Advancing Mixed Land Use Detection by Embedding Spatial Intelligence into Vision-Language Models

Meiliu Wu[1*], Qunying Huang[2], Song Gao[2]

[1*] School of Geographical and Earth Sciences, University of Glasgow, Glasgow, G12 8QQ, UK - meiliu.wu@glasgow.ac.uk
[2] Department of Geography, University of Wisconsin-Madison, 550 N Park St, Madison, WI 53706, USA -
{qhuang46, song.gao}@wisc.edu

Keywords: Mixed Land Use, Urban Analytics, Vision-Language Models, Spatial Intelligence, Spatially Explicit AI, GeoAI.

Abstract

Embedding spatial intelligence into vision-language models (VLMs) has offered a promising avenue to improve geospatial decision-making in complex urban environments. In this work, we propose a novel framework that augments the architecture of Contrastive Language-Image Pretraining (CLIP) with the techniques of spatial-context aware prompt engineering and spatially explicit contrastive learning. By leveraging a diverse set of geospatial imagery (e.g., street view, satellite, and map tile images), paired with contextual geospatial text generated and curated via GPT-4, our approach constructs robust multimodal representations that capture visual, textual, and spatial insights. The proposed model, termed GeospatialCLIP, is specifically evaluated for urban mixed land use detection, a critical task for sustainable urban planning and smart city development. Results demonstrate that GeospatialCLIP consistently outperforms traditional vision-based few-shot models (e.g., ResNet-152, Vision Transformers) and exhibits competitive performance with state-of-the-art models such as GPT-4. Notably, the incorporation of spatial prompts, especially those providing city-specific cues, significantly boosts detection accuracy. Our findings highlight the pivotal role of spatial intelligence in refining VLM performance and provide novel insights into the integration of geospatial reasoning within multimodal learning. Overall, this work establishes a foundation for future spatially explicit AI development and applications, paving the way for more comprehensive and interpretable models in urban analytics and beyond.

1. Introduction

As artificial intelligence (AI) evolves, the pursuit of more sophisticated, accurate, and intelligent models has led researchers to explore from single-modal to multimodal learning, typically with both visual and textual information as input (e.g., GPT-4.1[1] and Gemini[2]) (Lu et al., 2022). In light of this trend, recent studies have suggested improving the capabilities of vision-language models (VLMs) with spatial intelligence, an ability to interpret, analyze, and reason about spatial relationships in diverse data formats (e.g., images, maps, and 3D environments) (Gao, 2021, Li and Hsu, 2022, Chen et al., 2024). This capability has been integrated into many aspects of our daily lives, such as navigation and transportation networks (Iyer, 2021), emergency responses (Agbaje et al., 2024), smart cities (Wolniak and Stecuła, 2024), and precision agriculture (Akter et al., 2024). However, despite the growing interest in AI applications for spatial intelligence, many foundational aspects remain underdeveloped, from **data** and **methodology**, to **implementation** and **evaluation**. For example, how can we develop unbiased AI models while significant gaps remain in data coverage for many developing regions? What methods can we use to enable AI models to understand spatial relationships across various data modalities? Can we design a foundational framework for developing AI models to support diverse geospatial applications? And how should we assess the performances of these models?

To bridge these gaps, this study aims to embed spatial intelligence into VLMs to enhance their spatial thinking ability for better decision-making in complex real-world tasks. Theoretically, this innovative combination is designed to mimic the human ability to interpret the world through multiple senses, thereby enabling AI models to achieve a more comprehensive and interpretable understanding of our built environment. To validate the effectiveness of the proposed methods, urban mixed land use detection serves as an ideal case study, as it integrates complex geospatial patterns, multimodal data (e.g., satellite imagery, street view images, and textual urban descriptions and policies), and human activity dynamics to address real-world challenges in sustainable urban planning and smart city development (Wu et al., 2023). In summary, the key contributions of this work include:

- **Data**: Establishing benchmark datasets by curating high-quality, large-scale geospatial image-text paired datasets for training and testing VLMs in geospatial analytics;
- **Methodology**: Addressing the current methodological challenges and limitations in vision-language learning through innovative spatial-context aware prompt engineering and spatially explicit contrastive learning;
- **Implementation**: Developing an essential vision-language learning framework that integrates spatial knowledge effectively as well as demonstrating the enhanced capabilities of such a framework in urban mixed land use detection as a case study;
- **Evaluation**: Showcasing model advancement from baselines to the proposed VLMs, contributing to the theoretical and practical knowledge base towards enhancing VLMs by spatial intelligence, and thus shedding light on developing geospatial AI (GeoAI) multimodal foundation models in the future.

