

**Towards GeoAI Foundation Models: A Multimodal Learning  
Framework with Spatial Knowledge**

by

Meiliu Wu

A dissertation submitted in partial fulfillment of  
the requirements for the degree of

Doctor of Philosophy

(Geography)

at the

UNIVERSITY OF WISCONSIN–MADISON

2024

Date of final oral examination: April 8, 2024

The dissertation is approved by the following members of the Final Oral Committee:

Dr. Qunying Huang (Chair), Associate Professor, Geography

Dr. Song Gao, Associate Professor, Geography

Dr. Zhou Zhang, Assistant Professor, Biological Systems Engineering

Dr. Yong Jae Lee, Associate Professor, Computer Sciences

© Copyright by Meiliu Wu 2024  
All Rights Reserved

*I lovingly dedicate this dissertation to my dear mother, **Qiudi Ma**, whose unconditional support and love have been my constant source of strength, and in loving memory of my father, **Huode Wu**, whose spirit and guidance continue to inspire me.*

## ACKNOWLEDGMENTS

---

Embarking on this Ph.D. journey has been an extraordinary chapter in my life, filled with both immense challenges and invaluable learning. This journey would not have been possible without the support and encouragement of numerous individuals, to whom I owe a debt of gratitude.

At the forefront of this journey is my advisor/role model, Prof. Qunying Huang. Your guidance during my Ph.D. has been a beacon in my academic voyage, providing me with the freedom to delve into diverse research topics and experiences, and eventually cultivating me to become an independent scholar. You always give me dedicated advice and support in both my life and work, especially at my hardest times. Your brilliant inspiration, great passion, and heart-warming encouragement have keenly touched me and guided me through this challenging yet priceless journey. Working with you has profoundly enriched my life. Your mentorship has evolved into a lifelong friendship, for which I am truly grateful.

My heartfelt thanks are extended to the members of my committee, whose expertise and insights have been pivotal in shaping my research. Prof. Song Gao, your consistent encouragement and guidance, both in my research and career aspirations, have been precious and will forever be a cherished part of my life in the future. The moments when we skied, celebrated each of our milestones, and joined various academic and social events have brought tremendous joy to me over these years. Prof. Zhou Zhang, working with you on the NIFA project was a remarkable experience. Your support and mentorship have broadened my perspectives and played a crucial role in my academic achievements. Prof. Yong Jae Lee, your courses, expertise, and advice have been enlightening for me to establish interdisciplinary collaboration, and facilitate my exploration of cutting-edge research directions. My external advisor and co-author, Prof. David Wong, your mentorship in academic writing and critical thinking have

been fundamental to my scholarly growth, and your rigorous attitude and meticulous guidance in exploring varied ideas have markedly boosted my research development.

I also extend my genuine gratitude to the esteemed professors, dedicated colleagues, and supportive individuals at the Department of Geography, whose collective wisdom and good fellowship have been significantly important to my journey. Prof. Robert Roth, your expertise in cartography and thoughtful advice have not only enhanced my research but also inspired me to aim higher in my academic pursuits. Prof. A-Xing Zhu, our conversations during tea/coffee time and your helpful suggestions on my research and career development have been greatly appreciated. Prof. Jack Williams, our devoted department chair, working with you as the faculty/staff graduate representative on various departmental issues has been a catalyst for my personal and professional growth. Dr. William (Bill) Gartner, serving as your TA has been a sincerely gratifying experience, and your presence has been a true highlight in my graduate life here. Profs. Lisa Naughton, Sarah Moore, and Jenna Loyd, your insightful advice on navigating an academic career as a female faculty member has been my treasure of wealth. Mogi, throughout both my Master's and Ph.D. journeys in our department, you have always offered me crucial support in countless ways.

To my peers, friends, and colleagues, I am overwhelmed with gratitude. Drs. Bo Peng and Xinyi Liu, Chenxiao (Atlas) Guo, Jamp Vongkulsolkit, Chris Scheele, Songxi Yang, Ashmita Dhakal, Yuehan Qin, and Yang (Eric) Qiu, our collaboration and the infinite memories we created at the Spatial Computing and Data Mining Lab have been a treasured source of happiness and inspiration. I am blessed to have you all in my life, not just colleagues, but also caring friends. Drs. Yuhao Kang and Jinmeng Rao, Yunlei Liang, Yuhan Ji, Jake Kruse, Qianheng Zhang, Yanbing Chen, Lauren Gerlowski, Taylor Mcdowell, Becky Rose, Nora Schlenker,

Markia Silverman-Rodriguez, Kela Caldwell, Alicia Barceinas Cruz, Alicia Iverson, Gareth Baldrice-Franklin, Adrian George, Samto Wongso, Emily Diaz-Vallejo, Kallista Bley, Luke Leavitt, Gareth Baldrice-Franklin, Vignesh Ramachandran, Yuying Chen, Leanne Abraham, and many others, thank you for making my time at Science Hall unforgettable. All your talent and passion have elevated my experience beyond measure.

I am also grateful to my external colleagues and friends, Profs. Guiming Zhang, Zhe Zhang, and Jinwoo Park, your guidance and our teamwork serving as the Board Members for the AAG Cyberinfrastructure Specialty Group have been invaluable to me. In addition, being part of the GISphere GISalon International team has been an enriching experience, allowing me to make lasting friendships with many wonderful peers and scholars, including Zhou Bing, Dr. Hoeyun Kwon, Kejin Wang, Yikang Wang, Dr. Armita Kar, Ce Hou, and more.

Additionally, I warmly thank my dear friends – Dean Waldenmaier, Deb Payne, Barb and Tom Roe, Grace Yi Chen, Rahul Gulati, Judy and Ted Tang, and Nina Qin, among others. Your companionship throughout my Ph.D. journey has been a wonderful treasure. The joyful moments we have shared, the paramount support you provided during my years in Madison, and the countless times of our laughter and meals, have brought immeasurable delight to my life.

Finally, I reserve my deepest gratitude for my family: to my loving husband, Longjun Li, who has been my steadfast partner and fueled my ambition; to my parents and grandparents, whose endless love and wisdom have shaped who I am; to my big brother Yonghong Wu and sister-in-law Wanwan Hu, who have always been my anchors of strength; and to my adorable 4-year-old niece Jingsheng Wu, who brings immense joy and light into our lives. Your unconditional love, support, and belief in me have been the foundation of my journey. This achievement is as much yours as mine.

## CONTENTS

---

Contents v

List of Tables viii

List of Figures x

Abstract xiii

### **1 Introduction 1**

- 1.1 *The Rise of Multimodal Learning* 2
- 1.2 *A Plus from Spatial Knowledge Integration* 5
- 1.3 *Research Challenges and Research Questions* 6
- 1.4 *Dissertation Structure* 10

### **2 The Spatial Knowledge Enabled Multimodal Learning Framework and Building Blocks 11**

- 2.1 *A Foundational Multimodal Learning Framework Enhanced by Spatial Knowledge* 11
- 2.2 *Multimodal Geospatial Datasets* 15
  - 2.2.1 *Vision Input: Geospatial Imagery* . . . . . 15
  - 2.2.2 *Language Input: Generating Spatial and Non-spatial Context from Geospatial Images through GPT-4V* . . . 22
- 2.3 *Multimodal Feature Extraction and Embedding* 24
- 2.4 *Multimodal Learning Cores: From Backbone to Algorithm* 26
  - 2.4.1 *State-of-the-art Multimodal Foundation Models* . . . 26
  - 2.4.2 *Contrastive Learning: Algorithmic Foundations for Multimodal Learning* . . . . . 32

<b>3</b>	<b>Towards Zero-shot Learning of Multimodal Foundation Models : Establishing Benchmarks of Their Zero-shot Performances on Geospatial Image Reasoning</b>	37
3.1	<i>Introduction</i>	37
3.1.1	Zero-shot Learning . . . . .	37
3.1.2	Geospatial Image Reasoning . . . . .	38
3.2	<i>Methodology: Models and their Zero-shot Implementations</i>	41
3.2.1	Contrastive Language-Image Pre-training (CLIP) .	41
3.2.2	Generative Pre-trained Transformer 4 with Vision (GPT-4V) . . . . .	44
3.2.3	Workflow . . . . .	49
3.3	<i>Experiment #1 Global Image Geo-localization</i>	52
3.4	<i>Experiment #2 Urban Mixed Land Use Detection</i>	63
3.5	<i>Experiment #3 Urban Perception Prediction</i>	76
3.6	<i>Conclusions and Discussion</i>	82
<b>4</b>	<b>Towards Spatial-context Prompt Tuning for Multimodal Learning: GPT-4V Augmented Spatial-context Prompt Tuning on MFMs for Geospatial Applications</b>	86
4.1	<i>Introduction</i>	86
4.1.1	Prompt Tuning . . . . .	87
4.1.2	Prompt Ensembling . . . . .	89
4.1.3	Spatial-context Aware Prompts . . . . .	90
4.2	<i>Methodology</i>	92
4.2.1	Spatial-context Prompt Tuning on MFMs . . . . .	92
4.2.2	Workflow . . . . .	92
4.3	<i>Experiment #1 Global Image Geo-localization</i>	95
4.4	<i>Experiment #2 Urban Mixed Land Use Detection</i>	102
4.5	<i>Experiment #3 Urban Perception Prediction</i>	104
4.6	<i>Conclusions and Discussion</i>	107

<b>5</b>	<b>Towards GeoAI Multimodal Foundation Models: Developing GeospatialCLIP as a GeoAI Vision-Language Foundation Model</b>	<b>112</b>
5.1	<i>Introduction</i>	112
5.1.1	Vision-Language Foundation Models . . . . .	113
5.1.2	GeoAI for Vision-Language Learning . . . . .	114
5.2	<i>Methodology</i>	115
5.2.1	Geospatial Image-Text Pairwise Input . . . . .	115
5.2.2	Spatially Explicit Contrastive Learning on CLIP . .	116
5.2.3	Workflow . . . . .	117
5.3	<i>Experiment #1 Global Image Geo-localization</i>	118
5.4	<i>Experiment #2 Urban Mixed Land Use Detection</i>	124
5.5	<i>Experiment #3 Urban Perception Prediction</i>	128
5.6	<i>Conclusions and Discussion</i>	135
<b>6</b>	<b>Conclusions and Discussion: The Good, The Bad, and The Future</b>	<b>137</b>
6.1	<i>The Good</i>	137
6.2	<i>The Bad</i>	139
6.3	<i>The Future</i>	142
	References	144

## LIST OF TABLES

---

3.1	Example of GPT4-V’s performance on visual question answering (VQA) with spatial context based on street view images. . . . .	47
3.2	Example of GPT4-V’s performance on visual question answering (VQA) with spatial context based on satellite and map tile images. . . . .	48
3.3	Image-Text Pairwise Datasets Generated by GPT-4V . . . . .	49
3.4	CLIP’s Zero-shot Performance on Image Geo-localization . . . . .	55
3.5	CLIP’s Zero-shot Performance on the Im2GPS3k Benchmark Geo-localization (Compared with the State-of-the-art) . . . . .	61
3.6	Re-labeling Land Use Categories for OSM Data . . . . .	70
3.7	CLIP’s and GPT-4V’s Zero-shot Performances on Urban Perception Prediction . . . . .	80
3.8	GPT-4V’s Zero-shot Performances on Urban Perception Prediction Compared with the State-of-the-art . . . . .	80
4.1	Dimensions of Spatial Context-Aware Prompts . . . . .	92
4.2	CLIP’s Prompt Tuning Results of Street View Imagery Geo-localization . . . . .	95
4.3	CLIP’s Prompt Tuning Results of Esri Satellite Imagery Geo-localization (at 11 and 14 Zoom Levels in Years 2014 and 2023) . . . . .	98
4.4	CLIP’s Prompt Tuning Results of OpenStreetMap Map Tile Imagery Geo-localization . . . . .	100
4.5	CLIP’s Prompt Tuning Results of Mixed Land Use Detection . . . . .	103
4.6	CLIP’s Prompt Tuning Results of Urban Perception Prediction . . . . .	105
5.1	Zero-shot Learning Comparisons of Global Image Geo-localization between CLIP and GeospatialCLIP ("No prompt" & "Image type") . . . . .	119
5.2	Model Performance Comparison of the Im2GPS3k Benchmark Geo-localization (without prompt) . . . . .	121

5.3	GeospatialCLIP's Prompt Tuning Results of Mixed Land Use Detection . . . . .	125
5.4	GeospatialCLIP's Prompt Tuning Results of Beauty Prediction	129
5.5	GeospatialCLIP's Prompt Tuning Results of Boringness Prediction	130
5.6	GeospatialCLIP's Prompt Tuning Results of Depression Predic- tion . . . . .	131
5.7	GeospatialCLIP's Prompt Tuning Results of Liveliness Prediction	132
5.8	GeospatialCLIP's Prompt Tuning Results of Safety Prediction .	133
5.9	GeospatialCLIP's Prompt Tuning Results of Wealthiness Pre- diction . . . . .	134

## LIST OF FIGURES

---

2.1	A foundational multimodal learning framework with spatial knowledge. . . . .	12
2.2	The advantages of using street view images for multimodal learning. . . . .	16
2.3	Spatial distribution of the street view images in the Place Pulse 2.0 dataset. . . . .	17
2.4	Spatial distribution of the street view images in the Mapillary Public Dataset. . . . .	18
2.5	Spatial distribution of the satellite images collected from the Esri World Imagery Map Server (in 2014 and 2023). . . . .	19
2.6	Spatial distribution of the map tile images collected from the OpenStreetMap Raster Tile Server (in 2023). . . . .	21
3.1	CLIP pre-training and zero-shot prediction: an image encoder and a text encoder are pre-trained contrastively to predict which images should match which texts (left). After pre-training, CLIP can be used as a zero-shot classifier. For example, CLIP can estimate the most matching text for a given image and successfully predict the class of the image caption “a photo of a dog” for this dog image (right). . . . .	42
3.2	The workflow of zero-shot learning on CLIP and GPT-4V. . . . .	50
3.3	The development of global image geo-localization methods and a novel frontier method via vision-language multimodal learning based on image-text pairs with geo-location clues in both visual and textural formats. . . . .	52
3.4	Using t-SNE to visualize CLIP’s visual feature space of the cities in the Place Pulse 2.0 dataset. . . . .	57
3.5	Examples of attention maps from CLIP’s pre-trained image encoder to interpret street view imagery geo-localization. . . . .	58

3.6	Spatial distributions of (a) Im2GPS3k, a traditional benchmark dataset and (b-d) global imagery datasets curated for this dissertation (e.g., street view images, satellite images, and map tiles). Im2GPS3k for testing geo-localization methods is heavily concentrated in the U.S. and Western Europe, while the curated dataset is more spatially even and covers a global extent, obtaining a better spatial representation for image geo-localization.	62
3.7	Spatial distribution of Place Pulse 2.0 street view images in NYC.	68
3.8	Capturing mixed land uses of a location based on OSM land use data with a buffer radius of 50 meters (an example of Bedford Avenue in Brooklyn College), with OSM land use tags such as recreation and transportation. . . . .	71
3.9	Comparison of model zero-shot performances on mixed land use detection. Both CLIP and GPT-4V outperform few-shot ResNet and ViT. Note that the fluctuation of the ViT performance may be due to the few-shot sampling of each land use category, leading to unstable predictive results. . . . .	72
3.10	t-SNE visualization of the CLIP fusion space for different land use categories. . . . .	73
3.11	Examples of attention maps of CLIP's pre-trained image encoder for land use interpretation. . . . .	75
3.12	An example of a two-image comparison regarding the "safety" variable. . . . .	78
3.13	Boxplots of GPT-4's predicted scores versus human-rated scores of the six urban perception variables. . . . .	81
4.1	The workflow of spatial-context prompt tuning on CLIP, augmented by GPT-4V generated contexts. . . . .	93

5.1	The workflow of (1) developing GeospatialCLIP via spatially explicit contrastive learning based on the backbone of CLIP and (2) zero-shot learning on GeospatialCLIP for downstream tasks. . . . .	118
5.2	Zero-shot GeospatialCLIP significantly outperforms few-shot Linear-probing ResNet-152, and achieves a competitive performance compared with 4-shot Linear-probing CLIP on Place Pulse 2.0 image geo-localization. . . . .	122
5.3	Comparison of model performances on mixed land use detection.	127

## ABSTRACT

---

In recent years, the landscape of machine learning has been reshaped by the incorporation of multi-sourced, multimodal data, reflecting a concerted effort to attain a more holistic understanding of a given task. Specifically, the shift from single-modal to multimodal learning has emerged as a promising direction for emulating the human capacity to understand the world through a diversity of senses (e.g., vision, hearing, taste, smell, and touch).

This multimodal approach not only enables models to process and interpret various modalities of data, but also enriches their capacity to extract intricate patterns and produce more informed outcomes. Moreover, from both theoretical and practical perspectives, the evolution of multimodal learning can be further fueled by the integration of spatial knowledge, which is a distinctive and critical component of human cognition that can provide machine learning models with a deeper understanding of the spatial context and relationships between entities in their environment.

Therefore, this dissertation aims to develop a foundational multimodal learning framework strengthened by spatial knowledge. The significance of this framework will be demonstrated by its validity in strengthening multimodal foundation models (MFMs) with a stronger capability for various geospatial applications (e.g., image geo-localization, urban mixed land use detection, and urban perception prediction), by establishing high-quality, large-scale, geospatial multimodal datasets as benchmarks to evaluate their zero-shot performances, as well as integrating the techniques of spatial-context prompt tuning and spatially explicit contrastive learning to eventually develop geospatial artificial intelligence empowered MFMs (GeoAI MFMs).

## 1 INTRODUCTION

---

In the evolving field of machine learning, the pursuit of more sophisticated, accurate, and intelligent models has led researchers to explore beyond traditional approaches, e.g., single-modal learning. In light of this trend, this dissertation positions itself at the forefront of this evolution, aiming to bridge the gap between conventional single-modal learning approaches and more advanced multimodal learning paradigms. Specifically, to achieve an enhanced performance in various geospatial applications, this dissertation intends to study the integration of multimodal learning and spatial knowledge. This innovative combination is designed to mimic the human ability to interpret the world through multiple senses, thereby enabling machine learning models to achieve a more comprehensive understanding of our built environment. The exploration of integrating spatial knowledge into a foundational multimodal learning framework is not only academically intriguing, but also holds substantial potential for enhancing practical applications, particularly in the realm of geospatial analytics, marking a pivotal shift that promises to improve the interpretative power and applicability of machine learning models for geospatial applications.

Overall, this dissertation aims to explore and establish a foundational framework for multimodal learning, augmented with spatial knowledge. Primary objectives include:

- Developing an essential multimodal learning framework that integrates spatial knowledge effectively as well as demonstrating the enhanced capabilities of such a framework in various geospatial applications, such as image geo-localization, urban mixed land use detection, and urban perception prediction;
- Establishing benchmark datasets by curating high-quality, large-scale

geospatial multimodal datasets for training, evaluating, and testing multimodal foundation models in geospatial analytics;

- Addressing the current challenges and limitations in multimodal learning through innovative spatial context prompt tuning and spatially explicit contrastive learning;
- Contributing to the theoretical and practical knowledge base in the integration of multimodal learning and spatial knowledge, and thus shedding light on developing geospatial artificial intelligence (GeoAI) multimodal foundation models in the future.

## 1.1 The Rise of Multimodal Learning

Traditional machine learning models have primarily operated in single-modal learning, relying on a single data source for model training and driving the decision-making process (Huang et al., 2021). However, the world we aim to simulate is inherently multifaceted, with rich and diverse sensory cues that contribute to human perception and cognition (Mathias and von Kriegstein, 2023). The strength of multimodal learning lies in its design of fusing information from various sources (e.g., images, videos, text, and audio), which can potentially unlock a more in-depth level of learning and interpreting the intricate patterns of input. Specifically, multimodal learning, similar to the human learning process, incorporates a diversity of sensory modalities (e.g., vision, hearing, taste, smell, and touch) and thus can gain an enhanced comprehension of input to increase the accuracy, reliability, and comprehensiveness of the model outcomes in different tasks. By examining each data modality, multimodal learning can unravel the multifaceted dimensions of a given task, uncovering complex relationships between a wide range of tangible entities and conceptual dimensions (Ramachandram and Taylor, 2017), which are ubiquitous in

geospatial applications and beyond. Furthermore, integrating multiple data modalities can improve the robustness and generalization capabilities of machine learning models, particularly in scenarios with limited or noisy data, and thus can enable models to mitigate biases, handle ambiguities, and adapt to new domains more effectively (Qiu et al., 2022).

For example, Verma et al. (2019) used a smartphone-based camera and a handheld recorder to manually and simultaneously collect visual and auditory datasets on urban streets, and discovered that the extracted attributes from both modalities are useful for predicting greenness (i.e., tree cover), openness (e.g., buildings and sky), and sound classification (e.g., ambient road noise, birds, crowd, vehicle passing, and silence). Later, a follow-up study using the same visual and auditory datasets by Verma et al. (2020) derived a survey of eight auditory attributes (pleasant, chaotic, vibrant, uneventful, calm, monotonous, eventful, and annoying) (Axelsson et al., 2010; Kang et al., 2018a) and six perceptual attributes (safety, lively, beautiful, wealthy, depressing, and boring) (Dubey et al., 2016), and discovered that for auditory attributes, the sound of birds is perceived as pleasant and calm, while the sound of the crowd is observed to be chaotic, vibrant, and eventful. For visual attributes, the liveliness of the scene showed a positive correlation with the presence of markets, advertisement billboards, pedestrians, colors, and overall diversity of the scene, while beauty and wealthiness showed negative associations with the presence of markets and slums and positive associations with the presence of plazas, cars, nature, and greenery. In addition, the presence of slums and the absence of trees are perceived as depressing and boring, whereas markets have the opposite effect. Thus, with multimodal datasets, more comprehensive urban analytics can be achieved.

In recent years, the impressive ability of Large Language Models (LLMs), such as OpenAI's GPT-4V<sup>1</sup> and DALL·E-3<sup>2</sup>, Microsoft's BEiT-3 (Wang

---

<sup>1</sup><https://openai.com/gpt-4>

<sup>2</sup><https://openai.com/dall-e-3>

et al., 2022), and Google’s PaLI-3 (Chen et al., 2023), for image reasoning, and the photo-realistic images generated by Diffusion Models, e.g., image super resolution results produced by the Latent Diffusion Model (Rombach et al., 2022) and image inpainting results produced by RePaint (Lugmayr et al., 2022), have marked a significant milestone in the evolution of multimodal Generative AI. These models, with billions of parameters and trained on vast multimodal datasets, are often referred to as Multimodal Foundation Models (MFMs). For instance, in the field of text-conditioned visual generation, prominent examples include DALL·E-3, Stable Diffusion (Rombach et al., 2022), Imagen (Saharia et al., 2022), and Parti (Yu et al., 2022). Later, building on the success of text-to-image generation models, text-to-video generation models generate videos based on text prompts, such as Imagen Video (Ho et al., 2022), Make-A-Video (Singer et al., 2022), and OpenAI’s SORA<sup>3</sup>. They have captured the public’s imagination across industries, sparking a notably increasing interest in their potential applications. Thus, moving away from the traditional method of creating artificial intelligence (AI) models for specific purposes, more researchers now focus on crafting extensive and versatile models that are pre-trained on enormous data in different modalities. These models can be later tailored to a wide range of tasks through fine-tuning (e.g., few-shot learning (Wang et al., 2020)) or even zero-shot learning (Xian et al., 2018).

Despite these enriching findings attained by multimodal learning methods, integrating multimodal data requires careful consideration of data quality, compatibility in space and time, as well as potential biases among different areas and population groups (Boehm et al., 2022). Additionally, ethical considerations surrounding data privacy, ownership, and transparency must be addressed before applying multimodal data to model training (Katal et al., 2013). For example, this dissertation has collected geospatial images (e.g., street view images, satellite images, and map tiles)

---

<sup>3</sup><https://openai.com/sora>

from open-source platforms across the globe, representing a large-scale spatial coverage that is evenly distributed among populated urban areas.

## 1.2 A Plus from Spatial Knowledge Integration

While multimodal learning offers the promise of synthesizing diverse data streams, the incorporation of spatial knowledge may amplify its potential even further. Existing studies have shown that by integrating spatial knowledge, machine learning models are empowered to reason about spatial locations, proximity, and context (Janowicz et al., 2020).

For example, in the field of environmental studies, to address the spatial non-stationarity of the relationships between PM<sub>2.5</sub> concentrations and predictor variables such as aerosol optical depth and meteorological conditions, Zhan et al. (2017) developed a novel machine learning algorithm, i.e., Geographically-Weighted Gradient Boosting Machine Learning, by building spatial smoothing kernels to weigh the loss function. Similarly, Van Dao et al. (2020) achieved a significant improvement in landslide susceptibility prediction by using a spatially explicit neural network model. Moreover, Li et al. (2021a) developed a spatially explicit deep learning model referencing Tobler's first law of geography for weakly supervised terrain feature detection.

Furthermore, in the field of urban analytics, a review of spatially explicit GeoAI applications by Liu and Biljecki (2022) revealed that graph neural networks are promising solutions to incorporate spatial information, with challenges identified as data, scale, modifiable areal unit problem (MAUP), and lack of interpretation. As an instance, Liu et al. (2023b) proposed a spatially explicit GeoAI method that combines a graph convolutional network and extracted spatial information for the tasks of traffic volume prediction and population estimation. Additionally, Credit (2022) found that the inclusion of spatial lag parameters improved machine learn-

ing model accuracy in predicting post-construction employment density around newly built transit stations.

Yet, a disheartening fact is that most state-of-the-art multimodal foundation models still struggle with certain specialized, intricate, or abstract tasks, especially in the geospatial analytics domain (e.g., satellite image scene classification), partially due to data bias, meaning that their vast pre-training samples rarely contain geospatial multimodal datasets or specify the spatial context (e.g., geo-location) of visual features (Radford et al., 2021).

### 1.3 Research Challenges and Research Questions

To sum up, exciting opportunities have been rising due to the integration of spatial knowledge with multimodal learning (Janowicz et al., 2020), and many initial attempts have been made in geospatial applications (Liu and Biljecki, 2022). However, there have been insufficient endeavors in curating large-scale, multimodal geospatial datasets (e.g., geospatial image-text pairs worldwide) for training, fine-tuning, or evaluating MFMs “Of GeoAI, By GeoAI, For GeoAI.” This dilemma has also led to a lack of effort in benchmarking their performances on various geospatial applications. Therefore, the first step of this dissertation will focus on evaluating and benchmarking the performances of state-of-the-art MFMs (e.g., GPT-4V) on a wide range of applications in urban informatics, particularly via zero-shot learning. In this way, we can also effectively and efficiently obtain high-quality, large-scale, multimodal geospatial datasets for developing GeoAI MFMs as the ultimate goal.

Second, inspired by the progress achieved by spatial knowledge integration with machine learning, it is expected that multimodal learning likely benefits from several key modifications and optimizations, especially

involving the usage of MFMs for geospatial applications. Practically, these MFMs can be more competitive by leveraging spatial context, extracted from geospatial multimodal datasets, to guide the process of model fine-tuning or training. For example, incorporating text-based spatial contexts into image classification models has improved the accuracy of popular deep learning models on place type classification (e.g., restaurants, hotels, and museums) based on images by over 40% (Yan et al., 2018). Similarly, for zero-shot remote sensing scene classification, the locality-preservation deep cross-modal embedding networks (Li et al., 2021b) have been proven to outperform state-of-the-art methods, by employing transformer-based large language models (e.g., BERT (Devlin et al., 2018)) to extract semantic embeddings from expert-defined text descriptions for each scene class (e.g., river, lake, and wetland). Moreover, for zero-shot mixed land use detection, integrating street view images with spatial context-aware land use descriptions through vision-language multimodal learning can significantly outperform traditional vision-based methods and accurately capture the multiple functionalities of the ground features, benefiting from the incorporation of spatial context-aware prompts where the geographic scale of geo-locations matters (Wu et al., 2023).

Third, most MFMs are not spatially explicit regarding the design of their (pre-)training phrase (Janowicz et al., 2020). However, by highlighting spatially explicit features extracted from geospatial input (e.g., encoding the geo-locations as a type of spatial context embeddings, or fine-tuning model parameters based on geospatial image-text pairs instead of generic datasets), many studies have witnessed enhanced model performance across various geospatial applications (Janowicz et al., 2020). In fact, this potential could further enlighten the development of GeoAI MFMs.

To address the above limitations with the proposed ideas, three research questions (RQs) are raised and three corresponding research activities (RAs) are conducted:

- **RQ-1: How are MFMs' performances on geospatial image reasoning?** To answer this question, this dissertation proposes the research activity #1 (**RA-1**): **Towards Zero-shot Learning of MFMs**, e.g., Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) and GPT-4V (Achiam et al., 2023), aiming to establish benchmarks of their zero-shot performances on geospatial image reasoning. To encompass a diverse range of geospatial tasks, experiments of zero-shot learning will concentrate on three tasks using various geospatial datasets, including: (1) geo-localizing different types of global spatial images, such as street view images (SVI), satellite images, and map tiles; (2) urban mixed land use detection based on SVI; and (3) urban perception prediction of safety, wealthiness, beauty, depression, etc., based on SVI. In addition, the description, including the spatial context of a given geospatial image, can be later paired with the image and assembled as good-quality, large-scale, vision-language geospatial datasets for developing GeoAI MFMs, offering comprehensive comparisons, assessments, and insights of fine-tuning MFMs on a wide range of geospatial applications.
- **RQ-2: Will multimodal learning perform better with spatial context added to prompts?** To answer this question, this dissertation proposes the research activity #2 (**RA-2**): **Towards Spatial-context Prompt Tuning on MFMs**, which develops novel, GPT-4V augmented spatial-context prompt tuning methods applicable to MFMs. Experiments, also focusing on the three geospatial tasks mentioned in RA-1, will be conducted via three types of prompt comparisons: no prompt (e.g., the original MFM without prompt tuning), prompts with non-spatial context (e.g., generic image captions), and prompts with spatial context. In particular, the description of spatial context is categorized into five main types: (1) **geo-localization**, including clues and key features about geo-localization at different scales; (2)

spatial patterns, including detected objects and their spatial patterns as well as urban structure and spatial patterns; (3) Land use/land cover (LULC), including fine-grained (mixed) land use(s) description and coarse-grained LULC description; (4) urban perception description regarding safety, wealthiness, beauty, etc.; and (5) urban development, including urban growth description and prediction over 10 years (from 2014 to 2023). Experimental results will reveal the effect of each spatial context type on the advancement or deterioration in model performance towards the three tasks, and thus can contribute to the development of prompt tuning on MFMs for geospatial analytics.

- **RQ-3: Will GeoAI MFMs outperform generic MFMs?** In response, this dissertation proposes the research activity #3 (**RA-3**): **Towards GeoAI Multimodal Foundation Models** (GeoAI MFMs), which intends to develop a GeoAI vision-language foundation model by training and evaluating CLIP with over 150k pairs of high quality and large-scale (geo-image, geo-text) curated by GPT-4V. Specifically, the proposed model will integrate spatial context into its text embeddings, which is illuminated in RA-2, and then CLIP's pre-trained image encoder and text encoder will be fine-tuned contrastively based on geospatial image-text pairs curated in RA-1. This will enable the proposed model to learn the spatial context of the pairwise geospatial images and text, and thus to use it as an additional consideration for determining the similarity between the textual and visual representations. The effectiveness of the proposed spatially explicit contrastive learning method will be compared with the one of the vanilla baseline (e.g., the original CLIP) and evaluated on the three tasks in RA-1, and therefore can showcase the capabilities of GeoAI MFMs extended to incorporate location-based reasoning, spatial context understanding, geospatial semantics, and human perceptual

indications in urban areas.

## 1.4 Dissertation Structure

The dissertation is structured to provide a logical and thorough exploration of a foundational multimodal learning framework with spatial knowledge. Following this introductory chapter, Chapter 2 will delve into the theoretical and methodological underpinnings of multimodal learning. The subsequent three chapters will start with the evaluation of the current MFMs' effectiveness on geospatial image reasoning (Chapter 3), as well as the significance of leveraging spatial knowledge in multimodal learning for each of the three tasks via spatial-context prompt tuning and spatially explicit contrastive learning (Chapters 4 and 5, respectively). This dissertation will then close with a comprehensive conclusion of the pros and cons of the current GeoAI MFMs' development and an in-depth discussion of future directions (Chapter 6).

## 2 THE SPATIAL KNOWLEDGE ENABLED MULTIMODAL LEARNING FRAMEWORK AND BUILDING BLOCKS

---

The field of machine learning has undergone remarkable changes with the introduction of Multimodal Foundation Models (MFMs). These advanced models are at the forefront of AI research, with the ability to process, interpret, and generate content across different modalities, including text, audio, images, and video. Thus, this dissertation proposes novel approaches to harness the power of MFMs for boosting their power to support geospatial applications. More specifically, I intend to develop a foundational multimodal learning framework that incorporates spatial knowledge, thereby leveraging the capabilities of geospatial technologies. This framework will utilize **multimodal geospatial datasets** (RA-1), employ methods for **feature extraction and embedding** across different data modalities (RA-2), and leverage advanced **multimodal learning algorithms and techniques** (RA-3), demonstrated in Figure 2.1. Among these advancements are the state-of-the-art MFMs and cutting-edge self-supervised learning algorithms (e.g., contrastive learning). The ultimate goal is to significantly improve geospatial applications and contribute to the development of GeoAI MFMs. The following sections will start with the introduction of the proposed framework, followed by the elaboration of each of these core components.

