ReFound: Crafting a Foundation Model for Urban Region Understanding upon Language and Visual Foundations

Congxi Xiao[†]
School of Computer Science and Technology, University of Science and Technology of China
Hefei, China
xiaocongxi@mail.ustc.edu.cn

Jingbo Zhou[*]
Business Intelligence Lab, Baidu Research
Beijing, China
zhoujingbo@baidu.com

Yixiong Xiao
Business Intelligence Lab, Baidu Research
Beijing, China
xiaoyixiong@baidu.com

Jizhou Huang
Baidu Inc.
Beijing, China
huangjizhou01@baidu.com

Hui Xiong[*]
Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
Hong Kong SAR, China
xionghui@ust.hk

ABSTRACT

Understanding urban regional characteristics is pivotal in driving critical insights for urban planning and management. We have witnessed the successful application of pre-trained Foundation Models (FMs) in generating universal representations for various downstream tasks. However, applying this principle to the geospatial domain remains challenging, primarily due to the difficulty of gathering extensive data for developing a dedicated urban foundation model. Though there have been some attempts to empower the existing FMs with urban data, most of them focus on single-modality FMs without considering the multi-modality nature of urban region understanding tasks. To address this gap, we introduce **ReFound** – a novel framework for Re-training a Foundation model for urban region understanding, harnessing the strengths of both language and visual FMs. In this framework, we first invent a Mixture-of-Geospatial-Expert (MoGE) Transformer, to effectively integrate the embedding of multi-source geospatial data. Building on this, ReFound is enhanced by jointly distilling knowledge from language, visual, and visual-language FMs respectively, thus augmenting its generalization capabilities. Meanwhile, we design a masked geospatial data modeling approach alongside a cross-modal spatial alignment mechanism, to enhance the spatial knowledge of ReFound derived from geospatial data. Extensive experiments conducted on six real-world datasets over three urban region understanding tasks demonstrate the superior performance of our framework.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**.

KEYWORDS

Foundation model, Multimodal data, Urban region understanding

ACM Reference Format:
Congxi Xiao[†], Jingbo Zhou, Yixiong Xiao, Jizhou Huang, and Hui Xiong. 2024. ReFound: Crafting a Foundation Model for Urban Region Understanding upon Language and Visual Foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671992

1 INTRODUCTION

Urban region understanding, aimed to quantitatively explore urban regions' specific characteristics, is essential for generating insights crucial to informed, scientifically-based urban planning and management. Urban regions, fundamental to city life where people live, work, and entertain, are becoming increasingly complex and diverse due to accelerated urbanization. Consequently, leveraging publicly available urban data and machine learning methods to infer regions' attributes has gained significant research interest, encompassing tasks like urban village detection [9, 59], population prediction [3, 29], house price prediction [22, 49, 50, 54], community vibrancy estimation [51] and socioeconomic forecasting [1, 14, 30, 53, 66]. These problems usually require special domain expertise, enough labeled training data, and task-specific model designs.

In light of the successful application of Foundation Models (FMs) in Natural Language Processing (NLP) and Computer Vision (CV),

[*]Corresponding authors. [†]This work was done when the first author was an intern at Baidu Research under the supervision of Jingbo Zhou.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671992

a promising direction is to apply the pre-training paradigm for urban region understanding. This approach aims to develop an FM that can generate universal region representation, applicable to various downstream applications. However, general-purpose FMs, such as GPT-3 [7], ViT [15], and CLIP [40], are mainly trained with plain text and images, leading to underperformance in urban tasks due to a lack of spatial domain knowledge [36]. A practical necessity exists for developing specialized FMs tailored to urban region understanding.

However, it is challenging to effectively build an FM for urban region understanding, primarily due to the difficulty of acquiring the vast amount of corpus to pre-train an urban FM, in contrast to the general-purpose FMs. The ability of FMs, particularly large, state-of-the-art transformer-based models, to generalize effectively hinges on pre-training with extensive data. For instance, the GPT-3 [7] is trained with about 570 GB plaintext (45 TB before filtering). Similarly, ImageNet-21k [11] and JFT [44] datasets, which are commonly used for pre-training visual FMs, encompass over 14 million and 300 million images, respectively. Additionally, the CLIP model was trained by Radford et al. [40] using a vast collection of 400 million image-text pairs.