2. Geospatial Image-Text Pairwise Datasets

As discussed above, the first challenge lies in the training data. In this study, vision input includes geospatial imagery such as street view images, satellite images, and map tile images, while language input includes both spatial and non-spatial contexts of geospatial images generated and curated by GPT-4.

[1] `https://openai.com/index/gpt-4-1/`
[2] `https://deepmind.google/models/gemini/`

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W7-2025-121-2025

2.1 Vision Input: Geospatial Imagery

2.1.1 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 multimodal learning. In addition, using street view images has become a main research trend (Zemene et al., 2018, Wu and Huang, 2022), as these images have become largely available in public (Zhang et al., 2018), and are more likely to be concurrent with textual descriptions of urban environments, facilitating the training process of VLMs.

Place Pulse 2.0 This dataset, introduced by (Dubey et al., 2016) and consisting 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 VLMs. These images were collected with latitude-longitude coordinates uniformly sampled from grids spatially intersected with city boundaries.

Mapillary Public Dataset Additionally, 193,254 street view images across 430 most populated cities worldwide were also collected from the Mapillary API[3], each geo-located with latitude-longitude coordinates. 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.

As results, Figure 1 (a) and (b) display the spatial distributions of street view images collected in this study.

2.1.2 Satellite Images Another common geospatial image type 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 training VLMs for geospatial applications (e.g., urban planning, disaster response and management, and environmental monitoring and conservation). Specifically, this study collected 23,173 satellite images at different zoom levels (from 11 to 14) across 790 most populated cities worldwide from the 2023 Esri World Imagery Map Server[4], 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.

2.1.3 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 training VLMs to support geospatial tasks, by combining with textual data (e.g., geographic annotations, location-based social media posts, or descriptive map reports). This synthesis potentially allows for more accurate geographic information retrieval.

OpenStreetMap (OSM) provides a rich source of map tile images through its servers[5], 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. 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.

As results, the spatial distributions of satellite and map tile imagery are displayed in Figure 1 (c) and (d), respectively.

2.2 Language Input: Generating Spatial and Non-spatial Context from Geospatial Images through GPT-4

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, this work first conducted multiple visual question answering tasks on GPT-4 for geospatial image reasoning. In return, GPT-4 not only demonstrated remarkable zero-shot transfer capabilities for urban analytics, but also generated both spatially and non-spatially contextualized descriptions for each image.

As outputs, extensive geospatial image-text pairs have been curated, split into training and testing sets, and summarized in Table 1. Additionally, this study also generated geo-location text for images and used these pairs for model training (Section 3.3).

3. Methodology

This work develops two novel methods to embed spatial intelligence into VLMs, i.e., spatial-context aware prompt engineering and spatially explicit contrastive learning. A competitive, open-source VLM - Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) - is used as the model backbone and being fine-tuned to build GeospatialCLIP tailored for better supporting geospatial tasks.

3.1 Contrastive Language-Image Pre-training

CLIP (Radford et al., 2021) is 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.

3.2 Spatial-context Aware Prompt Engineering

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. 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. Intuitively, this method 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

[3] `https://www.mapillary.com/developer/api-documentation`

[4] `https://services.arcgisonline.com/ArcGIS/rest/services/World_Imagery/MapServer`