### 2.1 A Foundational Multimodal Learning Framework Enhanced by Spatial Knowledge

This dissertation will investigate the realm of multimodal learning enriched by spatial knowledge, which is exemplified as spatial context extracted from images and expressed in text. In particular, by utilizing

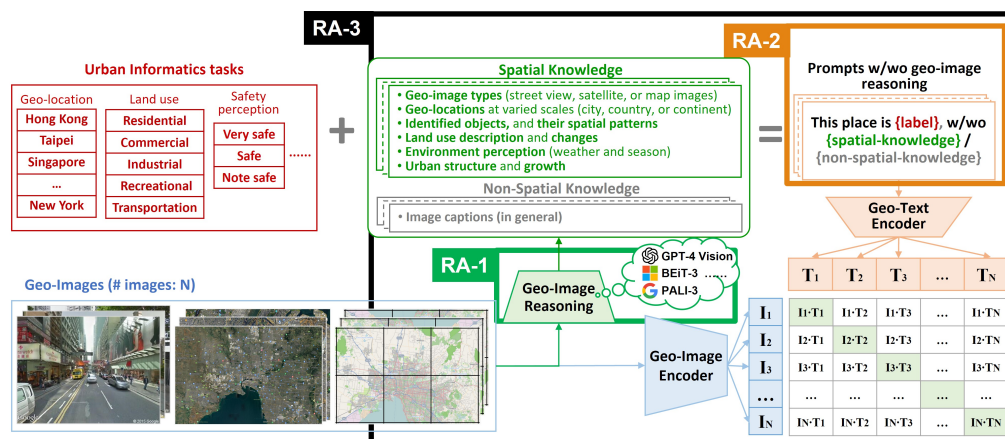


Figure 2.1: A foundational multimodal learning framework with spatial knowledge.

geospatial image-text pairs, a foundational multimodal learning framework with spatial knowledge (Figure 2.1) will be demonstrated via zero-shot learning, spatial-context prompt tuning, and spatially explicit contrastive learning on state-of-the-art MFMs (e.g., CLIP by Radford et al. (2021)), along with spatial context augmented by GPT-4V.

Building on this understanding, the integration of spatial knowledge into MFMs further enhances their ability to interpret and respond to the complexities of real-world environments. This approach leverages the intrinsic multi-sensory nature of human perception, which is crucial for navigating and understanding spatial contexts. In fact, the inception of multimodal learning was driven by the fact that real-world phenomena are inherently multi-sensory. The combination of different data streams, such as images, text, audio, and videos, allows models to mimic human capacity for a more holistic comprehension of a given task (Lu et al., 2022). Traditional machine learning methods, constrained by single-modal data or model architectures, struggle to encapsulate the richness of all relevant input in different modalities. On the contrary, the evolution of

multimodal learning traces its roots from data fusion approaches to the utilization of deep learning neural networks or transformers that apply the contrastive learning algorithm to process varied modalities. This multimodal learning practice examines how the synthesis of multimodal data can enhance model robustness and produce more comprehensive perspectives on the input. Moreover, the textual inclusion of spatial context within this multimodal learning framework can potentially introduce an even more enriching layer of understanding, enabling machines to grasp the geo-locations, spatial relationships, and/or local characteristics of the identified objects in both images and text.

The theoretical underpinnings lie in the fact that the inclusion of spatial context may further introduce a novel dimension that extends textual understanding in the process of multimodal learning. Humans intrinsically understand the natures and associations of entities, events, and concepts, partially by perceiving the explicit or implicit spatial context reflected by text. Thus, the adaptation of spatial context in textual input highlights its pivotal role in multimodal learning for enabling models to infer more intricate and meaningful relationships implied in the input, by enhancing their ability to learn the spatial context that can associate relevant entities captured in different modalities.

CLIP, as discussed in previous sections, is open-source and well-known for its ability to bridge the gap between textual descriptions and visual content via contrastive learning, through its pre-training on more than 400 million pairs of online images and their corresponding textual descriptions. With an enriched understanding of the content in both modalities, CLIP excels in associating images and text based on the similarity of their embedding space.

Generative Pre-trained Transformer 4 with Vision (i.e., GPT-4V), a state-of-the-art multimodal model known for its proficiency in image analysis and text generation, can be instructed to analyze images provided

by the user, representing one of the most groundbreaking advancements nowadays. Its integration of varied modalities (e.g., images) into LLMs has been considered a crucial frontier in AI research and development. GPT-4V has obtained the visual understanding capability by learning from an extensive corpus of multimodal data, including images, and thus demonstrated great potential in multimodal data comprehension.

With visual input of geospatial imagery such as street view images, satellite images, and map tiles (Section 2.2.1) that reflect various aspects of urban environments (e.g., buildings, roads, vehicles, pedestrians, greenness, urban structure, and their spatial patterns), MFMs are able to interpret and relate textual urban descriptors (e.g., "a bustling city street") with the associated visual features of urban scenes. In addition, MFMs potentially perform better via fine-tuning for various urban analytical applications, including urban image geo-localization, urban mixed land use detection, and urban perception prediction.

As for the subsequent RAs, RA-1 will first establish benchmarks of MFMs' zero-shot performance on geospatial image reasoning, which can also facilitate the curation of high-quality, large-scale, geospatial image-text pairs for spatial-context prompt tuning on MFMs in RA-2 as well as for developing GeoAI MFMs in RA-3. Next, RA-2 will tailor spatial context-aware prompts for geospatial images by utilizing the technique of prompt engineering, while RA-3 will employ the technique of spatially explicit contrastive learning by fine-tuning both the pre-trained text encoder and the pre-trained image encoder contrastively based on the curated geospatial image-text pairs. More methodological details will be described in each of the following RAs.

## 2.2 Multimodal Geospatial Datasets

As discussed above, the first component of the proposed framework is multimodal geospatial datasets. Specifically, vision and language input are utilized as the two data modalities, where vision input includes geospatial imagery such as street view images, satellite images, and map tile images, and language input includes both spatial and non-spatial contexts of geospatial images generated by GPT-4V.

### 2.2.1 Vision Input: Geospatial Imagery

#### Street View Images

Among various types of geospatial images, street view imagery stands out as a vital category, as these images can provide concrete and subtle visual features in urban environments, particularly from a human vision perspective, and thus are suitable for multi-modal learning. In addition, using street view images has become a main research trend (Zemene et al., 2018; Bansal et al., 2011), as these images have become largely available in public (Zhang et al., 2018a), and are more likely to be concurrent with textual descriptions of urban environments, facilitating the training process of language-image models (Figure 2.2).

**Place Pulse 2.0** The **Place Pulse 2.0** dataset, introduced by Dubey et al. (2016), which consists of 110,988 Google Street View images from 56 major cities across 28 countries worldwide captured between the years 2007 and 2012, will be used as the visual input of the proposed language-image models. These images were collected with latitude-longitude coordinates uniformly sampled from grids that are spatially intersected with city boundaries. The spatial distribution of the images in the Place Pulse 2.0 dataset is shown in the map of Figure 2.3.

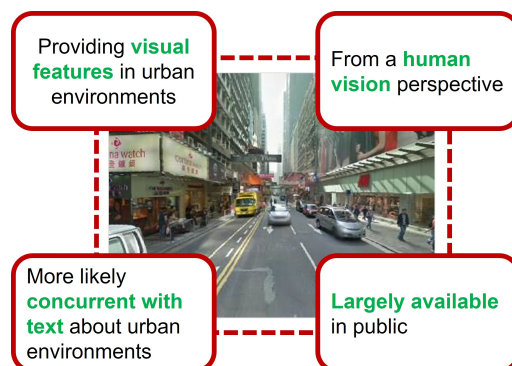


Figure 2.2: The advantages of using street view images for multimodal learning.

One goal of the creation of this dataset was to quantify the urban perception of different places around the world based on street views. It enables researchers in urban analytics to understand how people perceive urban spaces in terms of six perceptive variables, i.e., safety, liveliness, boringness, wealthiness, depression, and beauty. Participants in the study were then asked to view pairs of these images and answer questions related to urban perceptions, such as "Which place looks safer?" or "Which place looks more lively?". The responses were compiled to create a quantitative set of scores that provides insights into the subjective perceptions of urban environments, which has facilitated the development of machine learning algorithms that can automatically predict urban perceptions based on visual cues. Thus, this dataset is a significant contribution to the field of urban studies, computer vision, and machine learning.

**Mapillary Public Dataset** Additionally, 193,254 street view images across 430 most populated cities worldwide are also collected as the **Mapillary Public Dataset** from the Mapillary API<sup>1</sup>, each geo-located with latitude-longitude coordinates. The spatial distribution of this Mapillary dataset is

<sup>1</sup><https://www.mapillary.com/developer/api-documentation>



Figure 2.3: Spatial distribution of the street view images in the Place Pulse 2.0 dataset.

shown in the map of Figure 2.4.

One of the key advantages of the Mapillary dataset is that it is global, covering diverse urban environments worldwide and thus providing a more encompassing perspective of street views compared with the Place Pulse 2.0 dataset. Its global scope has supported studies on urban design, mobility, and accessibility by providing insights into the spatial characteristics of cities worldwide, bringing implications for policy-making, urban design, and the development of smart cities.

As crowd-sourced, volunteered geographic information (VGI), the Mapillary images were contributed by a large community of users who captured photos of streets and public places using smartphones, cameras, or other specialized imaging equipment. Therefore, this dataset is also characterized by its diversity and variability, in terms of geographic regions, urban landscapes, weather conditions, seasons, and times of day.

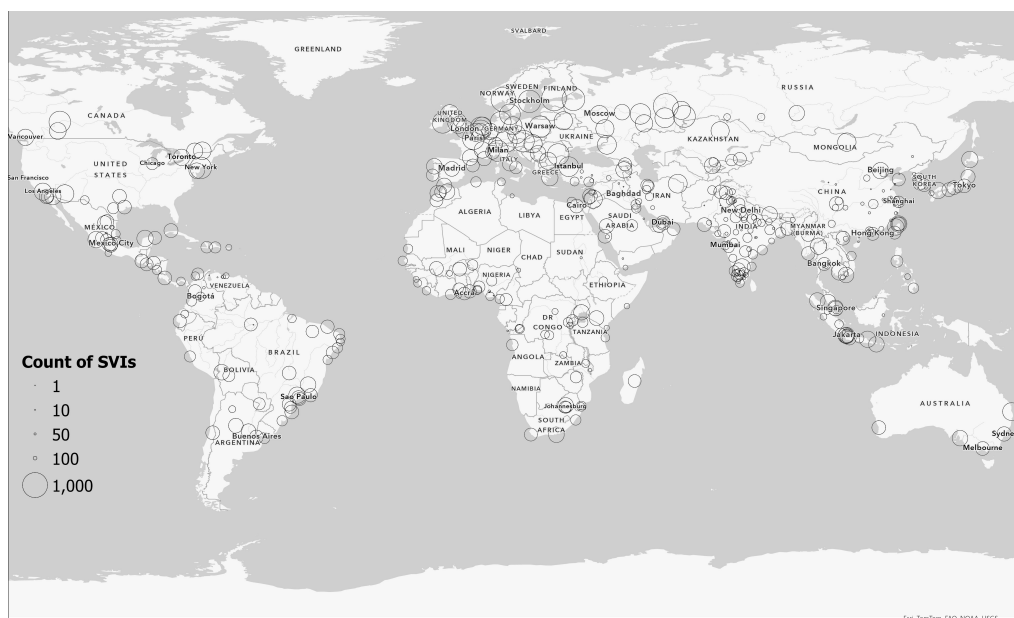


Figure 2.4: Spatial distribution of the street view images in the Mapillary Public Dataset.

Its diversity and variability make it an excellent resource for developing robust computer vision algorithms capable of generalizing across different urban environments, which is valuable for the tasks of semantic segmentation, object detection, and depth estimation in urban settings.

### Satellite Images

Another common geospatial image type for multimodal learning is satellite imagery. With its rich detail and comprehensive coverage of the Earth's surface, satellite imagery offers an unparalleled perspective on our planet. Furthermore, when combined with geospatial text, these images may unlock new potential in the domain of multimodal learning for geospatial applications (e.g., urban planning and development, disaster response and management, and environmental monitoring and conservation).



Figure 2.5: Spatial distribution of the satellite images collected from the Esri World Imagery Map Server (in 2014 and 2023).

**Esri World Imagery Map Server** Regarding the resources of large-scale satellite imagery, the **Esri World Imagery Map Server** provides an extensive, open-source collection of high-resolution satellite imagery from around the globe, which has been widely used in education and research, offering students and scholars detailed visual data that can enhance their studies and investigations into geographic and environmental phenomena.

One of the key features of the Esri World Imagery Map Server is its comprehensive coverage, offering detailed images of every part of the globe across different scales. This makes it an invaluable resource for projects that require a global perspective or for conducting analyses across different geographic regions. Moreover, this resource is updated annually, ensuring that users have access to the latest views of their areas of interest as well as the previous views for spatiotemporal change analysis.

In addition, the satellite images from the Esri World Imagery Map

Server are characterized by their high resolution, which allows for detailed visual inspections of specific sites, including urban areas, natural landscapes, and infrastructure projects, supporting a wide range of tasks such as monitoring changes in land uses and urban structures.

Specifically, 23,173 satellite images at different zoom levels (from 11 to 14) across 790 most populated cities worldwide were collected from the 2023 Esri World Imagery Map Server, each geo-located With latitude-longitude coordinates. To conduct temporal change analysis over the recent 10 years, 23,139 images with the same settings in terms of zoom levels and spatial coverage were also gathered from the 2014 Esri World Imagery Map Server. The spatial distribution of this satellite imagery dataset is shown in the map of Figure 2.5.

### **Map Tile Images**

Map tile images, the building blocks of digital maps that piece together to display detailed geographic information at various scales, may also find a unique place in the domain of image-text multimodal learning, by combining with textual data (e.g., geographic annotations, location-based social media posts, or descriptive map reports). This synthesis allows for more accurate geographic information retrieval. For instance, when training AI models to recognize specific landmarks or natural features from map tiles and associate them with relevant textual descriptions, a multimodal algorithm can potentially provide more spatial context-aware responses to queries about geo-locations of the map tiles. Furthermore, the integration of map tile images with text can potentially aid in map production and update in the future, by learning the descriptive text of (the changes of) map tiles. This innovative use of map tile images in multimodal learning underscores the vast potential of combining visual and textual data to extract deeper insights into our world's geography, enhancing both the accuracy and utility of digital mapping technologies.

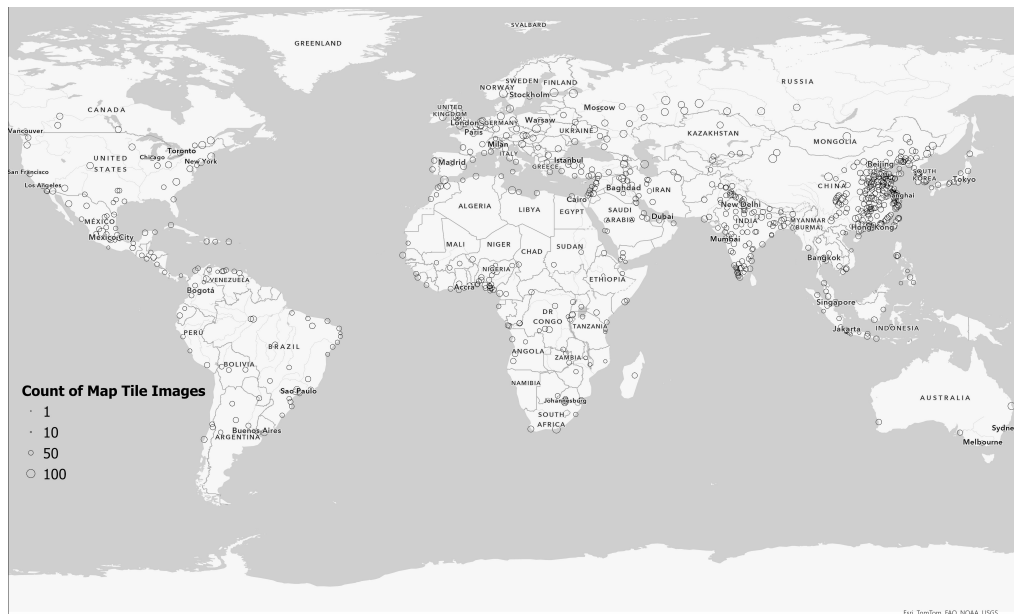


Figure 2.6: Spatial distribution of the map tile images collected from the OpenStreetMap Raster Tile Server (in 2023).

**OpenStreetMap Raster Tile Server** OpenStreetMap (OSM) provides a rich source of map tile images through its servers<sup>2</sup>, offering a detailed and dynamic view of the world's geography. These map tiles are essentially small, square bitmap images that represent different areas of the world map at various zoom levels. OSM's map tiles are particularly valuable because they are generated from a free, editable, crowd-sourced map of the world, maintained by a global community of volunteers, who make it incredibly detailed and up-to-date.

The map tiles from the OSM Raster Tile Server can be used in a wide range of applications, from web mapping services to mobile apps that can display and inquire about geographic information. The advantage of using OSM map tiles lies in their accessibility and the richness of the data they encapsulate, which includes not just geographical and topographical

<sup>2</sup><https://tile.openstreetmap.org/z/x/y.png>

information, but also metadata about different locations and landmarks.

Furthermore, the OSM server's map tile images support various academic and commercial projects, enabling the analysis of spatial patterns, infrastructure planning, and the study of urban development trends. Thus, when used in conjunction with map descriptive text, these images can significantly enhance image-text learning applications by providing a visual and spatial context to textual data, enriching the insights derived from analyses that combine geographic visualization with natural language processing.

Specifically, using the same configuration for the satellite imagery collection, 23,173 map tile images at different zoom levels (from 11 to 14) across 790 most populated cities worldwide were collected from the 2023 OSM Raster Tile Server, each geo-located With latitude-longitude coordinates. The spatial distribution of this map tile imagery dataset is shown in the map of Figure 2.6.

## **2.2.2 Language Input: Generating Spatial and Non-spatial Context from Geospatial Images through GPT-4V**

Given the vision input, there is a need for language input that matches the geospatial imagery. Note that all global imagery input, including street view images from the Place Pulse 2.0 dataset and the Mapillary public dataset, satellite images from Esri, and map tile images from OSM, do not contain textual descriptions of the urban scene reflected by themselves. To obtain geospatial image-text pairwise datasets for running the experiments of prompt tuning in Chapter 4 as well as for performing contrastively training on CLIP to develop a GeoAI language-image foundation model in Chapter 5, this dissertation will first conduct multiple VQA tasks on GPT-4V for geospatial image reasoning in Chapter 3, as displayed in Tables 3.1 and 3.2. In return, GPT-4V not only demonstrated remarkable zero-shot transfer capabilities for the tasks of mixed land use detection (Section 3.4)

and urban perception prediction (Section 3.5), but also generated both spatially and non-spatially contextualized textual descriptions for each image.

As outputs, extensive geospatial image-text pairs have been curated, split into training and testing sets, and summarized in Table 3.3.

- **Place Pulse 2.0 dataset:** 8,755 street view images are employed as a testing set, and each is tagged with its corresponding textual description, containing both non-spatial (e.g., image caption) and spatial contexts in four specific aspects: (1) urban perception predicted scores and description, (2) land use detection and reasoning, (3) detected objects and their spatial patterns, and (4) distinctive features with geo-location clues).
- **Esri Satellite Imagery dataset:** 4,755 satellite images are employed as a testing set, and each is also tagged with both non-spatial (e.g., image caption) and spatial contexts in five specific aspects: (1) land use detection and reasoning, (2) land use changes, (3) detected objects and their spatial patterns, (4) distinctive features with geo-location clues, and (5) urban growth description and prediction.
- **OSM Map Tile dataset:** 2,378 map tile images are employed as a testing set, and each is similarly incorporated with its non-spatial (e.g., image caption) and spatial contexts in three specific aspects: (1) land use detection and reasoning, (2) detected objects and their spatial patterns, and (3) distinctive features with geo-location clues.

## 2.3 Multimodal Feature Extraction and Embedding

With the first component of multimodal geospatial datasets, the next component of the proposed framework is multimodal feature extraction and embedding. In the tapestry of multimodal learning, the threads of different data modalities (e.g., text, images, audio, and environmental sensor records) are intricately woven together, presenting both a challenge and an opportunity in terms of extracting and embedding these multimodal features. The quest to understand and leverage multimodal data has led to significant advancements in feature extraction and embedding techniques.

Multimodal feature extraction and embedding is a sophisticated technique that stands at the crossroads of multiple disciplines, including computer vision, natural language processing (NLP), and potentially geographic information sciences if environmental or geospatial data are involved. At its core, it seeks to harness the complementary strengths of data from different modalities, enabling multimodal models to gain an in-depth understanding of complicated and intricate phenomena.

For text input, the analysis of text features begins with the extraction of meaningful descriptive information. Traditional NLP techniques, such as tokenization and part-of-speech tagging, lay the groundwork for more advanced processes. The advent of word embeddings, such as Word2Vec and GloVe, marked a significant leap forward, representing words in vector space and capturing their semantic relationships. Recently, the introduction of Transformer-based models (e.g., BERT and GPT) has revolutionized text analysis further, enabling context-aware embeddings that significantly enhance the understanding of language nuances.

As for image input, visual feature extraction has evolved from manual feature engineering to the use of deep learning models, particularly convolutional neural networks (CNNs). CNNs can automatically identify and

extract patterns and features from images, from simple edges to complex objects and scenes. These models have been instrumental in tasks ranging from image classification to object detection and scene understanding, providing a rich, visual context that complements textual information.

Particularly, regarding spatial context reflected in textual format, incorporating the meaningful spatial context into multimodal models can introduce an additional geographical dimension to data analysis, which is crucial for geospatial applications (e.g., location-based services and environmental monitoring). Techniques, including geo-tag extraction and spatial entity recognition, have explicitly converted noteworthy spatial information into a textual format, facilitating its integration with other data types and thus illuminating more comprehensive knowledge for multimodal model training. This convergence allows models to consider not only "what is happening" but also "where it is happening," adding an extra layer of context that can significantly enrich data interpretation for geospatial tasks.

Yet, the integration of multimodal data has posed a considerable challenge, demanding sophisticated methodologies that can merge information seamlessly. Fusion techniques (e.g., early, late, and hybrid fusion mechanisms) are employed to combine features at different stages of the model. Early fusion integrates raw data at the input level, while late fusion combines the outputs of separate models or modules. Hybrid approaches, often leveraging deep learning techniques, offer compromise and mixing features at intermediate levels to balance the benefits of both. Notably, deep learning Transformer-based models have played a pivotal role in this integration process. These models process, embed, and relate information across modalities, learning a unified representation that captures the essence of the data input.

## 2.4 Multimodal Learning Cores: From Backbone to Algorithm

After introducing the multimodal datasets and their feature extraction and embedding, the next component is multimodal learning cores, including the model backbone and self-learning algorithms for contrastive learning with multimodal datasets.

### 2.4.1 State-of-the-art Multimodal Foundation Models

MFMs are designed to process, integrate, and generate information across different modalities simultaneously. Unlike single-modal or unimodal models, which specialize in a single modality of data, multimodal models can comprehend complex queries that involve understanding relationships between different modalities of data. This capability is particularly relevant and desirable in today's digital age, where information is often presented in a combination of text, audio, images, and videos.

To date, several state-of-the-art MFMs have emerged, each revolutionizing the field of multimodal AI:

- **Contrastive Language–Image Pre-training (CLIP)** (Radford et al., 2021) by OpenAI in 2021, is designed to understand the relationship between text and images. CLIP can perform a wide range of vision tasks by interpreting natural language, enabling it to understand images in the context of textual descriptions from the internet so that it can recognize a vast number of visual concepts guided by text. CLIP is capable of performing a variety of tasks via zero-shot learning, including image classification and geo-localization.
- **DALL·E<sup>3</sup>**, another groundbreaking model from OpenAI in 2021, takes a step further by generating images from textual descriptions,

---

<sup>3</sup><https://openai.com/dall-e-3>

showcasing an impressive level of creativity and understanding of complex textual concepts. This is a series of models known for generating high-quality images from human-guided text. The latest version, DALL·E-3<sup>4</sup>, has greatly improved the quality and relevance of the generated images.

- **Gemini** (Team et al., 2023), developed by Google DeepMind, is a multimodal model capable of handling prompts in various formats, including images, text, code, and video. It has been meticulously crafted to facilitate smooth reasoning across multiple media types. When integrated with Vertex AI, Gemini demonstrates its capability to interpret text within images, transform image-based text into JSON format, and provide insightful responses about the images uploaded.
- Meta's **ImageBind** (Girdhar et al., 2023), represents a pioneering AI model that uniquely integrates data from six different modalities simultaneously, without requiring direct oversight. It understands the connections among various types of data, such as images, video, audio, text, depth, thermal, and inertial measurement units (IMUs). This innovation drives AI forward by empowering machines to analyze and interpret a diverse array of information more effectively.
- Meta's **SeamlessM4T** (Barrault et al., 2023), stands out as the first all-in-one AI solution for multilingual and multimodal translation and transcription. Capable of handling speech-to-text, speech-to-speech, text-to-speech, and text-to-text conversions, this model supports translations across up to 100 languages, with the specific capabilities varying by task.

---

<sup>4</sup><https://cdn.openai.com/papers/dall-e-3.pdf>

These models are "foundational" in the way that they provide a base and generic layer of understanding and generation capabilities across different domains due to their vast (pre-)training samples, which can then be fine-tuned or adapted for specific applications and tasks. As the field of AI continues to evolve rapidly, new models and updates to existing models are regularly introduced, pushing the boundaries of the capabilities of multimodal AI.

Correspondingly, the multimodal nature of these models has opened new avenues in varied AI applications, introducing profound and transformative implications across diverse sectors, including healthcare, education, entertainment, and beyond.

In healthcare, the applications of MFMs, e.g., Med-PaLM M (Tu et al., 2024) by Google DeepMind, are incredibly promising. These models can analyze and cross-reference information from various sources, such as medical imagery, patient histories, and genomic data, to aid in diagnosis, treatment planning, and personalized medicine. For example, Med-PaLM M could analyze radio-graphic images along with clinical notes to identify patterns and suggest diagnoses that might not be apparent through unimodal analysis. This holistic approach to the multimodal data from patients has demonstrated great potential to significantly improve the accuracy and efficiency of medical interventions.

Education is another domain where MFMs are making strides (Küchemann et al.). These models can create dynamic learning materials that integrate text, images, and audio-visual content, catering to different learning styles and needs. They can also be used to develop interactive learning platforms that can understand and respond to students' inputs, from text and voice to images and videos, along with tailored feedback, thus enhancing students' engagement. For instance, a student can take a picture of a math problem, and the platform, employing the techniques of image recognition and natural language processing (NLP), can not only solve

the problem but also provide a step-by-step explanation in both text and voice formats, allowing students to engage with the material in a way that suits them best. Also, students can collaborate on a project by sharing and discussing ideas through text, sketches, or video clips, and then these models can analyze these inputs, summarize discussions, and highlight key points, making collaborative learning more productive and engaging.

In particular, MFMs are making educational resources more inclusive and accessible to a wider audience. For visually impaired students, these models can convert text into speech, allowing the students to access written material. Similarly, for students with hearing impairments, these models can generate subtitles for videos or translate speech into sign language animations. Moreover, MFMs can create personalized learning experiences by continually assessing a student's understanding through quizzes, written assignments, and oral presentations, providing teachers with a more comprehensive view of each student's progress. Based on this multimodal analysis, these models can further recommend personalized content, exercises, and learning paths that both highlight the individual's strengths and address areas for improvement, making students' learning more efficient and effective. In addition, MFMs are being used to create immersive virtual and augmented reality (VR/AR) learning environments. These environments can simulate real-world scenarios, laboratory experiments, or historical events, allowing students to explore and interact with the content in a highly engaging and experiential manner. For example, these models can now make human interactions within simulated environments more natural and intuitive, so that a biology student can virtually dissect an animal, or a history student can experience being in a historical event.

Another captivating application of MFMs lies in the creative industries (Anantrasirichai and Bull, 2022). Models such as OpenAI's DALL·E, which generates images from textual descriptions, and Jukebox (Dhariwal et al., 2020), which creates music in various genres from textual prompts, have

demonstrated the potential to revolutionize the industry of content creation. These tools and models can assist artists, designers, and musicians in generating unique ideas and prototypes, significantly reducing the time and effort traditionally required in the process of creation. Moreover, they can democratize creativity, enabling individuals without intensive training from creative industries to express their ideas vividly and artistically.

Moreover, MFMs are also transforming how search engines operate as well as how we access and interact with information online (Huang et al., 2022; Chang et al., 2022). Traditional search engines have primarily relied on text-based queries and keyword matching to deliver results. However, the advent of MFMs has introduced a new paradigm, enabling search engines to understand and process queries across multimodal data, including text, images, and audio, and thus can yield more accurate or relevant results with more flexibility. For instance, MFMs can allow users to retrieve images by using text queries that describe the content of an image rather than relying on metadata or alt text. A user could search for "a photo of a yellow bird in the snow," and the MFM-empowered search engine can return accurate images matching the description, even if the image files do not contain those specific keywords in their metadata. Additionally, a user nowadays can ask a question by speaking into an advanced search engine combining audio and text modalities. Such a kind of search engine, utilizing an audio-text model, can comprehend the spoken question, contextually analyze it, and return results that are more precise and relevant. Furthermore, MFMs such as GPT-4V, can perform personalized searches by leveraging their comprehensive understanding of multimodal data to predict what users might be looking for, even with minimal input. Particularly, they can analyze a user's past search history, preferences, and even the content of uploaded images or spoken queries to personalize search results. This level of personalization is particularly useful for questions that have multiple, hierarchical layers of context,

which traditional text-based searches might struggle with, by ensuring that the information returned to users becomes more relevant and accurate, thereby reducing search time and enhancing overall user experience by catering to complex informational needs. In fact, the integration of state-of-the-art MFMs into search engines has been a game-changer, not only improving the richness and accuracy of search results in a more intuitive and efficient searching approach, which can significantly improve the usability and accessibility of digital services for a broader range of users, including those with disabilities, but also making the vast repository of online information more accessible and navigable than ever before. As these technologies continue to evolve, we can expect future search engines to become even more intelligent, context-aware, and capable of understanding the complexities of human queries in a way that feels natural and effortless.

While the applications of MFMs are varied, impactful, and promising, they also bring forth challenges and ethical considerations. Issues, including data privacy, biases in foundation models, and the misuse potential, need to be addressed to ensure that the development and deployment of these technologies are aligned with societal values and norms. Moreover, the computational and environmental costs associated with training MFMs demand a further exploration of more efficient and sustainable technologies. Thus, the journey ahead involves not only technological advancements but also navigating ethical and societal implications, making the development of MFMs benefit humanity as a whole (e.g., developing more efficient training methods, creating frameworks for ethical use, and designing models that are inclusive and capable of understanding human cultures and languages better).

## 2.4.2 Contrastive Learning: Algorithmic Foundations for Multimodal Learning

In the rapidly evolving field of machine learning, especially for training foundation models, the quest for models that can learn effectively with minimal supervision has fueled the development of the self-supervised learning paradigm, particularly for situations where labeled data is scarce or expensive to generate (Fei et al., 2022). Such a desirable paradigm has been illustrated by several notable algorithms:

- **Contrastive Learning** (Khosla et al., 2020), which is a popular approach in self-supervised learning, especially in the fields of computer vision and NLP. It operates on the learning principle that brings closer the representations of similar or "positive" pairs of data points while pushing apart those of dissimilar or "negative" pairs. For instance, in image processing, two different augmentations of the same image (e.g., cropped, resized, color-jittered versions) serve as positive pairs, and augmentations from different images serve as negative pairs during the process of contrastive learning. By optimizing the model to distinguish between these pairs, it learns meaningful representations of the input data, showing remarkable success in tasks such as image classification and sentence embedding without requiring labeled data.
- **Predictive Coding** (Huang and Rao, 2011), which is a framework that learns by predicting missing or future parts of the input data. In the context of NLP, models such as GPT (i.e., Generative Pre-trained Transformer) predict the next word in a sentence given the previous context, effectively learning the structure and semantics of the language through these predictions. In video processing, a model might predict the next frame in a video sequence, thereby learning about the dynamics and appearance of objects within the

video. Predictive coding harnesses the inherent structure of data to learn useful representations without the need for explicit labels.

- **Denoising Autoencoders** (Vincent et al., 2008), which is a type of neural network designed to learn representations by reconstructing the original input from a corrupted version. This process first adds noise to the input data and then trains the model to recover the original, uncorrupted input. Through this reconstruction process, the model learns to capture the most important features of the input. Denoising autoencoders have been used effectively for tasks such as image denoising, inpainting, and feature extraction.
- **Masked Language Modeling (MLM)** (Salazar et al., 2019), which is a self-supervised learning technique used primarily in NLP. In MLM, random words in a sentence are masked or replaced with a placeholder, and the model is trained to predict the original words based on the context provided by the remaining words. This approach enables the model to learn contextual relationships between words and better understand the syntax and semantics of the given language. For instance, BERT (i.e., Bidirectional Encoder Representations from Transformers) is an outstanding example that uses MLM to achieve state-of-the-art performance on various NLP tasks.
- **Clustering**, which involves grouping unlabeled data into clusters based on their similarities. This approach can help models discover the inherent structure of the input data and learn useful features or representations without explicit labels. Particularly, deep clustering algorithms (Caron et al., 2018), which combine traditional clustering techniques with deep learning, optimize both the feature representation learning and the clustering assignment simultaneously. This method has been applied successfully in domains such as image categorization and document classification.

