PM)	0.817	0.873	0.915
	Last Layer	0.169	0.606	0.732
EfficientNet-B0	Early Layer	0.099	0.169	0.239
	Middle Layer (PM)	0.437	0.732	0.845
	Last Layer	0.169	0.423	0.634

Figure 3.3 illustrates example matching results produced by different methods on the GTSRB dataset. We present the matching results for four target traffic signs, intentionally selected to cover diverse colors and shapes so as to highlight the effectiveness of the proposed method. The last target traffic sign is deliberately chosen to be blurred, enabling an assessment of how well each method copes with visually ambiguous traffic signs.

As illustrated in Figure 3.3, the four target traffic signs from the GTSRB dataset cover multiple colors, including red, black, and blue, and exhibit both circular and triangular shapes. Across the three CNN backbones—ResNet-50, DenseNet-121, and EfficientNet-B0, the proposed method consistently retrieves top-matched template signs that share similar color and shape properties with the targets. Compared with handcrafted descriptors such as HOG and SIFT, the proposed approach additionally maintains semantic consistency. For example, the central pattern of the “No Passing” sign is composed of vehicles, and our method is able to correctly capture such semantic content while still preserving shape and color information. Remarkably, even in the last case with a blurred traffic sign, the CNN-based midlevel feature approach successfully identifies the correct sign, highlighting its robustness. This property is especially important in practical

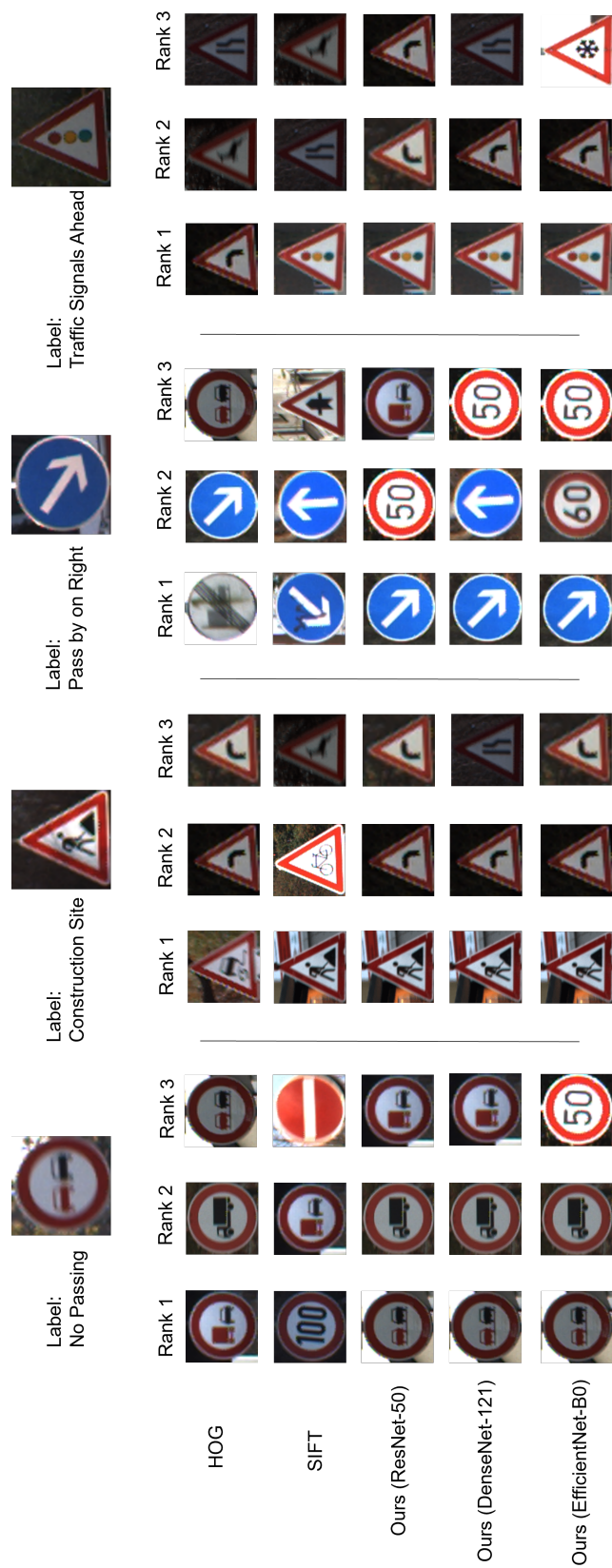


Figure 3.3: Examples of matching results for different methods on the GTSRB dataset.

Table 3.4: Analysis of inference latency per target image across distinct backbones and datasets. “Template Count” denotes the number of reference categories used for matching.

Benchmark	Template Count	Backbone Architecture	Avg. Latency (s)
GTSRB	43	ResNet-50	1.15
		DenseNet-121	1.26
		EfficientNet-B0	1.82
Sapporo Urban Road	111	ResNet-50	3.63
		DenseNet-121	4.01
		EfficientNet-B0	5.35

applications, where traffic signs are often degraded or blurred by challenging environmental conditions.

Figure 3.4 visualizes the matching results produced by various methods on the Sapporo urban road dataset. Given that this dataset consists of raw street scenes, we employed the ViT-Adapter combined with contour detection to isolate the traffic signs. To demonstrate the performance of our midlevel-feature TSM, four representative target signs were selected. These samples encompass diverse geometric shapes (rectangles, circles, triangles) and color profiles. As observed in Figure 3.4, across all three CNN backbones (ResNet-50, DenseNet-121, and EfficientNet-B0), our approach consistently retrieves templates that align with the target’s visual attributes. Furthermore, in contrast to handcrafted descriptors such as HOG and SIFT, the proposed method preserves semantic context. For instance, regarding signs containing text (e.g., “Stop Line” and “Temporary Stop”), our model successfully captures these semantic cues while maintaining consistency in color and shape.

Furthermore, Table 3.4 details the computational overhead associated with the matching process across different backbones. On the GTSRB benchmark, we observe per-image latencies of 1.15 s (ResNet-50), 1.26 s (DenseNet-121), and 1.82 s (EfficientNet-B0). Conversely, evaluations on the Sapporo dataset exhibit increased processing times, ranging from 3.63 to 5.35 seconds, respectively. This increase is directly attributable to the expanded size of the template gallery (111 classes versus 43). Despite this scaling factor, the framework maintains competitive efficiency across all architectures, validating its feasibility for practical deployment in real-world TSR systems.

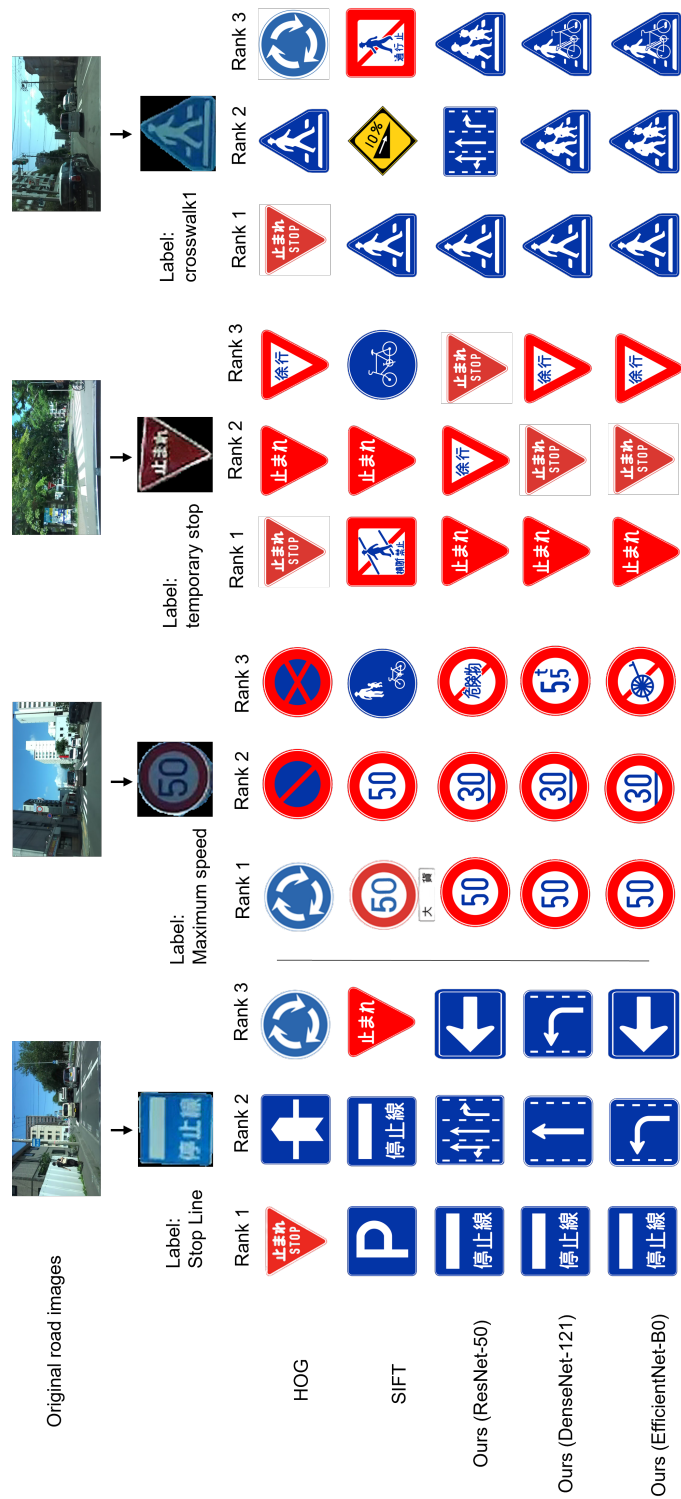


Figure 3.4: Examples of matching results for different methods on the Sapporo urban road dataset.

The proposed method yields consistent matching behavior on samples from both datasets, highlighting its effectiveness for TSR under varying conditions. On the GTSRB dataset, it accurately matches target signs with a wide range of colors and shapes, indicating strong capability in handling heterogeneous sign appearances. By retaining semantic context alongside color and shape features, the approach is well-aligned with the requirements of real-world deployment. Evaluations on the Sapporo urban road dataset demonstrate effective recognition of diverse sign geometries and colors, even amidst complex background clutter. Such robustness is essential for practical traffic management systems. In summary, the proposed midlevel-feature-based framework yields consistent matching performance across both datasets, validating its applicability to TSD and recognition tasks in varied environments.

3.4 Conclusion

In this chapter, we presented a zero-shot TSR framework built on a novel TSM scheme that exploits midlevel features. Through extensive experiments and analysis, we gained a deeper understanding of how midlevel features derived from CNNs behave. The findings underscore the importance of these representations, showing that they can effectively capture both semantic cues and shape properties that are characteristic of traffic signs. By relying on such midlevel feature representations, the proposed framework avoids the need for large-scale training or adaptation on country-specific datasets, thereby offering strong adaptability of TSR across different regions. Relative to prior studies, our work provides a new perspective on the fundamental challenges in TSR. The TSM mechanism based on midlevel features provides improved flexibility and computational efficiency, and by greatly reducing the reliance on large annotated training corpora, it overcomes a common limitation of many existing approaches. This renders the method particularly suitable for scenarios in which well-labeled datasets are scarce or difficult to obtain. The robustness and practicality of the proposed approach are further supported by its favorable results on two heterogeneous datasets, GTSRB and the Sapporo urban road dataset, where it delivers both accurate and efficient TSR.

Chapter 4

Cross-domain Traffic Sign Recognition via Few-shot In-context Learning

4.1 Introduction

In this chapter, we present a cross-domain few-shot in-context learning framework based on LMMs to enhance TSR performance. First, we construct a traffic sign detection network to isolate traffic signs from original road images. To reduce the need for extensive training data and improve robustness across different domains, we subsequently propose a few-shot in-context learning strategy grounded in the LMM. Specifically, our method generates descriptive texts derived from template signs to strengthen the fine-grained recognition capabilities of the LMM. These texts capture key attributes, such as shape, color, and structure, which effectively activate the LMM’s ability to distinguish subtle category differences. By integrating this information, the approach bridges the gap between template signs and real-world instances. Notably, the entire pipeline relies on concise, unified textual prompts and avoids the need for large-scale labeled image collections. We evaluate our method on the German, Belgium TSR benchmarks and two real-world datasets collected in Japan. The results confirm that the proposed approach achieves a significant improvement in TSR performance.

4.2 Methodology

As illustrated in Figure 4.1, our approach first conducts traffic sign detection (TSD) on original road images, followed by fine-grained traffic sign classification (TSC). By leveraging template

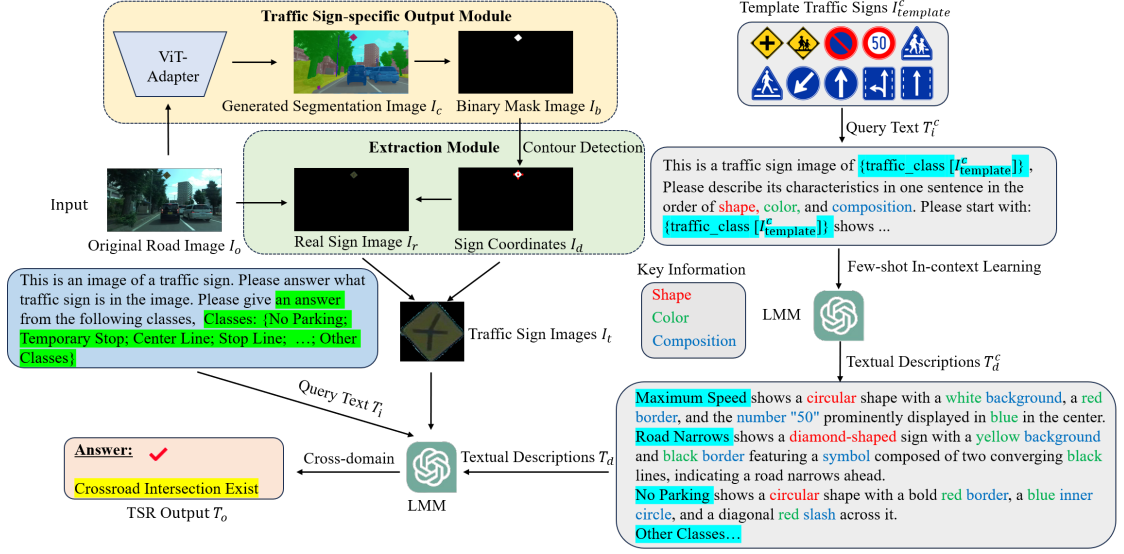


Figure 4.1: Overview of the proposed TSR framework. We first apply the proposed TSD network to detect and extract traffic signs from road images. Next, template traffic signs are fed into the LMM to generate description texts. These descriptions encode essential information about the shape, color, and structural components of traffic signs, thereby enhancing the LMM’s reasoning capability for TSR.

traffic signs from the national standard template database to build a cross-domain few-shot in-context learning framework with LMM, the method decrease the reliance on large-scale labeled training datasets.

4.2.1 Traffic Sign Detection

TSD constitutes the initial stage of the proposed approach and focuses on rapidly localizing and detecting traffic signs in original urban road images. Our design is inspired by ViT-Adapter [29]. ViT-Adapter extends the plain ViT [3] by incorporating image-specific inductive biases, thereby enabling accurate image understanding. To realize effective TSD, we develop a traffic sign-specific output module based on ViT-Adapter. In addition, we construct an extraction module to obtain real traffic sign images from the detected regions. Together, the traffic sign-specific output module and the extraction module form the proposed TSD network.

4.2.1.1 Traffic Sign-specific Output Module

We first take the original road image I_o as an input into ViT-Adapter. The vanilla ViT-Adapter can generate segmentation images with various object category labels for the original input image. In our scenario, traffic signs are the only objects of interest, and producing segmentation maps for all categories introduces redundant information that may adversely affect TSD. Therefore, we designed a specialized output module for traffic signs. Specifically, in the generated segmentation image I_c , each specific object category is encoded with a different color for recognition. Our specialized output module converts I_c into a binary mask image I_b . The method effectively distinguishes regions containing traffic signs from the rest of the image. The designed traffic sign-specific output module provides a clear delineation of the regions containing traffic signs and simplifies subsequent processing steps. The binary mask image I_b effectively separates traffic signs from the background and other objects in I_c , thereby enhancing the dedicated detection capability for traffic signs.

4.2.1.2 Traffic Sign Extraction Module

Next, we construct an extraction module to obtain real traffic sign images. The extraction module consists of two stages. In the first stage, we apply the contour detection algorithm [30] to the binary mask image I_b to precisely delineate and extract the boundary coordinates of traffic signs. Specifically, regarding I_b as a pixel grid in which each pixel is classified as foreground (traffic sign) or background (surroundings), the contour detection algorithm follows the borders of the foreground regions and efficiently outlines the shape of each detected traffic sign, yielding I_d . Here, I_d records the coordinate positions of the traffic signs. In the second stage, we use the original road image I_o together with the sign coordinates I_d produced by the traffic sign-specific output module of ViT-Adapter to obtain an image I_r that contains only real traffic signs. Since I_o and I_d share the same resolution, I_r is generated by scanning I_d for regions with white pixels and mapping these regions to the corresponding areas in I_o . Finally, we obtain the traffic sign image I_t from I_r by cropping according to the traffic sign coordinates in I_d computed by the contour detection algorithm. The image I_t thus serves as the final detected traffic sign output of our TSD network. It should be noted that although I_t could in principle be extracted directly from the original road image I_o using the coordinates in I_d , the resulting crop would still contain unnecessary

background regions. In contrast, our extraction module suppresses background information by setting it to black, thereby reducing potential interference and enabling more accurate TSD.

4.2.2 Cross-domain Few-shot In-context Learning for TSR

After completing TSD and obtaining the traffic sign image I_t , we perform cross-domain few-shot in-context learning TSR using LMM. In general, LMMs take an image $I \in \mathbb{R}^{H \times W \times 3}$ together with a text query $T_i = [t_i^1, \dots, t_i^{l_i}]$ of length l_i , and produce a textual output sequence $T_o = [t_o^1, \dots, t_o^{l_o}]$ of length l_o as follows:

$$T_o = \text{LMM}(I, T_i). \quad (4.1)$$

Since TSR is a fine-grained recognition task, different traffic sign classes often share very similar visual properties. If we directly feed the extracted traffic signs I_t into the LMM for recognition, the LMM can capture general visual features of the input, but it still struggles to reliably distinguish specific traffic sign categories. To address this issue and improve fine-grained discrimination, we adopt a new strategy: we first generate textual descriptions for each traffic sign category by leveraging template traffic signs, and then use these descriptions to guide the LMM in performing fine-grained TSC.

Specifically, in the proposed framework, we first feed the template traffic sign image I_{template}^c into the LMM to obtain a textual description T_d^c for each class in $I_{\text{template}}^c = [I_{\text{template}}^1, \dots, I_{\text{template}}^c]$ of traffic signs, formulated as follows:

$$T_d^c = \text{LMM}(I_{\text{template}}^c, T_i^c), \quad (4.2)$$

where T_i^c represents the corresponding input text query for the template traffic sign I_{template}^c . Thus, the textual descriptions of all classes T_d are expressed as

$$T_d = [T_d^1, \dots, T_d^c]. \quad (4.3)$$

The template traffic signs originate from the national standard traffic sign template database and exhibit canonical shapes, colors, and structural features. In contrast, real-world traffic sign images are affected by diverse factors such as illumination, viewing angle, and occlusion, leading to

substantial discrepancies from the template images and making cross-domain recognition challenging. To alleviate this issue, we exploit the strong visual understanding ability of LMM and convert images into textual descriptions, thereby narrowing the cross-domain gap. Moreover, since the proposed method relies solely on template traffic signs and the descriptions are generated only once for each template traffic sign class, our few-shot method avoids the reliance on large-scale labeled training datasets required by many previous approaches. We then input the extracted traffic sign image I_t , the generated textual descriptions for all template traffic sign classes T_d , and the query text T_i into the LMM to obtain the output T_o of the proposed cross-domain few-shot in-context learning TSR method as follows:

$$T_o = \text{LMM}(I_t, T_d, T_i). \quad (4.4)$$

As we explicitly account for the three fundamental attributes of traffic signs, i.e., shape, color, and composition when generating textual descriptions, the generated texts contain the key information for each category and thereby support LMM in performing fine-grained classification. Consequently, the textual descriptions produced by the proposed method, which encode the principal characteristics of traffic signs, effectively strengthen the fine-grained reasoning capability of LMM in TSR.

4.3 Experiments

4.3.1 Experimental Settings

In this subsection, we describe the experimental settings used in our study. We evaluated the proposed method on four datasets, including two benchmark datasets: the GTSRB dataset [34] and the Belgium traffic sign dataset (BTSD) [37], as well as two real-world road image datasets collected in Japan, namely the Sapporo urban road dataset and the Yokohama urban road dataset. We adopted GPT4-v as the LMM for TSR, and no additional model training was required in our framework. However, owing to the quota and traffic constraints of the GPT-4v API¹, we followed the protocol in [38] and constructed validation subsets from GTSRB and BTSD. Concretely, we selected 100 traffic sign images from 43 categories in GTSRB and 125 images from 62 categories

¹<https://platform.openai.com/account/limits>

in BTSD. The number of classes in each subset was kept identical to that in the corresponding full dataset, so as to thoroughly assess the fine-grained recognition capability of our approach. Since traffic signs are already cropped in these two benchmark datasets, we directly applied the proposed few-shot in-context learning scheme for fine-grained recognition. The Sapporo urban road dataset contains 68 images covering 18 types of traffic signs, while the Yokohama urban road dataset comprises 33 images from 13 traffic sign categories. For these two real-world Japanese datasets, we first employed the proposed TSD network to detect and extract traffic signs from the original road images, and then carried out TSR using the LMM-based method to verify the effectiveness of our approach in practical scenarios.

To assess the performance of the proposed TSR method, we employ Top- k accuracy as the evaluation metric, which offers a comprehensive measure of the method’s effectiveness. This metric is defined as follows:

$$\text{Top-}k = \frac{t_k}{\sum_t I_t}. \quad (4.5)$$

Here, t_k denotes the number of input traffic sign images that are correctly recognized within the Top- k predictions. Given the challenges of few-shot and zero-shot TSR, where no additional training data are required, Top- k accuracy serves as an effective indicator of recognition performance. Moreover, the Top- k results can offer reliable candidate traffic signs to assist drivers in understanding road conditions in advanced driver-assistance systems.

4.3.2 Experimental Results

Table 4.1 reports the Top- k recognition performance of different methods. We conduct a thorough evaluation on two benchmark traffic sign datasets, GTSRB and BTSD, as well as on two Japanese traffic sign datasets, to verify the effectiveness of the proposed approach. As shown in Table 4.1, our method achieves superior results compared with traditional techniques and CNN-based TSR models. On all four datasets, the Top- k accuracy of the proposed method surpasses that of the competing approaches by a clear margin, confirming its effectiveness. Moreover, the experimental results show that the LMM-based method we propose outperforms advanced transformer-based approaches such as MAE, CLIP, ViT-B, and ViT-L. It is noteworthy that our method attains a significantly higher Top-1 accuracy than both CNN- and transformer-based

Table 4.1: Top- k accuracy of different methods on four datasets.