[5] `https://tile.openstreetmap.org/{z}/{x}/{y}.png`

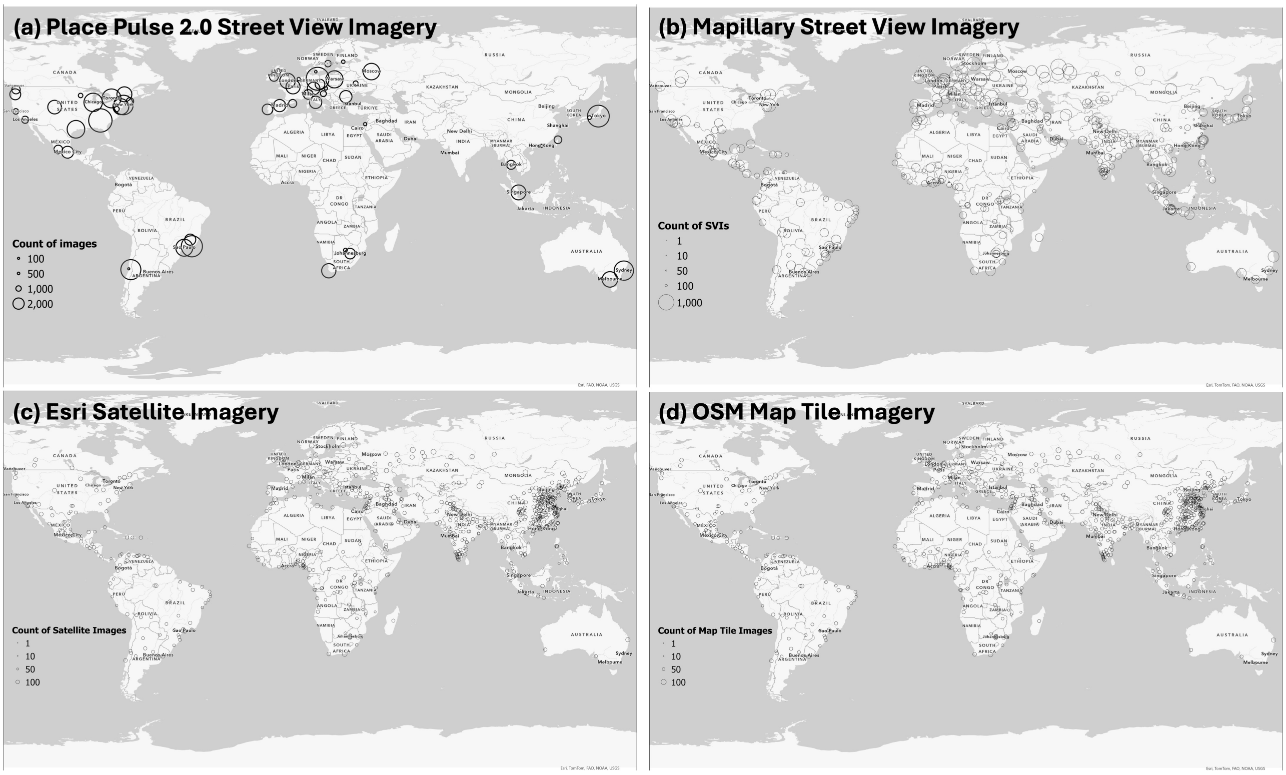
(a) Place Pulse 2.0 Street View Imagery
Count of images
100
500
1,000
2,000
(b) Mapillary Street View Imagery
Count of SVIs
1
10
50
100
1,000
(c) Esri Satellite Imagery
Count of Satellite Images
1
10
50
100
(d) OSM Map Tile Imagery
Count of Map Tile Images
1
10
50
100
Esri, TomTom, FAO, NOAA, USGS

Figure 1. Spatial distributions of (a) Place Pulse 2.0 street view imagery, (b) Mapillary street view imagery, (c) Esri satellite imagery, and (d) OSM map tile imagery. These datasets cover a global extent, obtaining a population-based spatial representation for model training in urban analytics.

Table 1. Geospatial Image–Text Pairwise Datasets Generated by GPT-4

Image dataset	**Num. training images**	**Num. testing images**	**Generated context types**						
			Image caption	**Urban perception**	**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	✓	–	✓	–	✓	✓	–

(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 (Figure 2).

Specifically, these prompts should at least reflect six dimensions, i.e., geospatial image types, geo-localization clues, spatial patterns, land use/land cover (LULC), urban perception, and urban development, which are outlined in detail in Table 2. This investigation scheme is particularly helpful for explaining how each type of spatial context can facilitate or impede the model performance in the case study (Section 4), 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.

3.3 Spatially Explicit Contrastive Learning

Within the CLIP architecture, this 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 (Figure 2). 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, an initial learning rate of $lr = 1e-7$

Table 2. 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*}.