These self-supervised learning algorithms have significantly expanded the capabilities of AI systems by leveraging the abundance of unlabeled data, reducing the reliance on costly labeled datasets, and enabling models to learn richer and more generalizable representations. As mentioned in the last section, the development of MFMs has also employed this innovative paradigm, allowing MFMs to learn from the data without explicit labels for every piece of information across different data modalities, thereby leading to highly sophisticated MFMs that demonstrate human-like understanding and creativity across various tasks.

In particular, among these popular self-supervised learning algorithms, contrastive learning stands out as a revolutionary approach, particularly within the domain of unsupervised representation learning across different data modalities for developing MFMs, while other algorithms are predominantly applied in single-modal or unimodal settings (e.g., computer vision or NLP tasks). As discussed, contrastive learning at its core is a technique used to learn generalizable representations by contrasting similar (positive) and dissimilar (negative) examples, which can be in different modalities. Its fundamental mechanism is achieved through a contrastive loss function, such as the Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010) or the more recent Triplet Loss (Schroff et al., 2015) and InfoNCE Loss (Oord et al., 2018) functions, which mathematically model the objective of bringing similar pairs closer and dissimilar pairs farther apart. Additionally, contrastive learning tends to learn more robust representations. This is due to its nature in learning patterns that can differentiate between similar and dissimilar examples (i.e., learning the links between examples instead of the examples themselves), and thus the learned representations are more resilient to overfitting and can be more generalizable across a variety of tasks and domains. This aspect of generalizability is crucial for developing MFMs that can be applied to real-world scenarios, where the testing data distribution may not perfectly

match the training set distribution.

In fact, the application of contrastive learning to MFMs has unlocked their new capabilities, enabling them to perform tasks that require an understanding of complex relationships across different data modalities. For instance, MFMs trained by the contrastive learning algorithm can now generate descriptive text for images (image captioning), answer questions based on combined textual and visual information (visual question answering), and even create images from textual descriptions (text-to-image generation), in which the development of sophisticated contrastive loss functions that can effectively measure similarities and differences across modalities has been crucial.

There have been several MFMs that employ contrastive learning in their model architectures, demonstrating its versatility and impact across various domains. For example, CLIP is a prominent instance designed to understand images along with text, by contrastively training on a vast dataset of images and corresponding textual descriptions. The success of CLIP (Radford et al., 2021) highlights the potential of contrastive learning in bridging the gap between visual and textual information, enabling applications such as zero-shot image classification and enhanced image search capabilities. Next, the ALIGN (i.e., A Large-scale Image and Noisy-text Embedding) model (Jia et al., 2021), similar to CLIP, leverages contrastive learning to align large-scale, noisy text and image data within a shared embedding space. This contrastive approach enables ALIGN to achieve remarkable performance in tasks requiring an understanding of the relationship between text and images (e.g., image captioning). Additionally, in the realm of audio and visual data, contrastive learning has been used to develop models that can learn representations by associating sounds with corresponding visual cues. For example, a model might be trained to match the sound of a dog barking with images of dogs, as opposed to images of other animals or unrelated objects. Such models can be used to find

relevant visual content based on audio queries and vice versa. Moreover, in domains where structured data is essential (e.g., knowledge graphs), contrastive learning has been adapted to work with text and graph data by learning to align textual descriptions with graph representations, which facilitates the task of entity linking where entities mentioned in the text must be matched with their corresponding nodes in a knowledge graph. By using contrastive learning to minimize the distance between text descriptions and their graph entity counterparts while maximizing the distance to unrelated entities, these models can significantly enhance information retrieval and question-answering systems that rely on structured knowledge.

Despite its impressive advancements in developing MFMs, the integration of contrastive learning with MFMs still faces challenges. One significant hurdle is the alignment and fusion of different modalities, which requires MFMs to understand not only each modality independently but also how they interact and complement each other. Also, the selection of positive and negative pairs is challenging yet crucial for the success of contrastive learning models, as suboptimal pairing can lead to poor performance. Furthermore, the computational complexity and cost associated with processing large batches of data for contrastive loss calculations can be substantially significant, posing scalability challenges. Therefore, as the field of contrastive learning continues to grow, future research should focus on addressing these challenges through more efficient fusion methods, better pair selection strategies, and the development of novel and scalable loss functions tailored for multimodal data.

### **3 TOWARDS ZERO-SHOT LEARNING OF MULTIMODAL FOUNDATION MODELS : ESTABLISHING BENCHMARKS OF THEIR ZERO-SHOT PERFORMANCES ON GEOSPATIAL IMAGE REASONING**

---

#### **3.1 Introduction**

In the advancing frontier of AI, multimodal foundation models (MFMs) represent a substantial breakthrough. These models, capable of understanding and processing information across multiple modalities of data (e.g., text, images, and audio), are redefining and pushing the boundaries of various AI applications, as discussed in Chapter 2. Particularly, their application in the geospatial domain has presented a novel arena, where the convergence of MFMs and spatial data, theories, and applications may lead to groundbreaking insights and innovations. Yet, nowadays there has remained insufficient effort in probing this line of research, starting from touching the base of exploring MFMs' zero-shot performance for geospatial image reasoning. Therefore, one major objective of this research activity is to fill this gap.

##### **3.1.1 Zero-shot Learning**

In the field of AI, particularly for LLMs, zero-shot performance refers to the model's ability to accurately perform tasks or make predictions on input data that it has never explicitly been trained on.

Mirroring the context of MFMs in geospatial applications, zero-shot learning implies leveraging the model's internal capacity to understand and analyze the multimodal input from the geospatial domain (e.g., street view images, satellite images, and map tiles, along with their spatial con-

text reflected in the images), based on its pre-trained multimodal representations without task-specific fine-tuning or training.

### 3.1.2 Geospatial Image Reasoning

Over decades, geospatial image reasoning has been standing as a crucial component in the ever-evolving landscape of AI. This task involves the extraction of meaningful insights from vast amounts of geospatial images, such as satellite imagery, street view images, and map tiles, facilitating a wide array of applications that have significant impacts on society and the environment, including urban planning and development (Biljecki and Ito, 2021), environmental conservation (Olea and Mateo-Tomás, 2013), disaster response and management (Adriano et al., 2021), and climate change monitoring (Cao et al., 2023). By identifying patterns, anomalies, and changes over time and space, geospatial image reasoning empowers decision-makers with actionable intelligence, driving informed policies and interventions. In a world where the dynamics of natural and human-made landscapes are constantly changing, the ability to accurately interpret and utilize geospatial images through sophisticated reasoning techniques is more critical than ever.

In particular, geospatial imagery reasoning often involves various facets such as geo-locations, object detection and their spatial arrangements, land use and land cover analysis, urban perception descriptions, and urban development forecasting. Consequently, downstream tasks or applications frequently include:

1. Image Geo-localization: Identifying the geographic location of an image by analyzing its visual features.
2. (Mixed) Urban Land Use Detection: Recognizing and distinguishing different land uses, which might occur in close proximity or mixed patterns, to understand the composition of urban environments.

3. **Urban Perception Prediction:** Predicting how people perceive different places in terms of safety, beauty, liveliness, etc., based on visual cues and spatial configurations.
4. **Change Detection and Monitoring:** Tracking and analyzing changes in landscapes over time, whether due to urban expansion, deforestation, or natural disasters.
5. **Infrastructure Assessment:** Assessing the condition and distribution of critical infrastructure such as roads, bridges, and buildings, particularly after natural calamities or for maintenance purposes.
6. **Agricultural and Environmental Monitoring:** Analyzing crop health, deforestation, and habitat loss, and monitoring the effects of climate change on diverse ecosystems.
7. **Urban Development Forecasting:** Predicting the growth or decline of urban areas, helping in planning and policy-making by anticipating future infrastructure and resource needs.

In the literature, these tasks typically rely on traditional methods such as manual feature engineering, image classification algorithms, and rule-based approaches. However, these methods face limitations like low scalability, difficulty in handling and fusing diverse data, and reliance on expert knowledge. Therefore, multimodal learning offers distinct advantages by leveraging the complementary strengths of different data modalities, which can enhance scalability via the fusion of various types of geospatial datasets, automate feature extraction via multimodal deep learning, and generalize better across diverse environments and patterns that might otherwise be missed with a single data modality.

Specifically, geospatial image reasoning requires the integration of complex spatiotemporal information. Theoretically, MFMs, by their nature and design, are adept at handling such diverse data streams, making them

particularly suitable for this domain. Their ability to encapsulate and process vast arrays of information, from the visual patterns in satellite imagery to the textual descriptions of geographic features, enables a more in-depth understanding of the Earth’s surface and human activities in urban areas.

Next, to harness the full potential of zero-shot learning of MFMs for this task, establishing comprehensive benchmarks is essential. These benchmarks should assess the accuracy and reliability of the models as well as their ability to generalize across different geographical scales, areas, and temporal dimensions. Additionally, these benchmarks should encompass a diverse set of tasks representing various challenges in geospatial image reasoning, including but not limited to:

- **Global Image geo-localization:** This task evaluates the model’s ability to determine the geographic location of an image in the world without explicit geographic markers. It requires the model to recognize and infer location-based landforms, architectural styles, vegetation types, and other geographically distinctive features, challenging the model’s understanding of spatial cues and cultural contexts.
- **Mixed land use detection:** Detecting mixed land use involves identifying areas where diverse human usages on lands or properties (e.g., residential, commercial, recreational, and industrial) overlap within a single image. This task tests the model’s ability to discern subtle differences in land patterns, building densities and styles, as well as the spatial arrangement of different land uses, reflecting the model’s understanding of the complex reality of urban landscapes.
- **Urban perception prediction:** Predicting urban perception involves assessing how people perceive various urban environments based on visual cues in images, such as safety, beauty, wealthiness, or liveliness. This requires the model to not only analyze physical attributes (e.g.,

building facades, street scenes, and the presence of green spaces), but also infer the social and psychological impressions that these features may convey to human observers.

To sum up, the first research activity of this dissertation aims to perform zero-shot learning on the state-of-the-art MFMs (e.g., CLIP and GPT-4V), emphasizing its transformative potential in geospatial image analytics and the necessity to establish rigorous benchmarks to evaluate their zero-shot performances. Moreover, curated from their desirable zero-shot results, high-quality, global-scale, geospatial image-text pairs can be effectively and efficiently obtained for developing GeoAI MFMs in the following research activities.

## **3.2 Methodology: Models and their Zero-shot Implementations**

As Chapter 2 states, this dissertation employs two MFMs via zero-shot learning, i.e., CLIP and GPT-4V, whose performances are evaluated on the three tasks related to geospatial image reasoning (i.e., global image geo-localization, mixed land use detection, and/or urban perception prediction).

### **3.2.1 Contrastive Language-Image Pre-training (CLIP)**

Our implementation follows the practice in (Radford et al., 2021) with three components as follows: 1) Text Encoding; 2) Image Encoding; and 3) Loss Calculation by Contrastive Learning, shown in Figure 3.1<sup>1</sup>. In the experiments, the labels of each application (e.g., "residential" and "commercial" for mixed land use detection) will be used as the textual

---

<sup>1</sup><https://openai.com/research/clip>

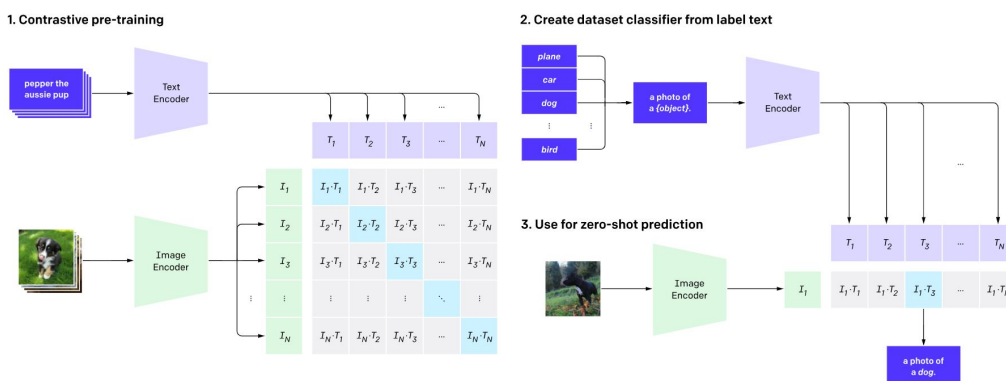


Figure 3.1: CLIP pre-training and zero-shot prediction: an image encoder and a text encoder are pre-trained contrastively to predict which images should match which texts (left). After pre-training, CLIP can be used as a zero-shot classifier. For example, CLIP can estimate the most matching text for a given image and successfully predict the class of the image caption “a photo of a dog” for this dog image (right).

input for a given geospatial image, so that CLIP, as a zero-shot classifier, can determine which label is best matching the image.

**Text Encoding.** The text encoder is a Transformer (Vaswani et al., 2017) modified by (Radford et al., 2019). It is a 12-layer, 768-width, and 12-attention-head model. The activations of the last layer of the Transformer are treated as the feature representation of the text input  $T_{[n,d_t]}^{\text{input}}$ , where  $n$  is the size of minibatch, and  $d_t$  is the feature dimensions of each text sample. This feature matrix is then layer-normalized and linearly projected into the multi-modal embedding space as  $T_{[n,d_e]}^{\text{output}}$ , where  $d_e$  is the dimensions of joint multi-modal embedding (Equation 3.1). Particularly, multiple spatial context-aware prompts can be constructed as textural input for prompt tuning for the next step in RA-2 (See Figures 2.1 and 4.1).

$$T_{[n,d_e]}^{\text{output}} = \left\| T_{[n,d_t]}^{\text{input}} \times W_{[d_t,d_e]}^T \right\|^2 \quad (3.1)$$

**Image Encoding.** The image encoder is based on Vision Transformer ViT-L/14 (Dosovitskiy et al., 2020), which we found performs best compared with other available image encoders (e.g., EfficientNet-style RN50x64, ViT-B/32, and ViT-B/16) in the CLIP model. It is a 24-layer, 1024-width, and 16-attention-head model. Similarly, the image feature matrix  $I_{[n,d_i]}^{\text{input}}$ , where  $n$  is the size of minibatch and  $d_i$  is the feature dimensions of each image sample, is then layer-normalized and linearly projected into the multi-modal embedding space as:

$$I_{[n,d_e]}^{\text{output}} = \left\| I_{[n,d_i]}^{\text{input}} \times W_{[d_i,d_e]}^I \right\|^2 \quad (3.2)$$

**Loss Calculation by Contrastive Learning.** After computing the feature embeddings of the images and the feature embeddings of the paired texts by the two encoders, the cosine similarity of these two sets of embeddings is then calculated and scaled by a temperature parameter  $t$  (Equation 3.3). Finally, the scaled pairwise cosine similarities are normalized by softmax, and used to measure the symmetric cross entropy loss (Equation 3.4). For more details of the pre-training updates, the Adam optimizer (Kingma and Ba, 2014) was used with decoupled weight decay regularization (Loshchilov and Hutter, 2017) for all weights (not gains or biases), and the learning rate (initialized as  $4 \times 10^{-4}$ ) was decayed using a cosine schedule (Loshchilov and Hutter, 2016).

$$\text{logits} = S_c(I_{[n,d_e]}^{\text{output}}, (I_{[n,d_e]}^{\text{output}})^T) \times e^t \quad (3.3)$$

where  $S_c$  is the cosine similarity and  $t$  is the temperature parameter. The symmetric loss function is defined as below:

$$\begin{aligned}
\text{loss}_T &= - \sum_{c=1}^M y_{i,c} \log(\text{logits}_{i,c}^T) \\
\text{loss}_I &= - \sum_{c=1}^M y_{i,c} \log(\text{logits}_{i,c}^I) \\
\text{loss} &= (\text{loss}_T + \text{loss}_I)/2
\end{aligned} \tag{3.4}$$

where  $M$  is the number of classes,  $\log$  is the natural log,  $y$  is a binary indicator (0 or 1) if class label  $c$  ( $c = [1, \dots, n]$ ) is the correct classification for image  $i$  ( $i = [1, \dots, n]$ ),  $\text{logits}_{i,c}^T$  is the predicted probability that image  $i$  is of class  $c$  using logits in Equation 3.3 normalized along the Text axis via a softmax, and  $\text{logits}_{i,c}^I$  is the predicted probability that image  $i$  is of class  $c$  using logits normalized along the Image axis via a softmax.

### 3.2.2 Generative Pre-trained Transformer 4 with Vision (GPT-4V)

Generative Pre-trained Transformer 4 with Vision, also denoted as GPT-4V, is a large multimodal model capable of processing both image and text inputs while generating textual outputs. Released in September 2023, GPT-4V represents a paradigm shift, endowing LLMs (e.g., GPT 3.5) with the ability to parse visual content alongside textual inputs, thereby transcending the limitations of language-centered AI models to emerge as a multifaceted analytical powerhouse. Despite being less proficient than humans in various real-world scenarios, GPT-4V has displayed a groundbreaking vision-language analytical power and demonstrated human-level performances across diverse professional and academic benchmarks (e.g., simulated human-designed exams, content moderation, and programming), affording users an enhanced level of interaction and comprehen-

sion. However, GPT-4V is not open-source, meaning that there is no access to the code, pre-training data, model architecture, or model weights.

Specifically, GPT-3.5 served as a precursor to GPT-4V, allowing for bug fixes, theoretical enhancements, and improved stability in subsequent training runs. Later, GPT-4V, built upon intricate deep learning frameworks (e.g., with a transformer-based model architecture), incorporates a vision encoder equipped with pre-training images, and then learns to align the encoded visual features with the underlying language modules, exhibiting remarkable efficacy in processing and interpreting complex visual and textual datasets and marking a notable achievement of its predictive performances.


The GPT-4V's training is facilitated by its exposure to a vast and diverse dataset encompassing textual and visual inputs sourced from myriad online repositories. Particularly, the training process incorporates reinforcement learning mechanisms, which serve to enrich and strengthen the model's adaptability and learned analytical insight. According to the GPT-4V technical report (Achiam et al., 2023), OpenAI's six-month iterative alignment process, drawing insights from adversarial testing initiatives and ChatGPT, has significantly enhanced GPT-4V's capacity, particularly in terms of factuality, steerability, and adherence to specified parameters. Of particular interest is the two-stage training protocol employed by GPT-4V. Initially aimed to assimilate vision-language associations, the model subsequently undergoes fine-tuning on a meticulously curated dataset, which fortifies its generation reliability and ensures the delivery of accurate and pertinent information to end-users.

OpenAI has also acknowledged GPT-4V's limitations, including inherent biases and occasional reasoning errors. Efforts to mitigate risks are ongoing, with enhanced safety protocols implemented during training to minimize harmful outputs. Nevertheless, the importance of exercising caution should be underscored whenever we utilize the model outputs,

especially in high-stakes contexts.

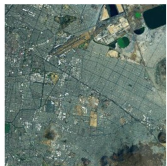

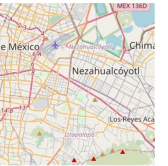
As for the usage of GPT-4V, its text input functionality is readily accessible through both `chat.openai.com` and the API, while its image input capability is currently under limited release in collaboration with select partners. OpenAI has also introduced OpenAI Evals, an open-source framework for automated AI model evaluation, encouraging community participation in identifying model shortcomings and guiding further improvements. Yet, there has been a lack of benchmarks in geospatial applications within the OpenAI Evals, which demands more effort in exploring its zero-shot performances in the geospatial domain. Thus, to perform zero-shot learning of GPT-4V for geospatial image reasoning, this research activity runs visual question answering (VQA) tasks via the GPT-4V API, utilizing the three types of geospatial images and focusing on the various aspects (or facets) of spatial context, such as distinctive features as geo-location clues, detected objects and their spatial patterns, LULC and its changes, urban perception prediction and corresponding description, urban growth description and prediction, etc., as well as the aspects of non-spatial context, e.g., an image caption (Tables 3.1 and 3.2). In general, the GPT-4V generated responses about geospatial image reasoning are constructive and insightful based on the qualitative evaluation, especially for the street view images, with successfully identified key objects and providing rich information about their spatial patterns that lead to reasonable or correct answers for several VQA tasks. Yet, GPT-4V also generates hallucinating results in a few cases, e.g., finding the distinctive features from an image that can indicate its geo-location. For example, when analyzing a satellite image from Mexico City, GPT-4V stated that the image may show areas with "colonial-style architecture" and "landmarks like the Angel of Independence or the Chapultepec Park," which does not exist in the image (or specifically, they are close to but still outside of the spatial extent of the image), as shown in Tables 3.2.

Table 3.1: Example of GPT4-V's performance on visual question answering (VQA) with spatial context based on street view images.

Image-text input with spatially contextualized questions	
GPT4-V's output*	<p>This image was taken in New York. Why do you think this image was taken in New York? Please give me the image caption of it. Then describe the objects in detail in this street view, including their spatial patterns. Next, based on your observation, what land use(s) are reflected (i.e., residential, commercial, industrial, recreational, transportation, and green space), and why? On a scale of 1 to 10 (with 1 being the lowest and 10 being the highest), please rate the image with an integer in the six perceptible aspects: safety, liveliness, boringness, wealthiness, depression, and beauty. Please also include the reasons for your ratings.</p> <ol style="list-style-type: none"> <li>Reasons why this image was taken in New York: The assumption that this image was taken in New York may be based on the <b>architectural style of the buildings</b>, the <b>density of residential structures</b>, and the <b>make and models of the cars</b>, which are indicative of urban areas in New York. Additionally, the <b>Google watermark</b> suggests it's a <b>street view</b> from a location-specific search.</li> <li>Image caption: "Urban Street Scene in New York with Residential and Commercial Buildings."</li> <li>Description of objects and their spatial patterns: The street view shows <i>tightly packed residential buildings with varying facades</i>, indicating mixed architectural styles and periods. <i>Cars line both sides of the street</i>, implying limited parking space. <i>Utility poles and cables crisscross above</i>, while the <i>road appears well-worn, with patches and signs of repair</i>.</li> <li>Land use(s) reflected in the image and why: Primary land use: <b>Residential</b>; the presence of multiple <b>mailboxes</b> and <b>entrance doors to buildings</b>, along with visible <b>apartment windows</b> and <b>fire escapes</b>, indicate residential use. Other potential land use(s): <b>Commercial</b>; <i>a storefront is visible on the right</i>, suggesting mixed-use with commercial activities at street level.</li> <li>Rating of safety: <b>Safety=4</b>; the area appears somewhat neglected, with <b>graffiti</b> visible and <b>no people</b> in sight, which may indicate safety concerns, although there's no direct evidence of immediate danger.</li> <li>Rating of liveliness: <b>Liveliness=3</b>; <b>the absence of people and visible activity</b> gives the impression of low liveliness, but <b>parked cars</b> suggest the presence of residents or visitors.</li> <li>Rating of boringness: <b>Boringness=7</b>; <b>the lack of visible commercial activity, people, or greenery</b> lends a monotonous feel to the scene.</li> <li>Rating of wealthiness: <b>Wealthiness=2</b>; <b>the disrepair and apparent age of the buildings</b>, along with <b>the condition of the street</b>, suggest a lower economic status.</li> <li>Rating of depression: <b>Depression=6</b>; <b>the overcast sky and worn infrastructure</b> may contribute to a <b>somber mood</b>, but this is subjective and not everyone may feel the same.</li> <li>Rating of beauty: <b>Beauty=3</b>; while there is a certain urban aesthetic, <b>the lack of green spaces</b>, <b>visible disrepair</b>, and <b>the utilitarian look of the buildings</b> lower the overall beauty rating.</li> </ol>

\* Note: **Green text**: reasonable/correct; **Orange text**: irrelevant; **Red text**: wrong or hallucinating; **Bold**: key objects for urban analytics and geospatial image reasoning; *Italic*: spatial patterns. This result was generated by gpt-4-vision-preview (GPT4-V) on November 16, 2023.

Table 3.2: Example of GPT4-V’s performance on visual question answering (VQA) with spatial context based on satellite and map tile images.

Image-text input with spatially contextualized questions			
	2014 Satellite	2023 Satellite	2023 Map Tile
	<p>These three images were taken at the same location in Mexico City. The first image is a remote sensing image taken in 2014, while the second one is a remote sensing image taken in 2023. The third image is an OpenStreetMap map tile in 2023. Based on the second and third images, are there any distinctive features or objects displayed in these images that can indicate they are from Mexico City? Please give me an image caption of all images. Describe the urban structure in detail, including the spatial patterns, based on the second and third images. Next, based on your observation, what land use and land cover (LULC) classes are displayed and why? Note that the LULC classes should be selected from the following nine options: urban or built-up land, agricultural, rangeland, forest, water areas, wetland, barren land, tundra, and perennial snow or ice. You should list the dominant LULC class as well as the subordinate ones. Based on the first and second images, what are the LULC changes from 2014 to 2023? Based on these identified changes, could you describe and predict the urban growth in detail?</p>		
GPT4-V’s output*	<ol style="list-style-type: none"> <li>1. Distinctive features in the second image: The second image may show <i>large, densely packed urban areas</i> with distinct <b>colonial-style architecture</b>, <b>major highways</b>, and <b>landmarks like the Angel of Independence or the Chapultepec Park</b> which are distinctive to Mexico City.</li> <li>2. Distinctive features in the third image: The OpenStreetMap tile should indicate the <b>street layout</b>, <b>public transportation lines</b>, and <b>named districts or landmarks</b>, such as <b>Zócalo</b>, which is the main plaza in Mexico City.</li> <li>3. Image caption of the first image: "Remote sensing image of Mexico City’s urban landscape in 2014 showing early stages of urban sprawl and land use."</li> <li>4. Image caption of the second image: "High-resolution remote sensing image of Mexico City in 2023, displaying dense urban development and sprawling city limits."</li> <li>5. Image caption of the third image: "Detailed OpenStreetMap tile of Mexico City in 2023, highlighting the intricate <b>street network</b> and urban planning."</li> <li>6. Urban structure and spatial patterns: Mexico City’s urban structure is characterized by <i>a dense central district with radial highways extending outwards, interspersed with green areas like parks and a mix of residential, commercial, and industrial zones.</i></li> <li>7. LULC classes reflected and why: Dominant LULC class: <b>Urban or built-up land</b>, due to visible <i>dense construction and road networks</i>; Subordinate LULC classes: <i>Agricultural on city outskirts, water areas along lakes and rivers, and green spaces that indicate parks or gardens.</i></li> <li>8. LULC changes from 2014 to 2023: <i>Transition from some agricultural and undeveloped land to urban areas, increase in built-up density, and possible reduction of green spaces due to urban expansion.</i></li> <li>9. Urban growth description and prediction: Growth is characterized by urban sprawl and infilling. Predicted future growth includes <i>further expansion into peripheral areas and vertical development</i>, with potential challenges to sustainability and green space conservation.</li> </ol>		
	<p>* Note: <b>Green text:</b> reasonable/correct; <b>Orange text:</b> irrelevant; <b>Red text:</b> wrong or hallucinating; <b>Bold:</b> key objects for urban analytics and geospatial image reasoning; <i>Italic:</i> spatial patterns. This result was generated by gpt-4-vision-preview (GPT4-V) on November 16, 2023.</p>		

As results, the training and testing datasets (i.e., geospatial image-text pairs) that have been generated by GPT-4V are summarized in Table 3.3, which will be utilized in the subsequent experiments to: (1) evaluate zero-shot performances of GPT-4V and CLIP on the tasks of mixed land use detection and urban perception prediction using the testing set of Place Pulse 2.0 street view images in this chapter; (2) evaluate the validity of spatial-context prompt tuning on CLIP for the three targeted tasks based on the testing sets in Chapter 4; and (3) perform spatially explicit contrastive learning on CLIP to develop GeoAI MFMs (i.e., GeospatialCLIP) using the training sets, as well as evaluating its zero-shot performances and spatial-context prompt tuning performances for the three targeted tasks using the testing sets in Chapter 5.

Table 3.3: Image-Text Pairwise Datasets Generated by GPT-4V

Image dataset	Num. training images	Num. testing images	Image caption	urban perception	Generated context types				
					land use	land use changes	spatial patterns	geo-features	urban growth
Place Pulse 2.0 street view images	6,489	8,755	✓	✓	✓	–	✓	✓	–
Esri Satellite images (2014 & 2023)	11,463	4,755	✓	–	✓	✓	✓	✓	✓
OSM map tiles (2023)	5,732	2,378	✓	–	✓	–	✓	✓	–

### 3.2.3 Workflow

Next, to evaluate the zero-shot performances of CLIP and GPT-4V, three tasks have been conducted: (1) global image geo-localization using street view images, satellite images, and map tile images by CLIP; (2) urban mixed land use detection using the subset of Place Pulse 2.0 street view images in New York City, each tagged with at least one OpenStreetMap land use label (e.g., residential, commercial, recreational, industrial, transportation, and greenfield) (Wu et al., 2023), tested by CLIP and GPT-4V;

and (3) urban perception prediction regarding six perceptive variables (i.e., beauty, boringness, depression, liveliness, safety, and wealthiness) using Place Pulse 2.0 street view images, each with human-rated ranking scores measured for the six variables (Dubey et al., 2016), tested by CLIP and GPT-4V.

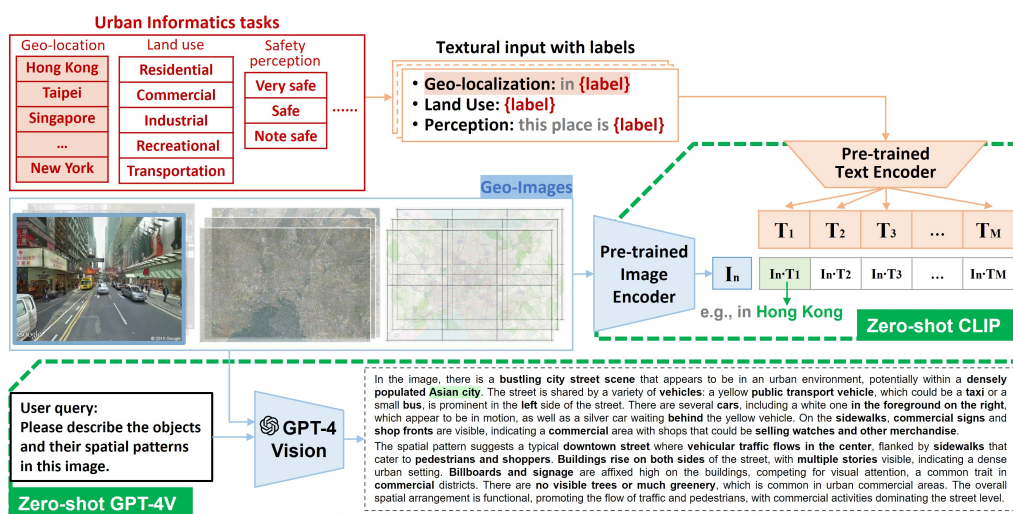


Figure 3.2: The workflow of zero-shot learning on CLIP and GPT-4V.

The zero-shot learning workflow of CLIP and GPT-4V is demonstrated in Figure 3.2. To run the evaluation of CLIP's zero-shot learning in the three tasks, the labels of each task will be first extracted, and then used as textual input for CLIP's pre-trained text encoder. Meanwhile, geospatial images will be the visual input for CLIP's pre-trained image encoder. For zero-shot prediction, given an image, the label that corresponds with the cell (i.e., the dot product of text and image embeddings) having the highest similarity value will be the model prediction outcome.

Notably, a previous research study (Wu and Huang, 2022) evaluating CLIP's zero-shot prediction on street view image geo-localization has discovered that simply combining city labels with prepositions that normally bind with city names (e.g., "at", "from", and "in") in the contextual prompt

can fairly boost CLIP's prediction performance by 1.51% compared with the case without such prepositions. Similarly, another study evaluating CLIP's zero-shot prediction on mixed land uses has found that simply using city labels as the textual input could achieve one of the best results among many crafted prompts (e.g., "[*land use label*] purpose" or "[*land use label*] use"), while the best result was achieved by prompt ensembling using the softmax function (Wu et al., 2023). Based on these findings, this research activity will construct the textual input for CLIP's pre-trained text encoder mainly based on the labels of each task, exerting the least amount of effort necessary in prompt engineering to achieve a more favorable outcome. For instance, in the task of image geo-localization, two types of textual input, i.e., "*in [geo-location label]*" and "*This [image type] image was taken in [geo-location label]*", will be used to evaluate CLIP's zero-shot performance, where "[*geo-location label*]" is formatted as "<*city, country, continent*>" and "[*image type*]" includes "*street view*", "*satellite*", and "*map tile*". Table 3.4 displays the zero-shot CLIP geo-localization accuracies of these two kinds of textual input, represented as "CLIP" and "CLIP + Image type", respectively.

As for running the evaluation of GPT-4V's zero-shot prediction on the tasks of mixed land use detection and urban perception prediction, both the image and spatial context-aware questions are assembled as an image-text multimodal inquiry for GPT-4V to perform VQA in each task. For instance, Figure 3.2 displays an example of GPT-4V's response when a user uploaded a street view image as visual input and requested "*Please describe the objects and their spatial patterns in this image.*" as textual input, and its response reflects multi-facet aspects of spatial context, e.g., vehicles and their spatial relationships, other visible objects and their positions in the image, urban functionalities, and detailed land use information and reasoning, which are informative and enlightening for both mixed land use detection and urban perception prediction.