Method	GTSRB			BTSD			Sapporo urban road			Yokohama urban road		
	Top1	Top5	Top10	Top1	Top5	Top10	Top1	Top5	Top10	Top1	Top5	Top10
Song et al. [39]	0.30	0.56	0.75	0.72	0.85	0.88	0.02	0.30	0.47	0.03	0.30	0.39
Ren et al. [36]	0.45	0.74	0.89	0.45	0.71	0.84	0.13	0.29	0.31	0.27	0.46	0.52
ResNet-50 [1]	0.63	0.83	0.83	0.90	0.95	0.96	0.34	0.72	0.79	0.49	0.68	0.70
DenseNet-121 [40]	0.60	0.90	0.97	0.66	0.89	0.93	0.66	0.80	0.82	0.67	0.76	0.79
EfficientNet-B0 [2]	0.63	0.94	1.00	0.81	0.94	0.99	0.37	0.74	0.79	0.36	0.64	0.70
Mobilenet-V3 [41]	0.58	0.95	1.00	0.79	0.98	1.00	0.65	0.82	0.84	0.64	0.76	0.79
MAE [42]	0.27	0.51	0.61	0.50	0.80	0.86	0.08	0.47	0.72	0.03	0.39	0.79
CLIP [43]	0.16	0.42	0.68	0.22	0.59	0.71	0.34	0.57	0.72	0.27	0.64	0.76
ViT-B [3]	0.36	0.68	0.86	0.60	0.77	0.91	0.24	0.72	0.88	0.24	0.73	0.91
ViT-L [3]	0.60	0.88	0.98	0.81	0.96	0.99	0.44	0.68	0.81	0.58	0.67	0.91
GPT-4v (Baseline)	0.80	0.86	0.87	0.61	0.70	0.75	0.28	0.43	0.54	0.46	0.70	0.76
Ours	0.88	0.97	1.00	0.91	0.98	1.00	0.75	0.84	0.91	0.88	0.94	1.00

baselines on all datasets, demonstrating its strong capability for precise TSR. We also evaluate a baseline that relies solely on LMM for TSR, whose Top- k accuracy remains relatively low across all four datasets, indicating that accurate fine-grained recognition of traffic signs is difficult when using only LMM. By contrast, our cross-domain template-based strategy supplies LMM with key information about traffic sign shape, color, and composition, thereby enhancing its fine-grained perception of traffic sign categories and leading to improved TSR performance.

Figure 4.2 shows examples of recognition results of the proposed method. We show two samples, one from a benchmark dataset and the other from a real-world dataset, to fully demonstrate the effectiveness of the proposed method. We show the original road images, input traffic signs, generated text descriptions, and cross-domain recognition results for LMM. As shown in Fig. 4.2, our method generates textual descriptions of traffic signs with simple and uniform prompts and template traffic sign images. Considering the characteristics of traffic signs, we design the input prompts for generating the text description so that the important features of the shape, color, and composition of traffic signs are fully considered. Then we use the generated text descriptions to enhance the fine-grained recognition capability of LMM for traffic signs. Due to the differences between the template traffic sign and the actual traffic sign images, we reduce this cross-domain difference by transforming the images into textual descriptions using LMM’s powerful recognition of image features, thus achieving effective TSR.

Figure 4.2 presents example recognition results obtained by the proposed method. We provide two representative cases, one from a benchmark dataset and the other from a real-world dataset, to comprehensively illustrate the effectiveness of the approach. For each case, we show the original road image, the input traffic sign, the generated textual descriptions, and the cross-domain recognition outputs of LMM. As shown in Fig. 4.2, our method produces textual descriptions of traffic signs using simple, unified prompts together with template traffic sign images. In designing the prompts for description generation, we explicitly incorporate the key attributes of traffic signs, shape, color, and composition, so that these crucial features are fully reflected in the resulting texts. The generated descriptions are then used to enhance the fine-grained recognition capability of LMM for traffic signs. Because there are inherent discrepancies between template signs and real-world sign images, we mitigate this cross-domain difference by leveraging LMM’s strong visual understanding to convert images into textual descriptions, thereby enabling effective

TSR.

4.3.3 Ablation Studies

In addition to the LMM-based baseline that inputs only cropped traffic sign images (baseline) in Table 4.1, another commonly used strategy in other vision tasks is to feed the original road images containing traffic signs directly into the LMM for recognition, which we denote as baseline-o. Since, to the best of our knowledge, there is no prior TSR work built on LMMs, the effectiveness of this strategy remains unclear. To further validate the advantages of the proposed method, we therefore conduct an ablation study. We compare the recognition outputs of three LMM-based settings on the same traffic sign: (i) baseline-o, which inputs the original road image; (ii) baseline, which inputs only the extracted traffic sign image; and (iii) our method. As illustrated in Fig. 4.3, in baseline-o the LMM receives the original road image directly. Although this image includes useful scene context, the traffic signs occupy only a small portion of the image and many fine-grained categories exhibit highly similar appearances (e.g., Crosswalk Alt. A and Crosswalk Alt. B), leading to poor TSR performance when relying solely on LMM. The baseline method, which first uses our TSD network to crop traffic signs and then feeds them into LMM, also performs unsatisfactorily, since the LMM still struggles to distinguish closely related traffic sign classes. In contrast, our method introduces template traffic signs to generate class-wise textual descriptions, thereby strengthening the LMM’s fine-grained discrimination ability and yielding more accurate TSR.

In addition, we present an example in which recognition fails because of an inaccurate generated textual description. As illustrated in Fig. 4.4, the traffic sign Direction-specific access zone 2 contains both straight and left-turn arrows, whereas the corresponding generated description includes only a straight arrow. Due to the missing of the key feature describing the left-turn arrow, LMM misclassifies this sign as Direction-specific access zone 4. This example highlights the critical role of precise textual descriptions in enabling fine-grained recognition with LMM. When expert knowledge is incorporated to revise and correct the description texts, accurate recognition can be achieved.

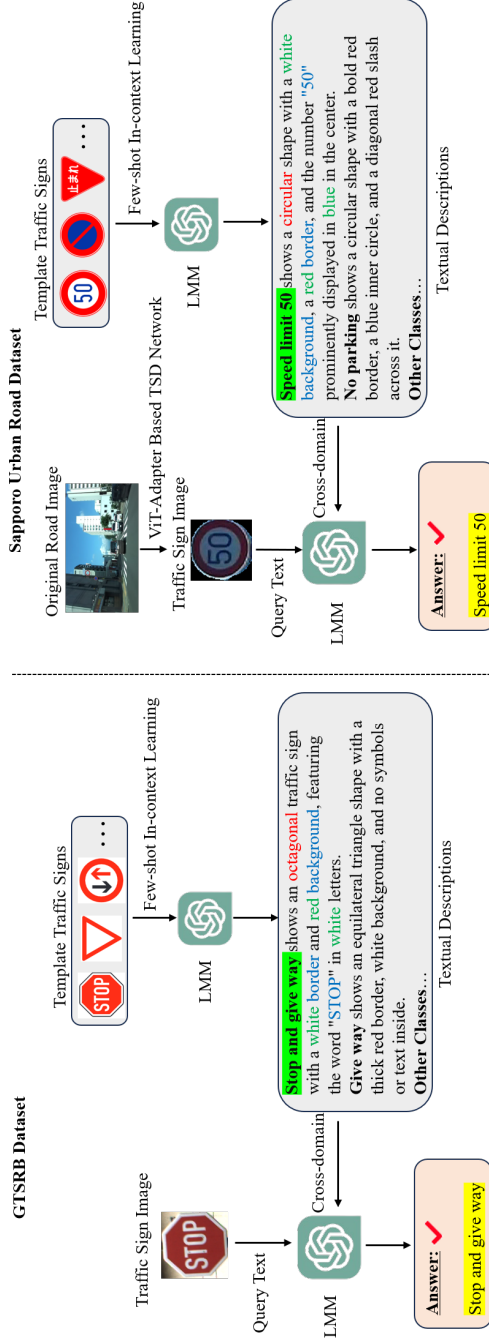


Figure 4.2: Top-1 recognition results of the proposed cross-domain few-shot in-context TSR method. Two representative examples are provided for the benchmark dataset (GTSRB) and the real-world Sapporo urban road dataset. For the Sapporo samples, we illustrate the full procedure, from traffic sign detection in original road images to final recognition by the LMM using the proposed approach.

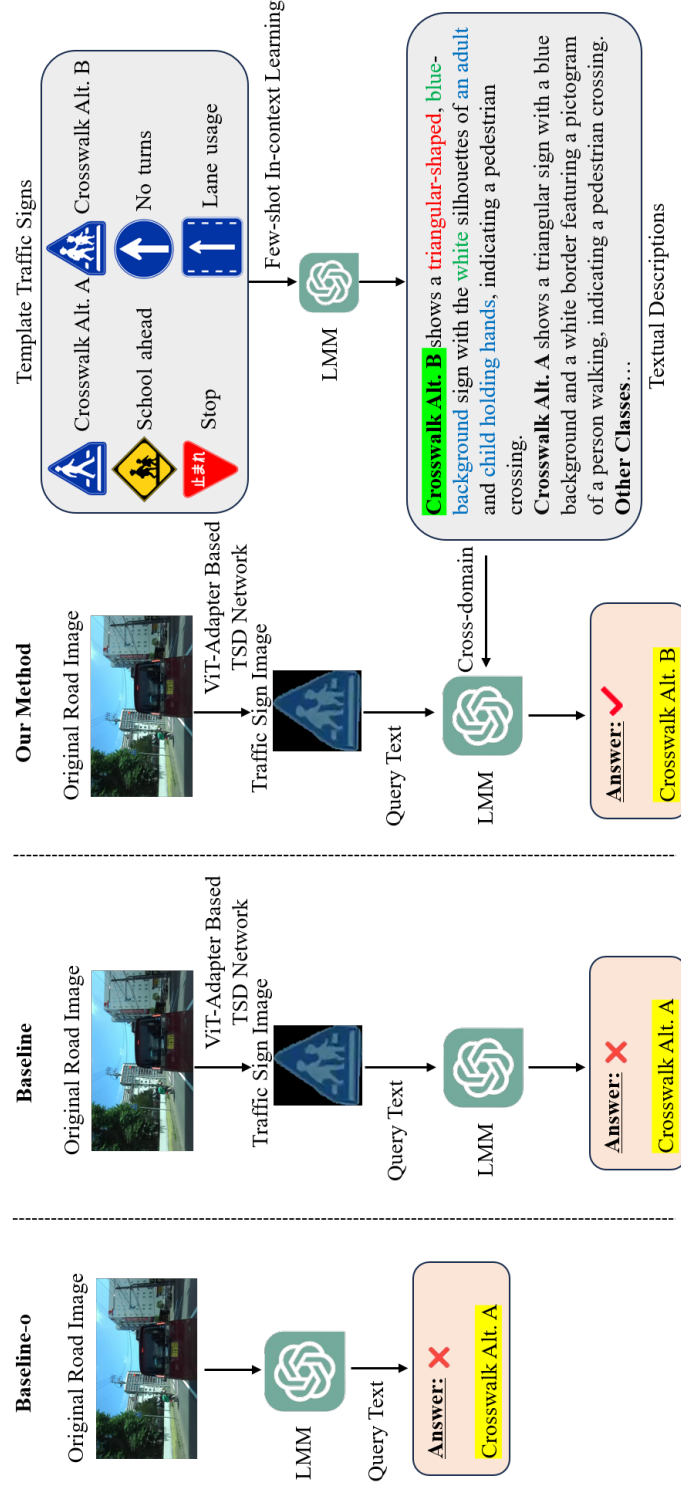


Figure 4.3: Top-1 recognition results of the baseline-o, baseline, and our method.

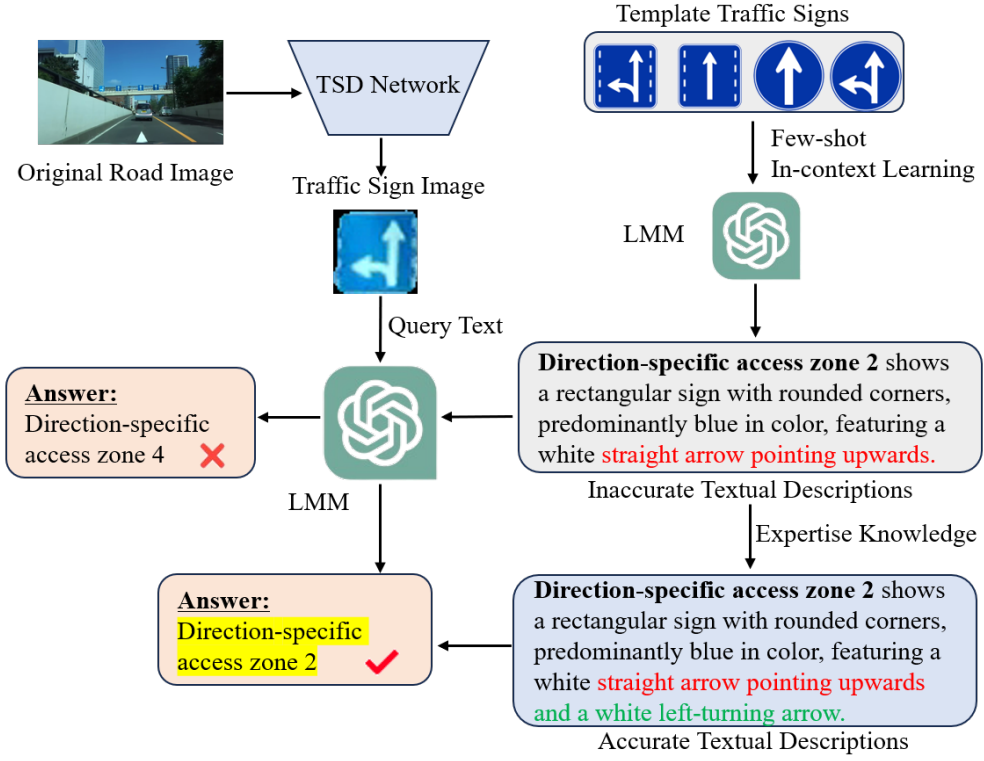


Figure 4.4: Example of incorrect recognition due to inaccurately generated description texts.

4.4 Conclusion

In this chapter, we introduced a cross-domain few-shot in-context learning approach based on LMM to improve TSR. We developed a new TSD network and leveraged template traffic signs to generate textual descriptions, thereby strengthening the fine-grained recognition capability of LMM for traffic signs. Extensive experiments on two benchmark datasets and two real-world datasets confirmed the effectiveness of the proposed approach. To the best of our knowledge, this is the first work that explicitly focuses on enhancing the fine-grained TSR ability of LMM, and we expect it to stimulate further TSR research built upon LMM in the future.

Chapter 5

Zero-shot Fine-grained Traffic Sign Recognition via Cross-domain Multi-step Thinking

5.1 Introduction

In this chapter, we introduce Cross-domain Multi-step Thinking (CdMT), a framework developed to support zero-shot fine-grained TSR under real-world conditions. The task remains difficult because of the discrepancy between ideal template signs and their real-world counterparts, and this challenge becomes even more pronounced in cross-country settings where traffic sign designs differ substantially across regions. CdMT tackles these issues by leveraging the multi-step reasoning capability of large multimodal models. Specifically, we construct three complementary types of descriptions, context, characteristic, and differential to guide the model through a structured reasoning process. Context descriptions, enhanced through center coordinate prompt refinement, help the model accurately locate target signs in cluttered road scenes and suppress irrelevant predictions by providing appropriate prior hypotheses. Characteristic descriptions are generated through in-context learning with template signs, thereby mitigating cross-domain discrepancies and strengthening fine-grained discrimination. Differential descriptions focus on subtle visual distinctions among similar signs, improving the model’s ability to resolve fine-grained ambiguities. A notable advantage of CdMT is that it does not require any model training and relies solely on a unified prompting strategy, making it suitable for cross-country TSR. We evaluate the framework on three public benchmark datasets and two real-world

datasets from different countries and the proposed CdMT shows promising performance on all five datasets.

5.2 Method

In this section, we outline the proposed cross-domain zero-shot TSR approach, illustrated in Fig. 5.1. The method begins by identifying and extracting traffic signs from the input road images using a tailored detection module. After obtaining the sign regions, a multi-step thinking procedure is introduced to guide the LMM through increasingly refined reasoning stages, thereby strengthening its ability to perform fine-grained TSR under cross-domain and real-world conditions.

5.2.1 Traffic Sign Extraction

5.2.1.1 Segmentation

In the proposed framework, the original road image $\mathcal{I}_o^i$, which contains multiple traffic signs ($i \in \{0, 1, 2, \dots, N\}$, where N is the total number of signs in the scene), is first processed by a segmentation module. The image $\mathcal{I}_o^i$ is passed through a segmentation network to generate a labeled output $\mathcal{I}_s^i$, where each pixel is assigned an object class index, allowing different elements in the scene to be distinguished. For TSR, traffic signs must be distinguished from all other objects. In $\mathcal{I}_s^i$, each category is represented by a unique color, enabling category identification. We then convert $\mathcal{I}_s^i$ into a mask image $\mathcal{I}_m^i$, which isolates traffic sign regions from the background and non-sign objects in $\mathcal{I}_s^i$. The proposed approach does not depend on any specific segmentation backbone, thus providing flexibility in implementation.

5.2.1.2 Extraction

After segmentation, a customized extraction detector is employed to isolate traffic signs. First, the detector applies a contour detection algorithm [30] to the mask image $\mathcal{I}_m^i$ to obtain the coordinates of traffic sign regions. Next, using the original road image $\mathcal{I}_o^i$ together with these coordinates, the detector extracts an image $\mathcal{I}_r^N$ that contains only real traffic signs. In $\mathcal{I}_r^N$, non-sign objects and background areas present in the original image are effectively removed. The

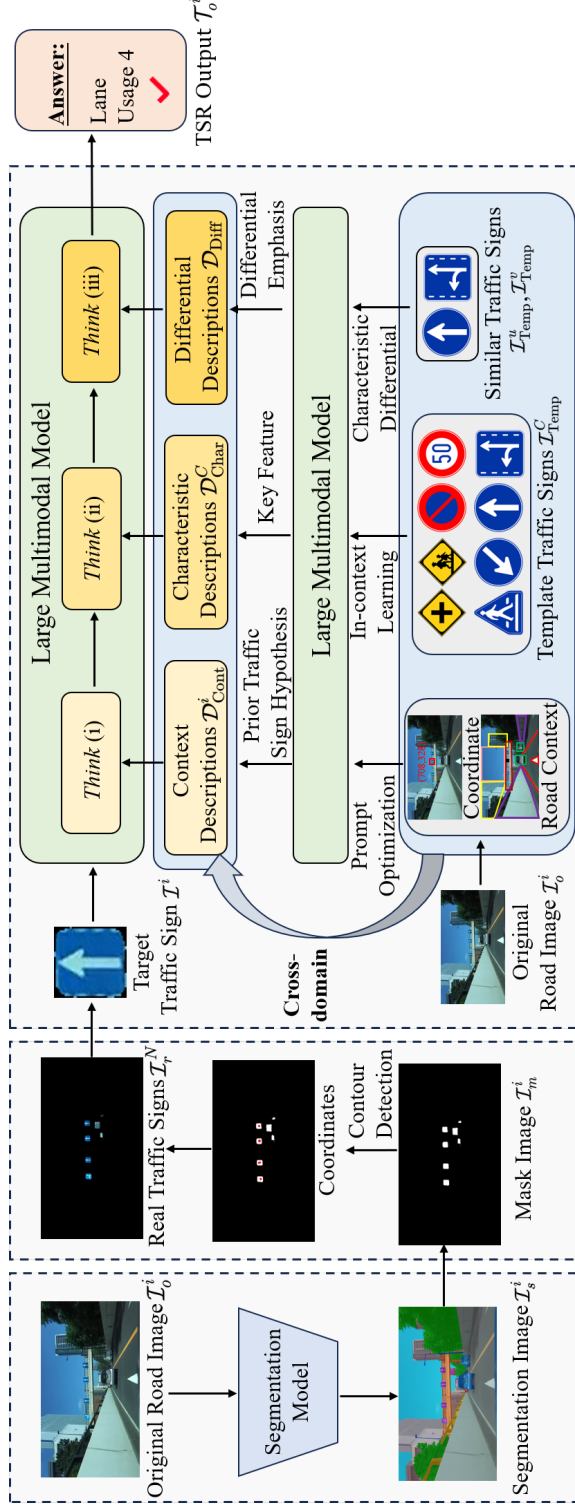


Figure 5.1: Schematic of the proposed CdMT framework. The approach initially localizes traffic signs and subsequently employs multi-step reasoning to achieve fine-grained TSR.

extraction detector subsequently obtains the traffic sign image $\mathcal{I}^i$ from $\mathcal{I}_r^N$ according to the corresponding traffic sign coordinates. Here, $\mathcal{I}^i \in \mathbb{R}^{H \times W \times 3}$ denotes the final extracted traffic sign image. It is worth noting that although $\mathcal{I}^i$ could also be cropped directly from the original road image $\mathcal{I}_o^i$ using these coordinates, the resulting patch could still include redundant background regions. In contrast, the extraction detector suppresses such background information, thereby reducing potential interference and benefiting the subsequent recognition process.

5.2.2 Multi-step Thinking

After extracting the traffic sign image $\mathcal{I}^i$, we apply a multi-step thinking strategy to fully utilize the LMM’s recognition capabilities, as detailed in the subsequent section.

5.2.2.1 Prior Knowledge Generation

In the proposed framework, the prior knowledge used for reasoning is formed from three complementary components: context descriptions derived from the original road images, characteristic descriptions generated from template traffic signs, and differential descriptions that emphasize fine-grained distinctions among visually similar signs. For each target sign, an LMM takes as input the extracted traffic sign image $\mathcal{I}^i$ together with a text prompt $\mathcal{T}^i = [t_1^i, \dots, t_{l_i}^i]$ of length l_i . The model then produces a corresponding output sequence $\mathcal{T}_{\text{out}}^i = [t_1^i, \dots, t_{l_o}^i]$ of length l_o , expressed as:

$$\mathcal{T}_{\text{out}}^i = \text{LMM}(\mathcal{I}^i, \mathcal{T}^i). \quad (5.1)$$

Context Descriptions: Since original road scenes provide a wealth of contextual details, we convert these images into context descriptions. This process allows the model to better utilize the available scene semantics. For a given road image $\mathcal{I}_o^i$, the context descriptions $\mathcal{D}^i\text{Cont} = [\mathcal{D}^i\text{Cont}, \dots, \mathcal{D}^N\text{Cont}]$ are generated as follows:

$$\mathcal{D}_{\text{Cont}}^i = \text{LMM}(\mathcal{I}_o^i, \mathcal{T}_{\text{Cont}}^i), \quad (5.2)$$

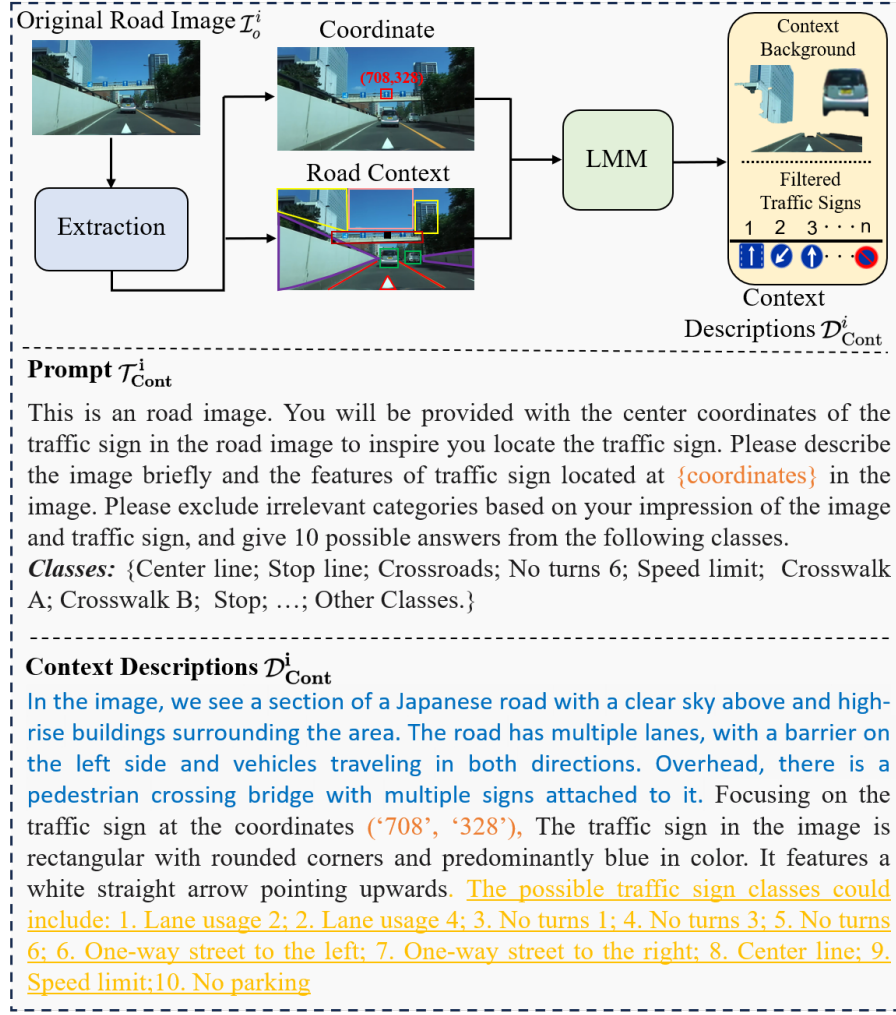


Figure 5.2: Generation of context descriptions. The sign’s coordinates, background context, and prior traffic sign hypothesis are encoded as (orange), (blue), and (yellow), respectively.

where $\mathcal{T}_{\text{Cont}}^i$ represents the textual prompt used to obtain the context descriptions. As illustrated in Fig. 5.2, $\mathcal{T}_{\text{Cont}}^i$ is designed to guide the LMM to extract background cues and scene-level information from the original road image. In addition, motivated by typical question–answering scenarios, we observe that narrowing the possible answer space helps reduce ambiguity for the LMM. To this end, a prior traffic sign hypothesis is introduced, enabling the model to filter out irrelevant categories and focus on a smaller set of plausible sign types.