Note: Content within "[ ]" is the assigned label for a given task.

was adopted. 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 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 (only with geo-location text), 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 $betas = (0.98, 0.999), eps = 1e-10, weight\ decay = 0.0$, 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.

3.4 A Vision-Language Learning Framework with Spatial Intelligence

The CLIP-based framework enhanced by spatial-context aware prompt engineering (Section 3.2) and spatially explicit contrastive learning (Section 3.3) to develop a GeoAI VLM (i.e., GeospatialCLIP) is demonstrated in Figure 2. To evaluate its performance, experiments have been conducted for urban mixed land use detection (Section 4). Specifically, the outcome of spatially explicit contrastive learning 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.

4. Experimental settings for Mixed Land Use Detection

Our work uses mixed land use detection as a case study to assess the proposed methods. Mixed land uses integrate various socioeconomic functionalities such as residential, commercial, industrial, and recreational spaces. Historically, mixed land use has been central to urban landscapes since the early 20th century (Moos et al., 2018), improving health conditions (Brown et al., 2009), housing values (Wu et al., 2018), crime reduction (Zahnow, 2018), and reducing automobile dependency (McCormack et al., 2001). Mixed land uses are recognized as essential for creating livable and sustainable communities (Ye et al., 2005).

However, mixed land use detection remains a challenge due to data limitations and processing methods. Most land use data assign a single label to each feature, while many support multiple functions (Pande et al., 2021). These data are typically gathered via visual interpretation of remote sensing imagery, which may not capture mixed land uses accurately, especially for multi-story properties (Helber et al., 2019). The persistent issue is the use of one-class classification for multi-class scenarios (Omrani et al., 2017). This approach relies on aggregating land uses within larger parcels, resulting in biased mixed land use detection (Tian et al., 2017). Recent studies have implemented multi-label concepts, but still face challenges such as coarse-grained aggregation and reliance on manual interpretation (Liang et al., 2021).

To address these issues, recent studies proposed using street view imagery, which offers detailed, side-view visuals of urban land uses (Zhu et al., 2019). Street view images can be geolocated at the point level, more informative than overhead-view imagery, and improving spatial resolution for land use classification (Castelluccio et al., 2015). Moreover, it has been demonstrated that VLMs combining street view images and contextual land-use prompts outperforms traditional vision-based methods (Wu et al., 2023).

4.1 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.

4.2 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-4 on mixed land use detec-

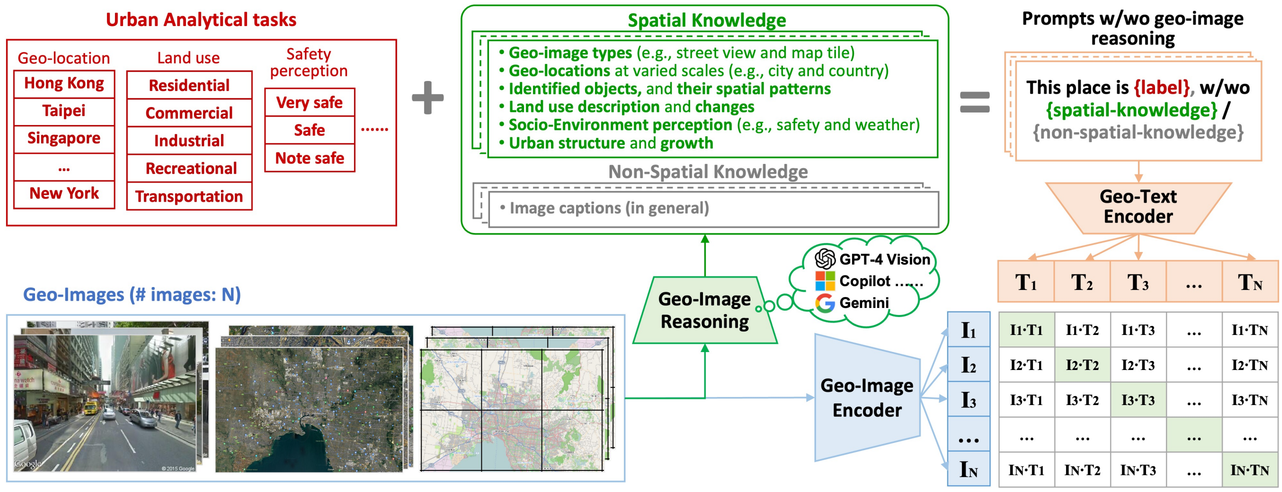
Urban Analytical tasks
Geo-location
Hong Kong
Taipei
Singapore
...
New York
Land use
Residential
Commercial
Industrial
Recreational
Transportation
Safety perception
Very safe
Safe
Note safe
......
Spatial Knowledge
• Geo-image types (e.g., street view and map tile)
• Geo-locations at varied scales (e.g., city and country)
• Identified objects, and their spatial patterns
• Land use description and changes
• Socio-Environment perception (e.g., safety and weather)
• Urban structure and growth
Non-Spatial Knowledge
• Image captions (in general)
Prompts w/wo geo-image reasoning
This place is {label}, w/wo {spatial-knowledge} / {non-spatial-knowledge}
Geo-Text Encoder
T1 T2 T3 ... TN
GPT-4 Vision
Copilot ......
Gemini
Geo-Image Reasoning
Geo-Images (# images: N)
Geo-Image Encoder
I1 I2 I3 ... IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
... ...
IN·T1 IN·T2 IN·T3 ... IN·TN