### 3.3 Experiment #1 Global Image Geo-localization

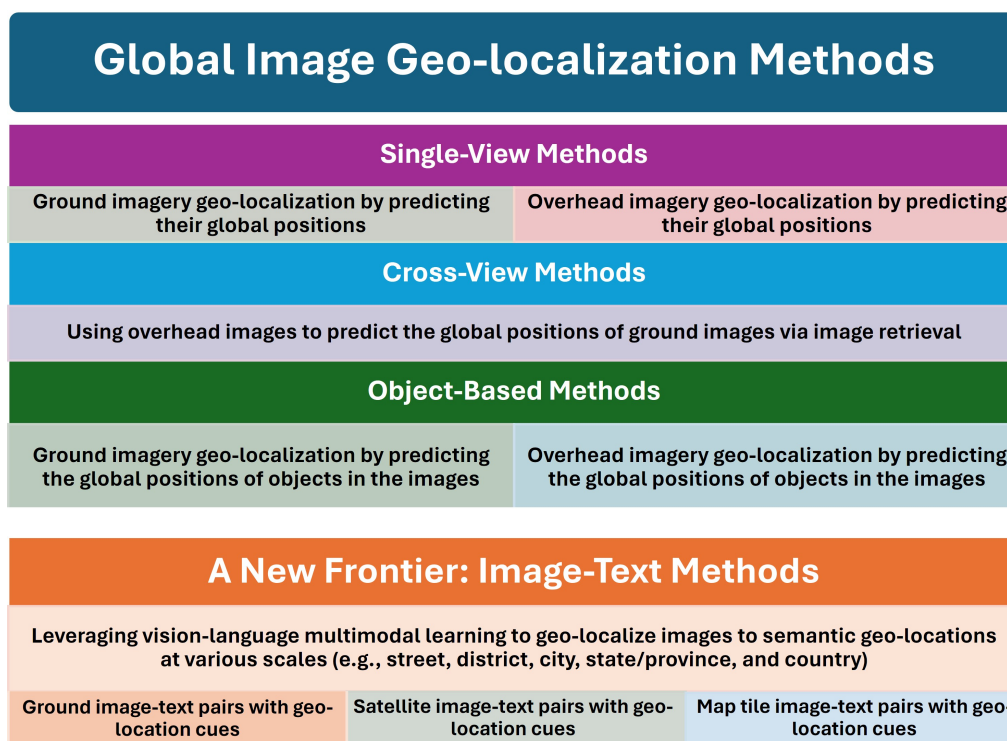


Figure 3.3: The development of global image geo-localization methods and a novel frontier method via vision-language multimodal learning based on image-text pairs with geo-location clues in both visual and textual formats.

The quest to determine the geographic origin of images, known as image geo-localization, has transcended simple curiosity to become a crucial aspect of various applications, from social media geo-tagging to image searching. Currently, there are two main approaches to address this task - one is image matching, and the other is classification-based. The former approach uses geo-tagged images as references and infers the

geo-location of a query image based on its most similar reference images (e.g., Hays and Efros (2008, 2015)), which are *cross-view methods* fraught with limitations (e.g., the sheer computational resources required and the matching difficulty for images taken from different perspectives or in varied conditions), while the latter approach partitions the world map into grids and trains a model to classify a given query image to one of the grids, known as *single-view methods* (e.g., Weyand et al. (2016); Seo et al. (2018)) (Figure 3.3). However, even for the images that are geo-localized correctly within the specified distance threshold using this classification approach, there is still uncertainty in determining their semantic geo-locations, especially when a grid intersects multiple administrative areal units.

Later, with the advent of deep learning, the focus has shifted towards creating models capable of understanding and extracting the fine-grained visual objects that correlate with geo-locations (i.e., *object-based methods*), by training convolutional neural networks (CNNs) and more complex architectures such as Generative Adversarial Networks (GANs), significantly improving the geo-localization accuracy (e.g., Chaabane et al. (2021); Wilson et al. (2022)).

Despite these advancements, the challenges of geo-localizing images remain across a wide range of scales, from the street to continental levels, especially in situations where distinct architectural or natural landmarks are sparse or when images are taken from different perspectives or viewpoints (e.g., ground view versus overhead view). Additionally, the scale at which geo-localization is accurate varies widely, with most methods excelling either at local or broad scales, but not both. Another significant limitation is the reliance on large datasets of geo-tagged images that are not uniformly available across the globe, leading to models or algorithms being biased towards well-represented areas.

To address these enduring challenges, this research activity proposes

a novel frontier "*image-text multimodal approaches*" for global image geo-localization, including different types of worldwide geospatial images (e.g., street view images, satellite images, and map tile images), by leveraging state-of-the-art MFMs (e.g., CLIP and GPT-4V). These vision-language MFMs have been trained on vast worldwide images and their corresponding textual descriptions, enabling them to understand and interpret an image within the context of natural language, including geo-location clues to a global extent. In particular, by associating images with semantic geo-locations across different scales (e.g., street, district, city, and country), this approach can automatically build the links between the visual and textual understanding of geospatial representations at different scales, allowing for more scalable geo-localization solutions. Moreover, by learning the variability of geospatial semantics described in natural language from visual input (e.g., images collected from different perspectives, lighting situations, weathers, and seasons), this method theoretically can abstract away from the specific visual representation of geospatial features, enabling more accurate and robust geo-localization across a broader range of conditions. Moving from theory to practice, the following subsections will focus on the experimental results of zero-shot learning on CLIP for global image geo-localization.

## **Results and Analysis**

First, there are several key geospatial factors to consider when experimenting with image geo-localization, which are critical for researchers to gain a more in-depth understanding of how the zero-shot model performs differently based on different settings of geospatial factors, including:

- Spatial variability across different cities and regions;
- Temporal variability across years;

Table 3.4: CLIP’s Zero-shot Performance on Image Geo-localization

Image Type	Dataset	Accuracy					
		City		Country		Continent	
		CLIP	CLIP+ Image type	CLIP	CLIP+ Image type	CLIP	CLIP+ Image type
Street view images	Place Pulse 2.0 Mapillary	63.0%	<b>63.9%</b>	79.2%	80.3%	93.6%	93.7%
		22.8%	<b>24.0%</b>	52.7%	55.8%	80.9%	81.2%
Satellite images	zoom 11 (2014)	8.7%	<b>12.1%</b>	34.7%	42.4%	63.6%	71.3%
	zoom 12 (2014)	8.2%	<b>9.5%</b>	33.4%	38.7%	63.0%	68.7%
	zoom 13 (2014)	5.3%	<b>5.8%</b>	24.2%	25.4%	53.4%	58.6%
	zoom 14 (2014)	10.1%	<b>11.4%</b>	39.1%	46.5%	71.8%	74.9%
	zoom 11 (2023)	10.4%	<b>12.1%</b>	40.1%	45.6%	70.0%	77.1%
	zoom 12 (2023)	9.9%	<b>11.6%</b>	36.2%	43.0%	61.9%	69.0%
	zoom 13 (2023)	10.0%	<b>11.5%</b>	38.9%	46.0%	68.1%	74.1%
	zoom 14 (2023)	9.6%	<b>11.1%</b>	43.5%	49.7%	73.4%	77.8%
Map tile images	zoom 11 (2023)	23.3%	<b>28.8%</b>	63.7%	66.2%	84.5%	87.2%
	zoom 12 (2023)	20.6%	<b>25.8%</b>	65.7%	65.9%	88.0%	89.4%
	zoom 13 (2023)	16.6%	<b>18.2%</b>	63.3%	59.3%	84.7%	86.0%
	zoom 14 (2023)	13.3%	<b>14.4%</b>	58.8%	54.0%	80.1%	82.7%

- Perspective variability in geographic scales (e.g., zoom levels), views (e.g., ground-view versus overhead-view), and rendering styles (e.g., a satellite image versus a map tile representing the same area).

In this task, unfortunately, the spatial variability of the model’s geo-localization performance has not been investigated, which deserves future effort, as foundation models have the common issue of systematic biases, including the one in the spatial representations learned during pre-training.

Similarly, temporal bias may also exist because recent data have been more likely used for pre-training, with historical data largely neglected.

Specifically, temporal bias will be evaluated within the satellite images between the years 2014 and 2023.

As for perspective biases, satellite images at different zoom levels will be compared to evaluate the biases in different geographic scales, (the same as map tiles at different scales). Next, street view images and satellite images will be compared to evaluate the biases in different views. Lastly, satellite images and map tiles at the same zoom level within the same year will be compared to evaluate the biases in rendering styles.

A notable finding is identified from Table 3.4: **CLIP + Image type**, with "*This [image type] image was taken in [geo-location label]*" as textual input, outperforms **CLIP**, with "*in [geo-location label]*" as the textual input, regardless of image types and datasets, or geo-localization scales (i.e., city, country, and continent). This indicates that geospatial image types (e.g., "street view", "satellite", "remote sensing", and "map tile") are valuable components for CLIP to consider when determining the geo-location of an image.

**Street View Imagery Geo-localization.** As for the evaluation of Place Pulse 2.0 street view imagery geo-localization, 13,254 images across 56 cities worldwide are used as a zero-shot testing dataset, versus 21,200 street view images across 405 cities worldwide as the Mapillary testing dataset. **CLIP + Image type** achieves the accuracies of 63.9%, 80.3%, and 93.7% for Place Pulse 2.0 at the city, country, and continent scales, respectively, and it has a fair performance on the Mapillary dataset with the accuracies of 24.0%, 55.8%, and 81.2%, considering the significant spatial variability of this dataset.

Specifically, CLIP's visual feature space of the cities in Place Pulse 2.0 is plotted in Figure 3.4, showing that distinctive visual representations for most cities have been captured by its pre-trained image encoder, which is probably because CLIP is self-supervised via contrastive pre-training

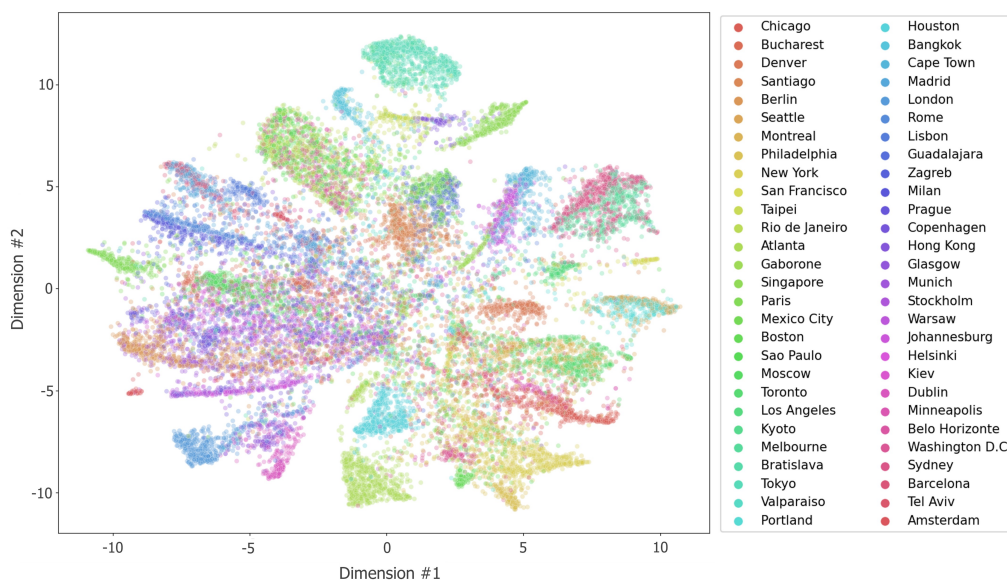


Figure 3.4: Using t-SNE to visualize CLIP’s visual feature space of the cities in the Place Pulse 2.0 dataset.

on enormous amounts of pairwise data, enabling visual features to be directly linked to city labels (in text) worldwide.

To further explore in detail how CLIP performs street view imagery geo-localization, some of its attention maps (i.e., weights of the last output layer) are visualized by averaging the weights over the 16 attention heads (Figure 3.5). Specifically, Figure 3.5(a) reveals that both buildings and vegetation can be main representations, while Figure 3.5(b) shows that the text in the image is strong signals (e.g., with characters in a specific language) and Figure 3.5(c) highlights highways and buildings. In street views where there are no such features, vegetation alone can be a distinctive representation for image geo-localization (Figure 3.5(d)). These results demonstrate that CLIP is able to leverage essential visual features for effective geo-localization.

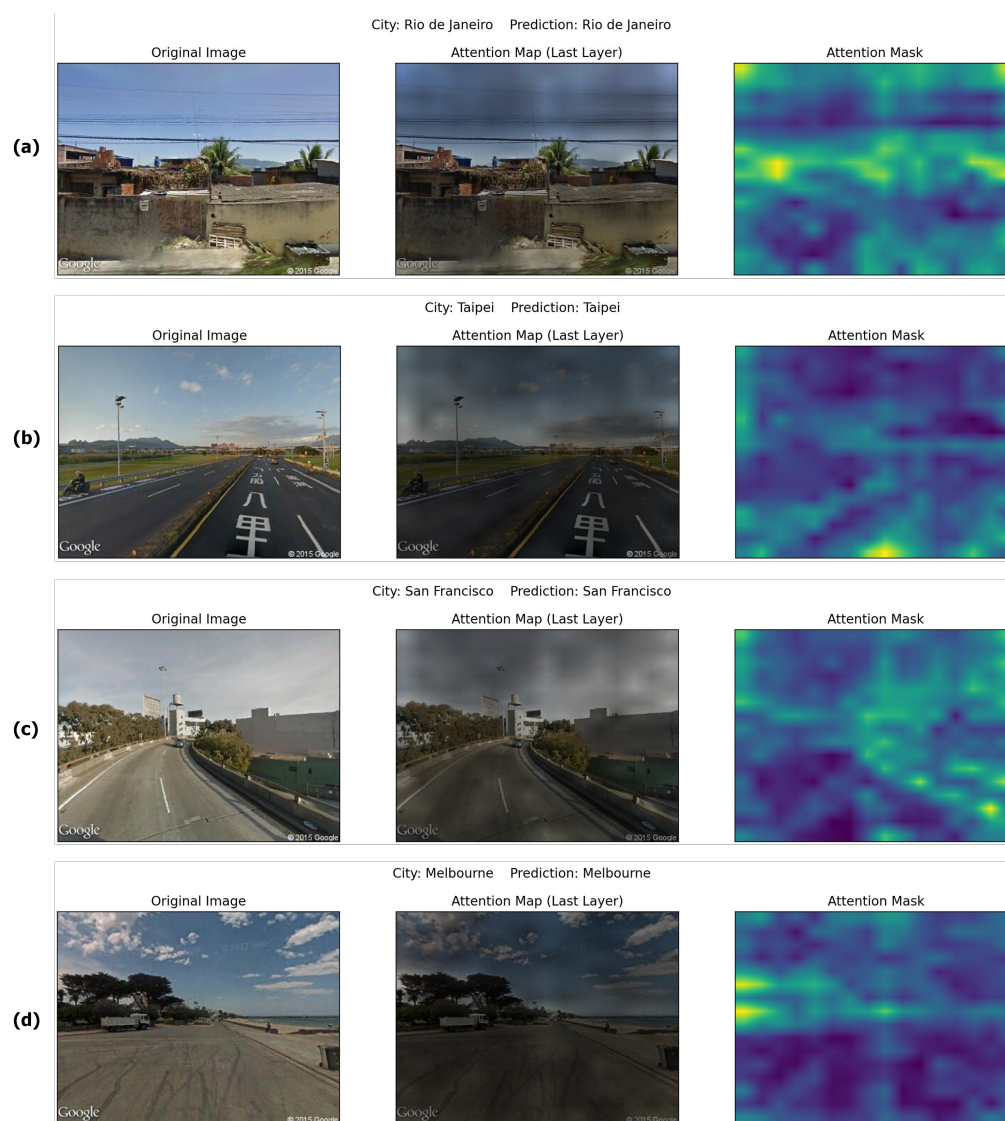


Figure 3.5: Examples of attention maps from CLIP's pre-trained image encoder to interpret street view imagery geo-localization.

**Satellite Imagery Geo-localization.** As for the evaluation of satellite imagery geo-localization, 43,894 images across 790 most populated cities worldwide from the years 2014 and 2023 are used as a zero-shot testing

dataset, with a range of image numbers from 3,104 to 25,010 at the zoom levels from 11 to 14.

Specifically, regarding the temporal differences between the two years, CLIP typically achieves better geo-localization performances in 2023 than in 2014 for most zoom levels (except for the 14-level), likely due to its pre-training images collected in more recent years. As for the spatial differences in zoom levels, CLIP generally performs better at coarse-grained levels (e.g., 11-level) than at finer-grained ones (e.g., 14-level), indicating that overhead imagery geo-localization may require a larger extent that can satisfactorily reflect the spatial pattern of urban structure (e.g., road networks and the arrangement of land uses) as a distinguishable visual clue, which leads to a more accurate prediction on the image’s city geo-location. As results, **CLIP + Image type** reaches fair accuracies of 12.1%, 45.6%, and 77.1% at the city, country, and continent scales, respectively, for the 11-level at 2023, considering the considerable amount of city labels being classified and the significant visual variability of this dataset (e.g., variation of urban landscapes). Yet, this also implies that there is more potential to improve CLIP’s geo-localization capability for satellite imagery in the future.

**Map Tile Imagery Geo-localization.** As for the evaluation of OSM map tile imagery geo-localization, 20,824 images across 790 cities worldwide at different zoom levels in 2023 are used as a zero-shot testing dataset, with a range of image numbers from 1,552 to 11,381 at the zoom levels from 11 to 14.

Specifically, regarding the spatial differences in zoom levels, similar to the results of satellite imagery geo-localization, CLIP also performs better at coarse-grained levels (e.g., 11-level) than at finer-grained ones (e.g., 14-level). This indicates that map tile imagery geo-localization also likely requires a larger extent that can satisfactorily reflect the overall spatial

pattern of urban structure to accurately determine the image’s city geo-location. As results, **CLIP + Image type** reaches fair accuracies of 28.8%, 66.2%, and 87.2% at the city, country, and continent scales, respectively, for the 11-level, which is significantly better than the prediction results of satellite imagery, achieving an accuracy improvement of 10.1% ~ 20.6%. This is probably because map tile images, compared with satellite images, can provide more visual clues for geo-localization, including map annotations (e.g., names of geospatial features and landmarks) and a unified symbolization of map features in terms of their land uses (e.g., urban areas, highway networks, water bodies, and vegetation) displayed by a clear and distinct color scheme. This also implies more opportunities to improve CLIP’s geo-localization capability for map tile imagery in the future.

**Online Imagery Geo-localization.** Another imagery dataset, i.e., Im2GPS3k (Vo et al., 2017), composed of 3,000 images collected from the Internet such as the social media platform Flickr<sup>2</sup>, is also utilized as a test set to evaluate CLIP’s image geo-localization ability, as this dataset has been widely accepted as a common benchmark due to its large size and high quality compared to other available datasets in 2017.

Accuracy is reported at two scales, i.e., city (25 Km) and region to country (200 ~ 750 Km). As results, CLIP achieves a competitive performance compared with the state-of-the-art models.

Yet, there has been a noteworthy argument in this field that the suitability of an image used for geo-localization tasks should be assessed Panagiotopoulos et al. (2022). For instance, an image of an ordinary mug or cat from the Internet probably provides few useful geo-location clues. Therefore, we may need to pause and revisit the suitability of using the Im2GPS3k dataset as a benchmark to evaluate a model’s geo-localization

---

<sup>2</sup><https://www.flickr.com/groups/flickrsocialmedia/>

Table 3.5: CLIP’s Zero-shot Performance on the Im2GPS3k Benchmark Geo-localization (Compared with the State-of-the-art)

Model	Accuracy	
	City	Region ~ Country
Zero-shot CLIP	32.4%	49.7%
ISN (Muller-Budack et al., 2018)	28.0%	36.6 ~ 49.7%
Translocator (Pramanick et al., 2022)	31.1%	46.7 ~ 58.9%
GeoDecoder (Clark et al., 2023)	33.5%	45.9 ~ 61.0%
GeoCLIP (Cepeda et al., 2023)	34.5%	50.7 ~ 69.7%
PIGEOTTO (Haas et al., 2023)	<b>36.7%</b>	<b>53.8 ~ 72.4%</b>

ability, especially for large-scale image geo-localization via deep learning. This typically involves training large models with massive random online images without considering the task-specific context. However, images that are irrelevant to a given task will be less likely to contribute to the model’s understanding of geo-locations and geospatial representation learning (e.g., animal images are probably not useful for geo-localization). As such, we need to ponder on the effectiveness and validity of this common but likely questionable practice. Do we need a model to geo-localize every single image in the world? Is it even practicable or reasonable? Is there any way we can improve data efficiency (e.g., using suitable input data such as street view images) instead of focusing on data quantity? The experiments of geo-localization based on global urban images (e.g., street view images, satellite images, and map tiles) rather than random online images in this study can at least prompt researchers to rethink these questions.

Another issue of the traditional Im2GPS3k benchmark lies in its unevenness of spatial representation. Heatmaps of Im2GPS3k’s spatial distribution compared to the ones of the curated urban image datasets used in this study are displayed in Figure 3.6. Note that a majority of Im2GPS3k images are located in North America and Europe, with very little repre-

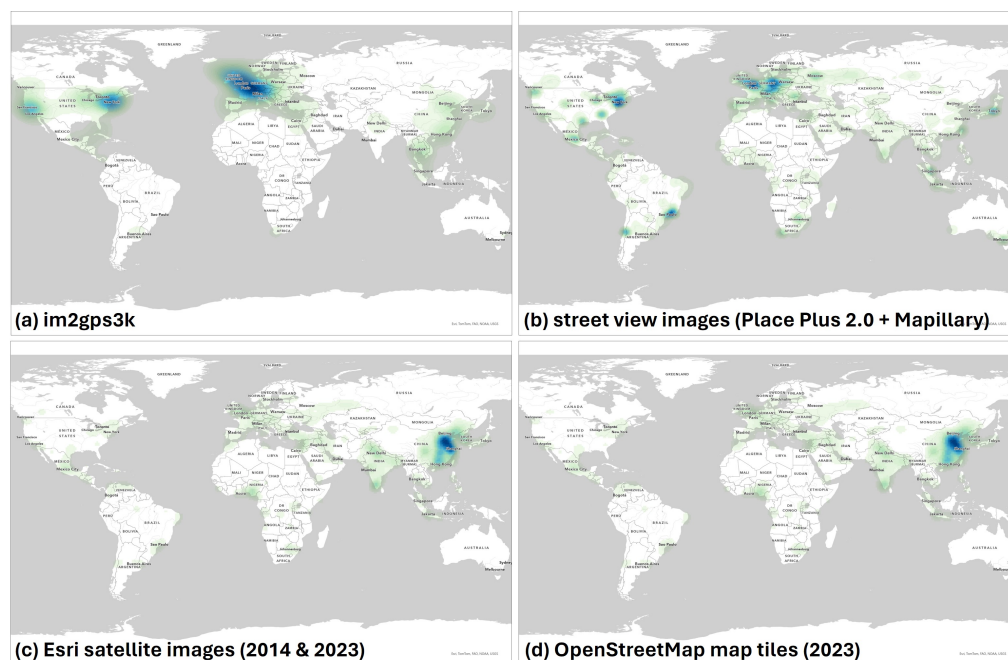


Figure 3.6: Spatial distributions of (a) Im2GPS3k, a traditional benchmark dataset and (b-d) global imagery datasets curated for this dissertation (e.g., street view images, satellite images, and map tiles). Im2GPS3k for testing geo-localization methods is heavily concentrated in the U.S. and Western Europe, while the curated dataset is more spatially even and covers a global extent, obtaining a better spatial representation for image geo-localization.

sensation in the other four populated continents (Figure 3.6 (a)), while the curated datasets provide a more spatially even sampling of the globe, with images on almost all populated urban areas, achieving a more comprehensive spatial representation (Figure 3.6 (b-d)), along with a variety of different image types.

### 3.4 Experiment #2 Urban Mixed Land Use Detection

Mixed land uses represent the integration of places with different socioeconomic functionalities, such as residential, commercial, industrial, transportation, recreation, and entertainment. Historically, mixed land use development has always been an essential part of US urban landscapes, and most cities have been developed with mixed land uses since the early twentieth century (Moos et al., 2018). Many existing studies have shown the important role of mixed land uses in urban development, such as improving neighborhood health conditions (e.g. body mass index (BMI) and obesity) (Brown et al., 2009), affecting housing values (Wu et al., 2018), reducing property and violent crimes (Zahnow, 2018), as well as alleviating automobile dependency (McCormack et al., 2001; Cervero, 1996). Over the past two decades, mixed land uses have been widely acknowledged to be essential for creating livable and sustainable communities, and are promoted by contemporary planning paradigms as an effective planning strategy for smart growth (Ye et al., 2005).

However, mixed land use detection has still remained a challenge because of the limitations of data sources as well as methods for data processing and analytics. Most land use data only record a single label (usually the most representative land use type) for each ground feature (e.g., a building), neighborhood block, or land parcel, while many of them possess multiple socioeconomic functionalities that support a variety of human activities (Gao et al., 2017; Pande et al., 2021). This is because land use data and maps are mainly obtained by visual interpretation based on overhead-view imagery such as remote sensing images (Zhang et al., 2018b; Helber et al., 2019; Ding et al., 2022; Yao et al., 2022; Yoo et al., 2022), which only provide the overhead view of the ground features and therefore hardly capture the detailed visual characteristics of mixed land

uses, especially for those multi-story properties. Moreover, this single-label issue has resulted in a persisting problem in mixed land use detection for decades – employing one-class classification methods for multi-class scenarios (Omrani et al., 2017). In other words, each target (e.g., a ground feature) is still considered as or predicted with only one land use class, and the estimated level of land use mixture relies on the spatial aggregation of the recorded or predicted land uses of the ground features within a larger land parcel (Abdullahi et al., 2015; Tian et al., 2017; Yue et al., 2017; Xing et al., 2018; Zhuo et al., 2019; He et al., 2020). Thus, using these incomplete data and inappropriate methods would produce biased, questionable results of mixed land use detection. A good sign has emerged that a few recent studies (Omrani et al., 2017; Liang et al., 2021; Wu et al., 2022) have realized and implemented the multi-label concept, although their land use validation datasets still suffer from: (1) being aggregated from fine-grained scales into coarse-grained ones (e.g., at a land patch or parcel level); (2) requiring tremendous human labors on visual interpretation from images; and/or (3) not being reusable for other cities or regions due to data unavailability.

To address these issues using overhead-view imagery, this study proposes to utilize ground-level street view imagery to predict mixed land uses due to its unique advantages. First, these images are mostly captured from side views and can provide more detailed, representative, and heterogeneous visual characteristics (e.g., the appearance of each floor of a property) related to urban land uses (Zhu et al., 2019), offering more useful information needed for achieving a more accurate detection of mixed land uses. Moreover, these images can be geo-located at a point level represented by geospatial coordinates (i.e., the finest-grained spatial scale), which can improve spatial resolution in identifying the mixture of land uses. In fact, street view images have demonstrated their effectiveness in differentiating and decomposing the compound uses of urban land (Li

et al., 2017; Zhang et al., 2017b; Kang et al., 2018b; Fang et al., 2022; Wei et al., 2022), emerging as popular research datasets due to their ability to offer more representative and fine-grained information compared to the overhead-view imagery, which is often used in coarse-grained scales (e.g., at the parcel or neighborhood level) (Castelluccio et al., 2015; Cheng et al., 2015; Tong et al., 2018).

Next, to effectively derive the mixed-use information, this study proposes a novel vision-language multimodal method that leverages street view images (vision part) and land-use contextualized prompts (language part, describing land uses with geo-locations at different scales), as this emerging multimodal technique has been proven to outperform the traditional image-based or text-based single-modal learning on various geospatial applications and tasks (He et al., 2020; Radford et al., 2021; Ji and Gao, 2023), including place type classification (Yan et al., 2018), (mixed) land use classification (Xing et al., 2018; Yokoya et al., 2018; Aswatha et al., 2020; Pastorino et al., 2022), remote sensing scene classification (Li et al., 2021b), and urban function and population mapping (Zhang et al., 2017b; Cao et al., 2018; Bai et al., 2023). The theoretical foundation of the proposed method is based on the fact that land-use scenarios serve as common descriptors for human land utilization, including categories such as residential and commercial. Hence, mixed land use detection can be equated to a task of Natural Language for Visual Reasoning (NLVR). In this context, it focuses on classifying an image by assessing the alignment between its descriptive land-use context and visual characteristics to assign the appropriate land use label(s). Next, the land-use context that has the largest similarity with the image scene can be predicted as the major land use type, and the similarity values of all land-use contexts can be used to evaluate the mixed degree of land uses for this image scene. As such, this method is also beneficial to the task of mixed land use mapping, enabling it at the point level (i.e., the geo-location of each ground image) and to

reflect land-use visual characteristics more comprehensively.

While there have been various multi-modal learning models and techniques that integrate natural language and computer vision and have shown significant promise for addressing classification-based challenges (El-Komy et al., 2022), the widely recognized model, i.e., CLIP (Radford et al., 2021), which was contrastively pre-trained on a vast and diverse dataset comprising 400 million text-image pairs, has stood out and showcased remarkable zero-shot classification performance on 16 distinct benchmarks. In addition, CLIP has demonstrated its capability to outperform traditional vision-based single-modal learning in mixed land use detection, by capturing the visual representation of street view images and successfully establishing a connection between this representation and the corresponding land-use textual representation (Wu et al., 2023). CLIP's superiority is supported by the fact that land-use texts are often part of the description of urban scenes, and some urban scenes are usually linked with certain descriptions of a land use type (e.g., a property with a chimney on the roof is more likely to be described as a "residential place"). In addition, street view images with a ground view tend to be more prevalent in conjunction with descriptive land-use texts in urban contexts, as they align more closely with human perception when contrasted with remote sensing imagery that offers a top-down bird's-eye view. This inherent advantage of the visual input further underscores CLIP's suitability for mixed land use detection via vision-language multi-modal learning.

Another model used for this experiment is GPT-4V, as it is currently one of the state-of-the-art MFMs on various downstream tasks (Achiam et al., 2023). Specifically, when performing spatial context-aware VQA tasks, GPT-4V has demonstrated a powerful capability in understanding the spatial context of various urban scenes, particularly for land use reasoning (Section 3.2.2), further highlighting its practicality for mixed land use detection.

## Study Area

This experiment is conducted within New York City (NYC), one of the most densely populated urban centers in the United States, distinguished by its pronounced prevalence of mixed land uses affirmed by the NYC Department of City Planning. A pivotal milestone in this context transpired in 1997, with the establishment of the mixed-use district designed to stimulate investments and enhance the vitality of neighborhoods, fostering a blend of residential and industrial activities in close proximity. Consequently, this initiative has culminated in the emergence of mixed land uses within designated zones, including land use purposes such as recreation, commerce, and industry.

## Datasets

**Place Pulse 2.0 Street View Images.** With this context, the street view images in NYC have been used to evaluate the zero-shot performance of CLIP and GPT-4V on mixed land use detection. Particularly, the applied dataset is a subset of the **Place Pulse 2.0** (Dubey et al., 2016) and consists of 3,398 Google Street View images captured from 2007 to 2012 (Section 2.2.1). These images were geo-located as coordinates with a latitude and longitude, sampled uniformly by grids that spatially overlap with the four primary boroughs of NYC, namely Brooklyn, Queens, Manhattan, and the Bronx. The spatial distribution of these images is mapped in Fig. 3.7, referencing Wu et al. (2023)'s study for mixed land use measurement and mapping using the same dataset.

**OpenStreetMap (OSM) Land Use Data.** OSM, a global platform for volunteered geographic information (VGI), provides public, cloud-sourced online maps. OSM data contain spatial geometries in vector format (i.e., point, line, or polygon) that are linked with various attributes, including land use information. Many previous studies have demonstrated the ef-



Figure 3.7: Spatial distribution of Place Pulse 2.0 street view images in NYC.

effectiveness of using OSM land use data to train models for urban land use classification (Fonte and Martinho, 2017; Fonte et al., 2020). In this work, the OSM shapefiles were downloaded from the BBBike website<sup>3</sup>. Specifically, after data preprocessing (e.g., removing the invalid features), the layers of "buildings" (277,875 polygons), "landuses" (21,362 polygons), and "natural" (9,233 polygons) in NYC were used as the land use validation data.

Additionally, since the original OSM land use types are defined as very detailed classes, we re-classified them into six categories that are more

<sup>3</sup><https://download.bbbike.org/osm/>

general, i.e., residential, commercial, industrial, greenfield, recreation, and transportation (Table 3.6), following similar data pre-processing procedures of previous studies (Zhang et al., 2017a; Wu et al., 2022, 2023) but focusing on more common classes for urban planning and design. Later, these six categories will be used to evaluate the model performance on mixed land use detection. Although the six categories do not appear mutually exclusive by concept (e.g., greenfield versus recreation), this proposed classification schema can differentiate between natural elements (e.g., greenfield, primarily denoting natural green spaces) and human-made ground features (e.g., recreation, principally representing human-made recreational facilities). Moreover, this approach aligns harmoniously with existing re-classification methodologies commonly employed within the field of urban planning and design, as discussed by Wu et al. (2022). Additionally, to maintain a balanced dataset across the re-classified land use categories, we analyzed the image distribution within each category, and the observed result exhibited a rational distribution, further substantiating the soundness and validity of the undertaken re-classification procedure.

Note that one common issue of using OSM land use data for validation is data sparsity. First, while OSM often provides rich geospatial information in developed regions (e.g., Europe and North America), the datasets are likely less comprehensive in many other regions. This unevenness of spatial representativeness should be acknowledged when using OSM in worldwide regions. Second, OSM land use data may not fully cover the entire study area, which means that mixed land use areas could not be captured in certain locations. Third, this data may not record multiple land use labels for every single multi-functional ground feature, leading to the missing of some land use(s) for certain features. Therefore, the model performance on mixed land use detection is possibly undermined due to the incompleteness of OSM land use data. However, we could still focus on the images that have been detected as mixed land use based on

Table 3.6: Re-labeling Land Use Categories for OSM Data

<b>Re-classified labels</b>	<b>Original OSM land use labels</b>
Residential	residential, garage, house, apartment, detached, terrace, shed, static caravan, shelter, dormitory
Greenfield	greenfield, grass, meadow, flowerbed, allotments, village green, farmland, farmyard, forest
Recreation	recreation, park, theatre, museum, attraction, grandstand, cinema, arts, sports, gallery, stadium, artwork, garden
Transportation	transport, bus station, bus depot, train station, traffic island, railway
Industrial	industrial, tank, warehouse, storage, brownfield
Commercial	commercial, nursing home, retail, restaurant, hotel, motel, office, food court, fast food, parking, bank, fuel, pharmacy, cafe, club, rental, studio, market, works, workshop, bar, pub

OSM labels, and then evaluate the model capacity in classifying mixed land use(s) by the degree of matching between the predicted labels and OSM ones. Specifically, the degree of matching equals the proportion of those correctly predicted images (i.e., the predicted land use is the OSM one, or one of them in the case of mixed uses).