Much like human thinking, which removes unlikely choices based on past knowledge, the candidate set from the prior hypothesis represents the LMM’s first understanding of the road scene. This preliminary understanding serves as a coarse guide for the subsequent fine-grained reasoning process. When multiple traffic signs coexist in a road image, however, producing reliable

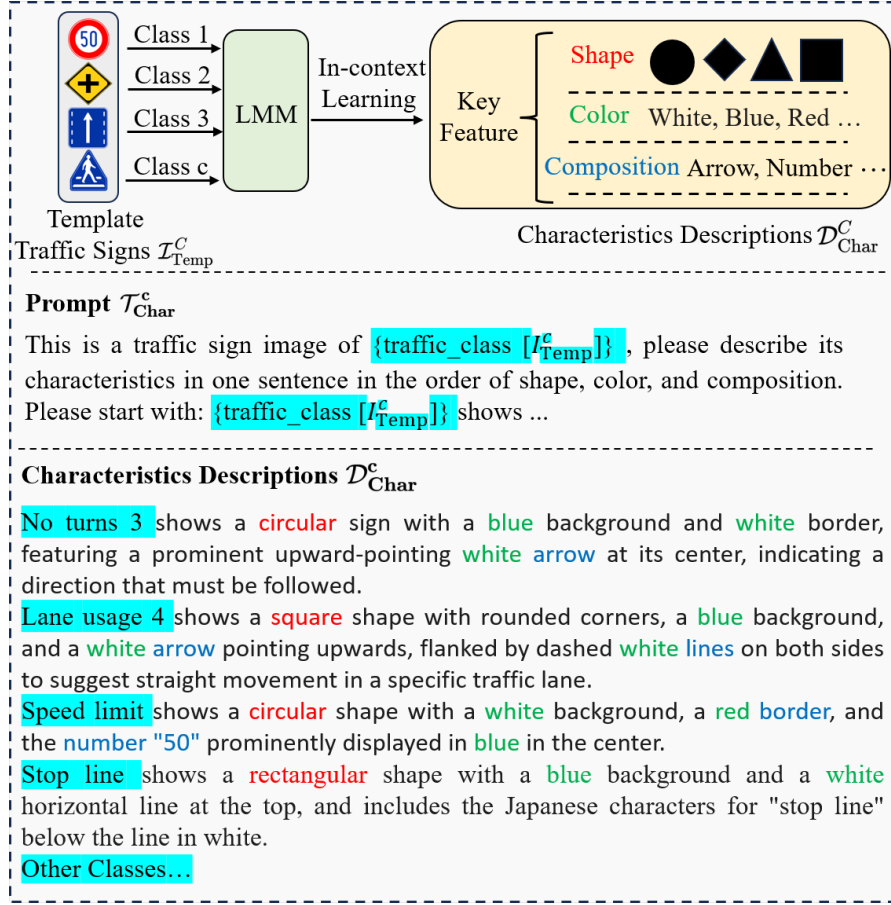


Figure 5.3: Generation of characteristic descriptions. In-context learning is introduced to guide the LMM in capturing the key features of traffic signs.

context descriptions and a reasonable prior hypothesis becomes more challenging, as the LMM may be uncertain about which sign should be referenced. To mitigate this issue, we introduce a center coordinate–based prompt optimization strategy. The center coordinates produced by the detection module, obtained at no extra computational cost, are incorporated into the prompt to guide the LMM in precisely locating the target traffic sign within the scene. By providing this spatial cue, the LMM is encouraged to attend to the correct sign region and, in turn, generate more accurate background descriptions and prior traffic sign hypotheses.

Characteristic Descriptions: Fine-grained TSR is challenging for LMMs because their built-in knowledge often fails to reliably distinguish specific traffic sign categories. Given that template traffic signs from national standard databases are readily available, we can alleviate the dependence on large-scale training data. In contrast to previous methods that perform template matching at the feature level, real-world traffic signs often exhibit variations caused by illumination,

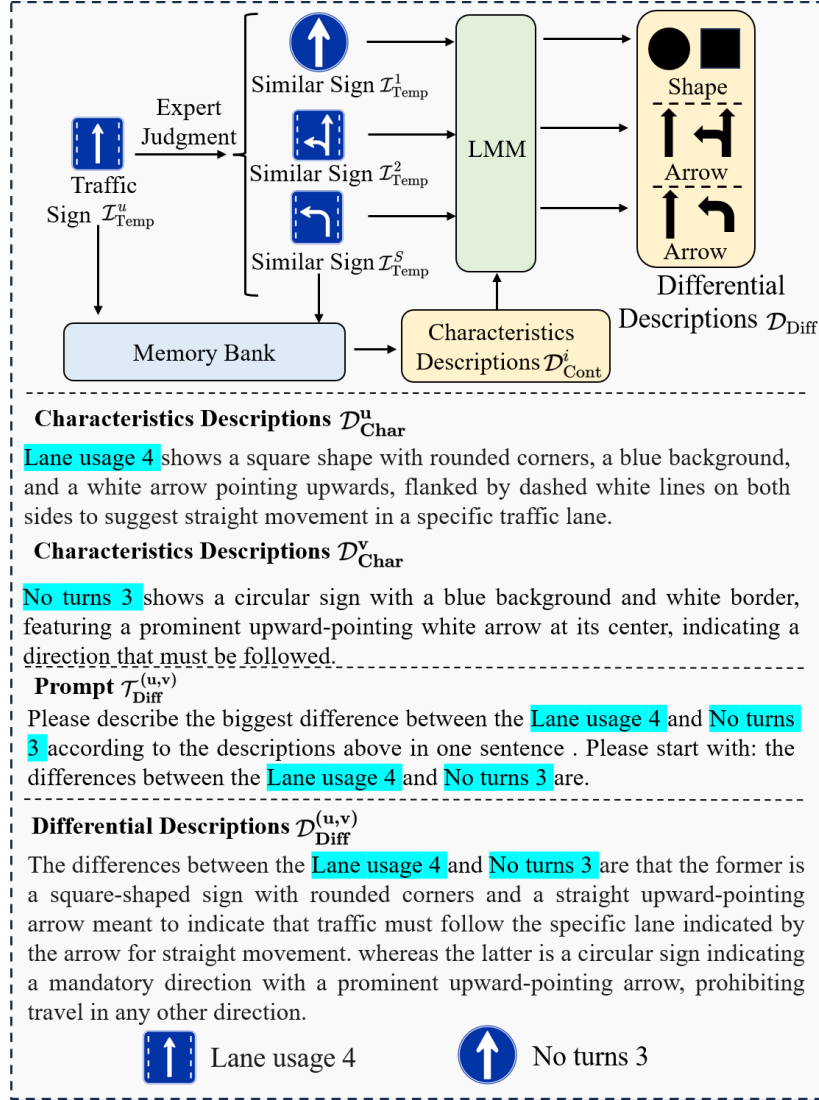


Figure 5.4: Generation of differential descriptions. Discriminative features among visually similar signs are emphasized to facilitate the fine-grained reasoning of the LMM.

viewpoint changes, and occlusions, which further increase the cross-domain gap. To address this issue, in-context learning is introduced to perform generation of characteristic descriptions $\mathcal{D}_{Char} = [\mathcal{D}_{Char}^1, \dots, \mathcal{D}_{Char}^C]$ for each class c of template traffic signs $\mathcal{I}_{Temp} = [\mathcal{I}_{Temp}^1, \dots, \mathcal{I}_{Temp}^C]$. This is achieved with prompts $\mathcal{T}_{Char} = [\mathcal{T}_{Char}^1, \dots, \mathcal{T}_{Char}^C]$ as follows:

$$\mathcal{D}_{Char}^c = \text{LMM}(\mathcal{I}_{Temp}^c, \mathcal{T}_{Char}^c), \quad (5.3)$$

where $\mathcal{T}_{Char}^c$ denotes the prompt tailored for $\mathcal{I}_{Temp}^c$.

As depicted in Fig.5.3, traffic signs are generally characterized by three core attributes: shape,

color, and composition. Our prompts (refer to Fig.5.3) are constructed to focus the attention of the LMM on these properties while filtering out extraneous information. Through this mechanism, the in-context learning module generates $\mathcal{D}_{\text{Char}}^c$ a single time and archives it within a memory bank. By avoiding feature-level matching, this strategy effectively mitigates the domain gap between template and real-world samples. Moreover, the prompt design remains consistent across categories, eliminating the need for class-specific tuning, thereby enhancing both efficiency and scalability.

Differential Descriptions: Some traffic signs exhibit very similar visual properties, which makes fine-grained discrimination difficult. To address this issue, we introduce differential descriptions that emphasize subtle differences between such signs. For a template sign $\mathcal{I}_{\text{Temp}}^u$ and a visually similar template sign $\mathcal{I}_{\text{Temp}}^v \in [\mathcal{I}_{\text{Temp}}^1, \dots, \mathcal{I}_{\text{Temp}}^S]$, we first derive their characteristic descriptions by using Eq. (5.3):

$$\mathcal{D}_{\text{Char}}^u = \text{LMM}(\mathcal{I}_{\text{Temp}}^u, \mathcal{T}_{\text{Char}}^u), \quad (5.4)$$

$$\mathcal{D}_{\text{Char}}^v = \text{LMM}(\mathcal{I}_{\text{Temp}}^v, \mathcal{T}_{\text{Char}}^v). \quad (5.5)$$

The differential description $\mathcal{D}_{\text{Diff}}^{u,v}$ is then derived as follows:

$$\mathcal{D}_{\text{Diff}}^{u,v} = \text{LMM}(\mathcal{D}_{\text{Char}}^u, \mathcal{D}_{\text{Char}}^v, \mathcal{T}_{\text{Diff}}^{u,v}), \quad (5.6)$$

where $\mathcal{T}_{\text{Diff}}^{u,v}$ represents the prompt formulated to highlight distinctions. The complete set of differential descriptions is given by:

$$\mathcal{D}_{\text{Diff}} = \bigcup_{u,v \in \{1, \dots, S\}} \mathcal{D}_{\text{Diff}}^{u,v}. \quad (5.7)$$

As illustrated in Fig. 5.4, experts manually determine pairs of similar traffic signs, and the corresponding characteristic descriptions stored in the memory bank are used to derive $\mathcal{D}_{\text{Diff}}$. These differential descriptions highlight subtle distinctions (e.g., lane usage vs. no turns), thereby further enhancing the fine-grained recognition capability of the LMM.

Algorithm 1 Cross-domain Multi-step Thinking (CdMT) for TSR

Input: Raw road image $\mathcal{I}_o^i$, template signs $\mathcal{I}_{\text{Temp}}$, prompts $\mathcal{T}_{\text{Cont}}$, $\mathcal{T}_{\text{Char}}$, $\mathcal{T}_{\text{Diff}}$, $\mathcal{T}_{\text{Multi}}$

Output: Recognized traffic sign type $\mathcal{T}_{\text{out}}^i$

Phase 1: Traffic Sign Extraction

- 1: $\mathcal{I}_s^i \leftarrow \text{Segment}(\mathcal{I}_o^i)$ ▷ Segment raw image
- 2: $\mathcal{I}_m^i \leftarrow \text{ConvertToMask}(\mathcal{I}_s^i)$ ▷ Generate mask
- 3: $\text{Coords} \leftarrow \text{ContourDetect}(\mathcal{I}_m^i)$ ▷ Extract coordinates
- 4: $\mathcal{I}_r^N \leftarrow \text{Extract}(\mathcal{I}_o^i, \text{Coords})$ ▷ Refine image
- 5: $\mathcal{I}^i \leftarrow \text{Retrieve}(\mathcal{I}_r^N, \text{Coords})$ ▷ Get target sign

Phase 2: Prior Knowledge Generation

- 6: $\mathcal{D}_{\text{Cont}}^i \leftarrow \text{LMM}(\mathcal{I}_o^i, \mathcal{T}_{\text{Cont}} + \text{Coords})$ ▷ Context with coords
- 7: **for** $c = 1$ to C **do** ▷ For each template class
- 8: $\mathcal{D}_{\text{Char}}^c \leftarrow \text{LMM}(\mathcal{I}_{\text{Temp}}^c, \mathcal{T}_{\text{Char}}^c)$
- 9: Store $\mathcal{D}_{\text{Char}}^c$ in memory bank
- 10: **end for**
- 11: **for** each similar pair (u, v) in $\mathcal{I}_{\text{Temp}}$ **do** ▷ Expert-identified
- 12: $\mathcal{D}_{\text{Diff}}^{u,v} \leftarrow \text{LMM}(\mathcal{D}_{\text{Char}}^u, \mathcal{D}_{\text{Char}}^v, \mathcal{T}_{\text{Diff}}^{u,v})$
- 13: $\mathcal{D}_{\text{Diff}} \leftarrow \mathcal{D}_{\text{Diff}} \cup \mathcal{D}_{\text{Diff}}^{u,v}$
- 14: **end for**

Phase 3: Multi-step Reasoning

- 15: $\mathcal{T}_{\text{out}}^i \leftarrow \text{LMM}(\mathcal{I}^i, \mathcal{D}_{\text{Cont}}^i, \mathcal{D}_{\text{Char}}, \mathcal{D}_{\text{Diff}}, \mathcal{T}_{\text{Multi}})$
 - 16: **return** $\mathcal{T}_{\text{out}}^i$
-

5.2.2.2 Multi-step Reasoning

Upon acquiring the context descriptions $\mathcal{D}_{\text{Cont}}^i$, characteristic descriptions $\mathcal{D}_{\text{Char}}^C$, and differential descriptions $\mathcal{D}_{\text{Diff}}$, the LMM executes a multi-stage reasoning protocol for each target instance as detailed below. **Step 1:** The model initially formulates a preliminary interpretation of the target image based on its intrinsic knowledge base. **Step 2:** It subsequently evaluates the environmental context by referencing $\mathcal{D}_{\text{Cont}}^i$ and constrains the candidate search space using prior traffic sign hypotheses. **Step 3:** By incorporating the characteristic descriptions, the LMM assimilates core attributes such as geometry, chromatic features, and structural patterns. These attributes are aligned with the visual representation of the target to trigger fine-grained recognition. **Final:** In the concluding phase, the model leverages differential descriptions to disambiguate the target from visually analogous categories, producing the refined prediction:

$$\mathcal{T}_o^i = \text{LMM}(\mathcal{I}^i, \mathcal{D}_{\text{Cont}}^i, \mathcal{D}_{\text{Char}}^C, \mathcal{D}_{\text{Diff}}, \mathcal{T}_{\text{Multi}}), \quad (5.8)$$

where $\mathcal{T}_{\text{Multi}}$ represents the structured multi-step prompt and $\mathcal{T}_o^i$ denotes the final classification output. Through this incremental inference mechanism, the LMM extracts discriminative features to determine the ground truth category. This multi-step reasoning strategy significantly strengthens the fine-grained recognition capacity of the LMM, thereby enhancing robustness in practical environments. The complete procedure of the CdMT framework is outlined in Algorithm 1.

5.3 Experiments

5.3.1 Experimental Settings

Extensive experiments were performed across five distinct datasets to validate our approach. These include three established benchmarks: GTSRB [34], BTSD [37], and Tsinghua-Tencent 100K (TT-100K) [44]. Among these, TT-100K is particularly notable for its complex environmental conditions, serving as a rigorous dataset. To further assess applicability in practical scenarios, we extended the evaluation to two real-world datasets collected in Japanese urban environments: Sapporo and Yokohama. The performance of fine-grained TSR was evaluated using both open-source and closed-source LMM architectures. Crucially, the proposed framework operates in a training-free manner. Due to API rate limits¹, we followed the experimental protocol of [38] and randomly sampled subsets from the validation sets. Importantly, we did not reduce the number of categories in these subsets; instead, we kept the class labels consistent with those of the full datasets to comprehensively verify the fine-grained TSR performance. To mitigate sampling bias, stratified random sampling was adopted to maintain balanced class distributions in each subset. For GTSRB and BTSD, traffic signs are already extracted; thus, multi-step thinking is directly applied to these sign images, and context descriptions are omitted as original road images are unavailable.

Regarding the TT-100K, Sapporo, and Yokohama datasets, traffic sign extraction is conducted directly on the original road images. To quantitatively assess the proposed CdMT framework,

¹<https://platform.openai.com/account/limits>

we employ the standard Top- k accuracy metric, defined as:

$$\text{Top-}k = \frac{C_k}{\sum_i I^i}. \quad (5.9)$$

In this formulation, C_k represents the count of target instances where the ground truth class is successfully retrieved within the Top- k candidates. Given the constraints of fine-grained TSR in a zero-shot setting, Top- k accuracy provides a robust and appropriate criterion for performance evaluation.

5.3.2 Experimental Results

Table 5.1 reports the comparative analysis of Top- k fine-grained TSR performance against state-of-the-art baselines. As evident from the tabulated results, existing methods yield suboptimal accuracy, underscoring the inherent challenges associated with zero-shot fine-grained TSR in unconstrained environments. Furthermore, methods proposed by Li et al. [15] and Zheng et al. (ViT-L) [46] exhibit significant performance variance across datasets from different nations. This trend indicates a limited generalization capability for cross-country TSR in the absence of training supervision.

Quantitative analysis reveals that the proposed framework establishes a comprehensive performance lead over existing baselines. Across all five datasets, our method achieves superior Top-1 and Top-3 accuracies compared to conventional paradigms, including hand-crafted descriptors (e.g., Song et al. [39]), CNN-based architectures (e.g., DenseNet-121 [40], EfficientNet-B0 [2]), and standard Transformer-based approaches (e.g., ViT-L [46], Luo et al. [47]). Furthermore, when benchmarked against cutting-edge vision transformers (such as Swin-Transformer V2 [49] and EVA-02 [54]) and specialized cross-domain adaptation models (like MaPLe [51] and CoOp [52]), our approach maintains a competitive edge. Crucially, the integration of CdMT consistently amplifies the capabilities of Large Multimodal Models. Regarding closed-source architectures, both CdMT-Gpt-4v and CdMT-Gpt-4o register distinct improvements over their vanilla counterparts. Notably, CdMT-Gpt-4o secures the state-of-the-art Top-1 accuracy across all evaluation benchmarks, proving that the framework effectively unlocks the fine-grained recognition potential of LMMs. This enhancement extends to open-source models as well; variants based on LLaVA-NeXT and VITA-1.5 exhibit significant accuracy gains, confirming that the benefits of

Table 5.1: Quantitative comparison of Top- k zero-shot fine-grained TSR performance across five diverse datasets. We benchmark the proposed approach against state-of-the-art methods. Bold and underlined values denote the best and second-best results, respectively. All metrics represent the mean accuracy calculated over five independent trials.

Method	GTSRB			BTSD			TT-100K			Sapporo			Yokohama		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Song et al. [39]	0.10	0.23	0.29	0.19	0.25	0.32	0.04	0.10	0.13	0.04	0.57	0.77	0.04	0.18	0.42
Ren et al. [36]	0.41	0.64	0.77	0.11	0.36	0.50	0.26	0.42	0.50	0.34	0.47	0.50	0.21	0.42	0.48
Gan et al. [45]	0.56	0.76	0.84	0.67	0.84	0.91	0.12	0.22	0.36	0.42	0.71	0.79	0.27	0.42	0.51
DenseNet-121 [40]	0.31	0.46	0.59	0.21	0.32	0.49	0.08	0.14	0.24	0.73	0.82	0.84	0.23	0.47	0.70
EfficientNet-B0 [2]	0.52	0.76	0.90	0.60	0.86	<u>0.93</u>	0.17	0.30	0.38	0.51	0.66	0.74	0.25	0.44	0.60
Li et al. [15]	0.75	0.83	0.89	0.82	0.91	0.94	0.27	0.46	0.60	0.70	0.80	0.83	0.29	0.45	0.69
Zheng et al. (ViT-L) [46]	0.44	0.58	0.70	0.39	0.57	0.64	0.09	0.16	0.21	0.54	0.63	0.75	0.19	0.36	0.44
Luo et al. [47]	0.15	0.35	0.48	0.22	0.27	0.34	0.14	0.28	0.41	0.39	0.57	0.70	0.18	0.35	0.56
MobileViT [48]	0.05	0.11	0.22	0.02	0.07	0.10	0.05	0.11	0.15	0.08	0.10	0.29	0.06	0.35	0.42
Swin-Transformer V2 [49]	0.14	0.26	0.37	0.06	0.17	0.32	0.09	0.17	0.23	0.06	0.10	0.18	0.09	0.27	0.58
MAE [42]	0.20	0.32	0.47	0.13	0.36	0.49	0.06	0.10	0.13	0.14	0.27	0.41	0.17	0.32	0.51
DeiT [50]	0.27	0.45	0.57	0.12	0.28	0.42	0.34	0.60	0.70	0.71	0.83	0.88	0.26	0.47	0.69
CLIP (ViT-B/32) [43]	0.24	0.35	0.48	0.20	0.30	0.38	0.29	0.50	0.62	0.27	0.50	0.57	0.14	0.48	0.60
MaPLe [51]	0.28	0.35	0.49	0.23	0.32	0.41	0.37	0.62	0.76	0.34	0.66	0.79	0.20	0.61	0.72
CoOp [52]	0.32	0.44	0.63	0.25	0.36	0.55	0.36	0.58	0.71	0.33	0.62	0.74	0.17	0.56	0.65
CLIP-Adapter [53]	0.37	0.52	0.71	0.32	0.43	0.61	0.43	0.69	0.83	0.41	0.69	0.83	0.26	0.63	0.75
EVA-02 [54]	0.41	0.67	0.75	0.30	0.51	0.66	0.32	0.61	0.76	0.48	0.53	0.62	0.29	0.46	0.70
LLaVA-1.5 [9]	0.32	0.45	0.48	0.28	0.33	0.42	0.13	0.21	0.38	0.09	0.20	0.46	0.11	0.32	0.43
LLaVA-NeXT [55]	0.39	0.48	0.57	0.31	0.38	0.46	0.20	0.31	0.52	0.10	0.23	0.51	0.13	0.38	0.47
VITA-1.5 [56]	0.45	0.56	0.63	0.39	0.48	0.56	0.25	0.42	0.63	0.18	0.31	0.40	0.20	0.45	0.72
Gpt-4v [57]	0.81	0.85	0.87	0.70	0.83	0.87	0.72	0.82	0.86	0.32	0.39	0.47	0.22	0.62	0.68
Gpt-4o [58]	0.89	0.89	0.90	0.83	0.86	0.87	0.74	0.83	0.86	0.57	0.69	0.78	0.49	0.71	0.83
CdMT-LLaVA-1.5	0.55	0.71	0.83	0.48	0.60	0.73	0.45	0.61	0.76	0.35	0.46	0.74	0.36	0.52	0.81
CdMT-LLaVA-NeXT	0.60	0.74	0.85	0.51	0.65	0.80	0.48	0.62	0.77	0.37	0.50	0.77	0.41	0.55	0.84
CdMT-VITA-1.5	0.68	0.83	0.90	0.65	0.76	0.82	0.52	0.67	0.80	0.45	0.63	0.82	0.61	0.75	0.87
CdMT-Gpt-4v	<u>0.91</u>	<u>0.96</u>	<u>0.97</u>	0.89	0.91	0.92	<u>0.90</u>	<u>0.97</u>	0.99	<u>0.77</u>	<u>0.86</u>	<u>0.89</u>	<u>0.83</u>	<u>0.91</u>	<u>0.95</u>
CdMT-Gpt-4o	0.93	0.97	0.98	<u>0.88</u>	0.91	0.91	0.97	0.99	0.99	0.89	0.95	1.00	0.85	0.96	0.97

CdMT are agnostic to the underlying architecture. Such universality, coupled with robust performance across geographically diverse datasets, underscores the reliability of our strategy. To ensure statistical rigor, all reported metrics represent the mean of five independent trials.