In contrast, the volume of available urban data is significantly less than that used to pre-train general-purpose FMs. An urban area (i.e., a city) is usually segmented into only a few thousand to tens of thousands of regions [3, 22]. Even if collecting data from multiple cities, it remains insufficient compared to the datasets used for general-purpose FM pre-training. This disparity necessitates developing a pre-training strategy that allows our model to achieve a comparable level of generality as FMs, but in a more data-efficient manner. Consequently, a compelling approach is to incorporate spatial knowledge of limited urban data, as well as the well-established FMs to craft an FM for urban region understanding.

While there have been some pioneering attempts to incorporate spatial knowledge into pre-trained FMs for urban applications, these studies typically focus on a single FM and have limited capacity to exploit multi-modal information for various urban tasks. A few studies have explored pre-training or post-training FMs using POI data with a language FM. For instance, SpaBERT [32] pre-trains a BERT model that encodes POIs, taking into account their relative positions, and GeoBERT [18] introduces a position embedding to reflect the distance of each POI to the region center. A recent study has pre-trained an FM from scratch using spatial entities from OpenStreetMap [4]. Another category of approaches construct geospatial visual FMs [37], using satellite imagery and based on a general-purpose visual FM pre-trained on ImageNet.

Our motivation stems from the premise that leveraging multiple well-established FMs with multi-modal urban data could be more advantageous than relying on a single FM. In other words, rather than depending solely on a language or a visual FM for training an urban FM, our goal is to combine several FMs into a cohesive framework. It would harness the textual data understanding capabilities of large language models, the image data understanding skills of visual FMs (such as ViT [15]), and the cross-modal data comprehension of visual-language FMs (e.g., CLIP [40]), to address the multi-modal nature of region understanding tasks. Recent surveys [36, 67] also highlight the necessity of combining and aligning different modalities, like POIs and satellite images, each containing

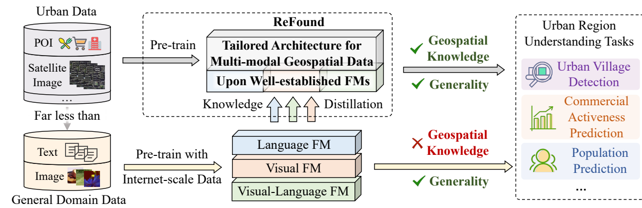


Figure 1: An illustration of the advantages of ReFound.

unique geospatial knowledge. It is identified as a significant challenge in developing FMs for urban applications. Consequently, there is a clear need to design a framework capable of effectively integrating multi-modal data (POIs and satellite images) and extracting generalized knowledge from different FMs.

In this paper, we present a framework that can craft Re-training a Foundation model for urban region understanding, termed as **ReFound**, based on existing language FMs, visual FMs, and visual-language FMs. As shown in Figure 1, we develop a tailored model architecture, which learns in-domain knowledge from multi-modal geospatial data while also leveraging the strong generality of existing FMs via knowledge distillation. This model is empowered with both general and specific domain knowledge, thereby capable of addressing a wide range of urban region understanding tasks.

Specifically, we propose to employ two primary data sources, which are POI and satellite image data. These data types are readily accessible via the Internet, which is a factor that greatly facilitates their use in both the pre-training phase and various downstream tasks. While other data types, such as human mobility [17, 64] and street-view data [8, 24], may also be useful but have limited coverage (e.g., difficulty in collection and privacy restrictions [26]).

To adapt to these geospatial data, we devise a multi-modal geospatial data embedding layer. It integrates the textual and visual content from POI and satellite image data within their respective spatial contexts. Subsequently, we employ a Mixture-of-Geospatial-Expert (MoGE) Transformer encoder. This encoder is specifically tailored to adapt to the unique characteristics of both types of urban data and facilitates a deep fusion between them. Upon this architecture, we formulate three distinct distillation objectives. These objectives are designed to transfer the extensive knowledge and generalization capabilities from well-established pre-trained language FM, visual FM, and visual-language FM, thereby augmenting ReFound's effectiveness. This approach also enables continual improvement of our model by leveraging advancements in general-purpose FMs. Additionally, to capture the nuances of the geospatial domain, we introduce two self-supervised learning tasks specifically tailored to urban data. The first is a unified objective for masking both POI and satellite image data; the second involves a cross-modal spatial alignment task, designed to align the semantics of the two modalities based on their spatial relationships.