Next, we need to carefully determine whether a location has mixed land uses, which will be considered as the validation to evaluate model performance. Fig. 3.8 displays the process of capturing mixed land uses of a location based on OSM land use data. For each location of an image, a buffer zone with a radius  $r$  (i.e., 50-meter or 100-meter) will be created and intersected with the polygons in the land use layer. If different land use types are identified in this buffer zone, then this location will be classified as mixed land uses.

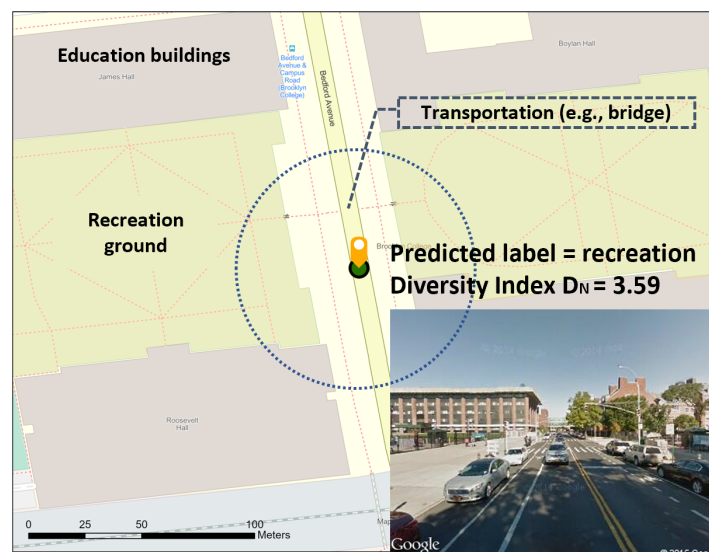


Figure 3.8: Capturing mixed land uses of a location based on OSM land use data with a buffer radius of 50 meters (an example of Bedford Avenue in Brooklyn College), with OSM land use tags such as recreation and transportation.

## Results and Analysis

To evaluate the effectiveness of the proposed zero-shot models, baseline model comparisons have been conducted using a traditional vision-based CNN model, ResNet-152 (He et al., 2016), pre-trained on ImageNet-1K\_V2 (Deng et al., 2009) as well as a vision transformer model, ViT (Dosovitskiy et al., 2020), pre-trained on ImageNet-1K\_V2 (Deng et al., 2009) and ImageNet2012 (Russakovsky et al., 2015), considering their well-established and strong performances across various computer vision benchmarks. These two baseline models are fine-tuned based on few-shot learning, and computed the cross-entropy loss after 32 training epochs, with a learning rate of  $10^{-3}$  and an optimizer of Stochastic Gradient Descent (SGD).

Fig. 3.9 illustrates the contrast in prediction performance of the proposed zero-shot models versus the few-shot ResNet-152 and ViT models.

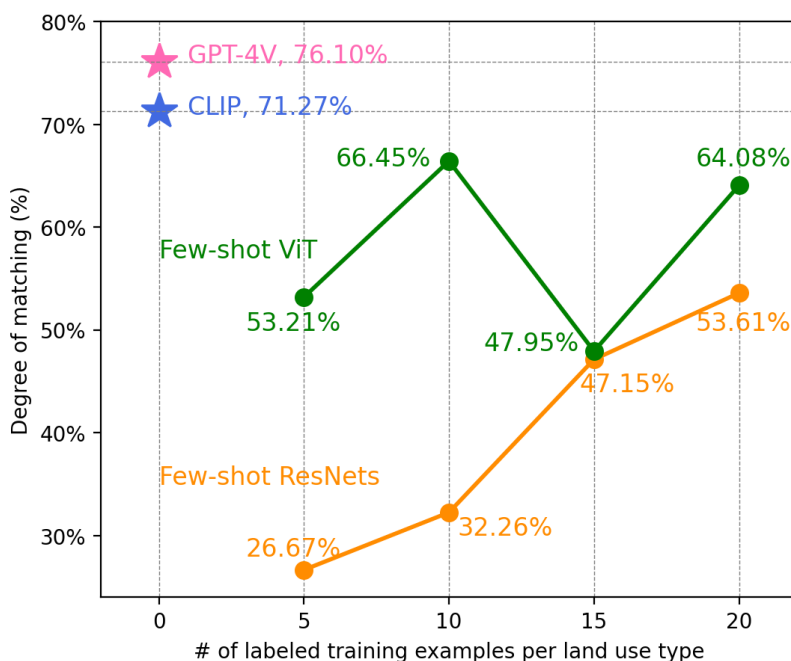


Figure 3.9: Comparison of model zero-shot performances on mixed land use detection. Both CLIP and GPT-4V outperform few-shot ResNet and ViT. Note that the fluctuation of the ViT performance may be due to the few-shot sampling of each land use category, leading to unstable predictive results.

As results, CLIP wins the few-shot ResNet-152 and ViT models with the degrees of matching as 71.27%, achieving 17.66% and 7.19% better than the 20-shot ResNet-152 and ViT, respectively. Moreover, GPT-4V outperforms CLIP by 4.83%, reaching a remarkable accuracy at 76.10%. These results underscore the competitive edge of language-image models in land use detection tasks, compared to traditional single-modal models that are only tailored for the input of imagery. For instance, CLIP yields visual features that are extracted in a contrastive learning manner and informed by the self-supervised pre-training on extensive text-image paired data, which also contain descriptions of land use scenarios or contexts. Consequently,

these features can be easily linked to distinct representations for each land use label (see Fig. 3.10). In contrast, traditional supervised single-modal models must derive visual features solely from input images, which poses the limitation that the labeled class of an image could be associated with many different visual objects detected from the image, in which the primary object(s) for the class may not be distinguished. This issue becomes especially pronounced in few-shot scenarios when the model has not yet established representative visual objects for each class. This result demonstrates the capacity of natural language to aid in referencing learnt visual objects, facilitating the feasibility of the model for land use detection via zero-shot transfer.

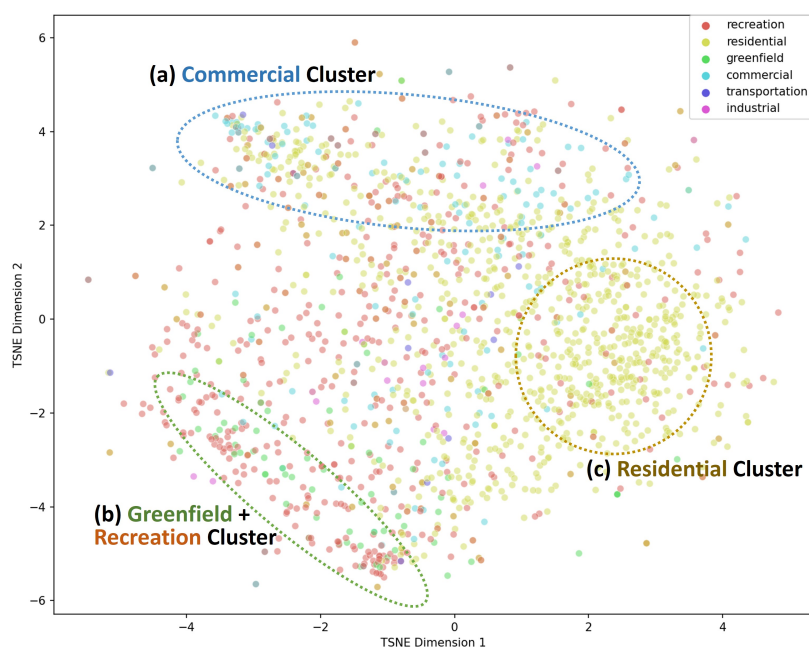


Figure 3.10: t-SNE visualization of the CLIP fusion space for different land use categories.

Specifically, CLIP’s visual feature space of land use labels is plotted in Figure 3.10, showing that a certain level of mixture in the visual repre-

sentations for land use labels has been captured by its pre-trained image encoder, as a result of CLIP's enormous amount of pre-training pairwise data that were learned contrastively, enabling the land-use visual features and their mixture representations to be linked to land use labels (in text).

Lastly, to obtain a better understanding of how CLIP interprets images and text together for predicting (mixed) land uses, some attention maps were plotted as examples to demonstrate the features that CLIP's attention was focused on. Specifically, Fig. 3.11 (a) shows a street view image located in residential areas with more diverse land uses (e.g., commercial and transportation), compared with the image shown in Fig. 3.11 (b) that is also located within residential areas but with a more monopolized view. Interestingly, the pattern of the attention mask is more scattered across the whole image in Fig. 3.11 (a), while CLIP exclusively focused on the chimney on the roof of a house in Fig. 3.11 (b), indicating that for those low-diversity areas CLIP is able to learn the key representation(s) of a given land use type (e.g., residential).

As for Fig. 3.11 (c), although the image is only labeled as industrial, yet we can see from the scene that there is also residential land use, which CLIP can identify. Similar situations can also be found in Fig. 3.11 (e) and (f), where the images have mixed land uses regardless of their OSM land use label(s), and CLIP is capable of interpreting their mixtures by putting attention to the representative features (e.g., learning commercial land use based on the front gates of stores and retails, or learning the residential land use based on the pattern of terrace buildings). In another scenario, Fig. 3.11 (d) and the attention mask show that recreational land use has been successfully detected based on the plants (green space) and certain features (e.g., the golf cart).

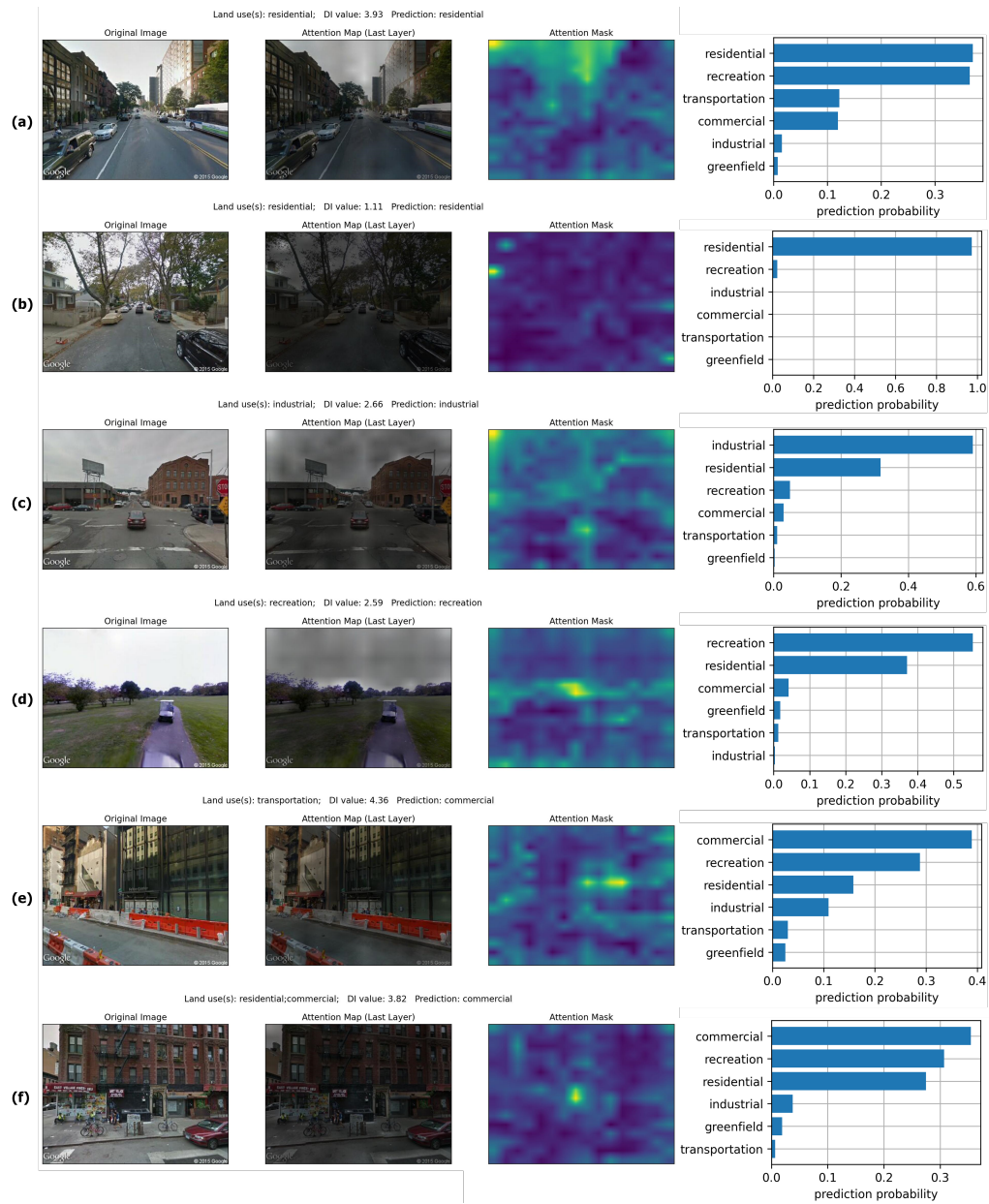


Figure 3.11: Examples of attention maps of CLIP's pre-trained image encoder for land use interpretation.

### 3.5 Experiment #3 Urban Perception Prediction

Urban perception prediction involves the use of computational models to predict human perceptions and subjective experiences of urban environments. This interdisciplinary field merges urban studies, psychology, and computer science to understand how people perceive urban spaces in terms of safety, beauty, liveliness, and other subjective variables. The ultimate goal is to inform urban planning and policy decisions, enhance urban design, and improve the quality of life in urban areas. Traditional approaches have relied on surveys and human observations, which are time-consuming, costly, and inherently limited in scope. With advancements in data availability and computational technologies, there has been a shift towards using data-driven approaches, leveraging large-scale urban data (images, videos, audio, social media content, etc.) to predict urban perceptions.

Recent approaches to urban perception prediction have predominantly focused on using machine learning techniques, especially deep learning, to analyze urban data. These methods typically aim to train models on datasets that are quantitatively annotated with human perceptions (e.g., ratings or scores of an urban perception variable), so that the models can learn to interpret the intricate patterns associated with these subjective ratings. Particularly, image-based models have been prominent, using street view imagery to predict how people perceive different urban environments, while text-based models, mostly trained on big data from social media or review platforms, have also been salient by their ability to capture perceptions expressed by urban residents and visitors.

While these approaches have significantly advanced the field, two research gaps and limitations have been identified. First, some existing models were trained on data from specific cities or regions, limiting their generalizability across dynamic urban contexts and diverse social and cultural circumstances. Second, most studies in this field focused on a sin-

gle data modality (e.g., images or text), overlooking the rich, multimodal nature of urban data that could provide a more holistic view of urban perception.

To address these limitations, this study proposes a novel multimodal learning method for urban perception prediction, by integrating multiple data modalities (e.g., images and text) at a global scale to gain a more comprehensive and robust prediction result via MFMs (e.g., CLIP and GPT-4V). Theoretically, this multimodal method is particularly suited for urban perception prediction for several reasons. First, by combining vast, global data from diverse sources and in different modalities, MFMs such as CLIP and GPT-4V can learn more generalizable features that are not specific to a particular type of data input (e.g., images or text), nor to a specific city or region. Second, multimodal data provides a richer and more comprehensive representation of a given urban environment, capturing different dimensions of urban life that a single modality might miss, and thus potentially uncovering unexpected, more integrated insights into urban perception by identifying factors that are not evident when analyzed in isolation. Furthermore, such multimodal input enables models to better accommodate the complexity of human perception by mimicking the human learning manner, which essentially leverages multiple sensory feelings to receive and interpret the multi-faceted urban experiences.

### **Datasets and Evaluation Metrics**

The **Place Pulse 2.0** dataset (Dubey et al., 2016) mentioned in Section 2.2.1, is a crowd-sourced dataset containing 110,988 images and 1,170,000 pairwise comparisons between these images provided by 81,630 online volunteers, and the comparisons were based on six human perceptual variables: safe, lively, boring, wealthy, depressing, and beautiful. For example, one comparison of a pair of two images can be "which place looks safer?" regarding the variable of "safety." The volunteers would

select one of the three options, the left image, the right image, or "equal," to indicate their perceptual judgment. Figure 3.12 displays an example of a two-image comparison.



Figure 3.12: An example of a two-image comparison regarding the "safety" variable.

Next, with the rating results in the form of two-image comparisons, the Microsoft Trueskill algorithm (Herbrich et al., 2006), a Bayesian skill rating system, is adopted to compute the rating score of each perceptual variable for each image (Dubey et al., 2016).

Particularly, to evaluate CLIP's zero-shot prediction performance on each perceptual variable, the distribution of the Trueskill scores will be used to create  $k$  classification categories based on  $k$ -means clustering algorithms (Hartigan and Wong, 1979; Likas et al., 2003; Sinaga and Yang, 2020). For example, when predicting the "safety" variable, the assigned label as part of textual input for CLIP's pre-trained text encoder can be one of three pre-defined categories, e.g., [*not safe*], [*safe*], or [*very safe*], if  $k = 3$ . Finally, for each image, the six generated labels will be the ground truth to examine the CLIP's zero-shot prediction performance on urban perception for the six perceptual variables (as a classification task).

As for the evaluation of GPT-4V's zero-shot performance, a specific VQA task has been run on GPT-4V to assess its ability for street view

image reasoning in Section 3.2.2 (Table 3.1), asking "On a scale of 1 to 10 (with 1 being the lowest and 10 being the highest), please rate the image with an integer in the six perceptive aspects: safety, liveliness, boringness, wealthiness, depression, and beauty. Please also include the reasons for your ratings." Therefore, the predicted perception scores from GPT-4V will be evaluated based on the regression metrics (e.g., Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE)) as well as the Spearman and Pearson Correlation Coefficients for examining whether the predicted scores are correlated with human-rated scores.

Note that the volunteers who participated in this **Place Pulse 2.0** project were from 162 countries, with a good mix of residents from both developed and developing countries. To explore the potential biases in the collection that may come from the demographics of the volunteers, a correlation significance test was conducted in a previous study (Salesses et al., 2013; Dubey et al., 2016). It is worth noting that individual preferences for urban appearance were not driven by participants' age, gender, or location, indicating that there is no significant demographic bias in the dataset. Furthermore, the internal consistency of the ratings in the dataset was also tested by the inter-user reproducibility and transitivity, both of which were found to be high (Salesses et al., 2013).

## Results and Analysis

As results, CLIP's zero-shot performance achieved a range of accuracies from 25.7% to 37.1% among the six perception variables, with the lowest accuracy for "liveliness" and highest accuracy for "beauty" (Table 3.7). These results will be used as the baseline to evaluate the performance of GeospatialCLIP in Chapter 5.

As for GPT-4V, the best zero-shot performance is for "wealthiness" with the lowest values of MAE, MSE, and RMSE, while its prediction

scores reflect a higher positively correlated relationship with the human-rated scores for "safety," "beauty", and "wealthiness," especially for "safety," reaching a Spearman’s correlation coefficient value of 0.384 (representing a positive monotonic relationship) and a Spearman’s correlation coefficient value of 0.404 (representing a positive linear relationship).

Table 3.7: CLIP’s and GPT-4V’s Zero-shot Performances on Urban Perception Prediction

Variable	CLIP	GPT-4V				
	Accuracy	MAE	MSE	RMSE	Spearman	Pearson
Beauty	<b>37.1%</b>	0.210	0.065	0.256	0.380	0.392
Boringness	28.2%	0.195	0.060	0.246	0.158	0.167
Depression	30.4%	0.242	0.084	0.291	0.216	0.232
Liveliness	25.7%	0.364	0.161	0.401	0.231	0.231
Safety	33.2%	0.179	0.047	0.217	<b>0.384</b>	<b>0.404</b>
Wealthiness	30.7%	<b>0.155</b>	<b>0.037</b>	<b>0.194</b>	0.352	0.360

Additionally, GPT-4V’s zero-shot performances are compared with one of the state-of-the-art models, i.e., Fine-tuned BERT (Ma and Wu, 2023) based on the metrics of MSE. As results, GPT-4V outperforms Fine-tuned BERT significantly for all six perception variables (Table 3.8), indicating GPT-4V’s powerful capability in zero-shot transfer learning.

Table 3.8: GPT-4V’s Zero-shot Performances on Urban Perception Prediction Compared with the State-of-the-art

Variable	GPT-4V’s MSE	Fine-tuned BERT’s MSE (Ma and Wu, 2023)
Beauty	<b>0.065</b>	0.43
Boringness	<b>0.060</b>	0.53
Depression	<b>0.084</b>	0.58
Liveliness	<b>0.161</b>	0.47
Safety	<b>0.047</b>	0.58
Wealthiness	<b>0.037</b>	0.55

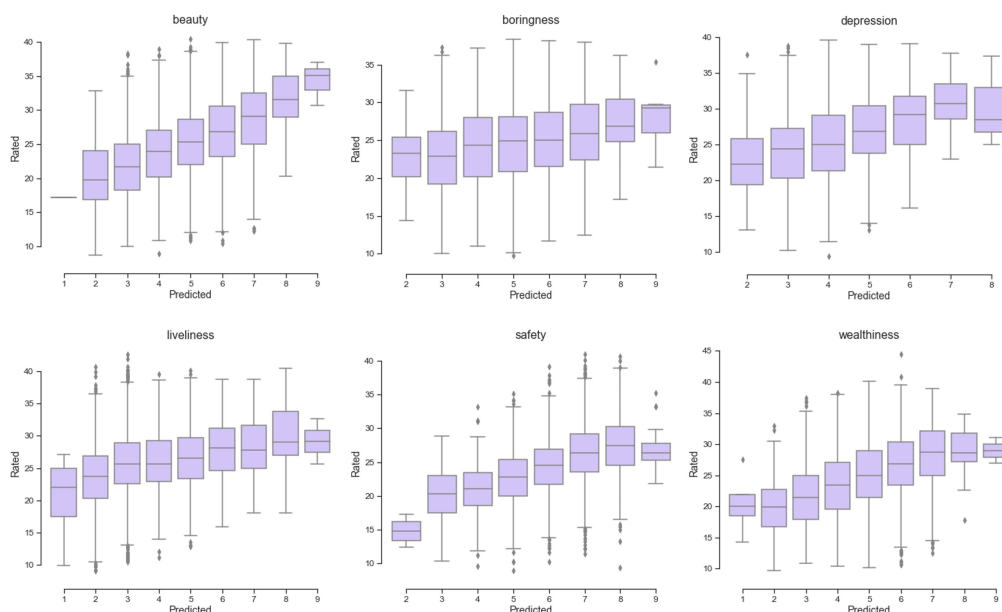


Figure 3.13: Boxplots of GPT-4's predicted scores versus human-rated scores of the six urban perception variables.

Lastly, boxplots of GPT-4's predicted scores versus human-rated scores of the six urban perception variables have been created and displayed in Figure 3.13, comparing the prediction values against human-rated ground-truth.

Specifically, the "beauty" plot spans from a predicted score of 1 to 9, with an overall gradual upward trend, suggesting that higher predicted scores are generally associated with higher human ratings. The spread of the boxes and whiskers indicates variability in this association, with a few outliers (depicted as individual points) indicating cases where this association deviates significantly from the median.

Similar to the "beauty" plot, the predicted scores for "boringness" are spread across a range from 2 to 9. The median lines are moderately level, which suggests that there is not a strong trend between the predicted scores and the human ratings. Meanwhile, the spread of the human ratings is

quite wide, especially for middle predicted scores, indicating a higher variability in human opinions on what is considered boring.

Next, the "depression" plot shows predicted scores from 2 to 8 and the median lines also appear level, similar to "boringness," indicating no strong upward trend. There is a wider variability in human ratings among the areas where the predicted "depression" scores are lower, in which humans likely have varied impressions related to depression.

The predicted "liveliness" scores range from 1 to 9, and the median lines slightly trend upward, implying that higher predictions may fairly correlate with higher human ratings of liveliness, although there's considerable variability for lower to middle predicted scores, and some outliers suggest that these areas were rated very differently from the median.

The "safety" boxplot has predicted scores from 2 to 9 and shows a clearly upward trend in the median lines, indicating that higher predicted safety scores tend to correspond with higher human ratings. However, there are numerous outliers, especially for middle to higher predicted scores, indicating some disagreement between the predictions and human ratings on safety.

Finally, the "wealthiness" boxplot ranges from predicted scores of 2 to 9, with an overall upward trend in median lines. The variability in human ratings seems to decrease with lower or higher predicted scores, suggesting a possible consensus on the perception of wealth in lower- or higher-rated areas.

### **3.6 Conclusions and Discussion**

This chapter has intensively explored the zero-shot learning capabilities of MFMs, specifically focusing on CLIP and GPT-4V, and their application in understanding and interpreting geospatial imagery. This investigation provides valuable insights into the data utilized, the methodologies adopted,

and the outcomes of three specific tasks: global image geo-localization, mixed land use detection, and urban perception prediction.

### **Data Insights**

The study utilized a comprehensive collection of geospatial imagery, including street view images, satellite images, and map tiles, sourced from geo-locations worldwide. This rich dataset enabled a robust examination of the MFMs' ability to generalize across different spatial contexts and interpret complex urban landscapes. The diversity and quality of the data played a critical role in benchmarking the models' performance, underscoring the importance of well-curated datasets in advancing GeoAI research.

### **Methodological Insights**

The methodology centered on zero-shot learning, allowing the MFMs to apply their pre-trained knowledge to new, unseen geospatial tasks without further training or fine-tuning. This approach highlighted the models' inherent capacity for cross-modal understanding and their remarkable potential to extract meaningful information from multimodal data. Specifically, this chapter showcased how advanced MFMs could navigate the complexities of geospatial image reasoning in a zero-shot learning scenario.

### **Task-specific Insights**

**Global Image Geo-localization** Both models demonstrated an outstanding ability to identify geo-locations from images alone, suggesting their profound understanding of visual cues related to urban landscapes, including architecture styles, road or highway networks, greenery and its spatial

arrangement, spatial patterns from both the ground view and overhead view, etc.

This task also highlighted the potential to enhance MFMs' spatial-context awareness to a more precise level, which is a promising direction in the future, especially for satellite images and map tiles. Also, while this research activity used the levels of city, country, and continent to geo-localize images, it would be intriguing to explore MFMs' ability to geo-localize images at a finer-grained level, e.g., the district level, or even the street level.

**Mixed Land Use Detection** The analysis revealed the models' capability to discern mixed land uses within urban environments, revealing their incredible understanding of urban areas and the spatial context of human activities. To continue this line of research, some potential directions are identified as the worth of a future effort. First, the current study area is NYC, while it could be extended to a global scale, given that the datasets applied (e.g., street view images and OSM land use information) are publicly available worldwide. Second, the current classification scheme focuses on only six major categories, while in the future more nuanced classes could be utilized to perform more precise mixed land use detection. For example, residential properties could be further classified as apartments, condos, single houses, etc. Similarly, commercial places could be further categorized as restaurants, supermarkets, cafés, bookstores, pharmacies, etc., which may be accomplished by leveraging object detection algorithms.

**Urban Perception Prediction** The models can fairly predict urban perceptions based on street view imagery, indicating a certain level of comprehension of the visual elements that influence human perceptions (e.g., safety, beauty, and liveliness). However, the outcomes imply the great potential to improve MFMs' understanding and prediction of these perception vari-

ables, by acquiring more insights into how urban spaces are perceived by both inhabitants and visitors. In fact, there has remained a huge black box for understanding how various environments represented in multimodal data affect humans' impressions and feelings towards different perception variables, which requires knowledge from multiple domains, such as socio-psychological sciences, urban planning, and computer sciences.

In conclusion, the exploration of the zero-shot learning of MFMs for geospatial image reasoning has revealed significant insights regarding the intersection of MFMs and GeoAI. The findings from this chapter not only demonstrate the models' proficiency in handling complex geospatial tasks, but also illuminate the path forward for applying multimodal learning in the geospatial domain. The intricate relationship between data quality, methodological rigor, and task specificity forms the cornerstone of this research activity, shedding light on future advancements in GeoAI.

## 4 TOWARDS SPATIAL-CONTEXT PROMPT TUNING FOR MULTIMODAL LEARNING: GPT-4V AUGMENTED SPATIAL-CONTEXT PROMPT TUNING ON MFMS FOR GEOSPATIAL APPLICATIONS

---

### 4.1 Introduction

The integration of spatial-context awareness into prompts for multimodal learning, particularly with the usage of MFMs, has opened new fields in current GeoAI research (Wu et al., 2023). However, one of the challenges lies in how to effectively integrate spatial context with the textual input of MFMs. Prompt tuning techniques have been applied to MFMs, which involve adjusting the textural inputs to a pre-trained model to achieve better performance on specific tasks. However, these mainstream techniques may not fully capture the various aspects of spatial context that could facilitate the model’s spatial understanding. This limitation underscores the need for the examination of spatial-context prompt tuning, a method designed to enhance the model’s ability to understand and utilize spatial information more effectively.

Spatial contexts are inherently ambiguous and can vary significantly depending on temporal dynamics, socio-cultural factors, and environmental differences. This variability presents a challenge for MFMs, as the models need to learn how to flexibly discern variations in spatial relationships and contexts during data interpretation. Current models may struggle to account for this variability, leading to less accurate or contextually inappropriate outputs for geospatial applications. For example, when geo-localizing satellite images in Section 3.3, CLIP typically achieves better performances in 2023 than in 2014 across zoom levels, as well as performing better at coarse-grained levels (e.g., 11-level) than at finer-grained

ones (e.g., 14-level). If such spatiotemporal dynamics can be presented in the prompts used as textual input for MFMs, then the model can learn more useful task-specific clues in textual features and connect these extra features with targeted visual features, potentially boosting the model's performance on geospatial applications.

Based on the identified challenges and the great potential of spatial-context prompt tuning, this research activity aims to investigate how to execute prompt tuning with spatial context, which is augmented by the VQA outputs from GPT-4V (Section 3.2.2), to enhance MFMs for geospatial applications, driving forward the capabilities of MFMs to better understand and interact with the spatial dimensions of our world. Regarding the rest of this section, first, the current status of prompt engineering with pros and cons is investigated, especially for geospatial applications, and then the theory and practice of designing spatial-context aware prompts is demonstrated, focusing on how they can address the limitations of prompt tuning techniques to enhance MFMs' performance for geospatial applications.

### **4.1.1 Prompt Tuning**

Prompt tuning, also known as prompt engineering, has represented a compelling frontier in the evolution of AI, especially within the domain of natural language processing (NLP) and the recent advent of LLMs and MFMs. This technique has emerged as pivotal for adapting large pre-trained models to specific tasks without the need for extensive re-training or fine-tuning of the entire model architecture. Instead, it is a means of shaping the textual input of language models by providing specific, contextualized information, and thus can enhance the interpretability of language models or multimodal foundation ones (Radford et al., 2019; Brown et al., 2020).

Specifically, prompt tuning by nature can be considered as an adapting

approach for pre-trained models (e.g., GPT series) to perform various or even unexplored tasks by carefully designing input prompts. Unlike formal fine-tuning, which adjusts the weights of a model across all or some layers to optimize model performance on a specific task, prompt tuning focuses on crafting the input prompt or adding a small number of trainable parameters tied to the prompt, aiming to better extract and leverage the immense knowledge encoded in pre-trained models and guiding them to apply their learned capabilities to different tasks. As a consequence, it can effectively reduce computation costs simply by manipulating the textual inputs of a model instead of updating its weights, which is especially practical for those recently published large models with millions of parameters. For example, ChatGPT 4<sup>1</sup> has an enormous count of 100 trillion parameters, and training such a model from scratch can take months. More importantly, the employment of prompt tuning can maintain the generalizability capability of the pre-trained model (Zhou et al., 2022).

As for its mechanisms, prompt tuning operates in two main types: manual and automated. Manual prompt tuning is to create handcrafted prompts that can navigate the model to obtain a better understanding of the input and the task. These prompts can include specific contexts, instructions, questions, or structures that align with desirable model output. Although the effectiveness of manual prompt tuning may depend on the expertise of the user in framing prompts that can resonate with the model's pre-training for a given task during the process of trial and error, yet manual prompt tuning is advantageous in terms of simplicity, customization, and transparency. Specifically, these prompts are devised explicitly to guide the model toward the preferred outcome, allowing a direct understanding of what information is provided to the model and how the model is instructed to process that information. In addition, these prompts can be easily understood and modified by humans, making the

---

<sup>1</sup><https://openai.com/product/gpt-4>

reasoning behind the model's responses more transparent. Moreover, users can iteratively refine their prompts based on observed model performance and insights into the task, and this iterative process fosters a deeper understanding of how different prompt formulations affect the model's outputs, contributing to the overall interpretability of the approach.