Figure 5.5 presents a specific recognition case from the Sapporo dataset, displaying the exact prompts and descriptions generated by the CdMT framework. As shown in the figure, we explicitly embed the center coordinates of the target sign (e.g., Center Line at (910, 271)) into the prompts during the context generation phase. This step assists the LMM in distinguishing the correct target from other objects. Furthermore, the prior traffic sign hypothesis enables the model to filter out unrelated classes from the complete candidate list. For characteristic description generation, the prompts are designed to ensure that the LMM explicitly attends to essential traffic sign attributes, including shape, color, and structural composition. Through in-context learning, the LMM produces concise characteristic descriptions for each traffic sign. By leveraging the LMM’s visual understanding to convert template images into textual descriptions, the proposed method reduces the cross-domain discrepancy between template and target traffic sign images. For differential description generation, visually similar traffic signs are provided to the LMM so that it can emphasize the distinctions between them and thereby enhance fine-grained recognition. All prompts used in the proposed framework are simple and unified, without the need for special adjustment for different target traffic signs. Figures 5.6–5.9 show recognition examples on the GTSRB, BTSD, TT-100K, and Yokohama datasets, respectively. Similar to the results on the Sapporo dataset, these findings indicate that the proposed CdMT framework is generalizable and does not require special adaptation for cross-country traffic signs.

5.3.3 Ablation Studies

5.3.3.1 Thinking Strategies

To validate the effectiveness of the CdMT strategy and analyze the impact of each description type, we assessed Top- k fine-grained TSR performance under various configurations across five datasets (see Table 5.2). The baseline method, which inputs target traffic signs directly into the LMM without any descriptive context, characteristic, or differential information, serves as a reference. As detailed in Table 5.2, this baseline consistently yields the lowest accuracy compared to any setup involving multi-step reasoning. Results demonstrate that performance steadily im-

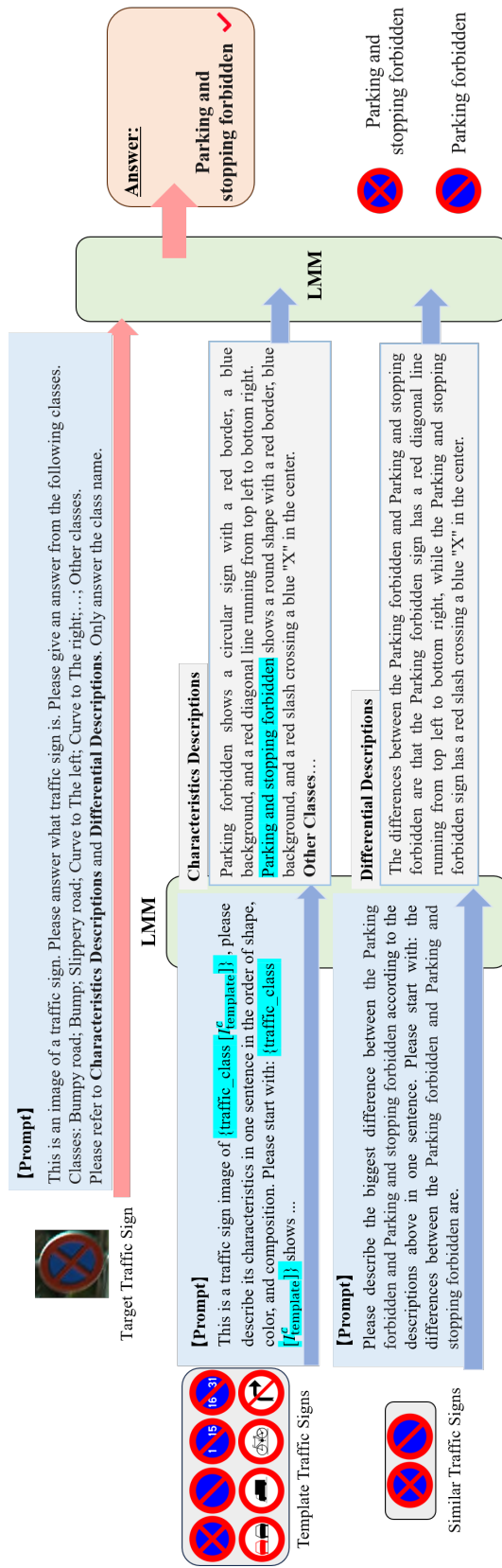


Figure 5.7: Qualitative recognition results generated by our method on the BTSD dataset.

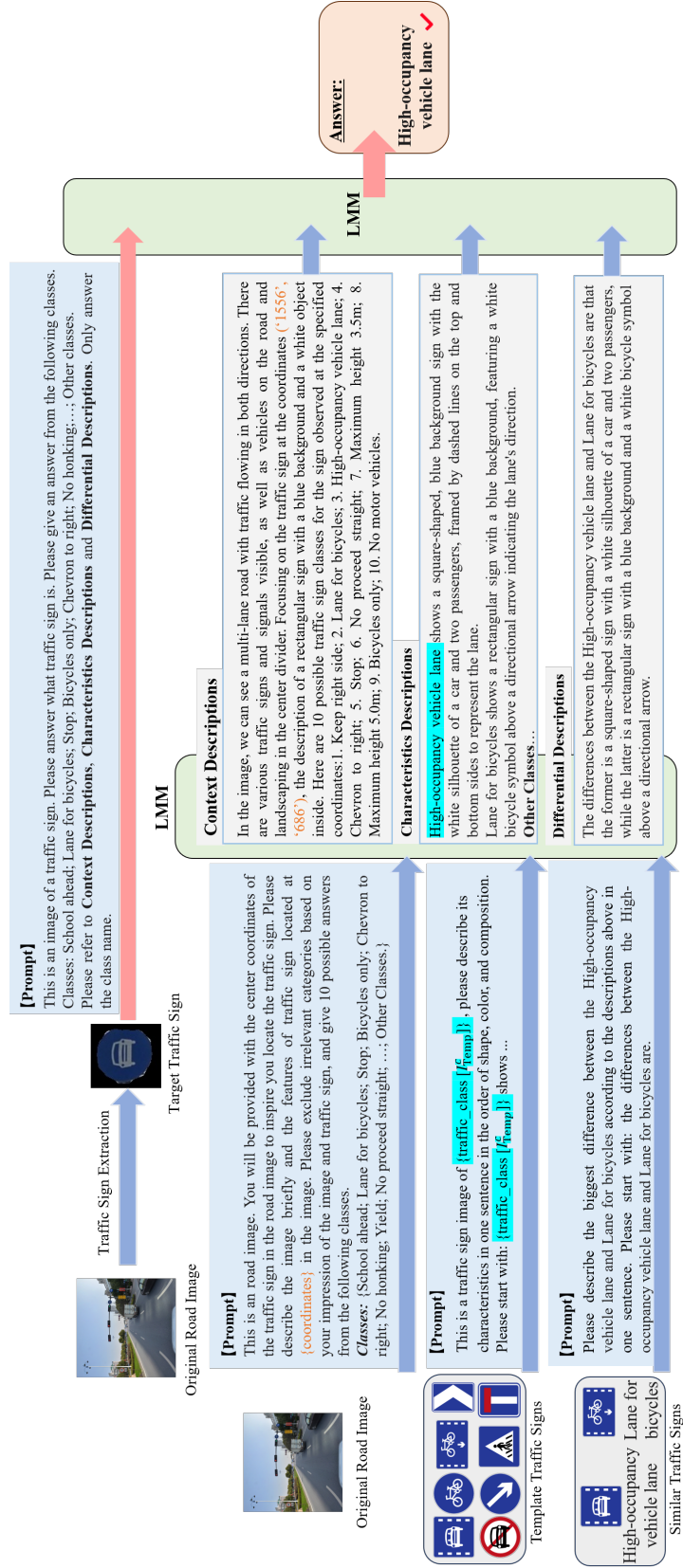


Figure 5.8: Qualitative recognition results generated by our method on the TT-100K dataset.

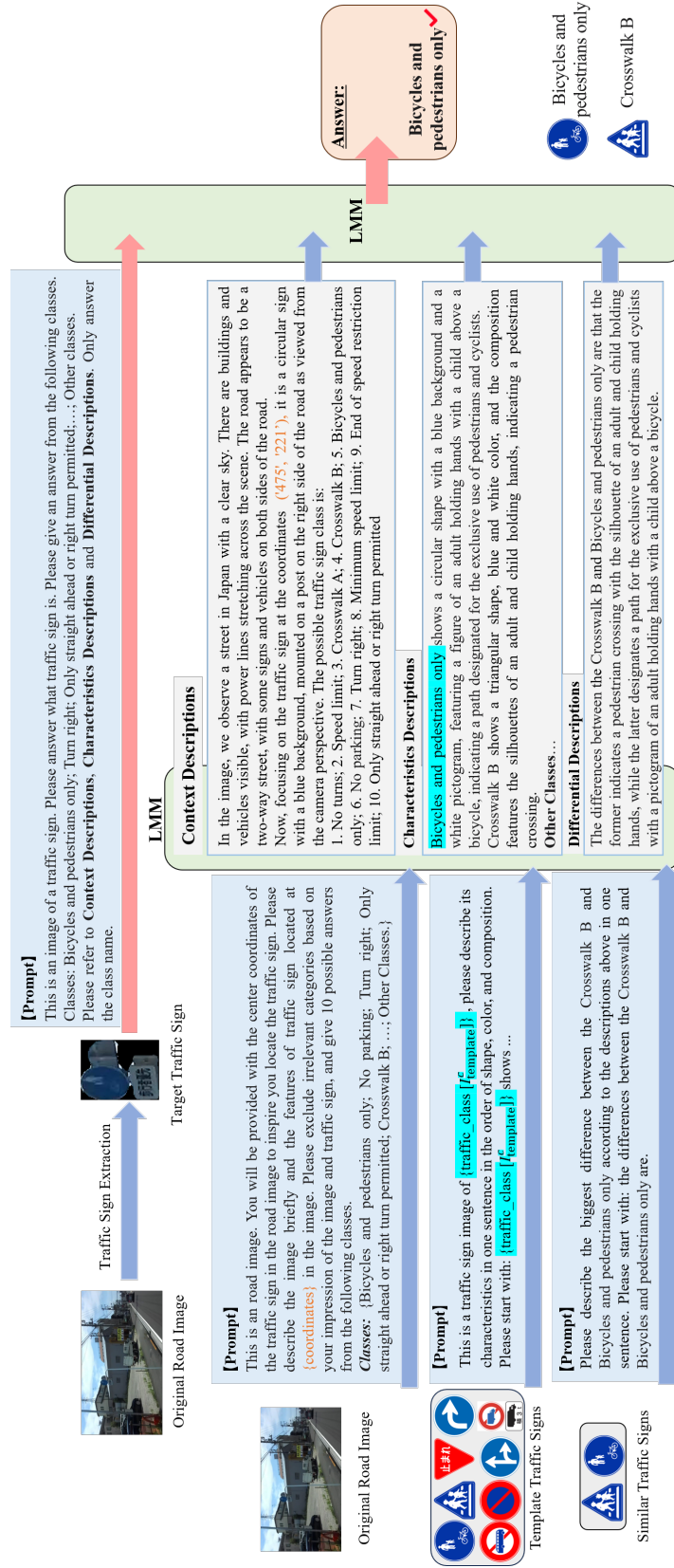


Figure 5.9: Qualitative recognition results generated by our method on the Yokohama dataset.

Table 5.2: Evaluation of Top- k zero-shot fine-grained TSR performance across distinct reasoning strategies. The abbreviations “Cont*”, “Char*”, and “Diff*” correspond to context, characteristic, and differential descriptions, respectively. Best results are highlighted in bold. The symbol “-” implies that context descriptions could not be generated due to the unavailability of original road images.

LMM	Cont*	Char*	Diff*	GTSRB			BTSD			TT-100K			Sapporo			Yokohama		
				Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Gpt-4v				0.81	0.85	0.87	0.70	0.83	0.87	0.72	0.82	0.86	0.32	0.39	0.47	0.22	0.62	0.68
	✓	✓		-	-	-	-	-	-	0.77	0.86	0.88	0.48	0.60	0.68	0.49	0.78	0.91
			✓	0.87	0.95	0.96	0.87	0.90	0.91	0.84	0.90	0.91	0.55	0.65	0.74	0.66	0.74	0.79
			✓	0.82	0.87	0.88	0.76	0.86	0.88	0.77	0.85	0.88	0.42	0.54	0.66	0.35	0.64	0.77
CdMT-Gpt-4v	✓		✓	-	-	-	-	-	-	0.76	0.85	0.89	0.62	0.74	0.78	0.55	0.83	0.91
	✓	✓		-	-	-	-	-	-	0.85	0.92	0.92	0.76	0.84	0.86	0.66	0.87	0.94
		✓	✓	0.91	0.96	0.97	0.89	0.91	0.92	0.88	0.94	0.95	0.68	0.82	0.87	0.81	0.88	0.94
	✓	✓	✓	-	-	-	-	-	-	0.90	0.97	0.99	0.77	0.86	0.89	0.83	0.91	0.95
Gpt-4o				0.89	0.89	0.90	0.83	0.86	0.87	0.74	0.83	0.86	0.57	0.69	0.78	0.49	0.71	0.83
	✓	✓		-	-	-	-	-	-	0.82	0.91	0.93	0.77	0.79	0.83	0.50	0.83	0.89
			✓	0.92	0.96	0.98	0.86	0.88	0.88	0.93	0.98	0.98	0.86	0.91	0.95	0.82	0.93	0.97
			✓	0.89	0.95	0.95	0.85	0.89	0.89	0.92	0.97	0.97	0.74	0.85	0.92	0.58	0.74	0.85
CdMT-Gpt-4o	✓		✓	-	-	-	-	-	-	0.93	0.97	0.97	0.85	0.91	0.93	0.68	0.85	0.90
	✓	✓		-	-	-	-	-	-	0.95	0.98	0.98	0.87	0.93	0.96	0.79	0.94	0.96
		✓	✓	0.93	0.97	0.98	0.88	0.91	0.91	0.96	0.98	0.99	0.89	0.94	0.99	0.82	0.94	0.96
	✓	✓	✓	-	-	-	-	-	-	0.97	0.99	0.99	0.89	0.95	1.00	0.85	0.96	0.97

proves as more reasoning steps are integrated, confirming the value of our strategy. Furthermore, the data indicates that each description type positively contributes to the LMM’s fine-grained recognition capabilities. When evaluated individually, the characteristic description provides the most significant gain. This is because it allows the LMM to jointly process the visual features of the target sign alongside the textual descriptions of templates, thereby improving discrimination. Additionally, adding context and differential descriptions leads to further improvements across all datasets, which is consistent with our original design expectations.

Figure 5.10 provides a qualitative comparison between the baseline (Gpt-4o) and our proposed CdMT-Gpt-4o framework across the five datasets. In contrast to the baseline, our approach demonstrates superior stability in real-world environments and exhibits stronger generalization capabilities for traffic signs from different nations. As observed in Fig. 5.10, the proposed strategy enables the LMM to effectively model both the diversity and similarity of traffic signs, resulting in more precise fine-grained TSR. Additionally, Figure 5.11 highlights representative failure cases. Primary errors occur when traffic signs suffer from severe blurring, which prevents the LMM from reliably interpreting the visual content.

5.3.3.2 Hypothesis and Coordinate

Using the optimal model configurations, we assessed the impact of the proposed prior traffic sign hypothesis and center coordinate–based prompt optimization. These evaluations were conducted across three real-world datasets: TT-100K, Sapporo, and Yokohama. We compared different context description generation strategies, and the corresponding results are summarized in Table 5.3. Note that all results in Table 5.3 are obtained under the full multi-step thinking setting, i.e., using context, characteristic, and differential descriptions simultaneously. As detailed in Table 5.3, excluding both the prior traffic sign hypothesis and center coordinate optimization results in performance comparable to the configuration in Table 5.2 (using only characteristic and differential descriptions). In this scenario, since the generic background context adds minimal value beyond the existing descriptions, the performance gain is negligible. A similar trend is observed when applying center coordinate optimization in isolation; while it assists in locating the target, the improvement remains marginal due to the dominance of the characteristic features. Conversely, employing the prior hypothesis without coordinate constraints actually degrades ac-

Dataset	Target Traffic Sign	TSR Results		Ground Truth
		Baseline	CdMT	
GTSRB		Bend to right 	Double bend (first to right) 	Double bend (first to right)
		Pedestrians in road ahead 	Road works 	Road works
		Town or city limit sign (front) 	Uneven road 	Uneven road
BTSD		Parking forbidden 	Parking and stopping forbidden 	Parking and stopping forbidden
		No entry for all drivers, in both directions 	Forbidden direction for all drivers 	Forbidden direction for all drivers
		Bump 	Hump 	Hump
TT-100K		Lane for bicycles 	High-occupancy vehicle lane 	High-occupancy vehicle lane
		No U-turns 	No left turn 	No left turn
		Pedestrian crossing ahead 	Merging traffic on right 	Merging traffic on right
Sapporo		Crosswalk B 	Crosswalk A 	Crosswalk A
		No turns 3 	One-way street to the right 	One-way street to the right
		No turns 3 	Keep left 	Keep left
Yokohama		Crosswalk B 	Bicycles and pedestrians only 	Bicycles and pedestrians only
		No turns 	Only straight ahead or right turn permitted 	Only straight ahead or right turn permitted
		Speed limit 	End of speed restriction limit 	End of speed restriction limit

Figure 5.10: Recognition examples of baseline and proposed CdMT (Gpt-4o).

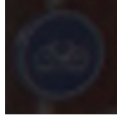
Target Traffic Sign	TSR Results		Ground Truth
	Baseline	CdMT	
	Cyclist and moped crossing 	Road reserved 	Compulsory cycleway 
	Curve left 	Right-of-way on upcoming intersection 	Slipperiness 
	Crosswalk A 	Crosswalk A 	Road closed to pedestrians 

Figure 5.11: Visualization of recognition failures yielded by the baseline and the proposed CdMT (Gpt-4o).

curacy. This occurs because the LMM struggles to distinguish the specific target in multi-sign images, leading to noisy and ambiguous outputs. Crucially, the joint application of both the prior hypothesis and coordinate optimization delivers the best performance, as it enables the LMM to precisely localize the target while effectively filtering out irrelevant candidates.

5.3.3.3 Thinking Orders

Table 5.4 investigates the impact of the reasoning sequence on TSR accuracy. For the GT-SRB and BTSD benchmarks, we inverted the order of characteristic and differential descriptions. Conversely, for the TT-100K, Sapporo, and Yokohama datasets, the sequence of context and characteristic descriptions was alternated. The quantitative outcomes of these permutations are detailed in the table. After changing the thinking order, the TSR performance remains almost unchanged compared with the original setting, indicating the robustness of the proposed method. This stability suggests that, although the order of reasoning steps is modified, the cues acquired by the LMM are essentially consistent, leading to no substantial difference in recognition accuracy.

Table 5.3: Evaluation of Top- k zero-shot fine-grained TSR performance employing diverse context description generation strategies. The best results are shown in bold.

LMM	Prior hypothesis	Center coordinates	TT-100K			Sapporo			Yokohama		
			Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
CdMT-Gpt-4v	✓		0.87	0.92	0.92	0.68	0.86	0.88	0.78	0.84	0.88
		✓	0.86	0.92	0.93	0.60	0.76	0.76	0.73	0.88	0.91
CdMT-Gpt-4o	✓		0.85	0.93	0.95	0.67	0.87	0.88	0.74	0.88	0.91
	✓	✓	0.90	0.97	0.99	0.77	0.86	0.89	0.83	0.91	0.95
CdMT-Gpt-4o			0.90	0.95	0.98	0.80	0.88	0.89	0.80	0.92	0.95
	✓		0.86	0.90	0.93	0.75	0.82	0.84	0.77	0.90	0.92
CdMT-Gpt-4o		✓	0.90	0.95	0.97	0.78	0.88	0.90	0.79	0.92	0.95
	✓	✓	0.97	0.99	0.99	0.89	0.95	1.00	0.85	0.96	0.97

Table 5.4: TSR performance under different reasoning configurations. The labels “w” and “w/o” denote the application or omission of the thinking steps, respectively.

LMM	Change Thinking	GTSRB			BTSD			TT-100K			Sapporo			Yokohama		
		Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
CdMT-Gpt-4v	w	0.91	0.96	0.97	0.89	0.91	0.92	0.89	0.97	0.99	0.77	0.87	0.89	0.83	0.92	0.95
	w/o	0.91	0.96	0.97	0.89	0.91	0.92	0.90	0.97	0.99	0.77	0.86	0.89	0.83	0.91	0.95
CdMT-Gpt-4o	w	0.93	0.97	0.98	0.88	0.91	0.91	0.96	0.98	0.99	0.89	0.95	1.00	0.87	0.97	0.98
	w/o	0.93	0.97	0.98	0.88	0.91	0.91	0.97	0.99	0.99	0.89	0.95	1.00	0.85	0.96	0.97

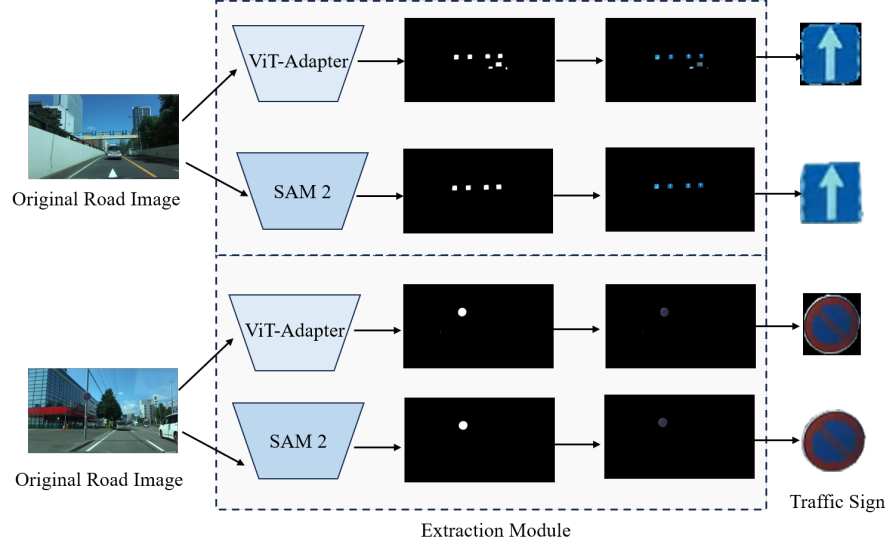


Figure 5.12: Visual results of traffic sign extraction achieved by the proposed module across various segmentation architectures.