We conducted comprehensive evaluations of our framework on three urban region understanding tasks across two cities. The experimental results demonstrate that our framework achieves significant improvements compared to other state-of-the-art methods. The major contributions of this paper are summarized as follows.

- We present a novel framework, termed as ReFound, marking the first endeavor to construct a special FM with multi-modal urban data as input for urban region understanding. This model leverages well-established language and visual FMs, and harnesses publicly available multi-modal urban data.

- To effectively build this model, we have carefully developed several novel components, including the multi-modal geospatial data embedding, the MoGE transformer, a distillation approach from FMs, masked geospatial data modeling, and a cross-modal spatial alignment mechanism.
- Comprehensive experimental evaluations have been conducted to validate the effectiveness of ReFound, demonstrating its superiority over state-of-the-art methods.

2 RELATED WORK

Foundation Models. In recent years, foundation models (FMs) have achieved great success across domains. Thanks to the powerful Transformer model [48] and self-supervised learning techniques, pre-trained FMs, e.g., BERT [12], GPT [41] and LLaMA [47], can capture the universal knowledge underlying massive unlabeled data, which can be employed in various tasks. Similarly, researchers in CV domain also build large-scale visual FMs competent in diverse vision tasks, where ViT [15], MAE [20] and BEiT [5] are well-known examples. Multi-modal FMs, such as CLIP [40], BLIP [27], BEiT-3 [52] and GPT-4 [2] have brought widespread attention.

In particular, some specific FMs are developed for geospatial domains. For example, SatMAE [10] and GFM [37] use masked image modeling on satellite images to pre-train FMs for geospatial applications. Some other studies also explore pre-training FMs for geo-entities or urban space representation with POI-related data, such as POI names [32], POI tags [18], map search history [21], geographic objects [13] and knowledge entities from OpenStreetMap [4]. Especially, CityFM [4] aims to pre-train a model from scratch to produce representations for different types of geo-entities. But it is not designed to consider the satellite image data. These FMs are mainly capable of modeling unimodal data, which cannot adapt to diverse urban region understanding. More recently, UrbanCLIP [60] combines multi-modal satellite images and LLM-generated textual descriptions for urban region profiling, however, it directly adopts the component trained for general language tasks, unlike our specially designed architecture which can effectively model the geographic data like POIs and capture its spatial characteristics. There is a recent survey [65] comprehensively summarizing the research efforts to constructing specific FMs for geospatial tasks. It also highlights the critical role of integrating multi-modal urban data and handling their spatial properties in building an FM for a wide array of urban applications.

Urban Region Embedding. Studies of urban region embedding, which focus on learning general urban region representation in a self-supervised manner, can be also viewed as applying the pre-training paradigm to obtain a model transferred to downstream urban region understanding tasks. Basically, these approaches first leverage uni-modal or multi-modal urban data to construct an urban region's attributes (e.g., POI categories [22], satellite image features [3], street views [30] and building groups [31]), and to characterize dependencies among regions (e.g., human mobility [57], functionality similarity [64], spatial proximity [31] and multiple relationships upon an urban knowledge graph [34]) from different views. Then, the region embeddings are learned by preserving certain region attributes and inter-region correlations, such as developing region relation reconstruction tasks and designing related contrastive learning objectives [63]. However, the typical practice of these methods is to train a specific model for an individual city without considering how to utilize the well-established FMs, which is hard to obtain high generalization ability like FMs.

3 PRELIMINARIES

In this section, we first introduce the basic concepts and data used in this study, then clarify the goal of our work.

Region. Regions refer to the geographical divisions of an urban area (e.g., a city) under a certain partition strategy. Different regions present different characteristics. In our work, without loss of generality, we obtain the region set $\mathcal{R} = \{R_1, R_2, ..., R_n\}$ by partitioning the urban area into non-overlapping grids of size $L_r m \times L_r m$.

Point of Interest. Points of interest (POIs) are venues offering a variety of services, such as restaurants and hospitals. Within a region R_i, there are usually a set of POIs $\mathcal{P}_i = \{P_{i1}, P_{i2}, ..., P_{im_i}\}$, where m_i denotes the number of POIs. Each POI has three attributes: textual name, category-id, and location (e.g., longitude and latitude), denoted as $name_{ij}$, c_{ij} and loc_{ij}, respectively. These attributes provide rich functional and spatial information of POIs, which help characterize potential human activities within the region.