Figure 2. A framework for embedding spatial intelligence into VLMs (e.g., CLIP), achieved by spatial-context aware prompt engineering and spatially explicit contrastive learning.

tion. Particularly, the applied dataset is a subset of the Place Pulse 2.0 and consists of 3,398 Google Street View images captured from 2007 to 2012, referencing (Wu et al., 2023)'s study for mixed land use measurement and mapping using the same dataset. These images were only used for model testing, sampled from the four primary boroughs of NYC, namely Brooklyn, Queens, Manhattan, and the Bronx.

OpenStreetMap (OSM) Land Use Data OSM 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 effectiveness of using OSM land use data to train models for urban land use classification (Fonte et al., 2020). In this work, the OSM shapefiles were downloaded from the BBBike website[6]. 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 general, i.e., residential, commercial, industrial, greenfield, recreation, and transportation, following similar data pre-processing procedures of previous studies (Wu et al., 2022, Wu et al., 2023).

Next, for each location of an image, a buffer zone with a radius r (i.e., 50-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. Specifically, accuracy is defined as the percentage of correctly predicted images (i.e., the predicted land use is the OSM one, or one of them in the case of mixed uses).

5. Results and Analysis

5.1 Zero-shot Learning on VLMs

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 illustrates the contrast in prediction performance of the proposed zero-shot models versus the few-shot ResNet-152 and ViT models. 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-4 outperforms CLIP by 4.83%, reaching a remarkable accuracy at 76.10%. These results underscore the competitive edge of VLMs 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. 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.

Specifically, CLIP's visual feature space of land use labels is plotted in showing that a certain level of mixture in the visual representations 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).

[6] https://download.bbbike.org/osm/

5.2 Spatial-context Aware Prompt Engineering on VLMs

Next, Table 3 shows the accuracies of CLIP's prompt tuning results 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**:
  - **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-4 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.

Table 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%
	City label	**71.27%**	**75.15%**

5.3 Spatially Explicit Contrastive Learning on VLMs

Table 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, 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.

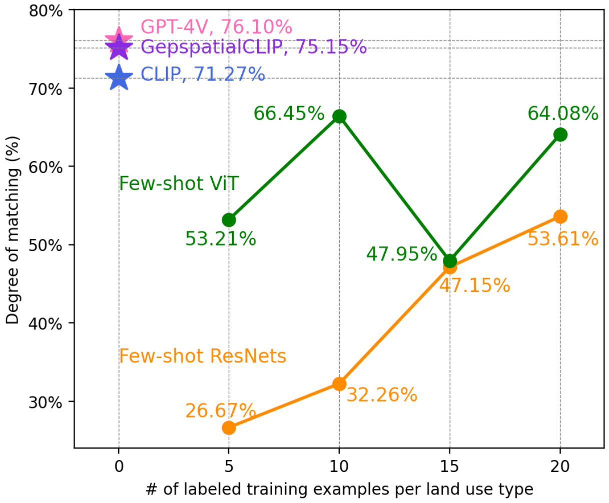

Figure 3. Comparison of model performances on mixed land use detection.