5.3.3.4 Flexibility and Extensibility

The results show that our multi-step strategy works well with various LMMs, supporting both open-source models (e.g., CdMT-LLaVA-NeXT, CdMT-VITA-1.5) and closed-source ones (e.g., CdMT-Gpt-4v, CdMT-Gpt-4o). It maintains stable performance across all these different architectures. Furthermore, the traffic sign extraction part is flexible; it does not rely on one specific network and can be replaced by more advanced models if necessary. Table 5.5 compares the TSR performance on the TT-100K, Sapporo, and Yokohama datasets using different extraction methods. As shown in the table, using the Segment Anything Model 2 (SAM 2) [59] leads to higher Top-1, Top-3, and Top-5 accuracy across all datasets compared to the ViT-Adapter [29] baseline. This suggests that better segmentation quality directly results in more reliable traffic sign extraction, which in turn improves the final TSR results. Figure 5.12 displays example outputs from the two segmentation models. It can be seen that while our method functions effectively with different backbones, SAM 2 produces much clearer and more accurate extraction results.

5.3.3.5 Inference Speed

Table 5.6 summarizes the inference speed of the proposed method with different combinations of LMMs and segmentation modules. When ViT-Adapter-base is used for segmentation, the

Table 5.5: Evaluation of extraction module impact on TSR performance.

Model	TT-100K			Sapporo			Yokohama		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
ViT-Adapter	0.93	0.96	0.98	0.84	0.91	0.96	0.80	0.88	0.93
SAM 2	0.97	0.99	0.99	0.89	0.95	1.00	0.85	0.96	0.97

Table 5.6: Inference speed for each traffic sign. “s” denotes seconds.

Extraction Module	LMM	Inference Speed
ViT-Adapter	LLaVA-v1.5	6.4 s
	LLaVA-NeXT	6.0 s
	VITA-1.5	2.3 s
	Gpt-4v	2.0 s
	Gpt-4o	1.6 s
SAM 2	LLaVA-v1.5	6.0 s
	LLaVA-NeXT	5.6 s
	VITA-1.5	1.9 s
	Gpt-4v	1.6 s
	Gpt-4o	1.2 s

average processing time is about 0.4 s per road image, indicating relatively high efficiency. Replacing it with the SAM 2-base extraction module further accelerates the segmentation stage and enables near real-time extraction. Among all evaluated LMM settings, Gpt-4o equipped with the proposed CdMT framework achieves the fastest end-to-end inference, requiring only 1.2 s per traffic sign. The CdMT variant built on VITA-1.5 achieves an inference time of 1.9 s per traffic sign. Owing to the modular design of the framework, it can be readily adapted to future LMM variants; thus, advances in model architectures and computational optimization can be quickly incorporated, with the potential to further reduce inference latency.

5.3.3.6 Significant Domain Shift

To examine the resilience of CdMT against significant domain shifts, we perform a qualitative case study on two complex real-world scenarios, presented in Fig. 5.13. The first instance, “Parking allowed,” is characterized by a distinct background color deviation from its template, while the second, “Overtaking vehicles forbidden,” suffers from partial occlusion by foliage, complicating the recognition process. Both examples represent common challenges in practical TSR.

Target Traffic Sign	Temple Traffic Sign	TSR Results		
	Parking allowed 	CdMT-LLaVA-NeXT	CdMT-VITA-1.5	CdMT-Gpt-4o
		Parking lot 	Parking lot 	Parking allowed 
	Overtaking vehicles forbidden 	No entry for the transportation vehicles 	Slipperiness 	Overtaking vehicles forbidden 

Figure 5.13: CdMT recognition results on samples with significant domain shifts.

We benchmark three model variants: CdMT-LLaVA-NeXT, CdMT-VITA-1.5, and CdMT-Gpt-4o. As observed, both CdMT-LLaVA-NeXT and CdMT-VITA-1.5 misclassified the “Parking allowed” sign as the visually similar “Parking lot” and failed to identify the occluded “Overtaking vehicles forbidden” sign. Conversely, CdMT-Gpt-4o correctly predicted both instances, demonstrating superior robustness to severe domain discrepancies. These findings highlight that the capability of the underlying LMM is a decisive factor in maintaining reliability under extreme conditions.

5.3.3.7 Impact of Description Granularity

We assess the influence of characteristic description length on TSR accuracy through an ablation study using CdMT-Gpt-4o, with quantitative results presented in Table 5.7. Three levels of verbosity are evaluated: short, medium, and long (exemplified in Fig. 5.14). Results across all datasets indicates that both medium and long configurations consistently outperform the short setting in terms of Top-1, Top-3, and Top-5 metrics. Notably, medium-length descriptions achieve performance parity with the long format while incurring significantly lower computational overhead. In contrast, the short setting suffers from a marked degradation in accuracy, likely attributed to the insufficient encoding of fine-grained visual details. These findings suggest that while excessive brevity risks omitting critical discriminative cues, highly detailed descriptions yield diminishing returns. Consequently, selecting an optimal description length is essential to balance the trade-off between recognition precision and computational efficiency.

Table 5.7: TSR results of CdMT-Gpt-4o with different characteristic description lengths.

LMM	Description Length	GTSRB			BTSD			TT-100K			Sapporo			Yokohama		
		Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
CdMT-Gpt-4o	Short	0.91	0.94	0.97	0.86	0.89	0.89	0.90	0.94	0.96	0.86	0.92	0.98	0.75	0.82	0.91
	Medium	0.93	0.97	0.98	0.88	0.91	0.91	0.97	0.99	0.99	0.89	0.95	1.00	0.85	0.96	0.97
	Long	0.92	0.97	0.98	0.88	0.91	0.92	0.97	0.98	0.99	0.89	0.94	0.99	0.85	0.95	0.96

Description Length	Template Traffic sign  Lane usage 4
Short	Lane usage 4 shows a white upward arrow on a blue square sign, indicating that traffic in this lane must go straight.
Midium	Lane usage 4 shows a square shape with rounded corners, a blue background, and a white arrow pointing upwards, flanked by dashed white lines on both sides to suggest straight movement in a specific traffic lane.
Long	Lane usage 4 shows a blue square sign with a thick white upward arrow in the center, flanked by vertical dashed white lane markings on both sides. The prominent arrow indicates that vehicles in this lane are only permitted to go straight ahead, and turning is not allowed. This sign is typically placed above or next to the relevant lane to clearly direct drivers to proceed forward without deviation.

Figure 5.14: Representative characteristic descriptions across short, medium, and long settings.

5.4 Discussion

5.4.1 Test Set Contamination

LMMs are trained on extensive web-scale data, which has led to concerns that these models may have inadvertently memorized public benchmarks [60]. To address this issue, our evaluation covers not only three widely used benchmark datasets (GTSRB, BTSD, and TT-100K) but also two private datasets collected in Japan (Sapporo and Yokohama). The proposed method achieves stable and reliable performance across all five datasets. Because the private datasets are not publicly released and therefore cannot appear in the training sources of any evaluated models, our experiments are free from test set contamination.

5.4.2 Significance and Practical Value

Our approach enables efficient TSR in diverse driving environments. It achieves consistent results across different countries without the need for training data, confirming its practical value. Collecting and annotating traffic sign datasets from multiple regions is costly and time-consuming, especially when accounting for heterogeneous data sources, privacy restrictions, and the difficulty of accessing data from less developed areas. By reducing the need for large-scale data collection, the proposed approach lowers deployment costs and also improves fairness. Cur-

rent advanced driving assistance and autonomous driving systems are often limited to specific regions and provide limited support in areas with fewer resources. By enabling robust cross-country TSR, the proposed method offers a way to expand these technologies to underdeveloped regions, helping promote broader accessibility and equity.

5.4.3 Limitation

5.4.3.1 Automated Identification of Similar Signs

Although this work validates the utility of differential descriptions for LMMs, the current process depends on manual curation by experts to identify similar sign pairs. This reliance on human judgment may introduce bias and limit the scalability of the method when applied to extensive datasets. To address this issue, future research will focus on developing automated algorithms to detect visually similar categories. Implementing such data-driven approaches will help ensure consistency and further expand the generalizability of the proposed framework.

5.4.3.2 Robustness under Diverse Weather Conditions

The experimental validation in this work relies on five distinct datasets, comprising public benchmarks (GTSRB, BTSD, TT-100K) and proprietary collections (Sapporo, Yokohama). It is important to note that these sources predominantly feature imagery captured under favorable weather conditions with minimal visual degradation. However, practical driving scenarios often involve adverse meteorological elements, such as rain, fog, or snow, which introduce artifacts or partial occlusions that can compromise recognition accuracy. Consequently, enhancing robustness against such environmental variability represents a critical objective for future research. A promising direction involves extending the thinking strategies to explicitly incorporate weather-related contextual cues, thereby facilitating reliable performance in challenging scenarios.

5.4.3.3 Computational Complexity and Latency

Table 5.8 outlines the computational overhead associated with each component of the CdMT-Gpt-4o framework. Specifically, traffic sign extraction consumes 0.1 s, while context generation and the multi-step reasoning phase require 0.4 s and 0.7 s, respectively. It is important to note

Table 5.8: Analysis of computational overhead for CdMT-Gpt-4o. The variables N_C and N_D represent the number of template classes and visually similar signs, respectively. “Phase Type” distinguishes whether a stage is performed online for inference or offline for preprocessing. All time measurements are reported in seconds.

CdMT Phase	Time Cost (s)	Phase Type
Traffic Sign Extraction	0.1	Online
Context Description	0.4	Online
Characteristic Description	$0.3 \times N_C$	Offline
Differential Description	$0.3 \times N_D$	Offline
Multi-step Reasoning	0.7	Online

that the generation of characteristic and differential descriptions is performed offline and cached, meaning these steps do not contribute to online inference time. However, the computational demand of the online reasoning process introduces additional latency, which may pose a bottleneck for real-time applications or deployment on resource-constrained edge devices. One possible solution is to apply model distillation to obtain smaller and faster LMM variants, reducing inference time while retaining accuracy. Another direction is to streamline the multi-step reasoning pipeline by removing redundant steps or adopting adaptive reasoning strategies that adjust reasoning depth based on input difficulty. These improvements would make the CdMT framework more practical for latency-sensitive applications in real-world environments.

5.4.3.4 Integration of Standardized Sign Taxonomies

While our current approach utilizes in-context learning to produce characteristic descriptions from templates, there is significant potential in leveraging standardized traffic sign taxonomies. Prominent examples include the regulations defined by the Vienna Convention and the design guidelines outlined in ISO 3864. Embedding these formal standards into the prompt design of CdMT could enhance both the interpretability and consistency of the generated descriptions. Furthermore, this integration may further improve the model’s generalization capabilities across diverse domains. Future research will investigate effective methods to incorporate such structured taxonomy information into the prompt generation process.

5.5 Conclusion

In this chapter, we proposed CdMT as a unified framework for fine-grained TSR. The framework is lightweight in design, broadly applicable, and can be extended to a wide range of LMMs. By introducing a structured multi-step thinking process, CdMT enables LMMs to perform fine-grained TSR without any model training. The evaluations conducted on three benchmark datasets, together with two real-world datasets show clear and consistent improvements across all settings. Looking ahead, several directions are worth exploring. One is to develop automatic methods for discovering groups of visually similar traffic signs, which would reduce manual effort in building description sets. Another is to improve robustness under different weather and illumination conditions. In addition, further reducing computation and latency will be important for applying CdMT in real-time driving systems.

Chapter 6

Multitask Remote Sensing Recognition Based on Large Multimodal Models

6.1 Introduction

While recent progress in LMMs has advanced multitask recognition in remote sensing, existing methods often fail to capture complex regional geospatial patterns, such as those found in Japan. Situated along a seismic belt, the country exhibits unique urban configurations and landscapes that diverge significantly from global standards. To overcome these limitations, we propose the **GeoJapan Fusion Framework (GFF)**, a multimodal architecture that combines an LLM with a Vision-Language Model (VLM). By utilizing in-context learning, this system improves multimodal alignment specifically for Japanese remote sensing scenarios. Furthermore, GFF incorporates a cross-modal feature fusion mechanism powered by Low-Rank Adaptation (LoRA), which facilitates efficient representation alignment and scalable adaptation. Complementing this framework, we construct the **GeoJapan** dataset, a large-scale benchmark constructed for multitask recognition. Experimental evaluations confirm that GFF consistently outperforms state-of-the-art LMMs across various tasks, highlighting its potential for advanced multimodal applications.

6.2 GeoJapan Data Construction

6.2.1 Study Region

The GeoJapan dataset comprises remote sensing imagery from two representative regions in Japan: Kawasaki and Sapporo. Kawasaki, situated in the Greater Tokyo Area, is characterized

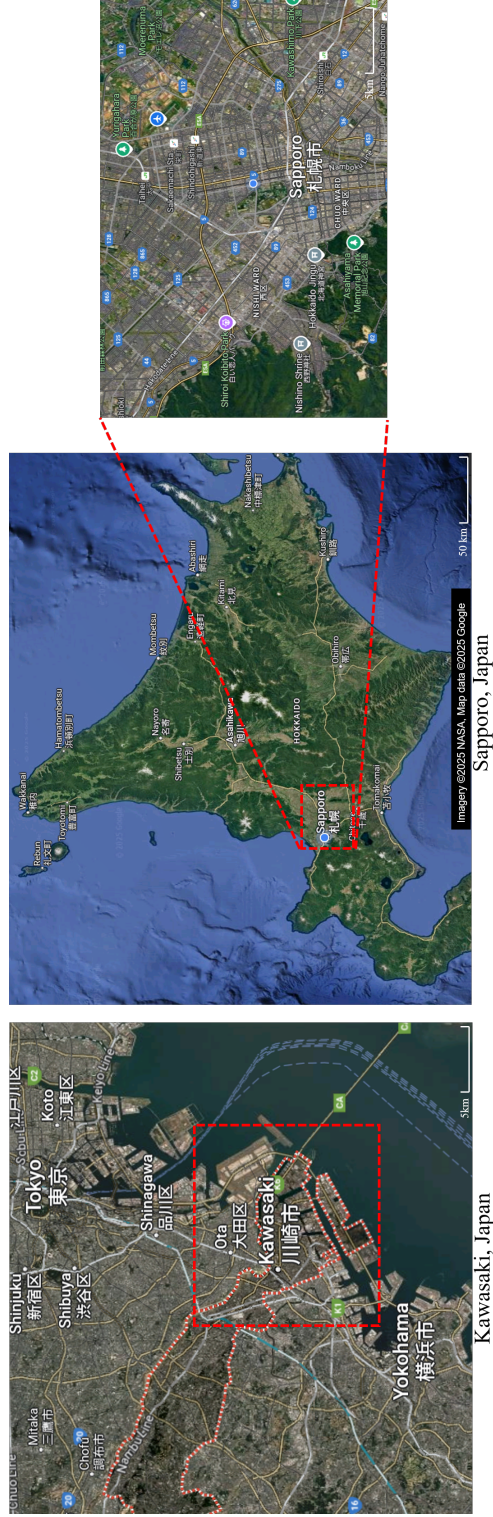


Figure 6.1: Overview of the study area. Kawasaki corresponds to an urban industrial zone, whereas Sapporo includes large natural regions with forests and farmland.

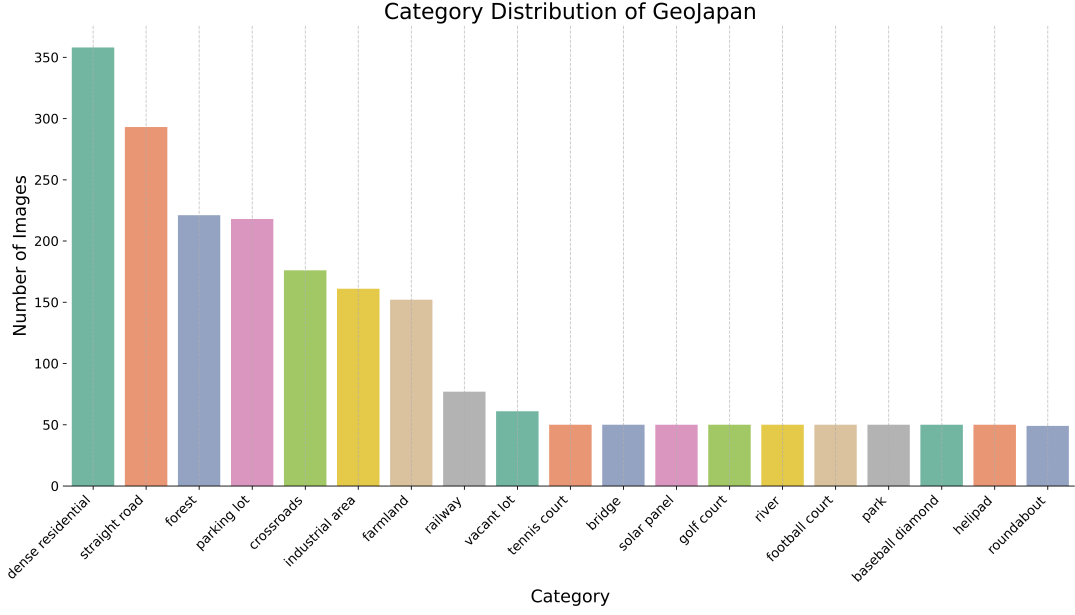


Figure 6.2: Distribution of geospatial object categories in the GeoJapan dataset.

by dense industrial and urban development. In contrast, Sapporo in northern Japan contains extensive forests and farmland, offering a mix of rural and natural landscapes. These regions were intentionally chosen to span a wide spectrum of geospatial object types commonly observed across Japan. The spatial coverage of the dataset is shown in Figure 6.1. By integrating imagery from both locations, GeoJapan includes urban, semi-urban, and natural scenes, providing a comprehensive basis for building and evaluating region-specific remote sensing LMMs.

6.2.2 Dataset Curation and Labeling Protocol

Within the GeoJapan dataset, we focus on two tasks: geospatial object classification and scene description. For the geospatial object classification task, we prepare a set of 2,216 image patches, each with a resolution of 256×256 pixels. This subset spans 19 land-cover classes frequently observed in Japanese urban and suburban regions. The categories include *baseball diamond*, *bridge*, *crossroads*, *dense residential*, *farmland*, *football court*, *forest*, *golf court*, *helipad*, *industrial area*, *park*, *parking lot*, *railway*, *river*, *roundabout*, *solar panel*, *straight road*, *tennis court*, and *vacant lot*, providing broad coverage of common geospatial structures.

The geospatial object categories are illustrated in Figure 6.2, highlighting the broad variability of object types across the target regions. All images in the dataset are paired with their corre-

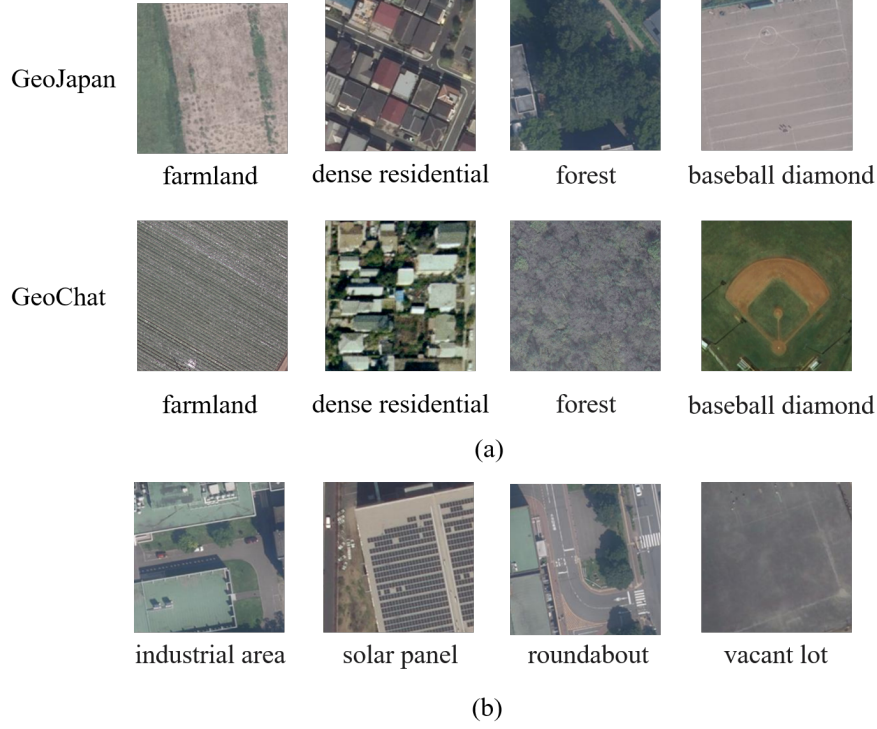


Figure 6.3: Examples of geospatial object categories. (a) Comparison between Japanese remote sensing imagery and general remote sensing imagery, showing the same object categories from both the GeoJapan and GeoChat datasets. (b) Illustration of additional geospatial object categories that are newly included in GeoJapan.

sponding geospatial object labels, forming a reliable basis for developing and evaluating models. Representative examples are provided in Figure 6.3. As shown in Figure 6.3(a), the appearance of several categories differs noticeably between Japanese remote sensing images and general remote sensing datasets. For example, the “farmland” category displays clear visual differences between GeoJapan and GeoChat. Similarly, categories such as “dense residential,” “forest,” and “baseball diamond” show distinct visual patterns that reflect the characteristics of Japanese landscapes. Figure 6.3(b) shows additional categories introduced in GeoJapan, including “industrial area,” “solar panel,” “roundabout,” and “vacant lot,” which further expand the range of geospatial objects represented in the dataset.

In addition, we create an instruction-tuning dataset that provides image instruction pairs designed to mirror practical user queries. An example constructed for a GeoJapan image is shown in Figure 6.4. This format allows LMMs to learn how to interpret visual content and produce suitable natural language outputs, helping narrow the gap between visual understanding and lan-

```

{
  "id": "Image_id",
  "image": "Image_path",
  "conversations": [
    {
      "from": "human",
      "prompt": "<image>\nClassify the given image in one of the following classes. Classes:
        baseball diamond, bridge, crossroads, dense residential, farmland, football court, forest,
        golf court, helipad, industrial area, park, parking lot, railway, river, roundabout, solar panel,
        straight road, tennis court, and vacant lot.\nAnswer in one word or a short phrase."
    },
    {
      "from": "gpt",
      "ground truth": "Name of the geospatial object"
    }
  ]
},

```

Figure 6.4: Example of instruction-tuning dataset for a GeoJapan image.

guage generation.

Furthermore, we define a scene description task to more fully reflect the complexity of Japanese remote sensing imagery. In this setting, we select 474 high-resolution images with a resolution of 800×800 pixels, each depicting detailed and realistic spatial layouts. In contrast to the geospatial object classification subset, these images usually contain multiple geospatial objects within a single view, making them more representative of real-world scenarios that demand comprehensive, scene-level interpretation. To facilitate LMM-based understanding of such complex scenes, we construct a dedicated scene description dataset. The annotation process begins with OpenAI o3 [61] generating initial drafts. To ensure the text is correct and meaningful, experts carefully verify these outputs (Figure 6.5). These curated texts act as ground truth labels, providing robust supervision for training.