Satellite Image. Each region R_i is covered by a satellite image S_i that captures its visual appearance from the over-head view. It contains rich geospatial information, such as spatial distributions of buildings and roads, as well as land-use types, which have been shown to be helpful in diverse region understanding tasks [25, 30].

The goal of this work is to pre-train a multi-modal foundation model for urban region understanding, based on existing FMs and urban data including POIs and satellite images. This model is expected to derive region representations with rich semantics to generally address various urban region understanding tasks.

4 METHODOLOGY

In this section, we detail our ReFound framework. We first introduce the model architecture design of ReFound. Then upon the model architecture, we propose how to empower ReFound with the ability to learn universal urban region representation with pre-training.

4.1 Architecture Design of the Framework

For the model architecture design, our ReFound mainly consists of two parts. First, we propose a Multi-modal Geospatial Data Embedding to transform POI data and satellite image data into a unified embedding sequence, which comprehensively integrates the textual, visual, and geospatial information within these data. Then, a Mixture-of-Geospatial-Experts Transformer performs the deep interaction and fusion among them to produce contextualized representations. For simplicity, we omit the subscript i, using $\mathcal{P} = \{P_1, P_2, ..., P_m\}$ and S to denote the POI and satellite image data in region R_i if without confusion.

4.1.1 *Multi-modal Geospatial Data Embedding*.

This module converts the raw POI and satellite image data into compact embedding with considering their geospatial context.

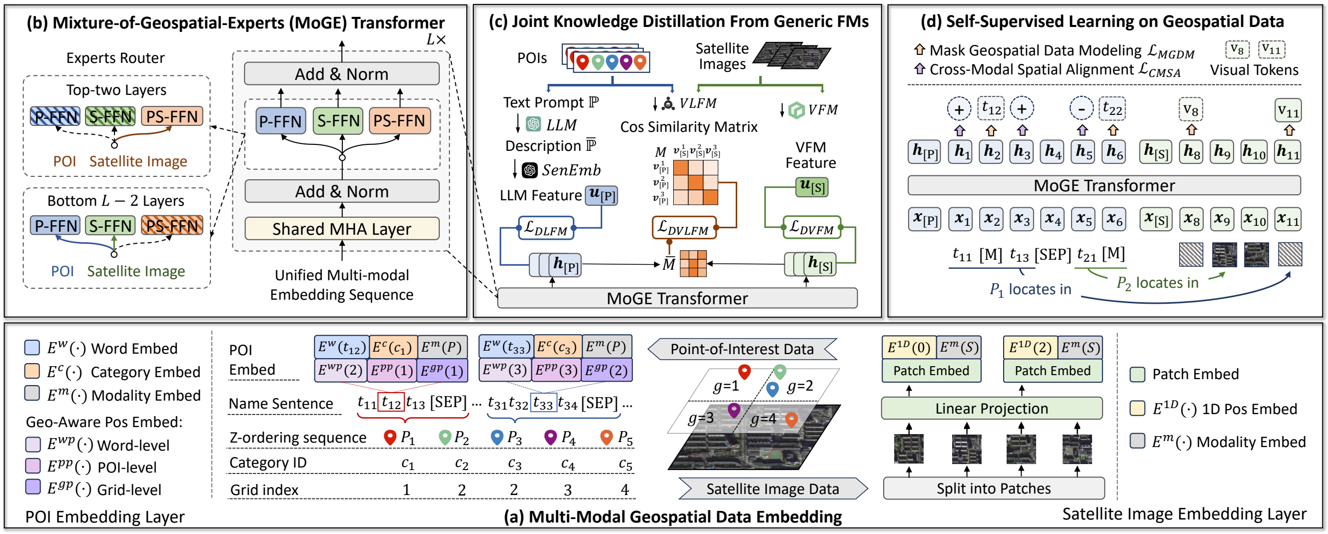


Figure 2: The architecture and pre-training framework of ReFound.