Automated prompt tuning, on the other hand, introduces an optimization process where a small set of parameters associated with the prompt are learned through back-propagation. This subset of parameters, often referred to as "soft prompts" or prompt embeddings, is optimized to guide the model towards desirable outputs for a specific task. Although this method allows for more scalable and adaptable prompt tuning, a huge drawback is its lack of interpretability. The model may discover prompts optimized for a given task, but cannot explain why it chose those embeddings. Similar to many deep learning models, "soft prompts" are opaque and cannot provide deep insights into the model's internal mechanisms. Thus, automated prompt tuning essentially creates a layer of abstraction between the user and the model, as these "soft prompts" or learned embeddings are not directly interpretable in human terms, making it difficult to discern how they guide the model's responses.

### 4.1.2 Prompt Ensembling

Furthermore, another notable technique of prompt tuning is known as prompt ensembling, which focuses on integrating multiple manual prompts to address the challenge of capturing the multifaceted aspects of a given task. By sophisticatedly combining a variety of prompts, this approach encourages the pre-trained models to generate rich, contextually relevant, and comprehensive textual output for the given task.

Existing studies have employed a variety of methodologies to implement prompt ensembling. For instance, DiVeRSe ("**D**iverse **V**erifier on **R**easoning **S**teps") (Li et al., 2022) is a method that improves the reliability

of answers in a threefold manner: (1) building multiple prompts to generate diverse completions; (2) training a voting verifier to distinguish good answers from bad answers; and (3) testing the voting verifier to check the correctness of reasoning steps. Another similar but more advanced method is Ask Me Anything (AMA) Prompting (Arora et al., 2022). The AMA approach uses an LLM to generate multiple prompts, instead of just using different few-shot exemplars, and estimates the dependencies between different prompts it creates, which will be used to weight prompts appropriately when aggregating their answers. In practice, majority voting may be the simplest way to weight prompts, while a softmax function is also a commonly used method (Radford et al., 2019; Brown et al., 2020).

Yet, determining the optimal combination of prompts and managing potential redundancy or conflicting information have remained concerns. The design of prompts within ensembles needs to ensure a balanced coverage of different aspects of a given task. Future research could explore the integration of automated methods for selecting and weighing prompts within ensembles. In addition, investigating the influence of prompt diversity and redundancy on model performance could provide insights into effective prompt-tuning strategies.

### **4.1.3 Spatial-context Aware Prompts**

Traditional prompts often lack spatially contextual awareness, leading to inefficiencies and inaccuracies in the model output, particularly for geospatial applications. To address this issue, spatial-context aware prompts have emerged as a promising approach, leveraging spatially contextual cues to enhance the outcomes for geospatial applications (Wu et al., 2023).

Spatial context refers to the geographical and physical environment-related information that can significantly influence the interpretation of data. For instance, in geospatial applications, understanding the location, topography, and other spatial factors is crucial for tasks such as environ-

mental monitoring, urban planning, and disaster response.

Correspondingly, spatial-context prompt tuning involves incorporating spatial metadata and geographical features directly into the prompts, enabling the pre-trained model to comprehend these spatial contexts. In other words, spatial-context aware prompts refer to prompts or queries that are generated with useful spatial contexts for geospatial applications. Unlike conventional prompts that rely solely on pre-defined or static templates, spatial-context aware prompts leverage spatial (or spatiotemporal) information and factors to tailor prompts so that they are more relevant and insightful to the given geospatial task.

Intuitively, prompt tuning with a focus on spatial-context awareness can offer strategic guidance to harness the power of pre-trained models for generating more insightful and spatial-context relevant descriptions, by tailoring prompts to capture domain-specific features and spatial contexts. Literature has shown that evaluating a constrained set of keywords and prompts can help better explain and interpret learned models (Caliskan et al., 2017). Thus, to design spatial context-aware prompts, initial approaches may involve formulating prompts that explicitly direct models to focus on a set of urban key objects and their spatial relationships, by providing spatial context cues or constraints (e.g., geo-locations at different scales, the detected objects and their spatial patterns, and land use description and reasoning) to ensure that the generated text aligns with desired outcomes for geospatial applications.

However, crafting prompts that strike a balance between specificity and creativity is a challenging task. Overly prescriptive prompts might hinder the generation of diverse and imaginative outputs, while vague prompts could lead to inaccurate or irrelevant descriptions. Therefore, more efforts should be put into ensuring that prompts effectively guide models in producing relevant, coherent, and spatially contextualized output.

## 4.2 Methodology

### 4.2.1 Spatial-context Prompt Tuning on MFMs

Based on the descriptions of these three types of geospatial images obtained in Section 2.2.2, contextualized prompts for MFMs can be designed and evaluated in the same three tasks in RA-1. Specifically, these prompts should at least reflect six dimensions, i.e., geospatial image types, geo-localization clues, spatial patterns, LULC, urban perception, and urban development, which are outlined in detail in Table 4.1. This investigation scheme is particularly helpful for explaining how each type of spatial context can facilitate or impede the model performance in the three tasks, further enhancing the interpretability of the prompt-tuned model as well as offering a more rigorous understanding of how prompts influence model behavior and decisions.

Table 4.1: Dimensions of Spatial Context-Aware Prompts

Dimension	Types of Spatial Context	Description	Prompt Examples
Geospatial image types	<image type>	"street view", "satellite", or "map tile."	This <image type> image is [label].
Geo-localization clues	<geo-location>	Formatted as [city] or [city, country, continent].	This place is [label] in <geo-locations>.
	<geo-features>	Distinctive features that can provide geo-location clues.	This place is [label], with <geo-features>.
	<geo-reasoning>	Explaining why this image is from its city.	This place is [label]. <geo-reasoning>.
Spatial patterns	<object patterns>	Objects and their spatial patterns.	This place is [label], showing <object patterns>.
	<urban patterns>	Urban structure and spatial patterns.	This place is [label], showing <urban patterns>.
LULC	<land use>	LULC description.	This place is [label] for <land use>.
	<land use changes>	LULC changes over 10 years.	This place is [label], with <land use changes>.
Urban perception	<perception>	Describing why this image looks [perception label].	This place is [label]. <perception>.
Urban development	<growth>	Urban growth description and prediction over 10 years.	This place is [label]. <growth>.

<sup>\*</sup>Note: Content within "[ ]" is the assigned label for a given task.

### 4.2.2 Workflow

To evaluate the performance of spatial-context aware CLIP, experiments have been conducted in the same three tasks in Chapter 3 (RA-1). The prompt-tuning workflow of CLIP is demonstrated in Figure 4.1, outlining

the process for integrating geospatial images and contextualized prompts to experiment with the tasks. In the workflow, GPT-4V is utilized to craft these contextualized prompts, as it has exhibited powerful capabilities to handle complex spatial reasoning and understanding in the experiments in Chapter 3.

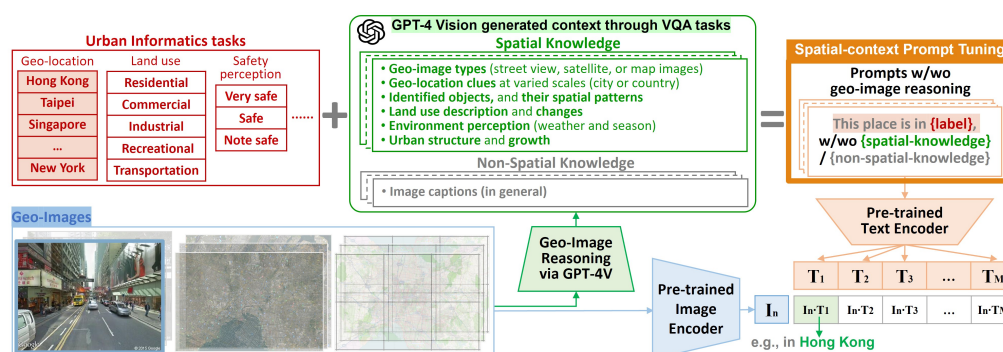


Figure 4.1: The workflow of spatial-context prompt tuning on CLIP, augmented by GPT-4V generated contexts.

Specifically, this workflow can be divided into several components, each representing a step or module in the process:

1. **Geospatial images:** This part presents the types of images for the three tasks, including geo-tagged street view images, satellite images, and map tiles, which provide visual input for CLIP's pre-trained image encoder as well as for GPT-4V to perform VQA tasks for geospatial image reasoning.
2. **Geospatial image reasoning via GPT-4V:** This step is originally part of RA-1 to evaluate the zero-shot performance of GPT-4V in the tasks of mixed land use detection and urban perception prediction. Specifically, the generated outcomes can be classified into two types of contextualized contexts (e.g., spatial versus non-spatial ones) and

used for spatial and non-spatial context-aware prompt tuning in this chapter.

3. **GPT-4V generated spatial and non-spatial contexts:** This component illustrates the specific types of spatial and non-spatial contexts that GPT-4V has generated through VQA tasks as mentioned above, including image caption, image type, geo-location clues at varied scales, identified objects and their spatial patterns, land use description and changes, urban perception, and urban structure and growth, etc.
4. **Targeted urban informatics tasks:** This step focuses on extracting labels for each task, which will be part of the textual input along with the GPT-4V generated contexts as listed above. For example, for the task of mixed land use detection, the classification scheme includes land use labels (e.g., residential, commercial, industrial, recreational, and transportation), and then a prompt, e.g., "This place is residential because <land use>" where <land use> contains the description of land use information and patterns generated by GPT-4V, can be constructed and used as the textual input for CLIP's pre-trained text encoder.
5. **Spatial-context prompt tuning:** This step involves the tuning of prompts to include or exclude the spatial and non-spatial contexts augmented by GPT-4V in Step 3, in order to generate the desired output. This process can show how prompts can be adjusted to focus on different aspects of the context.
6. **Pre-trained image and text encoders:** Given the textual input (i.e., various prompts) and visual input (i.e., various geospatial images), this part represents two steps: (1) the text encoder embeds the prompts, and the image encoder embeds the images; (2) the text and

image embeddings then will be fused as dot product to calculate their similarity. The cell with the highest similarity value will indicate the predicted label associated with the image.

## 4.3 Experiment #1 Global Image Geo-localization

### Street View Imagery Geo-localization

Table 4.2: CLIP’s Prompt Tuning Results of Street View Imagery Geo-localization

Prompt Type	Prompt	Accuracy on Place Pulse 2.0		
		City	Country	Continent
No prompt	–	63.0%	79.2%	93.6%
Non-spatial	Image Caption	31.5%	46.9%	64.7%
	Image type	63.9%	80.3%	93.7%
Spatial	Beauty	63.5%	78.5%	91.4%
	Boringness	64.6%	80.0%	90.4%
	Depression	65.9%	80.3%	92.5%
	Liveliness	65.0%	81.1%	91.8%
	Safety	63.9%	81.1%	90.4%
	Wealthiness	65.7%	81.4%	90.0%
	<b>Land use</b>	<b>67.5%</b>	<b>83.6%</b>	92.1%
	<b>Spatial patterns</b>	66.1%	79.3%	<b>94.3%</b>
	Geo-features	32.1%	48.2%	65.0%
	Geo-reasoning	18.2%	33.2%	47.9%

Table 4.2 provides the results from CLIP’s prompt tuning for street view imagery geo-localization. The accuracies are given for the identification of city, country, and continent based on different types of prompts.

- **No prompt:** Without any prompts, the model achieves 63.0% accuracy for the city, 79.2% for the country, and 93.6% for the continent. This indicates the model's inherent ability to geo-localize street view images to some extent based on the learned features during pre-training.
- **Non-spatial (Image Caption):** Using a non-spatial prompt like image captions significantly reduces the model's accuracy across all geographic scales compared to the baseline "no prompt." It decreases to 31.5% for the city, 46.9% for the country, and 64.7% for the continent. This implies that non-spatial captions can confound the model rather than assist in geo-localization.
- **Spatial Prompts:**
  - **Image type:** The prompt concerning the type of image marginally improves the model's accuracy for city and country identifications and slightly for continents, suggesting that knowing the image type can offer slight benefits for geo-localization tasks.
  - **Perception-based prompts (Beauty, Boringness, Depression, Liveliness, Safety, Wealthiness):** These prompts are subjective descriptions of human perceptions towards urban environments and they all show an improvement over "no prompt" for city-level geo-localization, indicating that urban perceptions can be tied to specific city characteristics. They generally show slight improvements for country-level geo-localization and slight decreases for continent-level geo-localization.
  - **Land use:** This prompt gives the highest increase in accuracy for city-level geo-localization to 67.5% and substantial improvements at the country level as well to 83.6%, suggesting that understanding land use is very effective for the model in determining geo-locations.

- **Spatial patterns:** This prompt significantly improves the accuracy for city-level and continent-level geo-localization to 66.1% and 94.3%, respectively, implying that at a city or continental scale, spatial patterns are more distinct and recognizable by the model, but they might not provide enough detail for accurate country-level identification.
- **Geo-features and Geo-reasoning:** These two prompts, which involve describing geographical features and the reasoning about the location based on those features, result in lower accuracies compared to "no prompt," especially for "geo-reasoning," which shows a substantial decrease. This could indicate that these specific prompts provided by GPT-4V are not as effective as others in aiding geo-localization. Based on a further investigation, when GPT-4V is asked to find geo-location clues, it is likely to provide such a response as "*I'm unable to confirm the specific reason why this image was taken in [geo-location] without more context, as there's no distinctive landmark or feature that would distinctly indicate it was taken there.*" Therefore, these two prompts may not provide much insight into street view imagery geo-localization tasks.

In conclusion, spatial context-aware prompts generally improve geo-localization accuracy, although the effectiveness of each prompt type varies by scale. In contrast, the non-spatial prompt "image caption" appears to hinder the model's performance. Notably, perception-based prompts are more useful at the city level; "land use" is the most effective prompt at smaller scales; and "spatial patterns" are particularly useful at the largest scale.

Table 4.3: CLIP’s Prompt Tuning Results of Esri Satellite Imagery Geo-localization (at 11 and 14 Zoom Levels in Years 2014 and 2023)

Prompt Type	Prompt	Accuracy		
		City	Country	Continent
No prompt	–	10.5%	43.7%	75.4%
Non-spatial	Image Caption	1.9%	17.7%	51.9%
Spatial	<b>Image type</b>	<b>12.1%</b>	<b>50.4%</b>	<b>78.4%</b>
	Land use	11.4%	48.4%	75.2%
	Land use changes	11.1%	48.7%	75.3%
	Spatial patterns	10.9%	48.0%	75.4%
	Geo-features	9.6%	44.3%	73.0%
	Urban growth	9.8%	44.8%	73.2%

### Satellite Imagery Geo-localization

Table 4.3 presents the results of CLIP’s prompt tuning on Esri satellite imagery geo-localization. The accuracy percentages are categorized based on whether the task is to identify the city, country, or continent of the imagery. Here’s the interpretation of the results with "no prompt" serving as the baseline:

- **No prompt:** The model’s accuracies without any prompt, as a baseline, are 10.5% for the city, 43.7% for the country, and 75.4% for the continent.
- **Non-spatial (Image Caption):** When using an image caption as a prompt, which normally contains a general description of an image without spatial information, there’s a noticeable drop in accuracy across all scales, with the most significant decrease at the city level (10.5% to 1.9%). This suggests that non-spatial prompts may confuse the model when attempting to geo-localize images, especially at finer-grained zoom levels such as the city level.
- **Spatial Prompts:** almost all spatial prompts improve the accuracy

over "no prompt" at the city and country levels, while almost all maintain or slightly decrease the accuracy at the continent level:

- **Image type:** The prompt of "satellite" has the highest accuracy improvements across all scales. It is particularly effective at the country level (50.4%) and continent level (78.4%), showing that knowing the type of image helps CLIP better contextualize the geo-locations of images, as "street view" also produces this improvement.
- **Land use:** The "land use" prompt also improves accuracy, especially at the country (48.4%) and continent levels (75.2%), indicating that knowing the land use context can aid in better geo-localization.
- **Land use changes:** A prompt detailing "land use changes" shows similar improvements to the "land use" prompt, suggesting that changes from historical to current land use situations can contribute to geo-localization accuracy.
- **Spatial patterns:** The "spatial patterns" prompt improves accuracy at the country (48.0%), and is slightly better than the baseline at the city level while being stable at the continent level, which could mean that recognizing patterns such as urban structures and layouts helps in identifying locations, albeit not as significantly as other prompts.
- **Geo-features:** This "geo-features" prompt shows a slight improvement in country identification but a minor reduction in accuracy at city and continent scales. Thus, for satellite images, distinctive geographic features are more indicative of country-level geo-localization.
- **Urban growth:** The country-level accuracy with the "urban growth" prompt is higher than "no prompt" but the accuracies of

the other two levels are lower. This suggests that understanding how urban areas have developed over time aids in the country-level geo-localization.

In summary, spatial prompts generally improve the model’s geo-localization accuracy, especially at more granular geographic scales (city- or country-level), compared to the baseline "no prompt." On the contrary, "Non-spatial" information such as image captions can actually decrease the geo-localization accuracy, which emphasizes the importance of spatial context in satellite imagery geo-localization tasks.

### Map Tile Imagery Geo-localization

Table 4.4: CLIP’s Prompt Tuning Results of OpenStreetMap Map Tile Imagery Geo-localization

Prompt Type	Prompt	Accuracy		
		City	Country	Continent
No prompt	–	12.7%	<b>59.7%</b>	83.5%
Non-spatial	Image Caption	2.6%	22.8%	61.1%
Spatial	Image type	<b>15.1%</b>	57.2%	<b>85.5%</b>
	Land use	14.5%	54.9%	85.2%
	Spatial patterns	13.8%	55.9%	85.0%
	Geo-features	11.3%	51.9%	81.9%

Table 4.4 presents the results of using different types of prompts to assist CLIP in geo-localizing OpenStreetMap Map Tile imagery, with accuracies measured at city, country, and continent level.

- **No prompt:** Without any prompt, CLIP achieves 12.7% accuracy for the city, 59.7% for the country as the best, and 83.5% for the continent. This sets the baseline to evaluate CLIP’s prompt-tuned performance without any additional contextual information.

- **Non-spatial (Image Caption):** Introducing a non-spatial prompt drastically reduces accuracy at all levels (2.6% for city, 22.8% for country, and 61.1% for continent), indicating that non-spatial information can mislead the model away from correct geo-localization.
- **Spatial Prompts:**
  - **Image type:** This "image type" prompt improves city identification (15.1%) most significantly and provides the greatest boost for continent identification (85.5%) compared to "no prompt," while the country identification (57.2%) slightly drops. This suggests that information about the image type has certain relevance in aiding map tile geo-localization.
  - **Land use:** The prompt that specifies "land use" offers an improvement for city-level accuracy (14.5%), a decrease for country-level accuracy (54.9%), and a boost for continent-level accuracy (85.2%) compared to "no prompt," highlighting its usefulness particularly at the city level.
  - **Spatial patterns:** With this "spatial patterns" prompt, there are slight improvements in city-level accuracy (13.8%) and continent-level accuracy (85.0%), yet with a marginal decrease for country accuracy (55.9%). It implies that spatial patterns are somewhat helpful in geo-localization but not as effective as other spatial prompts.
  - **Geo-features:** Similar to the situations of street view images and satellite images, this prompt's accuracy decreases at all levels (11.3% for city, 51.9% for country, and 81.9% for continent) compared to "no prompt," due to the issue of GPT-4V not being able to provide geo-localization clues.

In summary, spatial prompts generally offer an improvement over "no prompt" for city-level and continent-level geo-localization in the CLIP model, although with varying effectiveness, while there is an overall accuracy decrease for country identification. Notably, non-spatial prompts again hinder CLIP's performance, indicating that relevant spatial information is crucial for map tile imagery geo-localization with CLIP.

## 4.4 Experiment #2 Urban Mixed Land Use Detection

Table 4.5 shows the accuracies of CLIP's prompt tuning for detecting mixed land use from street view images, with various prompts applied:

- **No prompt:** The model achieves 69.36% accuracy without any prompts. This serves as the baseline for comparing the effectiveness of other prompts.
- **Non-spatial prompts:**
  - **Image Caption:** Using an image caption as a prompt decreases accuracy to 56.00%, suggesting that non-spatial captions may not contain useful information for land use detection and can actually degrade performance.
  - **Non-spatial ensemble:** Referring to the work by Wu et al. (2023), combining various non-spatial elements (e.g., "*for [land use]*" or "*[land use] purpose*") with a softmax function to form an ensemble prompt slightly improves the accuracy over the baseline to 69.97%, indicating that a well-structured combination of non-spatial information can be beneficial.
- **Spatial prompts:**

Table 4.5: CLIP’s Prompt Tuning Results of Mixed Land Use Detection

Prompt Type	Prompt	Accuracy
No prompt	–	69.36%
Non-spatial	Image Caption	56.0%
	Non-spatial ensemble	69.97%
Spatial	Image type	32.77%
	Beauty	45.8%
	Boringness	60.4%
	Depression	62.7%
	Liveliness	66.8%
	Safety	64.1%
	Wealthiness	53.5%
	Land use	31.7%
	Spatial patterns	51.5%
	Geo-features	44.6%
	<b>City label</b>	<b>71.27%</b>

- **Image type:** This prompt indicating the "image type" significantly reduces accuracy to 32.77%, which could imply that the type of image alone (i.e., "street view") is not informative for mixed land use detection.
- **Perception-based prompts (Beauty, Boringness, Depression, Liveliness, Safety, Wealthiness):** These prompts generally deteriorate accuracy over the baseline, with *Beauty* leading to the worst decrease to 45.8%. This suggests that subjective perceptions of a street view cannot provide useful contextual clues

about its land use.

- **Land use:** Very surprisingly, directly using a "land use" prompt lowers accuracy to 31.7%, indicating that CLIP may be confused by too complex land use descriptors, although its capability of street view imagery geo-localization is significantly improved by "land use" prompts.
- **Spatial patterns:** The prompt related to spatial patterns gives an accuracy of 51.5%, which is lower than the baseline, indicating that spatial patterns alone do not contribute to determining mixed land use.
- **Geo-features:** This prompt also decreases accuracy to 44.6%, probably due to the same reason mentioned earlier (i.e., GPT-4V cannot provide sufficient information in this prompt).
- **City label:** The highest increase in accuracy is observed with the "city label" prompt, jumping to 71.27%, which indicates that knowing the city where the image was taken from provides significant contextual information that aids in land use detection.

In conclusion, while most prompts do not boost CLIP's performance in this experiment, the "city label" prompt and the "non-spatial ensemble" prompt contain contextual information that can assist CLIP in achieving the best results over the baseline for mixed land use detection.

## 4.5 Experiment #3 Urban Perception Prediction

Table 4.6 shows the accuracies of different prompts used by CLIP in predicting urban perceptions from street view images. Each column represents a specific perception variable, and the results are compared against the "no prompt" baseline:

Table 4.6: CLIP's Prompt Tuning Results of Urban Perception Prediction

Prompt Type	Prompt	Accuracy					
		Beauty	Boringness	Depression	Liveliness	Safety	Wealthiness
No prompt	–	37.1%	28.2%	30.4%	25.7%	33.2%	30.7%
Non-spatial	Image Caption	33.9%	30.7%	32.5%	32.9%	28.2%	35.7%
Spatial	Image type	34.3%	28.2%	30.7%	34.3%	26.1%	35.7%
	Beauty	27.1%	31.4%	36.1%	36.4%	22.1%	32.5%
	Boringness	31.1%	25.0%	34.3%	32.5%	24.3%	30.4%
	Depression	32.9%	31.8%	33.6%	35.7%	25.7%	35.4%
	Liveliness	31.4%	29.3%	33.6%	30.4%	24.6%	32.1%
	Safety	36.1%	29.6%	32.9%	28.9%	27.1%	29.3%
	Wealthiness	27.9%	26.4%	32.5%	32.9%	23.6%	34.3%
	Land use	30.0%	31.4%	32.1%	40.0%	31.1%	27.1%
	Spatial patterns	32.1%	29.6%	31.8%	24.3%	30.4%	34.6%
Geo-features	35.4%	31.1%	36.1%	38.6%	26.4%	32.5%	
Geo-location	35.0%	31.8%	31.8%	32.1%	27.1%	35.7%	

- **No prompt:** CLIP achieves accuracies ranging from 25.7% for *Liveliness* to 37.1% for *Beauty* without any additional context. It reaches the best results for *Beauty* and *Safety* while yielding the worst result for *Depression*. This means that all prompts decrease accuracies for *Beauty* and *Safety* compared to "no prompt," especially for *Safety*.
- **Non-spatial (Image Caption):** the effectiveness of using the "image caption" prompt varies compared to "no prompt," with the improvements seen in *Wealthiness* (35.7% up from 30.7%, also the best result for *Wealthiness*), *Liveliness* (32.9% significantly up from 25.7%), *Depression*, and *Boringness*, but with decreases in *Beauty* and *Safety*.
- **Spatial prompts:**
  - **Image type:** This prompt leads to an increase in accuracy for *Wealthiness* (35.7%) similar to the "image caption" and "geo-

- location" prompts (as the best results for *Wealthiness*), but shows different effectiveness in other variables, with some accuracies slightly higher and others lower than the "no prompt" baseline.
- **Perception-based prompts (*Beauty, Boringness, Depression, Liveliness, Safety, Wealthiness*):** Each of these prompts leads to varied accuracy changes:
    - \* **Beauty:** Surprisingly, "Beauty" as a prompt has the largest drop to 27.1% for itself. It is also the worst case for *Safety* (22.1%), while achieving the best result for *Depression* (36.1%).
    - \* **Boringness:** The "Boringness" prompt shows different effectiveness across perception variables, with some accuracies slightly higher and others lower than the "no prompt" baseline.
    - \* **Depression:** The "Depression" prompt reaches the best result for *Boringness*, and significantly improves accuracies for *Depression* itself, *Liveliness*, and *Wealthiness*.
    - \* **Liveliness:** "Liveliness" as a prompt exhibits various effectiveness across perception variables, with accuracies of *Boringness, Depression, Liveliness* itself, and *Wealthiness* slightly higher and others lower than the "no prompt" baseline.
    - \* **Safety:** The "Safety" prompt has different effectiveness across perception variables, with some accuracies slightly higher and others lower than the "no prompt" baseline.
    - \* **Wealthiness:** Similarly, the "Wealthiness" prompt also has different effectiveness in other variables.
  - **Land use:** This prompt increases *Liveliness's* accuracy to the

- best level at 40.0%, but shows different results for others, and having the worse case for *Wealthiness* (at 27.1%).
- **Spatial patterns:** The "Spatial patterns" prompt demonstrates different effectiveness across perception variables, with some accuracies slightly higher and others lower than the "no prompt" baseline.
  - **Geo-features:** Similar to the "Spatial patterns" prompt, this "Geo-features" prompt leads to varied accuracy changes compared with "no prompt" across the perception variables. It reaches the best result for *Depression* at 36.1%.
  - **Geo-location:** This prompt improves accuracies across several categories, matching the highest increase in *Wealthiness* (35.7%) and *Boringness* (31.8%) and showing boosted results in *Depression* and *Liveliness*.

In summary, for *Beauty* and *Safety*, the model performs best without any prompt. For *Boringness*, the best results come from prompts "Depression" and "Geo-location." Next, *Depression* sees the best results with the "Beauty" and "Geo-features" prompts. *Liveliness* sees the most significant improvement with the "Land use" prompt. The *Wealthiness* variable has three prompts tied for best, "Image caption," "Image type," and "Geo-location," which may offer contextual clues about economic conditions.

## 4.6 Conclusions and Discussion

This chapter explored the transformative impact of spatial-context prompt tuning on MFMs, specifically digging into its efficacy across three geospatial tasks. The innovative approach of integrating spatial context into prompts has not only enhanced the models' performance but also opened

up new avenues for GeoAI research (e.g., how to better leverage MFMs' capabilities for geospatial applications). The conclusions and discussion in this section will synthesize the advancements achieved and outline the fertile grounds for future work.

### **Advancements by Spatial-context Prompt Tuning**

In the rapidly evolving field of AI, spatial-context prompt tuning has been one of the latest innovations for the GeoAI domain, aiming to optimize the performance of AI models by leveraging the spatial context of prompts to enhance their interpretive and generative capabilities. By learning the intricacies of this approach and its impact on multimodal models, its advancements and effects can be outlined as follows:

- **Enhanced Comprehension:** Spatial-context prompt tuning allows AI models to better interpret multi-faceted relationships, resulting in improved comprehension of complex inputs, which is particularly beneficial for multimodal models working with visual and textual data.
- **Improved Prediction Performance:** By learning and understanding spatial relationships, MFMs with spatial-context prompts demonstrate better prediction performance across varying geospatial inputs and tasks. This indicates that spatial-context prompts can effectively facilitate MFMs' capacities to identify and adapt to new patterns, enabling them to be more suitable for geospatial applications with significant variability.
- **Efficiency in Learning:** Incorporating spatial context into prompts reduces the data requirement and training cost for achieving more accurate predictions. That is, based on the spatial cues from textual prompts, MFMs can recognize contextualized patterns and indications better, without the need for exhaustive training datasets.

In general, spatial-context prompt tuning represents a promising leap forward in GeoAI research, offering models the ability to incorporate spatial and contextual cues into their comprehension tasks, holding the potential to revolutionize how models interpret complex relationships, as well as enabling them to generalize across diverse inputs and to provide profound insights in a range of geospatial applications.

Specifically, the experiments in this chapter were conducted for three geospatial tasks, including global image geo-localization, mixed land use detection, and urban perception prediction, and the subsequent paragraphs have highlighted the significance of spatial-context prompt tuning for these tasks.

**Global Image Geo-localization** Spatial-context prompt tuning significantly advanced the MFMs' capacity to geo-localize images accurately, including street view images, satellite images, and map tiles. By embedding spatial context directly into the prompts, CLIP showed an improved understanding of geographical clues, such as recognizing landmarks, architectural styles, and urban land uses across different scenes, leading to a more accurate geo-location prediction.

**Mixed Land Use Detection** The application of spatial-context prompts (e.g., "city label") and the non-spatial ensemble prompt improved CLIP's ability to identify and categorize mixed land uses within urban street view settings. This indicates that the prompt-tuned model received contextual enhanced clues that enabled a deeper understanding of human activities on urban land and the complex interplay of mixed land uses. These results also showcased how to guide the application of MFMs to benefit urban land use detection in the future.

**Urban Perception Prediction** Incorporating spatial context into the prompts resulted in a more in-depth interpretation of urban environments, as evi-

denced by the model's ability to predict urban perceptions with higher accuracies in most cases. This task highlighted the potential of integrating MFMs with spatial-context prompt tuning to contribute to urban design and policy, by providing a more profound understanding of how elements (e.g., the status of buildings and utilities, brightness, and green spaces) affect different perception variables.

### **Future Work and Promising Directions**

The advancements demonstrated in this chapter not only underscore the effectiveness of spatial-context prompt tuning, but also underline several areas ripe for further exploration:

- **Dynamic urban environments:** future work can encompass more dynamic aspects of urban environments (e.g., weather and seasonal variations), which may add extra layers to spatial-context aware consideration for prompt tuning in the field of urban analytics. In addition, a prospective examination might focus on refining the data structures that can represent spatial patterns and relationships.
- **Cross-modal spatial reasoning:** the integration of additional modalities, such as audio or environment sensor data, can be investigated, alongside how they could be converted into meaningful textual description as an addition to spatial-context prompt tuning.
- **Prompt ensembling:** a further investigation on prompt ensembling may open up new opportunities to use MFMs for geospatial tasks, e.g., assigning different weights to prompts that reflect different aspects of spatial context, by quantitatively evaluating the effectiveness of each individual prompt when combining them.
- **Newer applications:** spatial-context prompt tuning has unlocked possibilities in fields where spatial-context reasoning is crucial, such

as autonomous navigation and robotics. By understanding spatial patterns in input data, MFMs can offer insights that were previously challenging to obtain.

In conclusion, the exploration into spatial-context prompt tuning has enriched our understanding of its impact on MFMs' performance on multiple geospatial tasks. By gaining a better understanding of different aspects of spatial context and their various influences on MFMs' prompt-tuned performance, we can anticipate a future where MFMs play a more vital role in assisting urban researchers, designers, and planners in an informed decision-making process.

## 5 TOWARDS GEOAI MULTIMODAL FOUNDATION

### MODELS: DEVELOPING GEOSPATIALCLIP AS A GEOAI VISION-LANGUAGE FOUNDATION MODEL

---

#### 5.1 Introduction

As AI continues to advance, there is a growing recognition of the importance of multimodal models that can effectively process and understand both visual and textual information, the two most common modalities as data input. To date, the integration of vision and language has held immense potential for many geospatial applications (e.g., navigation, urban planning, and disaster response). To harness this potential, researchers are working towards the development of GeoAI multimodal foundation models (GeoAI MFMs), which can seamlessly fuse spatial knowledge with both vision and language inputs.

In particular, GeoAI MFMs would represent a paradigm shift in spatial computing, enabling MFMs to understand and interpret geospatial information more comprehensively. Similar to generic MFMs, GeoAI MFMs not only leverage insights learnt from multimodal data interchangeably, which promotes transfer learning, domain adaptation, and generalization across diverse geospatial tasks and datasets, but also enhance human-AI interaction in a more intuitive way, by accepting multimodal queries as inputs as well as supporting multimodal output. On top of these advantages, GeoAI MFMs possess an enhanced spatial understanding and reasoning, allowing themselves to comprehend spatial concepts and relationships expressed in both visual and textual formats, and thus showing huge potential to boost varied geospatial analytical tasks (e.g., scene understanding, perception prediction, and land use classification). In other words, the significance of GeoAI MFMs lies in their capability to capture rich and spatially contextual

relationships and semantics from visual cues and linguistic information inherent in geospatial data, which holds substantial implications for the future of spatial computing and GeoAI development and thus deserves more research efforts for further exploration.