6.3 Method

In this section, we present GFF, a system built to analyze remote sensing images. To ensure performance on multiple tasks, our method integrates a VLM based on Vision Transformers with

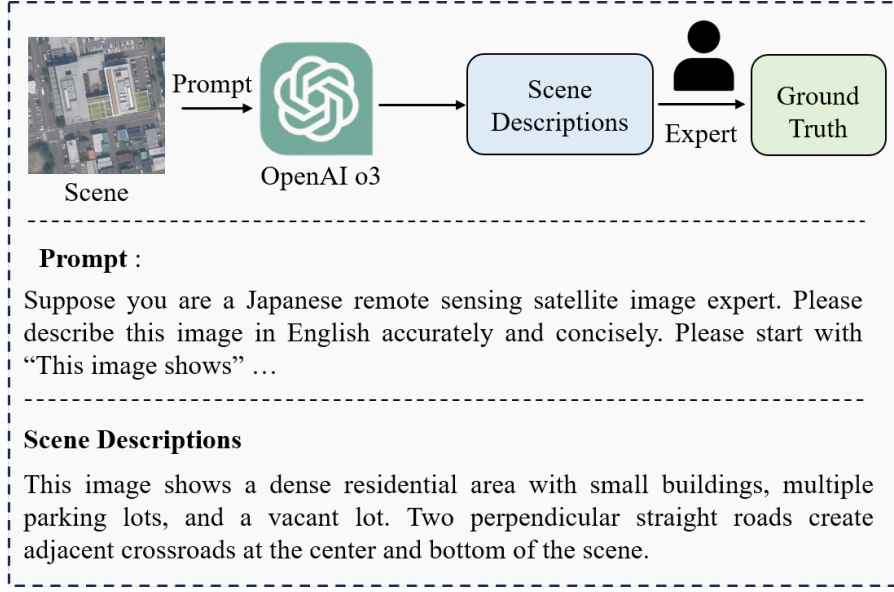


Figure 6.5: Scene description generation.

an LLM. We also use in-context learning, which helps the model better understand the specific details of Japanese remote sensing data. The architecture consists of several main parts. First, the system extracts and aligns visual features from the input image. Next, a fusion module combines these visual features with text inputs and context descriptions into a single sequence. Finally, the LLM uses this combined input to perform the recognition tasks.

6.3.1 Visual Feature Extraction and Projection

This subsection explains how the remote sensing image is converted into a sequence representation that can be processed by the multimodal model.

6.3.1.1 Linear Projection and Patch Embedding

Let $I \in \mathbb{R}^{H \times W \times C}$ be the input image. We first resize and normalize the image to obtain $I' \in \mathbb{R}^{H' \times W' \times C}$ so that the input resolution is consistent across samples. The image is then divided into N_p patches of size $P \times P$. Working with patches is more suitable for transformer-based encoders, as each patch can be treated as an independent token. Each patch p_i is flattened and mapped to a D_V -dimensional vector using a linear projection layer:

$$\mathbf{e}'_{patch,i} = \text{LinearProj}(p_i). \quad (6.1)$$

To retain spatial information that would otherwise be lost during flattening, a learnable two-dimensional positional embedding $\mathbf{e}_{pos,i}$ is added:

$$\mathbf{e}_{patch,i} = \mathbf{e}'_{patch,i} + \mathbf{e}_{pos,i}. \quad (6.2)$$

A trainable [CLS] token is placed at the beginning of the sequence so that the encoder has a dedicated position for aggregating global information. The final input to the vision transformer is

$$\mathbf{X}_{ViT} = [\mathbf{e}_{cls}, \mathbf{e}_{patch,1}, \mathbf{e}_{patch,2}, \dots, \mathbf{e}_{patch,N_p}] \in \mathbb{R}^{(N_p+1) \times D_V}. \quad (6.3)$$

This representation serves as the starting point for the subsequent transformer-based encoding.

6.3.1.2 Spatial Attention Calculation

Within the spatial attention architecture, the patch-level embedding sequence $\mathbf{X}_{ViT}$ serves as the input. This sequence traverses a stack of L_V Transformer encoder layers, which are designed to extract hierarchical visual features. Structurally, the l -th layer comprises two primary sub-components: a Multi-Head Self-Attention (MHSA) unit [62] and a point-wise Feed-Forward Network (FFN). To ensure training stability, Layer Normalization (LN) [63] is applied before each sub-layer. For the input features $\mathbf{Z}^{(l-1)}$ at layer l , the self-attention mechanism initiates by projecting the data into three distinct subspaces, Queries (Q), Keys (K), and Values (V) via learnable weight matrices:

$$\mathbf{Q}^{(l)} = \mathbf{Z}^{(l-1)} \mathbf{W}_Q^{(l)}, \quad (6.4)$$

$$\mathbf{K}^{(l)} = \mathbf{Z}^{(l-1)} \mathbf{W}_K^{(l)}, \quad (6.5)$$

$$\mathbf{V}^{(l)} = \mathbf{Z}^{(l-1)} \mathbf{W}_V^{(l)}, \quad (6.6)$$

These projections facilitate the modeling of global dependencies by determining the relevance of each patch token relative to the entire image context. The resulting attention map is computed using the standard scaled dot-product mechanism:

$$\text{Attention}(\mathbf{Q}^{(l)}, \mathbf{K}^{(l)}, \mathbf{V}^{(l)}) = \text{softmax}\left(\frac{\mathbf{Q}^{(l)}(\mathbf{K}^{(l)})^T}{\sqrt{d_k}}\right) \mathbf{V}^{(l)}, \quad (6.7)$$

where d_k represents the scaling factor corresponding to the key dimension. This operation effectively aggregates contextual information from all spatial locations. Subsequently, the output from the attention heads is processed through a residual connection and layer normalization to refine the feature representation:

$$\mathbf{Z}^{(l')} = \text{LayerNorm}(\mathbf{Z}^{(l-1)} + \text{MHSA}(\mathbf{Z}^{(l-1)})). \quad (6.8)$$

The refined features are then passed through the FFN, which is similarly encapsulated within a residual block to produce the final output for layer l :

$$\mathbf{Z}^{(l)} = \text{LayerNorm}(\mathbf{Z}^{(l')} + \text{FFN}(\mathbf{Z}^{(l')})). \quad (6.9)$$

Upon propagating through the complete stack of L_V layers, the encoder generates the final high-level visual feature sequence:

$$\mathbf{F}_V = \mathbf{Z}^{(L_V)} \in \mathbb{R}^{(N_p+1) \times D_V}, \quad (6.10)$$

This tensor $\mathbf{F}_V$ acts as the robust visual backbone, providing the necessary semantic embeddings for the subsequent multimodal fusion stage.

6.3.1.3 Visual Feature Projection

For visual feature projection, each visual token $\mathbf{f}_{V,i} \in \mathbb{R}^{D_V}$ in $\mathbf{F}_V$ is passed through an MLP adapter:

$$\mathbf{f}_{proj,i} = \text{MLP}(\mathbf{f}_{V,i}). \quad (6.11)$$

The resulting projected tokens are then arranged into the sequence

$$\mathbf{F}_{proj} = [\mathbf{f}_{proj,1}, \dots, \mathbf{f}_{proj,N_p+1}] \in \mathbb{R}^{(N_p+1) \times D_L}, \quad (6.12)$$

which will be combined with the textual features during the subsequent multimodal fusion stage.

6.3.2 Multimodal Fusion

We introduce the multimodal fusion process in this subsection.

6.3.2.1 In-Context Learning

In GFF, in-context learning is employed to enhance the model's ability to handle the distinctive properties of Japanese remote sensing scenes. To achieve this, we make use of the OpenAI o3 LMM and provide several example category descriptions as contextual cues. These examples guide the model in understanding the characteristics of various geospatial objects and scene types in Japan. During inference, each example image is processed separately by the LMM to produce a description $\mathbf{t}_i$. Collectively, these descriptions form a set $\{\mathbf{t}_i\}_{i=1}^N$ that reflects the visual and semantic patterns present in the examples. The final in-context description $\mathbf{T}_{icl}$ is constructed by concatenating all generated descriptions:

$$\mathbf{T}_{icl} = \text{Concatenate}(\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_N). \quad (6.13)$$

This aggregated textual summary $\mathbf{T}_{icl}$ is then used as the prompt in the subsequent encoding and fusion stages. By incorporating information distilled from multiple examples, GFF gains richer contextual signals, improving its ability to recognize geospatial objects and scenes specific to Japanese environments.

6.3.2.2 Text Prompt Encoding and Visual Feature Fusion

Once the visual features and the in-context learning descriptions are prepared, GFF combines them through a multimodal fusion process to form a single input sequence for the LLM. The fused sequence serves as the model's final input for reasoning and generation. Specifically, the multimodal input $\mathbf{X}_{multi}$ is constructed as:

$$\mathbf{X}_{multi} = [\mathbf{e}_{begin}, \mathbf{F}_{proj}, \mathbf{E}_T, \mathbf{E}_{icl}, \mathbf{e}_{end}], \quad (6.14)$$

where $\mathbf{e}_{begin}$ and $\mathbf{e}_{end}$ represent the special boundary tokens (start-of-sequence and end-of-sequence). $\mathbf{F}_{proj}$ denotes the projected visual feature sequence, $\mathbf{E}_T$ corresponds to the token embeddings of the GFF text prompt, and $\mathbf{E}_{icl}$ contains the embeddings of the in-context learning descriptions. By presenting these components as a unified sequence, the LLM can simultaneously process visual content, textual instructions, and the contextual cues associated with Japanese geospatial objects, enabling more accurate and context-aware multimodal predictions.

6.3.3 Large Language Model Inference

Within the architecture of the proposed GFF, the inference phase is executed by injecting the fused multimodal sequence $\mathbf{X}_{multi}$ into the Transformer decoder of the LLM. Functioning in an autoregressive manner, the decoder processes these inputs through a cascade of attention mechanisms and feed-forward networks. The fundamental operations governing this procedure are detailed below.

6.3.3.1 Self-Attention Calculation

Analogous to the visual encoding stage, the self-attention mechanism within each decoder layer is designed to attend globally to tokens across the multimodal sequence. This capability allows the LLM to synthesize and interpret combined textual and visual information. To strictly enforce the autoregressive nature of the decoder, a causal mask $\mathbf{M}_{causal}$ is introduced. This masking constraint ensures that the receptive field of each token is limited exclusively to preceding positions in the sequence. For the input representations $\mathbf{Y}^{(l-1)}$ at the l -th layer, the derivation of Query, Key, and Value matrices follows the linear projection logic established in Equations (6.4)–(6.6). Subsequently, the attention scores are computed as follows:

$$\text{Attention}(\mathbf{Q}^{(l)}, \mathbf{K}^{(l)}, \mathbf{V}^{(l)}) = \text{softmax}\left(\frac{\mathbf{Q}^{(l)}(\mathbf{K}^{(l)})^T}{\sqrt{d_k}} + \mathbf{M}_{causal}\right)\mathbf{V}^{(l)}. \quad (6.15)$$

Consistent with Equation (6.7), $(\mathbf{K}^{(l)})^T$ signifies the transpose operation, while d_k denotes the dimensionality scaling factor for queries and keys. Upon the completion of processing through all L decoder layers, the final output representation $\mathbf{Y}^{(L)}$ is generated following the protocols outlined in Equations (6.8)–(6.10).

Algorithm 2 GeoJapan Fusion Framework (GFF) for Multitask Remote Sensing Recognition

Input: Remote sensing image I , textual prompt E_T , in-context examples $\{E_i, t_i\}_{i=1}^N$

Output: Recognition output or generated description

Step 1: Input Preprocessing and Embedding

- 1: Rescale and normalize I to generate I' aligning with the encoder resolution.
- 2: Partition I' into a grid of N_p discrete patches p_i with dimensions $P \times P$.
- 3: flatten patches and execute linear projection:
- 4: $e_{patch,i} = \text{LinearProj}(p_i)$.
- 5: Inject positional information:
- 6: $e'_{patch,i} = e_{patch,i} + e_{pos,i}$.
- 7: Construct the initial ViT token sequence:
- 8: $X_{ViT} = [e_{cls}, e'_{patch,1}, \dots, e'_{patch,N_p}]$.

Step 2: Encoder Processing (Vision Transformer)

- 9: **for** $l = 1$ to L_V **do**
- 10: Derive projections for Queries, Keys, and Values:
- 11: $Q^{(l)} = Z^{(l-1)} W_Q^{(l)}$, $K^{(l)} = Z^{(l-1)} W_K^{(l)}$, $V^{(l)} = Z^{(l-1)} W_V^{(l)}$.
- 12: Calculate scaled dot-product attention scores:
- 13: $\text{Attention} = \text{softmax}\left(\frac{Q^{(l)} K^{(l)T}}{\sqrt{d_k}}\right) V^{(l)}$.
- 14: Update latent states via Normalization and Residual connections:
- 15: $Z^{(l')} = \text{LayerNorm}(Z^{(l-1)} + \text{Attention})$.
- 16: $Z^{(l)} = \text{LayerNorm}(Z^{(l')} + \text{FFN}(Z^{(l')}))$.
- 17: **end for**
- 18: Obtain final high-level visual features $F_V = Z^{(L_V)}$.

Step 3: Feature Alignment Projection

- 19: **for each** token $f_{V,i} \in F_V$ **do**
- 20: Map visual features to LLM space: $f_{proj,i} = \text{MLP}(f_{V,i})$
- 21: **end for**
- 22: Aggregate projected tokens: $F_{proj} = [f_{proj,1}, \dots, f_{proj,N_p+1}]$.

Step 4: Contextual Prompt Construction

- 23: **for each** example (E_i, t_i) **do**
- 24: Generate auxiliary description $t'_i = \text{LLM}(E_i)$
- 25: **end for**

- 26: Concatenate ICL components:

$$27: \quad T_{icl} = \text{concat}(t'_1, t'_2, \dots, t'_N).$$

Step 5: Sequence Assembly

- 28: Compose the unified multimodal input embedding:
- 29: $X_{multi} = [e_{begin}, F_{proj}, E_T, T_{icl}, e_{end}]$.

Step 6: Autoregressive Decoding

- 30: Process X_{multi} utilizing causal masking constraints:
 - 31: $\text{Attention} = \text{softmax}\left(\frac{Q^{(l)} K^{(l)T}}{\sqrt{d_k}} + M_{causal}\right) V^{(l)}$.
 - 32: **Return** Generated prediction sequence.
-

6.3.3.2 Instruction-Tuning Based on LoRA

To facilitate parameter-efficient adaptation, we inject LoRA modules into the linear transformations of the decoder. Throughout the instruction-tuning process, the frozen pretrained weights, denoted as $\mathbf{W}_0 \in \mathbb{R}^{d_{out} \times d_{in}}$, are kept static. Optimization is exclusively applied to a pair of rank-decomposition matrices: the down-projection $\mathbf{P}_{down} \in \mathbb{R}^{r \times d_{in}}$ and the up-projection $\mathbf{P}_{up} \in \mathbb{R}^{d_{out} \times r}$. Here, the rank hyperparameter r satisfies the condition $r \ll \min(d_{in}, d_{out})$. Consequently, the effective forward pass weight is reconstructed as:

$$\mathbf{W}_{adapted} = \mathbf{W}_0 + \mathbf{P}_{up}\mathbf{P}_{down}. \quad (6.16)$$

Since only the low-rank components are trainable, the number of parameters that require updating is greatly reduced, which in turn speeds up training and improves computational efficiency. After tuning, GFF can generate textual interpretations for input images and effectively handle geospatial objects and scene-level information in Japanese remote sensing imagery. Algorithm 2 provides an overview of the complete GFF workflow, from preprocessing to final output generation.

6.4 Experiments

6.4.1 Experimental Setup

In our implementation, the visual encoder is instantiated using the ViT-L backbone, coupled with Vicuna-1.5 as the foundational LLM for reasoning and text generation. Parameter optimization is conducted via the AdamW algorithm, configured with an initial learning rate of 2×10^{-5} , a cosine decay schedule, and a warm-up fraction of 0.03. The training process employs a per-device batch size of 16 with a gradient accumulation step of 1. All computational routines are executed on a cluster of four NVIDIA A100 GPUs, utilizing BF16 and TF32 mixed-precision acceleration to enhance efficiency. Besides, a unified cross-entropy function is applied to supervise both tasks.

6.4.2 Evaluation Metrics

To evaluate the geospatial object classification task, we employ both macro and weighted averages of Precision, Recall, and F1-score. For a given class k , these indicators are defined as:

$$\text{Precision}_k = \frac{N_{\text{TP}}^{(k)}}{N_{\text{TP}}^{(k)} + N_{\text{FP}}^{(k)}}, \quad \text{Recall}_k = \frac{N_{\text{TP}}^{(k)}}{N_{\text{TP}}^{(k)} + N_{\text{FN}}^{(k)}}, \quad (6.17)$$

$$\text{F1}_k = 2 \times \frac{\text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}, \quad (6.18)$$

where $N_{\text{TP}}^{(k)}$, $N_{\text{FP}}^{(k)}$, and $N_{\text{FN}}^{(k)}$ denote the counts of true positives, false positives, and false negatives for class k , respectively. The weighted versions of the metrics are obtained by averaging over all classes according to their sample frequencies, whereas the macro average is computed as the unweighted mean across classes so that each class contributes equally, regardless of its size.

To evaluate the quality of the generated scene descriptions, we employ a suite of standard text-generation metrics: BLEU [64], CIDEr [65], METEOR [66], and ROUGE [67]. Key characteristics of the primary metric are outlined as follows:

- **BLEU**: This metric evaluates the overlap of n-grams between the candidate and reference captions, introducing a penalty term for excessively concise outputs. Mathematically, it is formulated as the product of weighted precision scores adjusted by a brevity factor:

$$\text{BLEU} = \rho_{\text{BP}} \cdot \prod_{n=1}^N (P_n)^{w_n}, \quad (6.19)$$

where P_n represents the modified precision for n-grams, and ρ_{BP} serves as the brevity penalty coefficient.

- **CIDEr** evaluates how well the generated sentence agrees with a set of reference descriptions by comparing their n-gram representations. The metric computes a TF-IDF-weighted similarity, typically using cosine similarity, which enables CIDEr to capture semantic alignment beyond exact word matching.
- **METEOR** assesses the match between a candidate sentence and its references through a flexible alignment procedure that incorporates stemming, synonym expansion, and paraphrase matching. This design allows METEOR to provide a more detailed measure of

semantic consistency.

- **ROUGE-1**, **ROUGE-2**, and **ROUGE-L** quantify the n-gram overlap and structural similarity between generated and reference texts. ROUGE-1 and ROUGE-2 compute unigram and bigram matches, respectively, while ROUGE-L considers the longest common subsequence, offering an indication of overall fluency and sequence-level coherence.

6.4.3 Quantitative Results

In this subsection, we provide a detailed quantitative analysis of the proposed framework on the GeoJapan dataset. We report results for both geospatial object classification and scene description, covering all evaluation metrics introduced earlier. These experiments allow us to examine how well GFF handles fine-grained object categories as well as complex scene structures in realistic Japanese environments. In addition to overall performance comparisons with state-of-the-art LMMs, we also analyze class-level behavior and text-generation quality to offer a comprehensive view of the model’s strengths and limitations.

6.4.3.1 Geospatial Object Classification

Table 6.1 presents a quantitative evaluation of geospatial object classification performance on the GeoJapan benchmark. We compare the proposed GFF against a comprehensive suite of baselines, spanning general-purpose architectures (e.g., LLaVA-1.5 [9], Qwen-VL [10]) and specialized remote sensing models (e.g., GeoChat [22], VHM [28]). As evidenced by the results, GFF establishes a new state-of-the-art, obtaining the highest scores across all evaluation metrics. Specifically, regarding weighted averages, our framework achieves Precision, Recall, and F1-scores of 0.932, 0.929, and 0.929, respectively. This represents a substantial performance margin over the nearest competitor, VHM, which trails with corresponding scores of 0.705, 0.645, and 0.626. In addition, the high consistency between the weighted and macro-averaged metrics shows that GFF is robust to class imbalance. With a macro F1-score of 0.930, our method effectively handles data distribution issues where prior approaches often fail. This consistency underscores the effectiveness of the proposed multimodal fusion design in capturing diverse geospatial semantics within complex environments.

Table 6.1: Geospatial object classification performance on the GeoJapan dataset.

Method	Params	Published in	Weighted Average			Macro Average		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
General Large Multimodal Models								
LLaVA-1.5 [9]	7B	<i>NeurIPS 2023</i>	0.672	0.601	0.575	0.539	0.456	0.449
LLaVA-NeXT [55]	13B	<i>arXiv 2024</i>	0.656	0.487	0.489	0.665	0.547	0.536
Qwen-VL [10]	7B	<i>arXiv 2023</i>	0.558	0.479	0.472	0.569	0.442	0.445
MiniGPTv2 [68]	7B	<i>arXiv 2023</i>	0.589	0.395	0.380	0.477	0.335	0.320
MiniGPT-4-Vicuna [12]	7B	<i>ICLR 2024</i>	0.560	0.394	0.376	0.398	0.322	0.286
MiniGPT-4-LLaMA [12]	7B	<i>ICLR 2024</i>	0.568	0.482	0.481	0.503	0.398	0.410
Remote Sensing Large Multimodal Models								
GeoChat [22]	7B	<i>CVPR 2024</i>	0.651	0.553	0.552	0.614	0.507	0.501
GeoGround [23]	7B	<i>arXiv 2024</i>	0.401	0.353	0.271	0.393	0.302	0.262
RS-LLaVA [26]	7B	<i>RS 2025</i>	0.662	0.591	0.579	0.593	0.469	0.470
VHM [28]	7B	<i>AAAI 2025</i>	0.705	0.645	0.626	0.668	0.624	0.603
Falcon [69]	0.7B	<i>arXiv 2025</i>	0.673	0.583	0.600	0.646	0.518	0.544
GFF	7B	–	0.932	0.929	0.929	0.935	0.930	0.930

Table 6.2 details the class-wise breakdown of Precision, Recall, and F1-score on the GeoJapan dataset. The results show that GFF performs consistently well across most geospatial categories. Notably, visually distinct classes, specifically “bridge,” “football court,” “golf court,” and “solar panel” achieve perfect scores across all metrics, confirming the model’s ability to reliably identify unique features. Similarly, common land-cover types such as “dense residential,” “forest,” and “farmland” yield high accuracy, showing that the model captures natural and semi-urban patterns well. However, specific challenges remain for certain categories. Regarding the “industrial area” class, we observe a high precision of 0.964 but a lower recall of 0.844. This recall drop is likely caused by background clutter or confusion with visually similar classes like “dense residential” and “parking lot.” A similar trend appears in the “park” category, where recall drops to 0.833 despite perfect precision. We attribute this to the visual overlap between parks and other green areas like “forest” and “farmland.” Finally, road-related categories (“crossroads,” “roundabout,” and “straight road”) exhibit comparatively lower performance, with precision ranging from 0.800 to 0.920. The high visual similarity among these road types creates ambiguity, which leads to occasional misclassification.

To conclude, GFF demonstrates superior performance on the GeoJapan dataset compared to contemporary LMMs. It shows exceptional precision in identifying urban infrastructure and natural scenes, although challenges persist in recognizing objects with ambiguous or less distinct visual features.

6.4.3.2 Scene Description

Table 6.3 presents the results of the scene description task on the GeoJapan dataset. We compare GFF with recent general-purpose LMMs (e.g., Qwen-VL, LLaVA-1.5, MiniGPT variants) as well as leading remote sensing LMMs such as GeoChat, GeoGround, RS-LLaVA, VHM, and Falcon. Across all metrics, GFF delivers the strongest performance, achieving the highest scores among all evaluated methods (e.g., BLEU-1: 46.57, CIDEr: 25.31, METEOR: 41.41, ROUGE-L: 37.74). These results indicate that GFF generates more accurate and diverse scene descriptions on GeoJapan compared with existing LMMs. As a framework designed specifically for the characteristics of the newly constructed GeoJapan dataset, GFF can capture fine-grained visual and semantic cues in remote sensing images, highlighting its effectiveness for domain-specific scene

Table 6.2: Detailed classification performance across geospatial categories on GeoJapan, sorted by F1-Score.