POI Embedding. To derive the POI embedding of a region, we propose a joint encoding that integrates *Word Embedding*, *Geo-Aware Position Embedding*, and *Category Embedding* of all POIs in the region. The resulted embedding can effectively capture not only the textual toponym knowledge of POIs, but also their spatial distribution. Specifically, we **first** organize the names of POIs in the region into a *pseudo sentence*, making it compatible with the Transformer's input, and then encode this textual data into *Word Embedding*. **Second**, to consider the positional relation between words, we design a *Geo-Aware Position Embedding* to replace the conventional one as used in BERT [12] and GPT [41]. This is because words in this *pseudo sentence* come from names of different POIs, which are irregularly distributed in an urban region. They do not follow the rules in human language (e.g., grammar and discourse structures), but instead possess spatial relationships according to the geographical distribution of POIs, which hampers the use of the original sequence position embedding method. Our *Geo-Aware Position Embedding* is specially designed to handle such complex relationships without sacrificing the geospatial information. **Third**, the POI category id is further encoded by *Category Embedding*, as it embodies valuable functional semantics of a venue, whose benefits of characterizing an urban region have been extensively validated in previous studies [25, 64]. We formally define them as follows:

Word Embedding. For POIs $\mathcal{P} = \{P_1, P_2, ..., P_m\}$ in a region, they are first organized into the name sequence (pseudo sentence): $name_1\ name_2 \ldots name_m$. Then, we tokenize it and use [SEP] token to separate name tokens between different POIs:

$$t_{11}\, t_{12} \ldots t_{1n_1}\ \text{[SEP]}\ \mathrm{t}_{21}\, \mathrm{t}_{22} \ldots \mathrm{t}_{2n_2}\, \text{[SEP]} \ldots \mathrm{t}_{m1}\, \mathrm{t}_{m2} \ldots \mathrm{t}_{mn_m},$$

where t_{jk} denotes k-th token of P_j's name. Next, a word embedding table $E^w(\cdot)$ maps them into embedding space by: $E^w(t_{jk}) \in \mathbb{R}^d$.

Geo-Aware Position Embedding. We encode complex sequential and spatial relationships among POI name tokens into our model via carefully designed position embeddings from three levels:

- *Word-level.* For each POI P_j in the region, a word-level sequence position embedding is adopted to represent the token order within its name, which is same as the original position encoding in BERT [32]. We use $E^{wp}(k)$ to denote this embedding for token t_{jk} in $name_j$, derived by embedding function $E^{wp}(\cdot)$.
- *POI-level.* The POI-level position embedding is designed to account for the positional relationships of name tokens for different POIs in geographical space. Motivated by Tobler's First Law of Geography [46] that nearby things are more related, we encode the relative distance between POIs by serializing them based on geographic proximity, ensuring geographically closer ones are positioned nearer to each other. The serialization is achieved by Z-ordering strategy [39], a commonly used method to project 2-D geographical points into one dimension while preserving original locality [28]. Specifically, given a region, we rasterize it into $1m \times 1m$ fine-grained units. Each unit is associated with a Z-value yielded by Z-ordering function, and closer values indicate closer spatial distance between two units. Then, this value is assigned to POIs located in the corresponding unit, and thus, connecting POIs in order of their Z-values can produce the expected POI sequence for a region. Note that for those very close POIs in the same unit, we randomly set the order between them. Assuming that subscript j of P_j denotes POI's order in the resulted sequence, name token t_{jk} from $name_j$ will obtain the position embedding according to order j in the sequence: $E^{pp}(j) \in \mathbb{R}^d$, where $E^{pp}(\cdot)$ is the POI-level position embedding function.
- *Grid-level.* In addition to the distance between POIs, we further consider their 2D spatial distribution in the region, since this information is indicative of region's functionality [4, 36]. To accomplish this, we first discretize the region into non-overlapping grids, and assigned learnable embeddings to represent grids' relative positions. Then, the 2D position of a POI is defined as the grid it locates in. Formally, we split the region into $G = L_r^2/L_g^2$ uniform grids with size of L_g m $\times$ L_g m, and index them by $g = 1, 2, ..., G$. Then, the grid-level position embedding $E^{gp}(g) \in \mathbb{R}^d$ is produced for name tokens from POIs in grid g.

Overall, the geo-aware position embedding is finally obtained by the combination of three levels: $E^p(t_{jk}) = E^{wp}(k) + E^{pp}(j) + E^{gp}(g)$.

Category Embedding. For each POI P_j, a trainable embedding table $E^c(\cdot)$ maps its category-id into category embedding $E^c(c_j)$, which is shared to every name token t_{jk} of this POI.