In addition, Figure 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 5.1. 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 VLMs to these baselines, GPT-4 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 VLMs. As for GeospatialCLIP, with an accuracy of 75.15%, it also exceeds the few-shot models, gaining a comparable performance with GPT-4. CLIP's performance is impressive as well, topping the few-shot models.

Based on the findings, several implications may be inferred:

- All zero-shot VLMs (GPT-4, GeospatialCLIP, CLIP) demonstrate a higher degree of matching accuracy, showcasing the strength of VLMs in leveraging learned land use representations compared with single-modal image-based models.
- Both GPT-4 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-raining 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 VLMs' capabilities, highlighting the power of VLMs to generalize sufficiently from their learned representation space and perform well on land use tasks without additional labeled data.

6. Conclusions and Discussion

In this study, we have presented a comprehensive framework for embedding spatial intelligence into VLMs, evaluated on the task of urban mixed land use detection. Our proposed method integrates two core innovations, i.e., spatial-context aware prompt engineering and spatially explicit contrastive learning, within a CLIP-based architecture. The resulting model, GeospatialCLIP, harnesses multimodal data from diverse sources (e.g., street view imagery, satellite snapshots, and map tile images), alongside geospatial text generated and curated by GPT-4, to capture the complex spatial relationships inherent in urban landscapes.

6.1 Key Contributions

The primary contributions of this work can be summarized as follows. First, we introduce a novel methodology that fuses spatial intelligence with VLMs by designing spatially aware prompts that incorporate key dimensions of urban context. This includes geo-localization clues, spatial patterns, and urban development indicators, which collectively guide the model toward a more in-depth understanding of urban scenes. Second, the development of spatially explicit contrastive learning allows GeospatialCLIP to effectively align geospatial image-text pairs in a manner that emphasizes geographic relationships. Third, we construct large-scale geospatial image-text pairwise datasets by curating high-quality inputs from multiple sources, thereby addressing the data scarcity and bias issues prevalent in existing geospatial applications. Finally, our extensive experimental evaluation on mixed land use detection demonstrates that the proposed framework not only achieves superior accuracy relative to few-shot vision-based baselines, but also attains competitive performance with advanced models (e.g., GPT-4), particularly when leveraging carefully designed spatial prompts.

6.2 Pros and Cons of the Current Approach

One of the major advantages of our framework is its ability to explicitly model spatial relationships, a feature that is largely absent in existing VLMs. By integrating spatial-context cues directly into the prompt design and contrastive learning process, GeospatialCLIP is better equipped to discern subtle variations in spatial patterns. This leads to improved interpretability and enhanced performance in detecting mixed land use scenarios, a task that traditionally challenges single-modal models. Moreover, our zero-shot learning approach underscores the generalizability of the model, reducing reliance on extensive labeled datasets and thereby accelerating deployment in real-world settings.

However, the proposed methodology also has limitations. The performance of spatial-context aware prompts is highly sensitive to the choice and formulation of the prompt, which may introduce variability and require extensive empirical tuning. Additionally, while our approach effectively captures spatial relationships in urban environments, its applicability to rural or less densely populated areas remains to be fully explored. The computational complexity inherent in integrating additional spatial modules also poses challenges for scalability, especially when processing large volumes of high-resolution geospatial data. Furthermore, the reliance on geospatial text generated by GPT-4, although innovative, may lead to biases if the textual descriptions do not fully capture local details or are influenced by the inherent limitations of the language model.

6.3 Implications and Future Directions

The integration of spatial intelligence into VLMs opens up several promising directions for future research. One immediate avenue is the exploration of adaptive prompt strategies that can dynamically adjust spatial cues based on context or region-specific characteristics. This would not only enhance model robustness but also facilitate its application across a wider range of geographies. Future work could also investigate the incorporation of additional modalities, such as LiDAR or environmental sensor data, to further enrich the spatial representations in heterogeneous environments.

Another potential direction involves extending the spatio-temporal dimension of the proposed model. Urban environments are dynamic, and the ability to accurately predict mixed land uses across different cities or suburban areas as well as to capture land use changes over time through multi-temporal imagery could significantly advance our understanding of urban evolution and transitions, thereby providing valuable insights for urban planning and policy-making.