Within this context, this research activity aims to develop a vision-language foundation model tailored specifically for geospatial applications, i.e., GeospatialCLIP (i.e., GeoAI-empowered CLIP), which will emerge as a pioneering initiative. Building upon the success of CLIP, the proposed GeospatialCLIP will extend its capabilities to incorporate location-based reasoning, spatial context understanding, geospatial semantics, and human perceptual indications in urban areas.

The following subsections are organized in a structure that first highlights the importance and strengths of vision-language foundation models in the field of AI, along with its limitations and biases identified, especially for geospatial applications. Next, the advancements achieved by GeoAI for vision-language learning are then underlined, which leads to the integration of GeoAI knowledge and techniques for developing a GeoAI-empowered vision-language foundation model that can enhance various geospatial tasks.

### **5.1.1 Vision-Language Foundation Models**

The advent of vision-language foundation models (VLFMs) represents a significant milestone in the evolution of AI. These models, which comprehend and/or generate content that combines visual elements with natural language, have become instrumental in a variety of applications, from image captioning to sophisticated VQA tasks.

Basically, VLFMs leverage the synergistic potential of combining visual and textual data, offering a more comprehensive understanding of the given inputs and tasks. Models such as OpenAI's CLIP and Google's ALIGN have demonstrated remarkable capabilities in visual reasoning,

which is not possible for single-modal models (e.g., text-based or image-based ones). Similar to other foundation models, VLFMs are pre-trained on vast datasets with pairwise images and text, enabling them to learn a wide range of visual concepts and their linguistic descriptors and to be fine-tuned for various downstream applications.

However, due to their reliance on the diversity and quality of pre-training data, a significant concern is their inherent systematic biases, including the bias in their understanding of geospatial information. Specifically, VLFMs are often pre-trained on datasets that lack rich geospatial (meta)data, leading to challenges for applications that require precise geospatial understanding. For instance, in environmental monitoring or urban planning where the spatial distribution of features is critical, VLFMs may struggle to provide accurate insights due to the absence of evenly distributed geospatial information in their pre-training data. This limitation underscores the need for incorporating rich geospatial datasets in the pre-training phase to enhance VLFMs' spatial awareness more comprehensively.

### 5.1.2 GeoAI for Vision-Language Learning

To address these challenges, recent research studies have started to leverage GeoAI techniques to enhance VLFMs, by integrating geospatial (meta)data directly into the models' pre-training datasets or updating the models' architectures to better process geospatial information. For example, SkyScript (Wang et al., 2024) and RemoteCLIP (Liu et al., 2023a) have been developed and successfully equipped VLFMs (e.g., CLIP) with geospatial reasoning capabilities by incorporating location tags, satellite imagery, and geographically diverse text sources into their (re)pre-training process for remote sensing applications.

Another potential line of research can be focused on modifying the architectures of VLFMs to better accommodate geospatial information. A

vision paper by Mai et al. (2022) stated that a notable limitation of VLFMs is their insufficient integration with vector data, which is essential for spatial reasoning and facilitates the coordination between various modalities in GeoAI. To address this, the authors suggest modifying the model architectures to enrich multimodal geospatial representations with location encoding (e.g., geo-tagged texts and remote sensing images) before combining these multimodal data. By leveraging geographic footprints (vector data) embedded in location encoders, it is more straightforward to synchronize geo-tagged textual data with geospatial images, enabling VLFMs to improve their spatial reasoning capabilities and facilitate knowledge transfer between modalities.

Yet, one major concern is the scarcity of applicable geospatial image-text pairs for a wide range of geospatial tasks, which makes it difficult to develop a GeoAI VLFM that can generalize well. Moreover, the input datasets should also incorporate spatiotemporal dynamics of geospatial information, which is important for VLFMs to capture and adapt to changes over time and space.

## 5.2 Methodology

### 5.2.1 Geospatial Image-Text Pairwise Input

As discussed above, the process of geospatial dataset curation is the key to developing GeoAI VLFMs. In fact, this dissertation research has started this curation process since RA-1, in which GPT-4V has generated extensive high-quality, large-scale geospatial image-text pairs that hold satisfactory spatial representation across the globe and different tasks. Specifically, this dataset contains geotagged street view images, satellite imagery, and map tiles, along with their corresponding textual descriptions with diverse aspects of non-spatial and spatial contexts. Moreover, GPT-4V's remarkable zero-shot performances in RA-1 for the tasks of mixed land

use detection and urban perception prediction have manifestly validated the effectiveness of this geospatial multimodal dataset.

Therefore, this curated dataset sourced from RA-1 would be leveraged to develop GeoAI VLFMs. Specifically, the training sets defined in Table 3.3 shall be employed for model pre-training. Next, to evaluate the proposed GeoAI VLFMs' zero-shot learning and spatial-context prompt tuning performance, the testing sets would be utilized for the same three tasks in RA-1 and RA-2.

## 5.2.2 Spatially Explicit Contrastive Learning on CLIP

First, the foundation of our methodology is built on the CLIP architecture, known for its robust performance in associating images with textual descriptions through contrastive learning. With this backbone, this dissertation study introduces a spatially explicit textual module in the stage of text processing, designed to manipulate textual input for CLIP, allowing the model to understand and encode spatial relationships and contexts explicitly in the text embedding.

Specifically, the spatially explicit textual module is integrated with CLIP's text encoder. It first extracts different types of textual descriptions paired with a geo-tagged image, which are then concatenated with the labels of a given task to create contextualized text input. Later, the textual input becomes text embeddings after being encoded by the text encoder. Next, the image embeddings and the formulated text embeddings would be fused as dot products to measure their similarity, which is used to update the parameters in both text and image encoders based on the pre-defined contrastive loss computation. This integration allows the model to not only learn from the visual data, but also from the spatial context reflected by the textual data.

As for the training details, I adopted an initial learning rate of  $lr = 1e-7$ . The model was trained with 50 epochs, allowing sufficient time to learn

spatially contextualized text embeddings effectively and to integrate these insights with the image embeddings. Regarding the pre-training data, 180,119 pairs of geospatial image-text records are used, including 43,202 pairs of satellite images (14-level in 2023 and 2014), 30,626 pairs of map tiles (14-level), 20,334 pairs of Mapillary’s street view images, and 85,957 pairs of Place Pulse 2.0 street view images, to reach a balance of learned representations across different geospatial tasks. A batch size of 32 was used, optimized with Adam, with parameters as  $\text{betas} = (0.98, 0.999)$ ,  $\text{eps} = 1e - 10$ ,  $\text{weight decay} = 0.0$ , to achieve better computational efficiency and meet the need for a diverse set of inputs for effective contrastive learning. Lastly, CLIP’s original contrastive loss function was implemented, as its mechanism has already penalized incorrect geospatial associations between the contextualized text and the geospatial image. That is, the contrastive loss function brings closer the representations of "positive" pairs of geospatial text and images, while pushing apart those of "negative" pairs.

### 5.2.3 Workflow

To evaluate the performances of the proposed GeoAI VLFM (i.e., GeospatialCLIP), experiments have been conducted in the same three tasks as RA-1 and RA-2. The workflow of (1) spatially explicit contrastive pre-training CLIP to develop GeospatialCLIP and (2) zero-shot learning on GeospatialCLIP is demonstrated in Figure 5.1. Specifically, the outcome of spatially explicit contrastive pre-training is the spatially augmented text encoder and image encoder, which can better extract geospatial representations in both text and image formats and capture a more in-depth understanding of the geospatial relationships (e.g., similarity or dissimilarity) between visual and textual features. After pre-training in (1), GeospatialCLIP is further evaluated or tested on datasets it hasn’t seen during pre-training (zero-shot learning) in (2), further demonstrating its

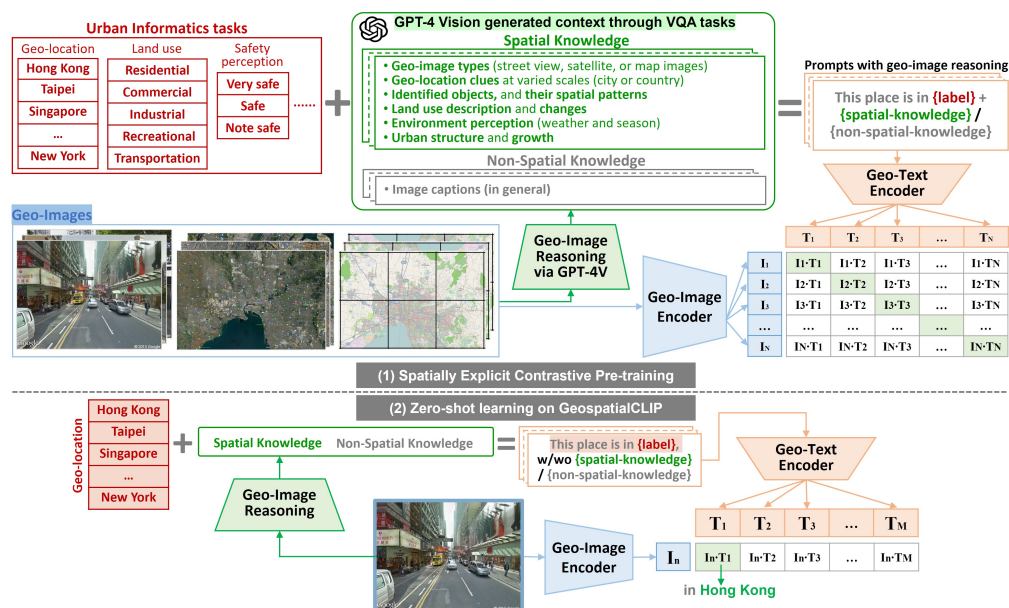


Figure 5.1: The workflow of (1) developing GeospatialCLIP via spatially explicit contrastive learning based on the backbone of CLIP and (2) zero-shot learning on GeospatialCLIP for downstream tasks.

ability to generalize from learned spatial knowledge to new geospatial data (e.g., data across different years and zoom levels).

## 5.3 Experiment #1 Global Image Geo-localization

As results, Table 5.1 compares the accuracy of the vanilla CLIP model, i.e., CLIP ("No prompt" or "Image type" prompt), with a GeoAI-empowered CLIP version, i.e., GeospatialCLIP ("No prompt" or "Image type" prompt), across different image datasets for the task of global image geo-localization. Several general patterns are identified, using CLIP as the baseline to evaluate the performances of GeospatialCLIP:

Table 5.1: Zero-shot Learning Comparisons of Global Image Geolocalization between CLIP and GeospatialCLIP ("No prompt" & "Image type")

Image Type	Dataset	Accuracy					
		City		Country		Continent	
		CLIP (no type)	Geospatial-CLIP (no type)	CLIP (no type)	Geospatial-CLIP (no type)	CLIP (no type)	Geospatial-CLIP (no type)
Street view images	Place Pulse 2.0	63.0%   63.9%	67.0%   <b>72.1%</b>	79.2%   80.3%	81.7%   <b>83.5%</b>	93.6%   93.7%	94.5%   <b>94.9%</b>
	Mapillary	22.8%   24.0%	24.9%   <b>28.3%</b>	52.7%   55.8%	54.5%   <b>58.3%</b>	80.9%   81.2%	82.0%   <b>82.9%</b>
Satellite images	zoom 11 (2014)	8.7%   12.1%	11.5%   <b>13.4%</b>	34.7%   42.4%	40.6%   <b>44.1%</b>	63.6%   71.3%	70.1%   <b>74.3%</b>
	zoom 12 (2014)	8.2%   9.5%	10.1%   <b>11.8%</b>	33.4%   38.7%	37.7%   <b>39.8%</b>	63.0%   68.7%	68.5%   <b>71.0%</b>
	zoom 13 (2014)	5.3%   5.8%	6.1%   <b>6.7%</b>	24.2%   25.4%	27.0%   <b>26.1%</b>	53.4%   <b>58.6%</b>	56.6%   58.5%
	zoom 14 (2014)	10.1%   11.4%	12.3%   <b>13.5%</b>	39.1%   46.5%	45.9%   <b>51.0%</b>	71.8%   74.9%	74.9%   <b>76.9%</b>
	zoom 11 (2023)	10.4%   12.1%	11.9%   <b>13.9%</b>	40.1%   45.6%	46.5%   <b>46.6%</b>	70.0%   77.1%	74.1%   <b>77.2%</b>
	zoom 12 (2023)	9.9%   11.6%	12.0%   <b>14.5%</b>	36.2%   43.0%	43.6%   <b>47.5%</b>	61.9%   69.0%	70.9%   <b>74.0%</b>
	zoom 13 (2023)	10.0%   11.5%	12.6%   <b>14.7%</b>	38.9%   46.0%	46.6%   <b>52.2%</b>	68.1%   74.1%	74.9%   <b>78.0%</b>
	zoom 14 (2023)	9.6%   11.1%	12.3%   <b>14.6%</b>	43.5%   49.7%	50.7%   <b>55.1%</b>	73.4%   77.8%	78.0%   <b>80.4%</b>
Map tile images	zoom 11 (2023)	23.3%   28.8%	26.9%   <b>32.1%</b>	63.7%   66.2%	67.2%   <b>68.2%</b>	84.5%   87.2%	86.6%   <b>89.1%</b>
	zoom 12 (2023)	20.6%   25.8%	24.5%   <b>29.2%</b>	65.7%   65.9%	69.2%   <b>69.7%</b>	88.0%   89.4%	89.8%   <b>91.5%</b>
	zoom 13 (2023)	16.6%   18.2%	18.8%   <b>21.1%</b>	63.3%   59.3%	<b>65.7%</b>   61.7%	84.7%   86.0%	86.0%   <b>88.0%</b>
	zoom 14 (2023)	13.3%   14.4%	<b>16.9%</b>   14.8%	58.8%   54.0%	54.4%   <b>59.7%</b>	80.1%   82.7%	<b>84.8%</b>   81.2%

- **Street View Images (Place Pulse 2.0 and Mapillary):**

- GeospatialCLIP shows an overall improvement over CLIP in both datasets across all geographic scales (i.e., city, country, and continent).

- The improvement is more pronounced in the Place Pulse 2.0 dataset, especially at the city level. This is probably because the Place Pulse 2.0 dataset is a higher quality dataset, containing Google street view images within the 56 most populated cities, while the Mapillary dataset contains citizen-sourced street view images captured and submitted by mobile devices from the 430 most populated cities worldwide.

- **Satellite Images (2014 and 2023, Zoom Levels 11 to 14):**

- GeospatialCLIP consistently outperforms CLIP across different zoom levels for the years 2014 and 2023.
- The margin of improvement is generally more significant in 2023 than in 2014 across different zoom levels.
- The margin of improvement is generally more significant at higher zoom levels (indicating a closer view) in 2023, while this pattern shows at lower zoom levels (indicating a more distant view) in 2014.

- **Map Tile Images (2023, Zoom Levels 11 to 14):**

- GeospatialCLIP again consistently outperforms the CLIP baseline at all zoom levels.
- The improvement tends to increase with a lower zoom level (a more distant view), indicating that after contrastive pre-training with geospatial image-text pairs, GeospatialCLIP learns better how to geo-localize map tiles at a more coarse-grained level.

Note that the "image type" prompt consistently outperforms "no prompt" for both CLIP and GeospatialCLIP across various image datasets, indicating that "image type" (e.g., "street view", "satellite", and "map tile") can enhance image geo-localization.

To sum up, GeospatialCLIP evidently enhances the accuracy of image geo-localization across different types of imagery and scales when compared to the baseline CLIP model. The improvements are more noticeable at both city and country levels, with considerable gains also observed at the continent level. This suggests that further pre-training with geospatial image-text pairs made on CLIP has equipped GeospatialCLIP with a better spatial understanding to interpret and geo-localize geospatial images. Particularly, the increased accuracy is more pronounced at higher zoom levels for 2023 satellite images, suggesting that GeospatialCLIP’s enhanced ability to discern useful geo-location details is more prominent in closer views. On the contrary, this improvement is more significant at lower zoom levels for map tiles, indicating that GeospatialCLIP’s ability to distinguish a larger map extent is improved more profoundly.

Table 5.2: Model Performance Comparison of the Im2GPS3k Benchmark Geo-localization (without prompt)

Model	Accuracy	
	City	Region ~ Country
GeospatialCLIP	32.7%	50.8%
CLIP	32.4%	49.7%
ISN (Muller-Budack et al., 2018)	28.0%	36.6 ~ 49.7%
Translocator (Pramanick et al., 2022)	31.1%	46.7 ~ 58.9%
GeoDecoder (Clark et al., 2023)	33.5%	45.9 ~ 61.0%
GeoCLIP (Cepeda et al., 2023)	34.5%	50.7 ~ 69.7%
PIGEOTTO (Haas et al., 2023)	<b>36.7%</b>	<b>53.8 ~ 72.4%</b>

Table 5.2 presents a comparison of various state-of-the-art geo-localization models’ performance on the Im2GPS3k Benchmark. At the city level, GeospatialCLIP outperforms CLIP, ISN, and Translocator, but slightly underperforms GeoDecoder, GeoCLIP, and PIGEOTTO. At the region and country scales, both GeospatialCLIP and CLIP show competitive performance, equaling GeoCLIP, while PIGEOTTO achieves the highest accuracy.

In summary, GeospatialCLIP exhibits competitive performance in the Im2GPS3k geo-localization task, particularly at broader scales. It slightly outperforms the baseline CLIP model. Although GeoCLIP and PIGEOTTO appear to set the standard for state-of-the-art performance in this task, the spatial representation of this benchmark remains an issue, as discussed in Section 3.3, and therefore evaluating model performance merely based on this dataset may be problematic.

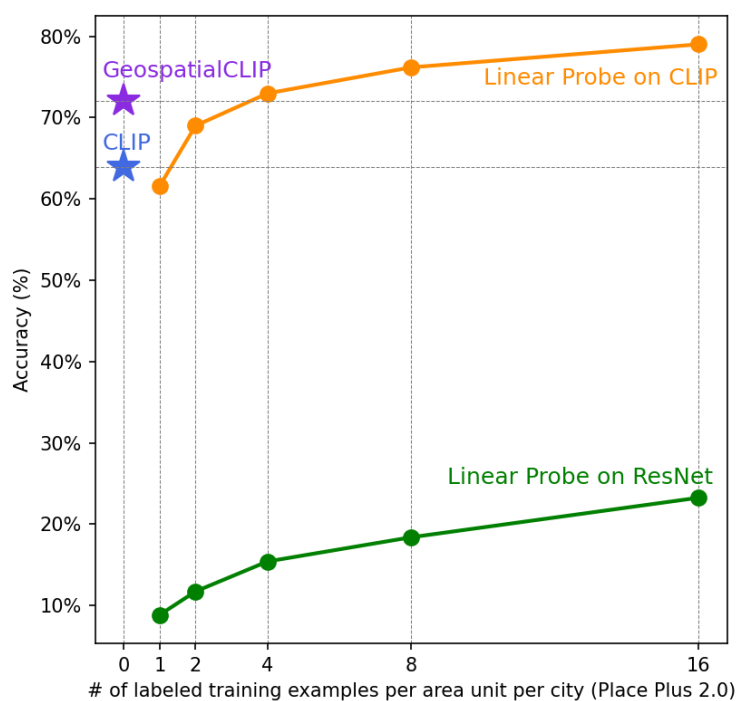


Figure 5.2: Zero-shot GeospatialCLIP significantly outperforms few-shot Linear-probing ResNet-152, and achieves a competitive performance compared with 4-shot Linear-probing CLIP on Place Pulse 2.0 image geo-localization.

Meanwhile, the linear-probing performance of CLIP (i.e., Linear-probing

CLIP) is also evaluated. Specifically, the image encoder from the CLIP-ViT-L/14 model (without the final classification layer) is used as the visual feature extractor. The output feature vector is then fed into a linear classifier (i.e., a logistic regression classifier in our case study based on the scikit-learn implementation) for model training on the Place Plus 2.0 dataset, with Cross-entropy Loss calculated, L-BFGS as the optimization strategy, and maximum 1,000 iterations. Likewise, a simple supervised baseline of a linear classifier is trained on top of ResNet-152 He et al. (2016) to extract features from the Place Plus 2.0 images, with Cross-entropy Loss calculated, Stochastic Gradient Descent (SGD) as the optimizer, and learning rate as 0.001 trained with 32 epochs.

Figure 5.2 visualizes prediction performance differences of zero-shot GeospatialCLIP and CLIP, and few-shot Linear-probing CLIP and ResNet-152, based on the testing set of the Place Plus 2.0 dataset. As results, zero-shot CLIP wins the other two 1-shot models with an accuracy of 63.9%, achieving 2.8% and 55.6% better than Linear-probing CLIP and ResNet-152, respectively, meaning that CLIP has captured distinctive representations for each city label (Figure 3.4). Although it might be intuitive to expect a zero-shot worse than a few-shot, yet this unexpected outcome can be explained by the different training approaches between zero-shot and few-shot. First, zero-shot CLIP is self-supervised by natural language, which enables visual features to be directly described or specified based on large amounts of pre-training pairwise data. By contrast, traditional supervised learning has to derive the visual features merely from input images, and such context-irrelevant image-based learning has the drawback that many different visual objects in one image can be linked to its class without knowing which object(s) should be primary, especially in the one-shot case, where the representative visual features of each class are not yet established. This finding proves that natural language significantly facilitates the reference of learned visual features, enabling zero-shot transfer

of CLIP for downstream geo-localization tasks. Furthermore, GeospatialCLIP's zero-shot performance is very strong, reaching an accuracy of 72.1%, significantly exceeding the 1-shot and 2-shot performances of CLIP and all few-shot ResNets, and obtaining a comparable level with 4-shot CLIP. This evidently validates the effectiveness of spatially explicit contrastive training, which has created a more sophisticated and powerful MFM by enabling the model to encode and learn geospatial information more effectively without the need for additional labeled data. Meanwhile, during few-shot learning, Linear-probing CLIP's predictability improves steadily, indicating that GeospatialCLIP will likely achieve a better result via few-shot learning with the same amount of training samples for downstream geospatial tasks.

## 5.4 Experiment #2 Urban Mixed Land Use Detection

Table 5.3 compares the accuracies between CLIP and GeospatialCLIP in detecting mixed land uses, with and without non-spatial and spatial prompts. A detailed analysis is described below:

- **No prompt:** This serves as the baseline accuracy for each prompt-tuned case, with CLIP at 69.36% and GeospatialCLIP at 70.81%.
- **Non-spatial prompts:**
  - **Image caption:** When using an image caption, CLIP's accuracy falls to 56.00%, and GeospatialCLIP to 57.2%, suggesting that non-spatial information might not be beneficial for mixed land use detection.
  - **Non-spatial ensemble:** Both models show an improvement over their "no prompt" baselines with a non-spatial ensemble.

Table 5.3: GeospatialCLIP’s Prompt Tuning Results of Mixed Land Use Detection

Prompt Type	Prompt	Acc. CLIP	Acc. GeospatialCLIP
No prompt	–	69.36%	70.81%
Non-spatial	Image caption	56.0%	57.2%
	Non-spatial ensemble	69.97%	72.41%
Spatial	Image type	32.77%	40.85%
	Beauty	45.8%	39.2%
	Boringness	60.4%	44.2%
	Depression	62.7%	46.7%
	Liveliness	66.8%	58.1%
	Safety	64.1%	47.2%
	Wealthiness	53.5%	41.4%
	Land use	31.7%	34.6%
	Spatial patterns	51.5%	48.0%
	Geo-features	44.6%	49.2%
	<b>City label</b>	71.27%	<b>75.15%</b>

ble, with CLIP reaching 69.97% and GeospatialCLIP reaching 72.41%. This indicates that a well-designed combination of non-spatial information can contribute positively.

- **Spatial prompts:**

- **Image type:** Introducing the image type as a prompt leads to a decrease in accuracy for both models, with CLIP falling to 32.77% and GeospatialCLIP to 40.85%. Despite the decrease, GeospatialCLIP maintains higher accuracy than CLIP with this prompt.

- **Perception-based prompts:** All perception-based prompts reduce the accuracy compared to the "no prompt" baseline for both models. However, for CLIP, the declines are less steep, and it consistently outperforms GeospatialCLIP in these categories.
- **Land use:** The "land use" prompt deteriorates the performance of GeospatialCLIP (to 34.6%) over its baseline, the same as CLIP (decreasing to 31.7%).
- **Spatial patterns and Geo-features:** Both prompts result in reduced accuracy for both models compared to the "no prompt" baseline.
- **City label:** This prompt provides the highest accuracy for both models, with CLIP reaching 71.27% and GeospatialCLIP achieving 75.15%. Both see an increase over the "no prompt" baseline, especially GeospatialCLIP, which suggests that it learns and leverages city-based knowledge very effectively.

To sum up, only "Non-spatial ensemble" and "City label" prompts can lead to higher accuracy in mixed land use detection tasks compared with the "no prompt" baseline.

In addition, Figure 5.3 compares the performance of different models, with few-shot ResNets and few-shot ViT as the baselines, which have the same model configuration as Section 3.4. Specifically, few-shot ResNets have the lowest performance, and few-shot ViT shows a considerable improvement over few-shot ResNets, displaying a variable performance as more training examples are fed.

Next, comparing the MFMs to these baselines, GPT-4V shows a remarkable performance of 76.10%, which not only significantly outperforms the few-shot performance of both ResNets and ViT but also surpasses the other MFMs. As for GeospatialCLIP, with an accuracy of 75.15%, it also exceeds the few-shot models, gaining a comparable performance with

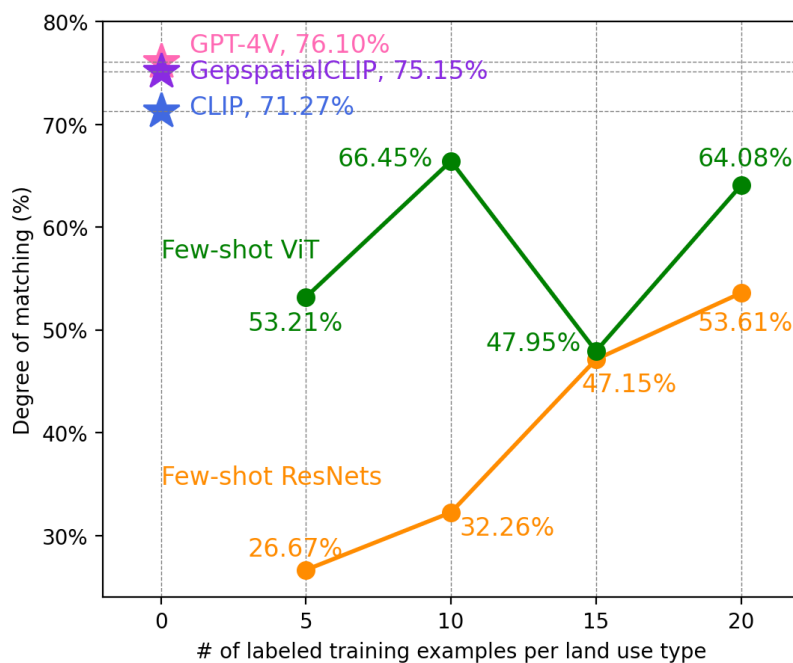


Figure 5.3: Comparison of model performances on mixed land use detection.

GPT-4V. CLIP’s performance is impressive as well, topping the few-shot models.

Based on the findings, several implications may be inferred:

- All zero-shot MFMs (GPT-4V, GeospatialCLIP, CLIP) demonstrate a higher degree of matching accuracy, showcasing the strength of MFMs in leveraging learned land use representations compared with single-modal image-based models.
- Both GPT-4V and GeospatialCLIP excel in this task and show very similar performances, indicating that they are particularly well-suited for mixed land use detection, likely due to their extensive pre-training on multimodal datasets that include urban scene and land use information.

- CLIP also performs quite well, suggesting that its vast image-text pairwise pre-training data endow it with satisfactory textual representations of visual concepts related to land uses.
- The few-shot image-based models require labeled examples to learn. While ViT shows a better performance than ResNets with a high degree of variance in its learning trajectory, both are outperformed by MFMs' capabilities, highlighting the power of MFMs to generalize sufficiently from their learned representation space and perform well on land use tasks without additional labeled data.

## 5.5 Experiment #3 Urban Perception Prediction

As for the results and analysis of urban perception prediction, Tables 5.4 - 5.9 showcase the performance of CLIP and GeospatialCLIP across the six perception variables. Below is the interpretation of the improvements achieved by GeospatialCLIP compared to CLIP, as well as the effectiveness of spatial prompts for GeospatialCLIP with "no prompt" as the baseline.

### Beauty Perception

In "No prompt," GeospatialCLIP starts off with slightly higher accuracy (37.9%) than CLIP (37.1%), and shows improved performance with spatial prompts such as "Boringness," "Liveliness," "Safety," and "Geo-location," exceeding its "no prompt" baseline, especially with the "Image type" prompt reaching the highest accuracy (41.8%).

### Boringness Perception

Both models have the same accuracy (28.2%) with "No prompt." Nevertheless, GeospatialCLIP benefits from the "Geo-location" prompt most

Table 5.4: GeospatialCLIP's Prompt Tuning Results of Beauty Prediction

Prompt Type	Prompt	Acc. CLIP	Acc. GeospatialCLIP
No prompt	–	37.1%	37.9%
Non-spatial	Image caption	33.9%	37.5%
Spatial	Image type	34.3%	<b>41.8%</b>
	Beauty	27.1%	25.0%
	Boringness	31.1%	38.2%
	Depression	32.9%	36.1%
	Liveliness	31.4%	38.2%
	Safety	36.1%	39.3%
	Wealthiness	27.9%	36.1%
	Land use	30.0%	31.1%
	Spatial patterns	32.1%	32.5%
	Geo-features	35.4%	36.4%
Geo-location	35.0%	39.6%	

significantly, jumping to 43.2%, indicating a remarkable improvement over the "no prompt" baseline.

### Depression Perception

GeospatialCLIP begins with a slightly higher accuracy (32.5%) than CLIP (30.4%) with "No prompt," and the "Geo-location" prompt boosts GeospatialCLIP's accuracy to the best at 41.1%, showcasing a notable improvement over the "no prompt" baseline.

Table 5.5: GeospatialCLIP’s Prompt Tuning Results of Boringness Prediction

Prompt Type	Prompt	Acc. CLIP	Acc. GeospatialCLIP
No prompt	–	28.2%	28.2%
Non-spatial	Image caption	30.7%	30.7%
Spatial	Image type	28.2%	28.6%
	Beauty	31.4%	37.1%
	Boringness	25.0%	26.1%
	Depression	31.8%	30.4%
	Liveliness	29.3%	35.7%
	Safety	29.6%	31.1%
	Wealthiness	26.4%	31.1%
	Land use	31.4%	34.6%
	Spatial patterns	29.6%	32.5%
	Geo-features	31.1%	36.1%
Geo-location	31.8%	<b>43.2%</b>	

### Liveliness Perception

GeospatialCLIP has a slightly lower baseline accuracy (24.6%) compared to CLIP (25.7%) with "No prompt," however, its performance is enhanced by all non-spatial and spatial prompts, especially with the "Depression" prompt, which boosts its accuracy to the best at 44.6%.

Table 5.6: GeospatialCLIP’s Prompt Tuning Results of Depression Prediction

Prompt Type	Prompt	Acc. CLIP	Acc. GeospatialCLIP
No prompt	–	30.4%	32.5%
Non-spatial	Image caption	32.5%	26.8%
Spatial	Image type	30.7%	31.8%
	Beauty	36.1%	36.4%
	Boringness	34.3%	36.8%
	Depression	33.6%	30.7%
	Liveliness	33.6%	36.4%
	Safety	32.9%	31.4%
	Wealthiness	32.5%	36.8%
	Land use	32.1%	28.9%
	Spatial patterns	31.8%	33.9%
	Geo-features	36.1%	38.2%
Geo-location	31.8%	<b>41.1%</b>	

### Safety Perception

GeospatialCLIP’s baseline accuracy (37.5%) is higher than CLIP’s (33.2%) with "No prompt." In fact, this is GeospatialCLIP’s best performance in safety prediction, meaning that prompt tuning with either non-spatial or spatial context does not improve its capability to understand the safety level in urban street scenes. Further exploration in the future may be required in order to gain a better understanding of how safety is perceived

Table 5.7: GeospatialCLIP’s Prompt Tuning Results of Liveliness Prediction

Prompt Type	Prompt	Acc. CLIP	Acc. GeospatialCLIP
No prompt	–	25.7%	24.6%
Non-spatial	Image caption	32.9%	28.6%
Spatial	Image type	34.3%	35.7%
	Beauty	36.4%	35.0%
	Boringness	32.5%	38.2%
	Depression	35.7%	<b>44.6%</b>
	Liveliness	30.4%	32.1%
	Safety	28.9%	33.2%
	Wealthiness	32.9%	32.9%
	Land use	40.0%	32.5%
	Spatial patterns	24.3%	26.8%
	Geo-features	38.6%	36.4%
Geo-location	32.1%	30.0%	

by humans in urban areas (e.g., objects and their spatial arrangement in the street), which will be insightful to guide the development of an MFM for enhanced performance.

### Wealthiness Perception

With "No prompt," GeospatialCLIP starts with a lower accuracy (28.9%) than CLIP (30.7%). Both non-spatial and spatial prompts enhance GeospatialCLIP’s accuracy across the board, with "Wealthiness," "Safety," and

Table 5.8: GeospatialCLIP’s Prompt Tuning Results of Safety Prediction

Prompt Type	Prompt	Acc. CLIP	Acc. GeospatialCLIP
No prompt	–	33.2%	<b>37.5%</b>
Non-spatial	Image caption	28.2%	31.8%
Spatial	Image type	26.1%	26.4%
	Beauty	22.1%	23.2%
	Boringness	24.3%	30.4%
	Depression	25.7%	32.1%
	Liveliness	24.6%	31.1%
	Safety	27.1%	25.4%
	Wealthiness	23.6%	25.7%
	Land use	31.1%	33.9%
	Spatial patterns	30.4%	25.0%
	Geo-features	26.4%	31.1%
Geo-location	27.1%	35.4%	

"Liveliness" yielding the most significant increases. Notably, the "Wealthiness" prompt provides the most positive information for its prediction.