Class Name	Samples	Precision	Recall	F1-Score
Bridge	19	1.000	1.000	1.000
Football court	19	1.000	1.000	1.000
Golf court	20	1.000	1.000	1.000
Solar panel	18	1.000	1.000	1.000
Dense residential	72	0.973	0.986	0.979
Forest	44	0.977	0.977	0.977
Helipad	19	0.950	1.000	0.974
River	18	1.000	0.889	0.941
Railway	15	0.882	1.000	0.938
Baseball diamond	20	0.870	1.000	0.930
Tennis court	17	0.850	1.000	0.919
Farmland	30	0.903	0.933	0.918
Park	18	1.000	0.833	0.909
Vacant lot	12	1.000	0.833	0.909
Industrial area	32	0.964	0.844	0.900
Straight road	59	0.855	0.898	0.876
Crossroads	35	0.838	0.886	0.861
Parking lot	44	0.897	0.796	0.843
Roundabout	10	0.800	0.800	0.800

understanding.

The GFF achieves the highest scores on BLEU-1 through BLEU-4, indicating that its generated descriptions are not only more informative but also exhibit stronger semantic consistency with the reference texts. This reflects the model’s enhanced capability to capture image details while preserving coherence in the narrative. Furthermore, the superior CIDEr and METEOR scores of GFF confirm that it can produce content-rich descriptions with high semantic faithfulness that closely correspond to the actual scene. In contrast, general LMMs such as Qwen-VL, LLaVA-1.5, and the MiniGPT series show noticeably weaker performance on this task. Although Qwen-VL achieves relatively strong results among the general models and outperforms the other general LMMs across all metrics, a clear performance gap remains compared with GFF, particularly in its ability to describe complex remote sensing scenes with sufficient detail and semantic coherence. Conversely, general-purpose architectures exemplified by LLaVA-1.5 and the MiniGPT family exhibit a marked performance deficit. Their suboptimal scores across

Table 6.3: Scene description performance on the GeoJapan dataset.

Model	Params	Published in	BLEU 1	BLEU 2	BLEU 3	BLEU 4	CIDEr	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
General Large Multimodal Models											
LLaVA-1.5 [9]	7B	<i>NeurIPS 2023</i>	36.61	21.82	15.02	10.58	7.37	34.43	40.08	16.52	31.05
LLaVA-NeXT [55]	13B	<i>arXiv 2024</i>	23.91	13.72	8.78	5.54	7.90	34.68	34.50	12.37	25.64
Qwen-VL [10]	7B	<i>arXiv 2023</i>	35.14	21.92	15.14	10.68	8.32	33.76	39.35	16.99	30.89
MiniGPTv2 [68]	7B	<i>arXiv 2023</i>	28.21	17.10	11.54	7.49	4.16	33.34	37.63	15.82	28.45
MiniGPT-4-Vicuna [12]	7B	<i>ICLR 2024</i>	20.23	10.42	5.83	3.06	0.38	25.68	24.46	6.81	17.85
MiniGPT-4-LLaMA [12]	7B	<i>ICLR 2024</i>	31.18	17.11	10.73	6.60	3.28	34.57	37.11	12.71	28.48
Remote Sensing Large Multimodal Models											
GeoChat [22]	7B	<i>CVPR 2024</i>	21.29	11.67	7.16	4.02	1.03	28.96	31.81	10.51	23.67
GeoGround [23]	7B	<i>arXiv 2024</i>	28.25	15.98	10.56	6.56	1.64	25.40	32.01	11.86	25.59
RS-LLaVA [26]	7B	<i>RS 2025</i>	28.65	13.47	5.19	1.45	2.15	32.77	35.33	9.63	24.98
VHM [28]	7B	<i>AAAI 2025</i>	33.82	19.53	12.99	8.16	4.32	29.88	38.86	14.38	30.76
Falcon [69]	0.7B	<i>arXiv 2025</i>	17.31	13.55	11.77	5.75	1.25	12.16	24.11	5.54	18.25
GFF	7B	–	46.57	30.16	21.15	15.07	25.31	41.41	48.68	22.62	37.74

all evaluation metrics suggest a lack of generalization capability when processing the unique visual semantics of Japanese remote sensing imagery. Regarding domain-specific baselines (e.g., GeoChat, RS-LLaVA, Falcon), while they offer marginal gains over general LMMs, they still lag significantly behind the proposed GFF. Even VHM, the strongest competitor among prior arts, fails to resolve region-specific nuances adequately. This limitation prevents these models from generating fine-grained descriptions that capture the distinctive structural features of complex scenes, further validating the necessity of our proposed fusion strategy.

Overall, the strong performance of GFF on the scene description task validates the effectiveness of the proposed architecture and underscores the utility of the GeoJapan dataset for practical remote sensing scene understanding.

6.4.4 Qualitative Results

In this subsection, we present the qualitative results for both the geospatial object classification and scene description tasks. Through representative examples, we illustrate how GFF interprets complex remote sensing scenes, identifies key geospatial objects, and generates coherent textual descriptions. These visual and textual examples complement the quantitative findings reported earlier and provide intuitive insights into the model’s behavior, highlighting its strengths as well as typical failure cases encountered in challenging Japanese environments.

Figure 6.7 presents qualitative classification results across four representative categories: “forest,” “roundabout,” “parking lot,” and “farmland.” The model yields accurate predictions for the first three instances, demonstrating full alignment with the ground truth. Specifically, GFF effectively captures distinct semantic cues, such as the dense vegetation of forests, the circular topology of roundabouts, and the regularized spatial patterns of parking lots. This robust performance stems from our in-context learning strategy, which integrates comprehensive semantic descriptions for all geospatial classes. By explicitly encoding these category-specific attributes, the model enhances its discriminative capability against visually confusing baselines. However, the final example reveals a failure case where “farmland” is erroneously classified as “vacant lot.” This error is primarily attributed to the high visual ambiguity between these two categories, particularly regarding parcel geometry and texture. Such instances highlight the inherent complexity of Japanese remote sensing imagery and the challenges associated with fine-grained classification

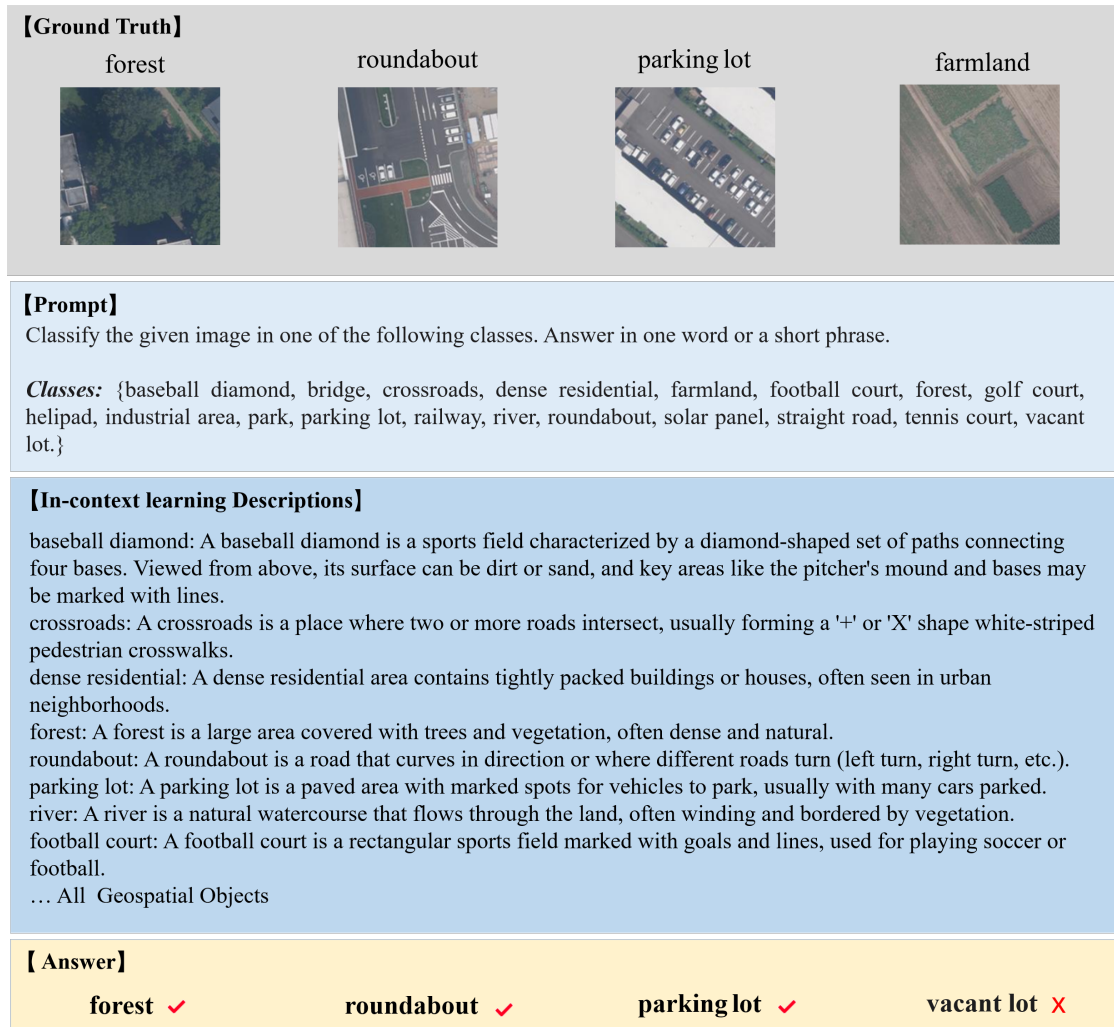


Figure 6.7: Examples of GFF outputs for geospatial object classification. Three correctly recognized cases and one misclassification are shown.

on the GeoJapan benchmark.

Figures 6.8 and 6.9 illustrate two representative examples of scene descriptions generated by different LMMs. In the first case, the reference annotation describes the scene as a dense residential area featuring crosswalks, a large building, and a parking lot on the right side. Among the evaluated models, LLaVA-NeXT and Qwen-VL produce descriptions that correctly mention major components such as residential buildings, roads, and surrounding vegetation, showing their ability to identify common urban elements. By comparison, the MiniGPT series generates more general descriptions and typically omits fine-grained scene details. The proposed GFF provides the most accurate and detailed caption. It faithfully identifies the scene as a “dense residential area with crosswalks” and explicitly names key structures including the “building,” “parking lot,”

and “crossroads,” closely aligning with the reference description. This demonstrates that GFF can robustly recognize essential elements in Japanese remote sensing imagery and effectively capture the nuanced characteristics of complex urban environments. Overall, these qualitative examples confirm that GFF delivers more precise and informative scene descriptions than existing LMMs, providing a clearer representation of salient features in Japanese remote sensing scenes. In the example shown in Figure 6.9, the description generated by GFF provides an accurate account of the scene, identifying key elements such as the “industrial building,” “parking lot,” and “river,” while also describing their relative positions (e.g., “to the left” and “to the right” ’). The model not only recognizes major man-made structures but also incorporates natural features and nearby infrastructure, demonstrating its ability to handle scenes containing multiple object types and spatial relationships. The caption produced by GFF is succinct yet descriptive, presenting the information in a clear and coherent manner. Compared with the other models, GFF offers a more detailed and faithful depiction of the scene, reflecting a stronger grasp of geospatial context. These observations further highlight the practical value of GFF for real-world remote sensing tasks, particularly those involving complex environments with numerous objects and inter-object interactions.

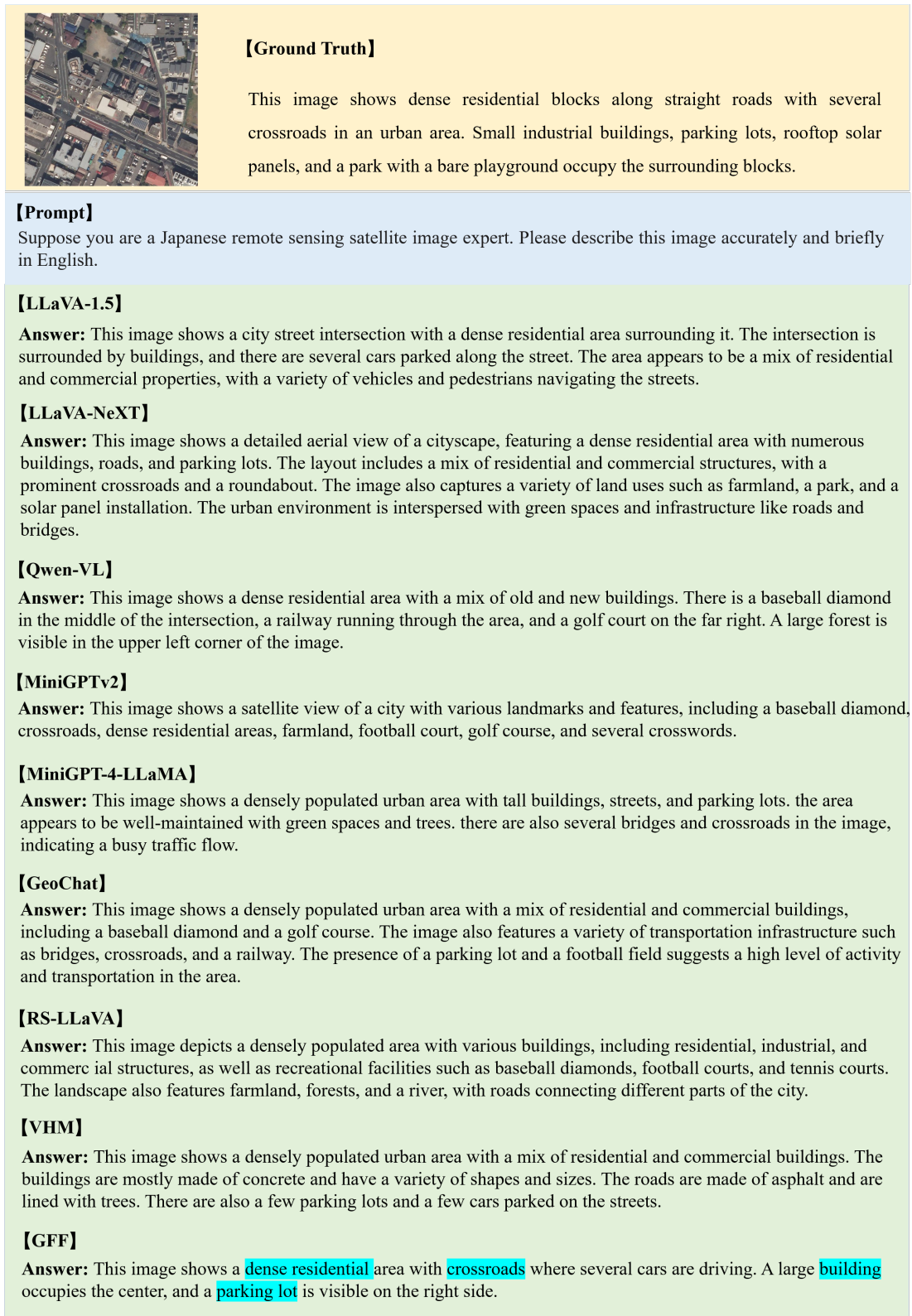
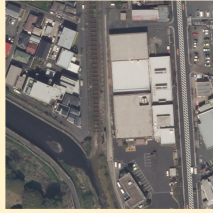


Figure 6.8: Qualitative results of GFF on the scene description task. Ground truth descriptions and outputs from different models are shown for comparison. Text highlighted in blue indicates portions of the GFF-generated descriptions that match the reference annotations and capture key scene information.



【Ground Truth】

This image shows dense residential blocks in the upper left, a river curving along the left edge, and an industrial area with large warehouses and a parking lot on the right. A straight road with marked lanes borders the far right side.

【Prompt】

Suppose you are a Japanese remote sensing satellite image expert. Please describe this image accurately and briefly in English.

【LLaVA-1.5】

Answer: This image shows a cityscape with a mix of residential and industrial areas. The residential area is characterized by a dense residential neighborhood with many houses, while the industrial area features a large white building surrounded by other buildings. The city is also home to a river that runs through the middle of the area, and a parking lot can be seen near the residential area.

【LLaVA-NeXT】

Answer: This image shows a satellite view of an urban area with a mix of residential and industrial structures. The layout includes a dense residential area with numerous buildings, a large industrial area with multiple structures, and a river running through the middle of the city. There are also parking lots, a road network, and a bridge crossing the river. The image provides a comprehensive view of the city's layout and infrastructure.

【Qwen-VL】

Answer: This image shows an aerial view of an industrial area with a large warehouse and multiple parking lots. There is a railway running through the center of the area and a river nearby. The image also shows a dense residential area with multiple buildings and a baseball diamond.

【MiniGPTv2】

Answer: This image shows a large, urban area with various land uses, including residential, industrial, and agricultural areas. there are several parks and green spaces throughout the area, as well as a river running through the center. the image also features several roads, including a straight road and a roundabout, as well as a helipad and a baseball diamond.

【MiniGPT-4-LLaMA】

Answer: This image shows a view of a city from a satellite perspective. the image depicts a dense residential area with many buildings, roads, and parking lots. there are also several bodies of water visible in the image, including a river and a lake. the image also shows a golf course and a park.

【GeoChat】

Answer: This image shows a satellite view of a city with various landmarks such as a bridge, crossroads, and a football field. There are also several buildings, a golf course, and a parking lot visible in the image. The city is located near a river, and there are two buildings close to each other at the top left corner of the image.

【RS-LLaVA】

Answer: This image depicts a diverse landscape with various structures and natural elements. It features a river, dense residential areas, industrial zones, sports facilities like baseball diamonds, football courts, and tennis courts, as well as recreational areas such as parks and forests. Additionally, there are bridges, crossroads, and parking lots, along with a railway and a helipad.

【VHM】

Answer: This image shows dense residential blocks in the upper left, a river curving along the left edge, a railway running vertically in the center, and an industrial area with large warehouses and a parking lot on the right. A straight road with marked lanes borders the far right side.

【GFF】

Answer: This image shows a large industrial building with a parking lot. A river flows along the left side of the scene, surrounded by dense residential areas. A vertical straight road runs through the right side.

Figure 6.9: Qualitative results of GFF on the scene description task. Ground truth descriptions and outputs from different models are shown for comparison. Text highlighted in blue indicates portions of the GFF-generated descriptions that match the reference annotations and capture key scene information.

Table 6.4: Impact assessment of in-context learning on GFF classification metrics using the GeoJapan dataset.

ICL Setting	Weighted Average			Macro Average		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Baseline (No ICL)	0.816	0.814	0.805	0.771	0.769	0.758
GFF (With ICL)	0.932	0.929	0.929	0.935	0.930	0.930

Table 6.5: Impact assessment of in-context learning on GFF scene description metrics using the GeoJapan dataset.

ICL Setting	BLEU 1	BLEU 2	BLEU 3	BLEU 4	CIDEr	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Baseline (No ICL)	46.09	29.39	20.28	13.92	18.94	41.69	46.89	20.13	35.39
GFF (With ICL)	46.57	30.16	21.15	15.07	25.31	41.41	48.68	22.62	37.74

6.4.5 Ablation Studies

We performed a set of ablation studies to examine how each component and design choice affects the overall performance of GFF. The findings of these analyses are presented in this subsection.

6.4.5.1 In-Context Learning

To evaluate the specific contribution of the in-context learning module, Table 6.4 details the classification performance of GFF. The inclusion of in-context examples results in a significant performance gain, driving the weighted-average Precision from 0.816 up to 0.932. Consistent improvements are also observed across the Recall and F1-score metrics. These results confirm that providing contextual cues is essential for strengthening both the accuracy and stability of the model.

Table 6.5 reports the scene description results of GFF under different in-context learning settings. With in-context examples incorporated, BLEU-1 and METEOR show a slight improvement (from 46.09 to 46.57 and from 41.41 to 41.69), while the other metrics, including BLEU-2, BLEU-3, BLEU-4, CIDEr, ROUGE-1, ROUGE-2, and ROUGE-L, exhibit noticeably larger gains. These improvements suggest that providing additional contextual cues helps the model produce more accurate and detailed scene descriptions.

Table 6.6: Temporal evolution of classification performance on the GeoJapan benchmark throughout the training process.

Training Stage	Weighted Average			Macro Average		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Epoch 1	0.824	0.795	0.792	0.817	0.748	0.755
Epoch 5	0.916	0.910	0.910	0.916	0.912	0.910
Epoch 10	0.926	0.913	0.915	0.925	0.921	0.921
Epoch 15	0.931	0.921	0.928	0.931	0.929	0.930
Epoch 20	0.932	0.929	0.929	0.935	0.930	0.930

Synthesizing these results, it is evident that the proposed framework derives substantial advantages from the integration of in-context information. The application of in-context learning equips the model with a robust understanding of scene semantics. Consequently, this enhancement leads to superior accuracy in geospatial object classification and facilitates the generation of more detailed, informative scene descriptions.

6.4.5.2 Training Epochs

Table 6.6 illustrates the evolution of classification metrics for GFF across a span of 20 training epochs. The data reveals a consistent upward trend in model performance. Specifically, the weighted average F1-score rises from an initial value of 0.792 at Epoch 1 to a final peak of 0.929 at Epoch 20, while Precision sees a comparable gain from 0.824 to 0.932. This positive trajectory is similarly reflected in the macro-averaged scores, confirming that prolonged training enhances the model’s capability to discriminate among diverse object categories. Notably, performance metrics begin to stabilize around Epoch 15, suggesting that model convergence is largely achieved by this phase.

A comparable trend is observed for the scene description task, as shown in Figure 6.10. GFF exhibits clear improvements across all standard text generation metrics. For example, the CIDEr score increases from 3.73 to 25.31, reflecting a substantial boost in the relevance and fidelity of the generated captions. The BLEU scores also rise steadily, with BLEU-4 improving from 9.22 to 15.07. These results indicate that continued training enables the model to more effectively capture complex spatial structures and convey detailed scene information.

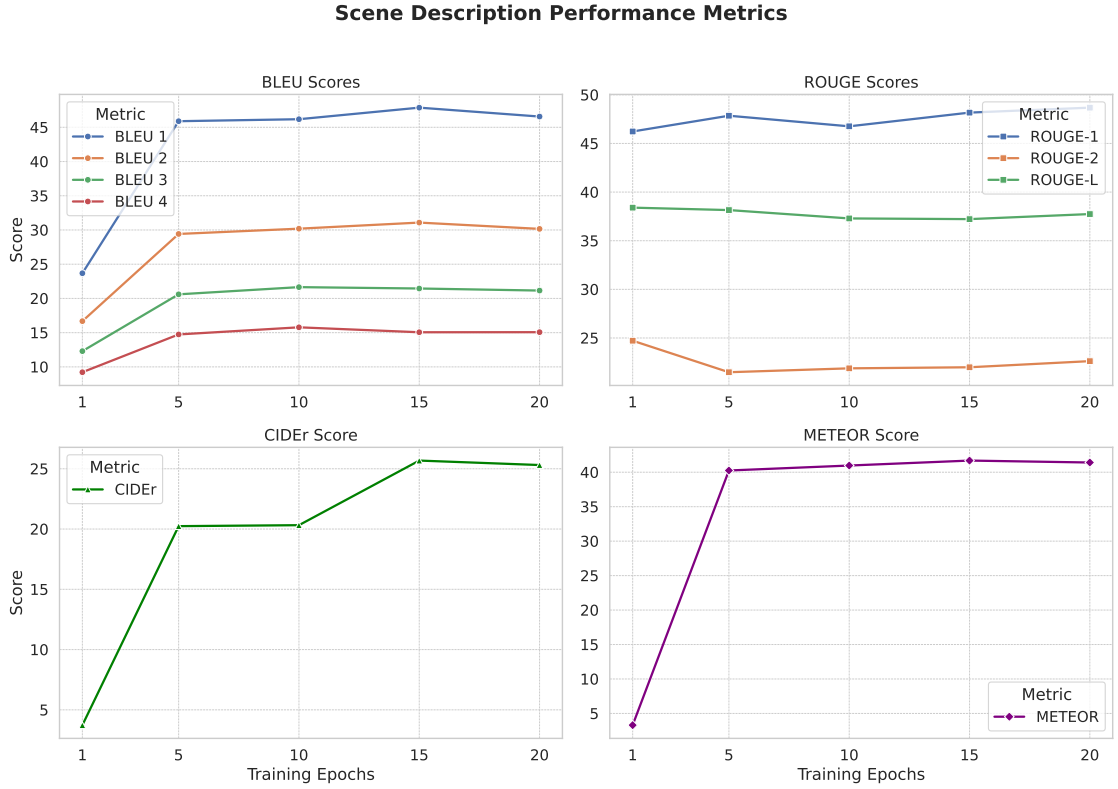


Figure 6.10: GFF scene description results on the GeoJapan dataset over training epochs.