Additionally, the development of more interpretable models remains a key priority. Although GeospatialCLIP demonstrates improved interpretability through spatial-context aware prompts, further research is needed to disentangle the specific contributions of various spatial cues to the overall decision-making process. For example, the impact of different prompt dimensions could be visualized by performing t-SNE based on the textual embedding space. Developing methods for model interpretability will be critical for building trust and facilitating the adoption of GeoAI systems in operational settings.

References

Agbaje, T. H., Abomaye-Nimenibo, N., Ezeh, C. J., Bello, A., Olorunnishola, A., 2024. Building Damage Assessment in Aftermath of Disaster Events by Leveraging Geoai (Geospatial Artificial Intelligence). *World Journal of Advanced Research and Reviews*, 23(1), 667–687.

Akter, J., Kamruzzaman, M., Hasan, R., Khatoon, R., Farabi, S. F., Ullah, M. W., 2024. Artificial intelligence in american agriculture: A comprehensive review of spatial analysis and precision farming for sustainability. *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, IEEE, 1–7.

Brown, B. B., Yamada, I., Smith, K. R., Zick, C. D., Kowaleski-Jones, L., Fan, J. X., 2009. Mixed land use and walkability: Variations in land use measures and relationships with BMI, overweight, and obesity. *Health & place*, 15(4), 1130–1141.

Caliskan, A., Bryson, J. J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.

Castelluccio, M., Poggi, G., Sansone, C., Verdoliva, L., 2015. Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092*.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F., 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14455–14465.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C. A., 2016. Deep learning the city: Quantifying urban perception at a global scale. *European conference on computer vision*, Springer, 196–212.

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

Gao, S., 2021. *Geospatial artificial intelligence (GeoAI)*. 10, Oxford University Press New York.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Iyer, L. S., 2021. AI enabled applications towards intelligent transportation. *Transportation Engineering*, 5, 100083.

Li, W., Hsu, C.-Y., 2022. GeoAI for large-scale image analysis and machine vision: recent progress of artificial intelligence in geography. *ISPRS International Journal of Geo-Information*, 11(7), 385.

Liang, X., Guan, Q., Clarke, K. C., Chen, G., Guo, S., Yao, Y., 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.

Lu, H., Zhou, Q., Fei, N., Lu, Z., Ding, M., Wen, J., Du, C., Zhao, X., Sun, H., He, H. et al., 2022. Multimodal foundation models are better simulators of the human brain. *arXiv preprint arXiv:2208.08263*.

McCormack, E., Rutherford, G. S., Wilkinson, M. G., 2001. Travel impacts of mixed land use neighborhoods in Seattle, Washington. *Transportation Research Record*, 1780(1), 25–32.

Moos, M., Vinodrai, T., Revington, N., Seasons, M., 2018. Planning for mixed use: affordable for whom? *Journal of the American Planning Association*, 84(1), 7–20.

Omrani, H., Tayyebi, A., Pijanowski, B., 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.

Pande, C. B., Moharir, K. N., Singh, S. K., Varade, A. M., Elbeltagi, A., Khadri, S., Choudhari, P., 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.

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 211-252.

Tian, L., Liang, Y., Zhang, B., 2017. Measuring residential and industrial land use mix in the peri-urban areas of China. *Land Use Policy*, 69, 427–438.

Wolniak, R., Stecuła, K., 2024. Artificial Intelligence in Smart Cities—Applications, Barriers, and Future Directions: A Review. *Smart Cities*, 7(3), 1346–1389.

Wu, J., Song, Y., Liang, J., Wang, Q., Lin, J., 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, M., Huang, Q., 2022. Im2city: image geo-localization via multimodal learning. *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 50–61.

Wu, M., Huang, Q., Gao, S., Zhang, Z., 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, X., Liu, X., Zhang, D., Zhang, J., He, J., Xu, X., 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.

Ye, L., Mandpe, S., Meyer, P. B., 2005. What is "smart growth?"—Really? *Journal of Planning Literature*, 19(3), 301–315.

Zahnow, R., 2018. Mixed land use: Implications for violence and property crime.

Zemene, E., Tesfaye, Y. T., Idrees, H., Prati, A., Pelillo, M., Shah, M., 2018. Large-scale image geo-localization using dominant sets. *IEEE transactions on pattern analysis and machine intelligence*, 41(1), 148–161.

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., Ratti, C., 2018. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180, 148–160.

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