To sum up, in almost all cases, spatial prompts can effectively improve GeospatialCLIP’s performance in urban perception prediction tasks, constantly enabling GeospatialCLIP to outperform CLIP and its own baseline ("No prompt") results. Remarkably, the "Geo-location" prompt consistently stands out as quite effective in enhancing GeospatialCLIP’s accuracy across different perception variables, suggesting that knowing the geo-

Table 5.9: GeospatialCLIP’s Prompt Tuning Results of Wealthiness Prediction

Prompt Type	Prompt	Acc. CLIP	Acc. GeospatialCLIP
No prompt	–	30.7%	28.9%
Non-spatial	Image caption	35.7%	33.6%
Spatial	Image type	35.7%	32.1%
	Beauty	32.5%	39.3%
	Boringness	30.4%	37.1%
	Depression	35.4%	37.1%
	Liveliness	32.1%	41.1%
	Safety	29.3%	41.1%
	Wealthiness	34.3%	<b>41.4%</b>
	Land use	27.1%	38.2%
	Spatial patterns	34.6%	35.7%
	Geo-features	32.5%	37.9%
Geo-location	35.7%	34.3%	

location (e.g., a city label) of a street view can assist GeospatialCLIP in extracting the local characteristic features, which helps GeospatialCLIP better understand the link between these features and the social and psychological impressions that these features may bring to humans across different perception variables.

## 5.6 Conclusions and Discussion

This chapter has performed the innovative practice of spatially explicit contrastive learning/pre-training on CLIP to develop a GeoAI MFM, i.e., GeospatialCLIP, and its capabilities have been evaluated on the three geospatial tasks, including global image geo-localization, mixed land use detection, and urban perception prediction. This research marks a significant methodological evolution in GeoAI, bringing to light new possibilities and challenges.

### Methodological Advancements

The introduction of spatially explicit contrastive pre-training has set a new benchmark for the development of GeoAI MFMs capable of comprehending and interpreting complex geospatial datasets. This approach has significantly enriched the models' understanding of spatial relationships, enabling more precise predictions across varied geospatial tasks. By encoding spatial context directly into the learning process of textual embedding, a deeper level of geospatial insight has been unlocked for MFMs, demonstrating the untapped potential of GeoAI, incorporated with MFMs, in addressing complex spatial challenges in the future.

### Future Work from a Methodological Perspective

Meanwhile, several promising directions have been pinpointed:

- Enhancing spatial feature extraction: future research could further explore the modification of model architectures for extracting and encoding more diverse spatial features into the pre-training process. This includes processing and integrating spatial topological and morphological information (e.g., from vector data) to deepen the models' understanding of complex spatial dynamics, as well as developing

spatial context encoders that can be integrated with multimodal learning model architectures.

- Cross-Modal data integration: expanding the scope of data modalities (e.g., audio, video, or even olfactory data) could potentially enhance the models' ability to interpret spatial environments more comprehensively. Research into efficient cross-modal data representation and fusion techniques is particularly crucial for this advancement.
- Developing more interpretable and explainable GeoAI: enhancing the interpretability and explainability of GeoAI MFMs is crucial for their application in the geospatial domain, especially for sensitive areas (e.g., environmental health analysis). Future work could develop both quantitative and qualitative methods to make the decision-making process of these models transparent and understandable to human users (Liang et al., 2022).

The exploration into spatially explicit contrastive learning/pre-training for GeoAI MFMs has illuminated the methodological advancements necessary for the next generation of GeoAI tools. The identified future research directions emphasize the need for continued innovation in advancing model architectures, integrating more multimodal data, as well as enhancing the interpretability of GeoAI MFMs. Moving forward, we can further harness the promising potential of GeoAI and MFMs to revolutionize our understanding and interaction with the geospatial fields, reaching a future where geospatial sciences and AI converge to address global challenges more effectively, redefining the boundaries of what GeoAI models can achieve.

## 6 CONCLUSIONS AND DISCUSSION: THE GOOD, THE BAD, AND THE FUTURE

---

This dissertation study has proposed a foundational multimodal learning framework enhanced by spatial knowledge. Specifically, the significance of the framework has been evidenced by demonstrating its validity in establishing benchmarks of the zero-shot performance of multimodal foundation models (RA-1), as well as empowering these models with a stronger capability for various geospatial tasks (e.g., image geolocalization and mixed land use detection in urban areas), via the methods of spatial-context prompt tuning (RA-2) and spatially explicit contrastive pre-training (RA-3), addressing the current challenges and limitations in the realm of multimodal learning for the geospatial domain.

### 6.1 The Good

The journey began with the ambition of extending the capabilities of MFMs to process and comprehend geospatial information with an enhanced understanding of spatial context across different geospatial tasks. Throughout the dissertation, we've seen how this objective was systematically addressed through innovative data curation and methodologies, including spatial-context prompt tuning and spatially explicit contrastive pre-training. The detailed experiments have demonstrated the enormous potential of MFMs in geospatial applications, setting new benchmarks for the development of GeoAI MFMs.

Specifically, this dissertation has demonstrated that GeoAI, an innovative intersection of geospatial information sciences and artificial intelligence (AI), has emerged as a transformative force in the integration with multimodal learning for urban informatics. MFMs empowered by GeoAI,

which leverage diverse data modalities such as images, text, and geospatial context, are pioneering and advanced solutions for challenges in urban studies.

One of the remarkable applications of GeoAI MFMs is in the field of image geo-localization. This technology now allows for better identification of the geo-location of an image by analyzing its visual and textual contents, enabling the automatic cataloging of urban imagery without explicit metadata, and aiding in the efficient management and retrieval of geographic information as well as enhancing the richness of spatial databases. The integration of this technology into mobile applications may also help users navigate and discover urban areas, enriching their interaction with city spaces.

Additionally, detecting and analyzing mixed land use in urban environments is another area where GeoAI MFMs excel. By processing street view images alongside urban land use depiction, these models can now identify patterns and trends in land use that are not easily discernible through traditional methods. For urban planners and developers, this is crucial to design more efficient and sustainable urban spaces with mixed land uses and to facilitate the understanding of how residential, commercial, recreational, transportation, industrial, and greenfield zones should coexist and interact, which is essential for traffic management, environmental planning, and zoning regulations. Therefore, this method assists in promoting a more balanced urban development, reducing commute times, and enhancing the quality of life for city dwellers.

Furthermore, GeoAI MFMs have been instrumental in urban perception prediction, by providing more insights into how different urban environments are perceived in terms of safety, aesthetics, livability, etc. Through the analysis of urban scenes and descriptions, these models can now accurately predict public perceptions and sentiments about different urban areas. This capability is particularly beneficial for city councils and

marketing agencies aiming to improve public spaces and for real estate developers interested in investment decisions. Predictive analytics in urban perception also allows for proactive urban design that enhances public satisfaction and engagement, aiding in monitoring and responding to the changing dynamics of urban life, and contributing to more engaging and responsive urban governance.

In summary, the integration of GeoAI MFMs in urban informatics has marked a significant leap forward in managing and understanding urban environments. With applications spanning image geo-localization, mixed land use detection, and urban perception prediction, these models offer a powerful toolkit for enhancing urban planning, development, and management. The convergence of AI, geospatial information sciences, and multimodal data not only enriches our analytical capabilities but also paves the way for more informed and sustainable urban futures. As these technologies continue to evolve, their potential to transform urban landscapes grows, promising smarter, more efficient, and more humane urban environments.

Methodologically, the dissertation stands out for pioneering the incorporation of spatial knowledge into the learning and operational mechanisms of MFMs, achieving remarkable improvements in model performance across the evaluated tasks listed above. These advancements highlight the capabilities of GeoAI MFMs in interpreting complex geospatial data, suggesting a promising direction to leverage multimodal learning for future GeoAI research.

## **6.2 The Bad**

Despite the progress, this journey witnessed its own challenges. One of the primary concerns has been the computational cost associated with training a large GeoAI MFM. For example, the size of the pre-training geospatial

datasets used in this dissertation is approximately 90 GB, and thus the pre-training process was time-consuming with a considerable level of uncertainty in hit-or-miss. Moreover, the issue of model interpretability remains a significant hurdle. These challenges underscore the necessity for continued effort in data efficiency, model robustness, and an enhanced explainability of GeoAI.

Another major consideration is ethical issues. For instance, one of the primary ethical concerns with large foundation models is the bias embedded in their vast (pre-)training data. These models often rely on internet-sourced datasets that reflect historical prejudices, societal stereotypes, and unequal representation across different groups and areas. As such, they can inherit and even amplify biases related to race, gender, and religion, potentially leading to AI outputs that are discriminatory or reinforce harmful stereotypes, thus perpetuating existing social inequalities. Note that this ethical challenge is not only identifying these biases, but also developing methodologies to mitigate them. This involves curating more balanced datasets, applying de-biasing techniques (e.g., data augmentation, re-sampling, and incorporating fairness constraints) during training, and continuously monitoring the outputs for biases. Moreover, it requires a commitment from AI developers to prioritize fairness and inclusivity in their AI systems.

Another ethical concern lies in the fact that large foundation models are often associated with significant privacy concerns. The vast amount of data used in (pre-)training these models likely include personal information, which can be inadvertently learned and reproduced by the models, which might generate text that discloses real personal data, potentially leading to privacy violations. Therefore, privacy protection in AI demands stringent data handling and processing protocols, involving anonymizing datasets, obtaining consent from data subjects, and implementing robust data security measures. For example, regulatory frameworks such as the

General Data Protection Regulation (GDPR)<sup>1</sup> have provided guidelines and obligations for AI developers to ensure that user data are protected against misuse and exploitation.

Furthermore, an additional ethical issue is the potential that the capabilities of large foundation models could be exploited for malicious purposes. The ease with which these models can generate convincing fake content, such as deepfakes<sup>2</sup> or misleading news and articles, has posed significant risks. Such fake content can be used to spread misinformation, manipulate public opinion, or even conduct social engineering attacks. To address this issue, we need to develop detection tools to identify AI-generated content, foster media literacy among the public, and create legal and ethical frameworks to govern the use of AI in content creation. Therefore, it is essential for policymakers, AI developers, and civil society to work together to mitigate the risks and dangers posed by the misuse of AI.

Lastly, the deployment of large foundation models could impact our society in many profound ways. While they offer potential benefits such as increased efficiency and new capabilities, they also raise questions about job displacement, power dynamics, and the control of AI technologies. The concentration of AI development in a few large corporations exacerbates concerns about monopolistic practices and the lack of accountability in AI governance. To ethically navigate these challenges, it is critical for us to engage in broad societal dialogues about the role of AI, including discussions on who benefits from AI technologies, who might be harmed, and how to ensure equitable access to the benefits of AI. Moreover, this process should involve implementing policies that encourage competition, regulate monopolistic practices, and ensure that AI developments are aligned with societal values and needs.

In summary, the ethical issues surrounding large foundation models,

---

<sup>1</sup><https://gdpr-info.eu/>

<sup>2</sup><https://deepfakesweb.com/>

including GeoAI MFMs, are complex and multifaceted. Addressing these concerns requires a collaborative effort among AI researchers, developers, ethicists, policymakers, and the broader public. By fostering ethical frameworks that promote fairness, privacy, accountability, and inclusivity, the development and deployment of AI technologies can be directed towards benefiting society as a whole. Ensuring ethical AI usage is not just a technical challenge but a societal imperative that calls for vigilant oversight, thoughtful regulation, and active engagement from all stakeholders.

### **6.3 The Future**

With the pros and cons discussed above, this dissertation has identified several promising directions for future work. From enhancing spatial feature extraction to developing more reliable, interpretable, and explainable GeoAI models, the potential for growth and innovation within this field is immense. Particularly, the exploration of cross-modal data integration and the refinement of model architectures for geospatial context embeddings offer exciting avenues for enriching the spatial understanding and capabilities of MFMs in the future.

Moreover, the evolution of GeoAI MFMs towards achieving finer-grained image geo-localization (e.g., at a street or district level), mixed land use detection with a more refined classification scheme, and more accurate urban perception prediction underscores the potential impact of GeoAI technologies in shaping our understanding of urban landscapes. As GeoAI continues to evolve, its integration with emerging technologies and data modalities holds the promise of revolutionizing the field of geospatial sciences, urban studies, computer sciences, and beyond.

In conclusion, this dissertation marks a significant step forward in the intersection of AI and geospatial sciences for urban informatics, laying a solid foundation for future research while highlighting the importance of

addressing the accompanying computational, methodological, and ethical challenges. As we move forward, it is imperative to continue this exploration with a commitment to innovation, responsibility, and the broader goal of harnessing AI for the betterment of our society and environment.

## REFERENCES

---

- Abdullahi, Saleh, Biswajeet Pradhan, Shattri Mansor, and Abdul Rashid Mohamed Shariff. 2015. Gis-based modeling for the spatial measurement and evaluation of mixed land use development for a compact city. *GIScience & Remote Sensing* 52(1):18–39.
- Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Adriano, Bruno, Naoto Yokoya, Junshi Xia, Hiroyuki Miura, Wen Liu, Masashi Matsuoka, and Shunichi Koshimura. 2021. Learning from multi-modal and multitemporal earth observation data for building damage mapping. *ISPRS Journal of Photogrammetry and Remote Sensing* 175:132–143.
- Anantrasirichai, Nantheera, and David Bull. 2022. Artificial intelligence in the creative industries: a review. *Artificial intelligence review* 55(1): 589–656.
- Arora, Simran, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.
- Aswatha, Shashaank Mattur, Jayanta Mukherjee, Prabir K Biswas, and Subhas Aikat. 2020. Unsupervised classification of land cover using multi-modal data from multi-spectral and hybrid-polarimetric sar imageries. *International Journal of Remote Sensing* 41(14):5277–5304.

- Axelsson, Östen, Mats E Nilsson, and Birgitta Berglund. 2010. A principal components model of soundscape perception. *The Journal of the Acoustical Society of America* 128(5):2836–2846.
- Bai, Lubin, Weiming Huang, Xiuyuan Zhang, Shihong Du, Gao Cong, Haoyu Wang, and Bo Liu. 2023. Geographic mapping with unsupervised multi-modal representation learning from vhr images and pois. *ISPRS Journal of Photogrammetry and Remote Sensing* 201:193–208.
- Bansal, Mayank, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. 2011. Geo-localization of street views with aerial image databases. In *Proceedings of the 19th acm international conference on multimedia*, 1125–1128.
- Barrault, Loïc, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.
- Biljecki, Filip, and Koichi Ito. 2021. Street view imagery in urban analytics and gis: A review. *Landscape and Urban Planning* 215:104217.
- Boehm, Kevin M, Pegah Khosravi, Rami Vanguri, Jianjiong Gao, and Sohrab P Shah. 2022. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer* 22(2):114–126.
- Brown, Barbara B, Ikuho Yamada, Ken R Smith, Cathleen D Zick, Lori Kowaleski-Jones, and Jessie X Fan. 2009. Mixed land use and walkability: Variations in land use measures and relationships with bmi, overweight, and obesity. *Health & place* 15(4):1130–1141.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

- Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.
- Cao, Rui, Cai Liao, Qing Li, Wei Tu, Rui Zhu, Nianxue Luo, Guoping Qiu, and Wenzhong Shi. 2023. Integrating satellite and street-level images for local climate zone mapping. *International Journal of Applied Earth Observation and Geoinformation* 119:103323.
- Cao, Rui, Jiasong Zhu, Wei Tu, Qingquan Li, Jinzhou Cao, Bozhi Liu, Qian Zhang, and Guoping Qiu. 2018. Integrating aerial and street view images for urban land use classification. *Remote Sensing* 10(10):1553.
- Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the european conference on computer vision (eccv)*, 132–149.
- Castelluccio, Marco, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. 2015. Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092*.
- Cepeda, Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *arXiv preprint arXiv:2309.16020*.
- Cervero, Robert. 1996. Mixed land-uses and commuting: Evidence from the american housing survey. *Transportation Research Part A: Policy and Practice* 30(5):361–377.
- Chaabane, Mohamed, Lionel Gueguen, Ameni Trabelsi, Ross Beveridge, and Stephen O’Hara. 2021. End-to-end learning improves static object geo-localization from video. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, 2063–2072.

Chang, Yingshan, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*, 16495–16504.

Chen, Xi, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Cheng, Gong, Junwei Han, Lei Guo, Zhenbao Liu, Shuhui Bu, and Jinchang Ren. 2015. Effective and efficient midlevel visual elements-oriented land-use classification using vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 53(8):4238–4249.

Clark, Brandon, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. 2023. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 23182–23190.

Credit, Kevin. 2022. Spatial models or random forest? evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in los angeles. *Geographical Analysis* 54(1):58–83.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhariwal, Prafulla, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Ding, Qing, Zhenfeng Shao, Xiao Huang, Orhan Altan, and Bin Hu. 2022. Time-series land cover mapping and urban expansion analysis using openstreetmap data and remote sensing big data: A case study of guangdong-hong kong-macao greater bay area, china. *International Journal of Applied Earth Observation and Geoinformation* 113:103001.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dubey, Abhimanyu, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *European conference on computer vision*, 196–212. Springer.

El-Komy, Amir, Osama R Shahin, Rasha M Abd El-Aziz, and Ahmed I Taloba. 2022. Integration of computer vision and natural language processing in multimedia robotics application. *Information Sciences* 7:6.

Fang, Fang, Linyun Zeng, Shengwen Li, Daoyuan Zheng, Jiahui Zhang, Yuanyuan Liu, and Bo Wan. 2022. Spatial context-aware method for urban land use classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing* 192:1–12.

Fei, Nanyi, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications* 13(1):3094.

Fonte, Cidália C, and Nuno Martinho. 2017. Assessing the applicability of openstreetmap data to assist the validation of land use/land cover maps. *International Journal of Geographical Information Science* 31(12):2382–2400.

Fonte, Cidália C, Joaquim Patriarca, Ismael Jesus, and Diogo Duarte. 2020. Automatic extraction and filtering of openstreetmap data to generate training datasets for land use land cover classification. *Remote Sensing* 12(20):3428.

Gao, Song, Krzysztof Janowicz, and Helen Couclelis. 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS* 21(3):446–467.

Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15180–15190.

Gutmann, Michael, and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, ed. Yee Whye Teh and Mike Titterton, vol. 9 of *Proceedings of Machine Learning Research*, 297–304. Chia Laguna Resort, Sardinia, Italy: PMLR.

Haas, Lukas, Silas Alberti, and Michal Skreta. 2023. Pigeon: Predicting image geolocations. *arXiv preprint arXiv:2307.05845*.

Hartigan, John A, and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28(1):100–108.

Hays, James, and Alexei A Efros. 2008. Im2gps: estimating geographic information from a single image. In *2008 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

———. 2015. Large-scale image geolocalization. In *Multimodal location estimation of videos and images*, 41–62. Springer.

He, Jialyu, Xia Li, Penghua Liu, Xinxin Wu, Jinbao Zhang, Dachuan Zhang, Xiaojuan Liu, and Yao Yao. 2020. Accurate estimation of the proportion of mixed land use at the street-block level by integrating high spatial resolution images and geospatial big data. *IEEE Transactions on Geoscience and Remote Sensing* 59(8):6357–6370.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Herbrich, Ralf, Tom Minka, and Thore Graepel. 2006. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems* 19.

Ho, Jonathan, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Huang, Pei-Chi, Ejan Shakya, Myoungkyu Song, and Mahadevan Subramaniam. 2022. Biomdse: A multimodal deep learning-based search engine framework for biofilm documents classifications. In *2022 IEEE*

*international conference on bioinformatics and biomedicine (bibm)*, 3608–3612. IEEE.

Huang, Yanping, and Rajesh PN Rao. 2011. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(5):580–593.

Huang, Yu, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems* 34: 10944–10956.

Janowicz, Krzysztof, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. 2020. Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond.

Ji, Yuhan, and Song Gao. 2023. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations. *arXiv preprint arXiv:2307.03678*.

Jia, Chao, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Kang, Jian, Francesco Aletta, Efstathios Margaritis, and Ming Yang. 2018a. A model for implementing soundscape maps in smart cities. *Noise Mapping* 5(1):46–59.

Kang, Jian, Marco Körner, Yuanyuan Wang, Hannes Taubenböck, and Xiao Xiang Zhu. 2018b. Building instance classification using street view images. *ISPRS journal of photogrammetry and remote sensing* 145:44–59.

Katal, Avita, Mohammad Wazid, and Rayan H Goudar. 2013. Big data: issues, challenges, tools and good practices. In *2013 sixth international conference on contemporary computing (ic3)*, 404–409. IEEE.

Khosla, Prannay, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33:18661–18673.

Kingma, Diederik P, and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Küchemann, Stefan, Karina E Avila, Yavuz Dinc, Chiara Hortmann, Natalia Revenga, Verena Ruf, Niklas Stausberg, Steffen Steinert, Frank Fischer, Martin Fischer, et al. Are large multimodal foundation models all we need? on opportunities and challenges of these models in education.

Li, Wenwen, Chia-Yu Hsu, and Maosheng Hu. 2021a. Tobler’s first law in geoi: A spatially explicit deep learning model for terrain feature detection under weak supervision. *Annals of the American Association of Geographers* 111(7):1887–1905.

Li, Xiaojiang, Chuanrong Zhang, and Weidong Li. 2017. Building block level urban land-use information retrieval based on google street view images. *GIScience & Remote Sensing* 54(6):819–835.

Li, Yansheng, Zhihui Zhu, Jin-Gang Yu, and Yongjun Zhang. 2021b. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing* 59(12):10590–10603.

Li, Yifei, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. On the advance of making language models better reasoners. *arXiv preprint arXiv:2206.02336*.

Liang, Paul Pu, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2022. Multiviz: Towards visualizing and understanding multimodal models. *arXiv preprint arXiv:2207.00056*.

Liang, Xun, Qingfeng Guan, Keith C Clarke, Guangzhao Chen, Song Guo, and Yao Yao. 2021. Mixed-cell cellular automata: A new approach for simulating the spatio-temporal dynamics of mixed land use structures. *Landscape and Urban Planning* 205:103960.

Likas, Aristidis, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition* 36(2):451–461.

Liu, Fan, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. 2023a. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*.

Liu, Pengyuan, and Filip Biljecki. 2022. A review of spatially-explicit geospatial applications in urban geography. *International Journal of Applied Earth Observation and Geoinformation* 112:102936.

Liu, Pengyuan, Yan Zhang, and Filip Biljecki. 2023b. Explainable spatially explicit geospatial artificial intelligence in urban analytics. *Environment and Planning B: Urban Analytics and City Science* 23998083231204689.

Loshchilov, Ilya, and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

———. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, Haoyu, Qiongyi Zhou, Nanyi Fei, Zhiwu Lu, Mingyu Ding, Jingyuan Wen, Changde Du, Xin Zhao, Hao Sun, Huiguang He, et al. 2022. Multi-modal foundation models are better simulators of the human brain. *arXiv preprint arXiv:2208.08263*.

- Lugmayr, Andreas, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.
- Ma, Haoran, and Dongdong Wu. 2023. A natural language processing-based approach: mapping human perception by understanding deep semantic features in street view images. *arXiv preprint arXiv:2311.17354*.
- Mai, Gengchen, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. 2022. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th international conference on advances in geographic information systems*, 1–4.
- Mathias, Brian, and Katharina von Kriegstein. 2023. Enriched learning: Behavior, brain, and computation. *Trends in Cognitive Sciences* 27(1):81–97.
- McCormack, Edward, G Scott Rutherford, and Martina G Wilkinson. 2001. Travel impacts of mixed land use neighborhoods in seattle, washington. *Transportation Research Record* 1780(1):25–32.
- Moos, Markus, Tara Vinodrai, Nick Revington, and Michael Seasons. 2018. Planning for mixed use: affordable for whom? *Journal of the American Planning Association* 84(1):7–20.
- Muller-Budack, Eric, Kader Pustu-Iren, and Ralph Ewerth. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the european conference on computer vision (eccv)*, 563–579.
- Olea, Pedro P, and Patricia Mateo-Tomás. 2013. Assessing species habitat using google street view: a case study of cliff-nesting vultures. *PLoS One* 8(1):e54582.

- Omrani, Hichem, Amin Tayyebi, and Bryan Pijanowski. 2017. Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based land transformation model: an integrated ml-ca-ltm modeling framework. *GIScience & Remote Sensing* 54(3):283–304.
- Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Panagiotopoulos, Apostolos, Giorgos Kordopatis-Zilos, and Symeon Papadopoulos. 2022. Leveraging selective prediction for reliable image geolocation. In *International conference on multimedia modeling*, 369–381. Springer.
- Pande, Chaitanya B, Kanak N Moharir, Sudhir Kumar Singh, Abhay M Varade, Ahmed Elbeltagi, SFR Khadri, and Pandurang Choudhari. 2021. Estimation of crop and forest biomass resources in a semi-arid region using satellite data and gis. *Journal of the Saudi Society of Agricultural Sciences* 20(5):302–311.
- Pastorino, Martina, Federico Gallo, Angela Di Febbraro, Gabriele Moser, Nicola Sacco, and Sebastiano B Serpico. 2022. Multimodal fusion of mobility demand data and remote sensing imagery for urban land-use and land-cover mapping. *Remote Sensing* 14(14):3370.
- Pramanick, Shraman, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. 2022. Where in the world is this image? transformer-based geo-localization in the wild. In *European conference on computer vision*, 196–215. Springer.
- Qiu, Sen, Hongkai Zhao, Nan Jiang, Zhelong Wang, Long Liu, Yi An, Hongyu Zhao, Xin Miao, Ruichen Liu, and Giancarlo Fortino. 2022. Multi-sensor information fusion based on machine learning for real applications

in human activity recognition: State-of-the-art and research challenges. *Information Fusion* 80:241–265.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Ramachandram, Dhanesh, and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* 34(6):96–108.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35:36479–36494.

- Salazar, Julian, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.
- Salesses, Philip, Katja Schechtner, and César A Hidalgo. 2013. The collaborative image of the city: mapping the inequality of urban perception. *PloS one* 8(7):e68400.
- Schroff, Florian, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.
- Seo, Paul Hongsuck, Tobias Weyand, Jack Sim, and Bohyung Han. 2018. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European conference on computer vision (ECCV)*, 536–551.
- Sinaga, Kristina P, and Miin-Shen Yang. 2020. Unsupervised k-means clustering algorithm. *IEEE access* 8:80716–80727.
- Singer, Uriel, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.
- Team, Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multi-modal models. *arXiv preprint arXiv:2312.11805*.
- Tian, Li, Yinlong Liang, and Bo Zhang. 2017. Measuring residential and industrial land use mix in the peri-urban areas of China. *Land Use Policy* 69:427–438.
- Tong, Xin-Yi, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. 2018. Learning transferable deep

models for land-use classification with high-resolution remote sensing images. *arXiv preprint arXiv:1807.05713*.

Tu, Tao, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI* 1(3): A10a2300138.

Van Dao, Dong, Abolfazl Jaafari, Mahmoud Bayat, Davood Mafi-Gholami, Chongchong Qi, Hossein Moayedi, Tran Van Phong, Hai-Bang Ly, Tien-Thinh Le, Phan Trong Trinh, et al. 2020. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *Catena* 188:104451.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Verma, Deepank, Arnab Jana, and Krithi Ramamritham. 2019. Machine-based understanding of manually collected visual and auditory datasets for urban perception studies. *Landscape and Urban Planning* 190:103604.

———. 2020. Predicting human perception of the urban environment in a spatiotemporal urban setting using locally acquired street view images and audio clips. *Building and Environment* 186:107340.

Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on machine learning*, 1096–1103.

- Vo, Nam, Nathan Jacobs, and James Hays. 2017. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, 2621–2630.
- Wang, Wenhui, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.
- Wang, Yaqing, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53(3):1–34.
- Wang, Zhecheng, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. 2024. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 5805–5813.
- Wei, Jingxian, Wenze Yue, Mengmeng Li, and Jiabin Gao. 2022. Mapping human perception of urban landscape from street-view images: A deep-learning approach. *International Journal of Applied Earth Observation and Geoinformation* 112:102886.
- Weyand, Tobias, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *European conference on computer vision*, 37–55. Springer.
- Wilson, Daniel, Thayer Alshaabi, Colin Van Oort, Xiaohan Zhang, Jonathan Nelson, and Safwan Wshah. 2022. Object tracking and geolocalization from street images. *Remote Sensing* 14(11):2575.
- Wu, Jiayu, Yan Song, Junhui Liang, Qingxi Wang, and Jian Lin. 2018. Impact of mixed land use on housing values in high-density areas: Evidence from Beijing. *Journal of Urban Planning and Development* 144(1):05017019.

- Wu, Meiliu, and Qunying Huang. 2022. Im2city: image geo-localization via multi-modal learning. In *Proceedings of the 5th acm sigspatial international workshop on ai for geographic knowledge discovery*, 50–61.
- Wu, Meiliu, Qunying Huang, Song Gao, and Zhou Zhang. 2023. Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning. *International Journal of Applied Earth Observation and Geoinformation* 125:103591.
- Wu, Xinxin, Xiaoping Liu, Dachuan Zhang, Jinbao Zhang, Jialyu He, and Xiaocong Xu. 2022. Simulating mixed land-use change under multi-label concept by integrating a convolutional neural network and cellular automata: A case study of huizhou, china. *GIScience & Remote Sensing* 59(1):609–632.
- Xian, Yongqin, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence* 41(9):2251–2265.
- Xing, Hanfa, Yuan Meng, and Yan Shi. 2018. A dynamic human activity-driven model for mixed land use evaluation using social media data. *Transactions in GIS* 22(5):1130–1151.
- Yan, Bo, Krzysztof Janowicz, Gengchen Mai, and Rui Zhu. 2018. xNet+SC: Classifying Places Based on Images by Incorporating Spatial Contexts. In *10th international conference on geographic information science (giscience 2018)*, ed. Stephan Winter, Amy Griffin, and Monika Sester, vol. 114 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 17:1–17:15. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- Yao, Yao, Xiaoqin Yan, Peng Luo, Yuyun Liang, Shuliang Ren, Ying Hu, Jian Han, and Qingfeng Guan. 2022. Classifying land-use patterns by

integrating time-series electricity data and high-spatial resolution remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation* 106:102664.

Ye, Lin, Sumedha Mandpe, and Peter B Meyer. 2005. What is “smart growth?”—really? *Journal of Planning Literature* 19(3):301–315.

Yokoya, Naoto, Pedram Ghamisi, Junshi Xia, Sergey Sukhanov, Roel Heremans, Ivan Tankoyeu, Benjamin Bechtel, Bertrand Le Saux, Gabriele Moser, and Devis Tuia. 2018. Open data for global multimodal land use classification: Outcome of the 2017 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11(5):1363–1377.

Yoo, Suhong, Jisang Lee, Mohammad Gholami Farkoushi, Eunkwan Lee, and Hong-Gyoo Sohn. 2022. Automatic generation of land use maps using aerial orthoimages and building floor data with a conv-depth block (cdb) resu-net architecture. *International Journal of Applied Earth Observation and Geoinformation* 107:102678.

Yu, Jiahui, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2(3):5.

Yue, Yang, Yan Zhuang, Anthony GO Yeh, Jin-Yun Xie, Cheng-Lin Ma, and Qing-Quan Li. 2017. Measurements of poi-based mixed use and their relationships with neighbourhood vibrancy. *International Journal of Geographical Information Science* 31(4):658–675.

Zahnaw, Renee. 2018. Mixed land use: Implications for violence and property crime.

Zemene, Eyasu, Yonatan Tariku Tesfaye, Haroon Idrees, Andrea Prati, Marcello Pelillo, and Mubarak Shah. 2018. Large-scale image geolocalization using dominant sets. *IEEE transactions on pattern analysis and machine intelligence* 41(1):148–161.

Zhan, Yu, Yuzhou Luo, Xunfei Deng, Huajin Chen, Michael L Grieneisen, Xueyou Shen, Lizhong Zhu, and Minghua Zhang. 2017. Spatiotemporal prediction of continuous daily pm<sub>2.5</sub> concentrations across china using a spatially explicit machine learning algorithm. *Atmospheric environment* 155:129–139.

Zhang, Fan, Bolei Zhou, Liu Liu, Yu Liu, Helene H Fung, Hui Lin, and Carlo Ratti. 2018a. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning* 180:148–160.

Zhang, Pengbin, Yinghai Ke, Zhenxin Zhang, Mingli Wang, Peng Li, and Shuangyue Zhang. 2018b. Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery. *Sensors* 18(11):3717.

Zhang, Weixing, Weidong Li, Chuanrong Zhang, Dean M Hanink, Xiaojiang Li, and Wenjie Wang. 2017a. Parcel-based urban land use classification in megacity using airborne lidar, high resolution orthoimagery, and google street view. *Computers, Environment and Urban Systems* 64:215–228.

———. 2017b. Parcel feature data derived from google street view images for urban land use classification in brooklyn, new york city for urban land use classification in brooklyn, new york city. *Data in brief* 12:175–179.

Zhou, Kaiyang, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130(9):2337–2348.

Zhu, Yi, Xueqing Deng, and Shawn Newsam. 2019. Fine-grained land use classification at the city scale using ground-level images. *IEEE Transactions on Multimedia* 21(7):1825–1838.

Zhuo, Yuefei, Hongyu Zheng, Cifang Wu, Zhongguo Xu, Guan Li, and Zhenning Yu. 2019. Compatibility mix degree index: A novel measure to characterize urban land use mix pattern. *Computers, Environment and Urban Systems* 75:49–60.