Table 6.7: Analysis of resource utilization and latency for GFF across distinct tasks.

Target Task	VRAM Usage (GB)	Processing Latency (s)
Geospatial Object Classification	3.75	0.128
Scene Description	3.75	0.887

6.4.5.3 Resource Efficiency Analysis

To evaluate the operational feasibility of GFF, we analyzed its resource overhead, specifically focusing on memory footprint and processing speed (refer to Table 6.7). Regarding memory, the system maintains a consistent usage of approximately 3.75 GB across both functional modes. In terms of speed, the model achieves a per-sample latency of 0.128 s for the classification track, while the generation of scene descriptions requires 0.887 s. These figures indicate that the framework operates within a reasonable computational budget, thereby validating its potential for practical deployment in remote sensing workflows.

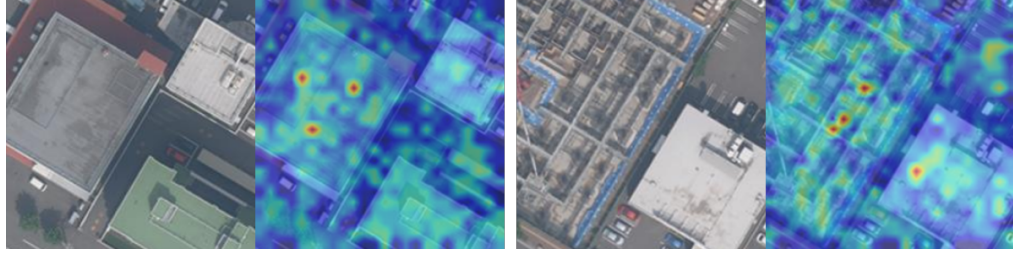


Figure 6.11: Visualization of attention maps, with heatmaps showing regions emphasized by GFF.

6.4.5.4 Attention Map

Figure 6.11 visualizes the attention maps for two samples from the “industrial area” category. In both images, GFF predominantly attends to factory-related structures, which constitute the most informative features for distinguishing this category from others. In the second example, the highlighted regions align closely with the key industrial facilities, showing that the model can focus on task-relevant areas for both classification and scene description. These visualizations indicate that GFF is capable of automatically identifying discriminative regions, thereby improving the interpretability of the framework.

6.5 Discussion

6.5.1 Methodological Analysis

Methodologically, the success of the proposed approach is driven by a unified framework that couples robust multimodal fusion with multitask learning objectives. This strategy is indispensable for modeling the non-trivial dependencies between visual content and textual semantics. The extensive experimental results demonstrate that GFF can accurately recognize a wide range of geospatial objects while generating detailed and reliable scene descriptions. In particular, the introduction of in-context learning allows the model to adaptively exploit prior knowledge about Japanese urban characteristics, leading to substantial gains across multiple evaluation metrics. These findings indicate that the architectural and algorithmic choices in GFF effectively promote a holistic understanding of complex scenes, which in turn yields superior performance in both recognition and description tasks.

6.5.2 Key Challenges and Proposed Solutions

The design of GFF required us to address several important technical issues. A primary challenge was processing high-resolution images together with multiple text tasks. This demanded a fusion method that could combine different data types without causing conflicts between them. To achieve this, we developed an adaptive fusion module that integrates visual and text information into a shared representation space. Another key concern was computational efficiency. To make the model practical for real-world use, we applied LoRA. This technique significantly reduces the number of trainable parameters, lowering the computational load while maintaining high accuracy. Furthermore, the large variation in terrain and scene complexity across different regions made it difficult to extract features stably. We solved this problem by introducing in-context learning, which improves the model’s ability to understand region-specific details in remote sensing data. By combining these strategies, GFF effectively overcomes these technical barriers, enabling accurate and efficient recognition across diverse scenarios.

6.5.3 Implications

The GeoJapan dataset and the proposed GFF provide meaningful contributions to LMM-based remote sensing research. First, the dataset is constructed specifically to support multitask recognition and regional scene understanding, offering a comprehensive benchmark for developing models capable of fine-grained interpretation across different geographic settings. Second, the results obtained with GFF illustrate how multitask learning can be effectively leveraged to improve recognition performance in complex remote sensing scenes. The model’s robustness and generalization capability indicate strong potential for practical deployment in real-world applications such as GIS updates, urban planning, environmental surveillance, and disaster response.

6.5.4 Limitations

Despite the robust general capabilities of GFF, challenges persist in differentiating fine-grained categories that share high visual resemblance, as evidenced by the confusion between “vacant lot” and “farmland.” This observation highlights the need to refine the model’s discriminative power for classes exhibiting overlapping visual and textual features. Furthermore, while the framework

excels in analyzing densely structured urban environments, we observe a slight performance degradation in settings characterized by weaker structural cues or limited heterogeneity. These findings underscore the importance of advancing scene-level reasoning mechanisms to resolve ambiguity in complex scenarios more reliably.

6.5.5 Future Directions

Future avenues of research will prioritize two key objectives. The first involves evolving GFF into a system capable of real-time, multi-turn conversational reasoning. Although the existing architecture excels at analyzing static imagery, adapting it for continuous dialogue requires robust mechanisms to preserve contextual history across interactions. By integrating advanced dialogue management and continuous inference protocols, we aim to enable dynamic scene understanding, thereby producing responses that are more coherent and context-aware.

The second objective focuses on scaling up the GeoJapan benchmark beyond the current scope of Kawasaki and Sapporo. We intend to include a broader spectrum of Japanese environments, such as diverse seismic regions and rural-to-urban transition zones. This augmentation is designed to ensure that the dataset reflects the heterogeneity of real-world landscapes. Coupled with rigorous evaluation protocols, specifically cross-region validation, this initiative serves to bolster the model’s generalization capabilities, ensuring robust interpretation across varying geographical distributions. Additionally, we will introduce human-in-the-loop evaluations to qualitatively measure the geographical accuracy and fluency of the generated text. Finally, we aim to explore the synergy between classification and description tasks, conducting a granular analysis to determine how performance gains in one domain translate to the other.

6.6 Conclusion

In this study, we presented GFF, a multimodal large-model framework designed to handle multiple remote sensing tasks within a unified architecture. Alongside the model, we constructed GeoJapan, a dedicated benchmark that reflects the diversity of Japanese geospatial environments and provides high-quality annotations for both object-level classification and scene-level description. Through comprehensive experiments, GFF demonstrated clear advantages over a wide

range of recently developed LMMs. GFF’s performance improvements are obtained from the combination of an effective multimodal fusion design and the complementary nature of its multitask learning strategy. Incorporating in-context learning proved particularly beneficial, as the additional contextual information helped the model better capture regional characteristics and improved the quality of the generated outputs across several evaluation metrics. These results highlight the importance of coupling domain-tailored datasets with frameworks that explicitly model cross-modal interactions, enabling more reliable interpretation of complex and heterogeneous remote sensing scenes. In future work, we plan to extend the reasoning capability of GFF to multiturn, real-time settings and expand the scope of GeoJapan to cover a broader set of geographic regions. These efforts aim to develop a more adaptable and application-ready remote sensing LMM for practical geospatial analysis scenarios.

Chapter 7

Conclusions

As conclusions of this thesis, this chapter provides an overview of the propositions and findings presented throughout the study and suggests promising directions for future research.

7.1 Overview of the Proposition in This Thesis

This section presents a comprehensive overview of the propositions proposed in this thesis, building upon the background, motivations, and previous chapters.

The core proposition of this thesis is to explore how LMMs can be effectively adapted and extended for visual recognition tasks in real-world scenarios. The research particularly focuses on enhancing domain generalization, interpretability, and data efficiency under zero-shot and few-shot conditions. To achieve these objectives, this thesis introduces a progressive series of multimodal visual understanding methods that progressively advance from structural feature alignment to adaptive reasoning and large-scale multimodal fusion.

The studies presented in this thesis are conducted in four main stages. The first stage investigates feature-level alignment for zero-shot recognition, where a midlevel feature correspondence space is constructed to improve cross-domain generalization without labeled supervision. The second stage explores few-shot adaptation through in-context learning, utilizing multimodal contextual cues to achieve efficient knowledge transfer with minimal annotated data. The third stage introduces cross-domain multi-step reasoning to enhance semantic interpretability and reasoning consistency of LMMs for fine-grained recognition. Finally, the fourth stage applies large-scale multimodal fusion for remote sensing recognition, integrating visual imagery with geographic

and textual information to achieve semantic-level understanding across complex geographic environments.

Each of these stages contributes toward enhancing the adaptability and reasoning ability of LMMs in visual recognition. Through extensive experiments performed on traffic sign datasets and remote sensing imagery, the proposed methods demonstrate superior performance in zero-shot, few-shot, and cross-domain settings. Collectively, these studies provide a unified research trajectory that bridges representation learning, semantic reasoning, and multimodal application.

The contributions of this thesis can be summarized as follows:

- This thesis develops several multimodal learning methods for visual recognition that extend LMMs to real-world data-scarce and cross-domain scenarios. These methods improve feature-level alignment, facilitate data-efficient learning, and enhance multimodal reasoning and interpretability.
- The proposed models are validated in both traffic sign recognition and remote sensing applications, demonstrating their effectiveness in achieving robust, generalizable, and context-aware visual understanding. The findings offer valuable theoretical and empirical insights into adapting large multimodal models for domain-specific visual recognition tasks.
- The thesis contributes to bridging the gap between large-scale pre-training and practical applications by presenting generalized multimodal learning strategies applicable across diverse visual domains.

Below, the basic structure of this thesis is briefly reviewed.

Chapter 1 provides an introduction to the research background, motivation, and objectives of the study. Chapter 2 reviews the related work on multimodal learning and visual recognition, clarifying the key challenges that this thesis aims to address. Chapter 3 presents the method of feature-level alignment for zero-shot cross-domain recognition. Chapter 4 describes the few-shot in-context adaptation strategy for data-efficient learning. Chapter 5 introduces the cross-domain multi-step reasoning framework for enhancing interpretability and generalization in LMMs. Chapter 6 constructs and evaluates a multimodal fusion framework for remote sensing recognition, demonstrating the applicability of the proposed approaches to large-scale real-world

data. Finally, Chapter 7 concludes the thesis with an overview of the findings and prospects for future research.

7.2 Future Directions

This section outlines the potential directions for future development of multimodal visual recognition research and the continued evolution of LMMs.

Future work can focus on refining and extending the proposed multimodal learning methods. On the methodological side, exploring more efficient alignment mechanisms and adaptive fine-tuning strategies will be important for reducing dependence on large-scale data and computation. Further studies can also develop advanced reasoning modules that combine multimodal chain-of-thought inference with explicit knowledge integration, thereby improving interpretability and consistency in complex reasoning tasks.

From the application perspective, the proposed approaches can be extended to additional visual domains beyond traffic and remote sensing, such as medical imaging, autonomous driving, and industrial inspection. Each of these domains presents distinct visual characteristics and data modalities, and investigating how LMMs adapt to such domain shifts will deepen the understanding of generalizable multimodal learning.

Another promising research direction lies in the integration of multiple adaptation paradigms, such as feature alignment, few-shot learning, and multimodal reasoning into unified training frameworks. Designing multitask or continual learning systems that dynamically balance these components can further improve data efficiency and model stability. Moreover, the interpretability and trustworthiness of LMMs should be examined in safety-critical applications, ensuring transparency and reliability of multimodal reasoning results.

As LMMs continue to evolve toward general artificial intelligence, it is essential to consider ethical and societal implications in their deployment. Future research should address issues of data privacy, fairness, and the responsible use of large-scale multimodal data. Developing transparent evaluation standards and ethical guidelines will play a vital role in ensuring that advancements in multimodal visual recognition contribute safely and equitably to scientific and societal progress.

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my supervisors, Prof. Miki Haseyama and Prof. Takahiro Ogawa. Throughout my five-year tenure at the Graduate School of Information Science and Technology, Hokkaido University, their invaluable guidance and unwavering support have played a crucial role in fostering my growth as a researcher. They are respected and charismatic mentors who have exerted a profound influence on my scientific career. I would also like to express my sincere appreciation to Prof. Yoshinori Dobashi for his insightful comments and constructive suggestions, which significantly refined my research.

I also extend my sincere appreciation to Associate Professor Ren Togo, Associate Professor Keisuke Maeda, Specially Appointed Assistant Professor Guang Li, and Assistant Professor Naoki Saito. I am truly grateful for their constant support and professional advice regarding both my research endeavours and academic life. In particular, I would like to pay high tribute to Associate Professor Ren Togo. His patience and strong sense of responsibility have deeply shaped my character and my attitude towards life and research. I also owe a special debt of gratitude to Dr. Guang Li for his generous assistance and insightful suggestions during my doctoral studies.

Finally, I would like to thank my family for their unsparing support. Their silent dedication and selfless assistance have been the bedrock of my life's journey. It is their love that enabled my twenty-one-year academic pursuit, taking me from a small town to the wider world. My thanks also go to all members of the Laboratory of Media Dynamics for their care and camaraderie, and to Hokkaido University for providing the excellent platform and resources that made this work possible.

References

- [1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [2] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 6105–6114.
- [3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186.
- [5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, 2019.
- [6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., “Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.
- [7] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [8] O. AI, “Gpt-5 system card,” *OpenAI Blog*, vol. 1, no. 8, 2019.

- [9] H. Liu, C. Li, Y. Li, and Y. J. Li, “Visual instruction tuning,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023, vol. 36, pp. 34892–34916.
- [10] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” *arXiv preprint arXiv:2308.12966*, 2023.
- [11] W. Dai, Z. Chen, Z. Jiang, J. Li, and J. Feng, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 25859–25870.
- [12] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024, pp. 1–17.
- [13] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, Q. Ye, and F. Wei, “Grounding multi-modal large language models to the world,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- [14] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, “Multimodal-gpt: A vision and language model for dialogue with humans,” *arXiv preprint arXiv:2305.04790*, 2023.
- [15] J. Li and Z. Wang, “Real-time traffic sign recognition based on efficient cnns in the wild,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 975–984, 2018.
- [16] Y. Yang, H. Luo, H. Xu, and F. Wu, “Towards real-time traffic sign detection and classification,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2022–2031, 2015.
- [17] A. Shustanov and P. Yakimov, “Cnn design for real-time traffic sign recognition,” *Procedia Engineering*, vol. 201, pp. 718–725, 2017.
- [18] H. Luo, Y. Yang, B. Tong, F. Wu, and B. Fan, “Traffic sign recognition using a multi-task convolutional neural network,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 4, pp. 1100–1111, 2017.

- [19] S. Zhou, C. Deng, Z. Piao, and B. Zhao, “Few-shot traffic sign recognition with clustering inductive bias and random neural network,” *Pattern Recognition*, vol. 100, pp. 107160, 2020.
- [20] Z. Huang, Y. Yu, J. Gu, and H. Liu, “An efficient method for traffic sign recognition based on extreme learning machine,” *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 920–933, 2016.
- [21] Y. Guo, W. Feng, F. Yin, and C.-L. Liu, “Signparser: An end-to-end framework for traffic sign understanding,” *International Journal of Computer Vision*, vol. 132, no. 3, pp. 805–821, 2024.
- [22] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, “Geochat: Grounded large vision-language model for remote sensing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27831–27840.
- [23] Y. Zhou, M. Lan, X. Li, Y. Ke, X. Jiang, L. Feng, and W. Zhang, “Geoground: A unified large vision-language model. for remote sensing visual grounding,” *arXiv preprint arXiv:2411.11904*, 2024.
- [24] J. Luo, Z. Pang, Y. Zhang, T. Wang, L. Wang, B. Dang, J. Lao, J. Wang, J. Chen, Y. Tan, et al., “Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding,” *arXiv preprint arXiv:2406.10100*, 2024.
- [25] M. Noman, N. Ahsan, M. Naseer, H. Cholakkal, R. M. Anwer, S. Khan, and F. S. Khan, “Cdchat: A large multimodal model for remote sensing change description,” *arXiv preprint arXiv:2409.16261*, 2024.
- [26] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, and F. Melgani, “Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery,” *Remote Sensing*, vol. 16, no. 9, 2024.
- [27] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, J. Li, and X. Mao, “Earthmarker: A visual prompting multimodal large language model for remote sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–19, 2025.

- [28] C. Pang, X. Weng, J. Wu, J. Li, Y. Liu, J. Sun, W. Li, S. Wang, L. Feng, G. Xia, and Others, “Vhm: Versatile and honest vision language model for remote sensing image analysis,” in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025, pp. 6381–6388.
- [29] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, “Vision transformer adapter for dense predictions,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [30] S. Suzuki et al., “Topological structural analysis of digitized binary images by border following,” *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.
- [31] D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3642–3649.
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2012, vol. 25, pp. 1–9.
- [33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [34] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition,” *Neural Networks*, vol. 32, pp. 323–332, 2012.
- [35] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.
- [36] F. Ren, J. Huang, R. Jiang, and R. Klette, “General traffic sign recognition by feature matching,” in *Proceedings of the International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2009, pp. 409–414.

- [37] M. Mathias, R. Timofte, R. Benenson, and L. Van Gool, “Traffic sign recognition—how far are we from the solution?,” in *Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.
- [38] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, “Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v,” *arXiv preprint arXiv:2310.11441*, pp. 1–23, 2023.
- [39] S. Yucong and G. Shuqing, “Traffic sign recognition based on hog feature extraction,” *Journal of Measurements in Engineering*, vol. 9, no. 3, pp. 142–155, 2021.
- [40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.
- [41] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, and Others, “Searching for mobilenetv3,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1314–1324.
- [42] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16000–16009.
- [43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, vol. 139, pp. 8748–8763.
- [44] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, “Traffic-sign detection and classification in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2110–2118.
- [45] Y. Gan, G. Li, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, “Zero-shot traffic sign recognition based on midlevel feature matching,” *Sensors*, vol. 23, no. 23, pp. 9607, 2023.

- [46] Y. Zheng and W. Jiang, “Evaluation of vision transformers for traffic sign classification,” *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–14, 2022.
- [47] Q. Luo and W. Zheng, “Pre-locator incorporating swin-transformer refined classifier for traffic sign recognition,” *Intelligent Automation & Soft Computing*, vol. 37, no. 2, 2023.
- [48] S. Mehta and M. Rastegari, “Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer,” *arXiv preprint arXiv:2110.02178*, 2021.
- [49] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., “Swin transformer v2: Scaling up capacity and resolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12009–12019.
- [50] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 10347–10357.
- [51] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, “Maple: Multi-modal prompt learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19113–19122.
- [52] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [53] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, “Clip-adapter: Better vision-language models with feature adapters,” *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.
- [54] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, “Eva-02: A visual representation for neon genesis,” *arXiv preprint arXiv:2303.11331*, 2023.
- [55] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, “Llava-next: Improved reasoning, ocr, and world knowledge,” *arXiv preprint arXiv:2401.02436*, 2024.
- [56] C. Fu, H. Lin, X. Wang, Y.-F. Zhang, Y. Shen, X. Liu, Y. Li, Z. Long, H. Gao, K. Li, et al., “Vita-1.5: Towards gpt-4o level real-time vision and speech interaction,” *arXiv preprint arXiv:2501.01957*, 2025.

- [57] “Openai. gpt-4v(ision) system card,” 2023.
- [58] “Openai. gpt-4o system card,” 2024.
- [59] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, “Sam 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024.
- [60] Y. Oren, N. Meister, N. Chatterji, F. Ladhak, and T. B. Hashimoto, “Proving test set contamination in black box language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024, pp. 1–19.
- [61] “Openai o3 and o4-mini system card,” 2025.
- [62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017, vol. 30.
- [63] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.
- [64] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.
- [65] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.
- [66] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshop (ACL-Workshop)*, 2005, pp. 65–72.
- [67] C. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshop (ACL-Workshop)*, 2004, pp. 74–81.

- [68] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, “Minigpt-v2: large language model as a unified interface for vision-language multi-task learning,” *arXiv preprint arXiv:2310.09478*, 2023.
- [69] Y. kelu, X. Nuo, Y. Rong, X. Yingying, G. Zhuoyan, K. Titinunt, R. yi, Z. Pu, W. Jin, W. Ning, and L. Chao, “Falcon: A remote sensing vision-language foundation model,” *arXiv preprint arXiv:2503.11070*, 2025.

Achievements of the Author

(A) Journal

- [A-1] Y. Gan, G. Li, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, “Cross-domain multi-step thinking: Zero-shot fine-grained traffic sign recognition in the wild,” *Knowledge-Based Systems* (IF: 7.6), vol. 327, 114172, 2025.
- [A-2] Y. Gan, G. Li, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, “GeoJapan fusion framework: A large multimodal model for regional remote sensing recognition,” *Remote Sensing* (IF: 4.1), vol. 17, no. 17, 3044, 2025.
- [A-3] Y. Gan, G. Li, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, “Zero-shot traffic sign recognition based on midlevel feature matching,” *Sensors* (IF: 3.5), vol. 23, no. 23, 9607, 2023.

(B) International Conference

- [B-1] Y. Gan, G. Li, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, “Cross-domain few-shot in-context learning for enhancing traffic sign recognition,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2564–2570, 2024.
- [B-2] Y. Gan, R. Togo, T. Ogawa, and M. Haseyama, “Transformer based multimodal scene recognition in soccer videos,” in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2022.
- [B-3] Y. Gan, R. Togo, T. Ogawa, and M. Haseyama, “Scene retrieval in soccer videos by spatial-temporal attention with video vision transformer,” in *Proceedings of the IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pp. 453–454, 2022.
- [B-4] Y. Gan, R. Togo, T. Ogawa, and M. Haseyama, “Multi-class similar scene retrieval in soccer videos: a scene confusion reduction method based on combination of long and short

frame sequences,” in *Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE)*, pp. 117–118, 2021.

(C) Domestic Conference

[C-1] Y. Gan, G. Li, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, “Fine-grained traffic sign recognition via cross-domain few-shot in-context learning,” *Meeting on Image Recognition and Understanding (MIRU)*, pp. 1–5, 2024.

[C-2] Y. Gan, G. Li, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, “A note on traffic sign recognition based on vision transformer adapter using visual feature matching,” *ITE Technical Report*, vol. 47, no. 6, pp. 208–211, 2023.

[C-3] Y. Gan, R. Togo, T. Ogawa, and M. Haseyama, “A note on transformer-based scene recognition in soccer videos using different length of clips,” *ITE Technical Report*, vol. 46, no. 6, pp. 167–170, 2022.

(D) Scholarship

[D-1] Next Generation AI Doctoral Fellowship, 2024.

[D-2] Hokkaido University Ambitious Doctoral Fellowship, 2023.

(E) Academic Visit

[E-1] Invited by the University of Cambridge to give a lecture on “Construction of Intelligent Agent Technology Based on Multimodal Large Language Model.”