Finally, we encode POI data by summing up the word embedding, geo-aware position embedding, category embedding and an additional modality embedding $E^m(P)$:

$$E^P(t_{jk}) = E^w(t_{jk}) + E^p(t_{jk}) + E^c(c_j) + E^m(P) \quad (1)$$

The resulted POI embedding sequence can be denoted by: $\boldsymbol{X}^P = \{\boldsymbol{x}_{[\mathrm{P}]}, \boldsymbol{x}_1^P, \boldsymbol{x}_2^P, ..., \boldsymbol{x}_{L^P-1}^P\}$, where $\boldsymbol{x}_i^P$ is computed based on Eq.(1). L^P denotes the max sequence length, and $\boldsymbol{x}_{[\mathrm{P}]}$ is the embedding of the CLS token [P] inserted to the head of sequence.

Satellite Image Embedding. Following ViT [15] and a recent geospatial FM [37], we represent the satellite image by directly splitting it into patches and encoding the patches with linear projection. Formally, satellite image $S \in \mathbb{R}^{H\times W\times 3}$ is reshaped into a sequence of $s \times s$ patches with length $L^S = HW/s^2$, which are linearly projected into d-dimensional patch embeddings. Then, we prepend a learnable CLS token [S] to sequence, and insert learnable 1D position embeddings by $E^{1D}(\cdot)$ and modality embedding $E^m(S)$ to each patch to get the satellite image embedding: $\boldsymbol{X}^S = \{\boldsymbol{x}_{[\mathrm{S}]}, \boldsymbol{x}_1^S, \boldsymbol{x}_2^S, ..., \boldsymbol{x}_{L^S}^S\}$.

Finally, sequences of POIs and satellite image embeddings are concatenated to obtain the unified multi-modal embedding sequence of a region: $\boldsymbol{X} = [\boldsymbol{X}^P; \boldsymbol{X}^S]$, whose length is $L = L^P + L^S + 1$.

4.1.2 *Mixture-of-Geospatial-Experts Transformer.*

After the embedding module, we propose a Mixture-of-Geospatial-Experts (MoGE) Transformer encoder that generates the contextualized region representation upon multi-modal inputs. Following the multiway transformer approach [6, 52], the core concept of MoGE transformer is applying specialized sub-networks to adapt to different types of geospatial data. This strategy effectively addresses both modality-specific patterns and the complex cross-modal dependencies observed in POIs and satellite images of regions.

As shown in Figure 2(b), the MoGE Transformer replaces the single feed-forward network (FFN) of the standard Transformer[48] with a collection of sub-networks, each possessing distinct parameters. These are designated as geography experts and are specifically designed to process different data types: POI data (P-FFN), satellite image data (S-FFN), and both data types (PL-FFN). As indicated in previous work, different forms of geospatial data exhibit special structures and unique characters [36]. It's hard for a single network to effectively represent them. Whereas, when applying MoGE Transformer, different parts of the input sequence are routed to corresponding specialized experts, according to their modality. These expert sub-networks can adjust to different modalities to handle their specific patterns.

Moreover, the MoGE Transformer adopts a one-tower architecture with multi-head self-attention (MSA) shared across modalities at each layer. It enables the deep fusion between POI and satellite image data, as well as their spatial information. The shared parameters foster the semantics alignment [6] and knowledge transfer [38] across modalities, which are critical for multi-modal region representation learning as highlighted by extensive research [59, 64]. Note that at the top-two layers, we also use PL-FFN for both POI and satellite image data to facilitate the modality fusion.

Taking multi-modal embedding $\boldsymbol{X}$ as input, the MoGE Transformer produces contextualized representations $\boldsymbol{H} = \{\boldsymbol{h}_0, \boldsymbol{h}_1, ..., \boldsymbol{h}_L\}$, where $\boldsymbol{h}_{[\mathrm{P}]} = \boldsymbol{h}_0$ and $\boldsymbol{h}_{[\mathrm{S}]} = \boldsymbol{h}_{L^P}$ correspond to two CLS tokens that pool the POI and satellite image representation, respectively.

4.2 Pre-training Objectives

Upon the above architecture designs, we pre-train ReFound with two series of objectives. At first, we propose three objectives that jointly distill abundant knowledge from multiple general-purpose pre-trained FMs. This transfers existing FMs' generalization capacities to ReFound for addressing diverse tasks. Second, as spatial domain knowledge is essential for urban region understanding, we design two self-supervised learning tasks, to capture in-domain feature from multi-modal geospatial data.

4.2.1 *Joint Knowledge Distillation from Generic FMs.*

To empower ReFound with universal effectiveness in diverse tasks, we acquire strong representing ability from pre-trained FMs. As ReFound is expected to effectively handle multi-modal information, including textual (POI name) and visual data (satellite image), we design three distilling objectives to enhance it by simultaneously taking advantage of the abundant knowledge of language foundation models (LFMs), the representing power of visual foundation models (VFMs), as well as the semantic alignment ability of visual-language foundation models (VLFMs). Basically, the knowledge distillation follows the teacher-student paradigm, where the student model ReFound is jointly guided by three teacher models: LFM, VFM and VLFM. An illustration is shown in Figure 2(c).

Distillation of Language Foundation Model (DLFM). We first distill large language models (LLMs, which are definitely LFMs) to enhance ReFound's understanding of region functionalities based on POI data, from a natural language perspective. Trained on extensive datasets, LLMs possess rich real-world knowledge and powerful reasoning capabilities [56]. They can provide valuable insights not originally captured in POI data, such as the availability of life services and potential resident activities within an area. To leverage this, we propose prompting the LLM to generate supplemental descriptions of a region's functionality based on the POI data. The knowledge from the LLM teacher is then encoded into an LLM-feature, guiding ReFound in capturing functional semantics from POI data through feature-based distillation [19].

In detail, given POIs $\mathcal{P}$ of a region, we derive textual prompt $\mathbb{P}$ based on POI names. The procedure of generating and encoding LLM-based region function description can be expressed as:

$$\boldsymbol{u}_P = SenEmb(\overline{\mathbb{P}}), \;\; \overline{\mathbb{P}} = LLM(\mathbb{P}) \quad (2)$$

where $\overline{\mathbb{P}}$ denotes region's function description based on POI data, generated by $LLM(\cdot)$, and $\boldsymbol{u}_P$ is the LLM-feature obtained by sentence embedding model $SenEmb(\cdot)$. A specific example of this augmentation is provided in Appendix A.4. Then, to infuse the rich semantic information within LLM-feature into ReFound, the knowledge distillation objective is formed with the cosine similarity between $\boldsymbol{u}_P$ and POI representation derived by our model:

$$\mathcal{L}_{DLFM} = -Cos(\, \sigma_{POI}(\boldsymbol{h}_{[\mathrm{P}]}),\, \boldsymbol{u}_P\,), \quad (3)$$

where $Cos(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{a}\cdot\boldsymbol{b}/(||\boldsymbol{a}||_2 \cdot ||\boldsymbol{b}||_2)$ denotes the cosine similarity between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\boldsymbol{h}_{[\mathrm{P}]}$ is the pooled POI representation derived from CLS token [P], and $\sigma_{POI}(\cdot)$ denotes a linear to transform it into LLM-feature space. This objective enables our model to capture the functionality semantics of regions from POIs, under the guidance of the informative LLM-feature.

Distillation of Visual Foundation Model (DVFM). We enhance ReFound's semantic representation capability for satellite images by distilling visual foundation models (VFMs). Trained on expansive datasets like ImageNet-22k [11], VFMs possess exceptional image representation capabilities. Prior studies have highlighted VFMs' superior performance in certain geospatial tasks over models pre-trained exclusively on satellite imagery [10, 37].

Specifically, given the satellite image S of a region, we adopt the pre-trained visual foundation model $VFM(\cdot)$ as teacher model to extract its semantic feature by: $\boldsymbol{u}_S = VFM(S)$. Then, this feature guides the satellite image representation of our student model, via the cosine similarity objectives:

$$\mathcal{L}_{DVFM} = -Cos(\,\sigma_{Sate}(\boldsymbol{h}_{[\mathrm{S}]}),\, \boldsymbol{u}_S\,), \tag{4}$$

where $\boldsymbol{h}_{[\mathrm{S}]}$ is the pooled satellite image representation learned by our ReFound model, and $\sigma_{Sate}(\cdot)$ denotes the linear projection head for this distillation task. In this way, ReFound learns from VFM about how to extract semantic features from satellite images.

Distillation of Visual-Language Foundation Model (DVLFM). As noted in recent studies [25], encoding multi-modal geospatial data into a semantically aligned space is essential for consistently commendable performance across various region understanding tasks. Accordingly, we further enhance ReFound's semantic alignment between text-based POI and satellite image data, via a knowledge distillation of visual-language foundation models (VLFMs), since they have demonstrated impressive power to jointly understand text and image contents in a range of general [27, 40] and domain-specific [43, 55] applications.

Following [45], we accomplish this by matching the POI-satellite image cross-modal cosine similarity matrix from ReFound with that derived from a VLFM teacher. Recent VLFMs (e.g., CLIP [40]) are typically pre-trained to be able to compare the semantic similarity between samples from different modalities, using cosine similarity in the latent space. Thus, the cosine similarity matrix, formed by a batch of POI and satellite image representations from VLFMs, reflects their semantic comparison relationships. Building on this idea, if the matrix generated by ReFound matches the one derived from VLFMs, we can transfer the VLFMs' powerful cross-modal alignment ability to our model.

Specifically, given a batch of regions $\{R_i\}_{i=1}^B$ with batch size B, we use a VLFM teacher model to encode their POI and satellite image data into semantic representations: $\{\boldsymbol{v}_P^i\}_{i=1}^B$ and $\{\boldsymbol{v}_S^i\}_{i=1}^B$, which form the cosine similarity $M \in \mathbb{R}^{B\times B}$, where $M_{i,j} = Cos(\boldsymbol{v}_P^i, \boldsymbol{v}_S^j)$. In this matrix, the i-th row (column) reflects the semantic relationships between POIs (satellite image) of region R_i and satellite images (POIs) of all regions in the batch. Then, our model also generates representation vectors $\{\boldsymbol{h}_{[\mathrm{P}]}^i\}_{i=1}^B$ and $\{\boldsymbol{h}_{[\mathrm{S}]}^i\}_{i=1}^B$ for two modalities, and calculate the matrix $\overline{M}$ with $\overline{M}_{ij} = Cos(\mu_{POI}(\boldsymbol{h}_{[\mathrm{P}]}^i), \mu_{Sate}(\boldsymbol{h}_{[\mathrm{S}]}^j))$, where μ_{POI} and μ_{Sate} are linear projection heads for two modalities. The knowledge distillation from VLFM to ReFound (i.e. matching $\overline{M}$ to M) is achieved by minimizing the KL-divergence of every corresponding row and column between these two matrices:

$$\mathcal{L}_{DVLFM} = \sum_{1\le i\le B} KL(\rho(\overline{M}_i)||\rho(M_i)) + \sum_{1\le j\le B} KL(\rho(\overline{M}_j^T)||\rho(M_j^T)) \tag{5}$$

where ρ denotes the softmax function that transforms the row and column of cosine similarity matrix into a probability distribution.

Note that in practice, for the VLFM teacher model side, we use its satellite image embeddings as pseudo POI embeddings to form the matrix M, i.e. $M_{i,j} = Cos(\boldsymbol{v}_S^i, \boldsymbol{v}_S^j)$, rather than directly applying its text encoder on POI data. This is because VLFMs' text encoder are generally pre-trained on visually-grounded text, such as the caption of the paired image. While for POI name sequence that does not directly describe the satellite image content, VLFMs' text encoder may be unreliable in aligning its semantics to the paired satellite image. As indicated in [45], embedding $\boldsymbol{v}_S^i$ of the satellite image, can be also viewed as the embedding $\boldsymbol{z}_P^i$ whose semantic is perfectly aligned with $\boldsymbol{v}_S^i$ in VLFMs' latent space: $\boldsymbol{z}_P^i = \boldsymbol{v}_S^i$. Thus, it's reasonable to replace $\boldsymbol{v}_P^i$ with $\boldsymbol{v}_S^i$ to guarantee capturing correct semantic relationships in M for effective knowledge distillation.

4.2.2 *Self-Supervised Learning on Geospatial Data.*

To learn in-domain features underlying two kinds of geospatial data, we pre-trained ReFound with two objectives: Masked Geospatial Data Modeling and Cross-modal Spatial Alignment. They allow ReFound to understand the semantics of POI and satellite image data, and align two modalities via their spatial relationships in the region.

Masked Geospatial Data Modeling (MGDM). The mask-then-predict paradigm has shown promising performance in pre-training FMs to learn semantic representation for texts [12], images [5], and multi-modal (e.g., text-image pairs) data [52]. Following this line, we pre-train ReFound via a masked prediction objective on multi-modal geospatial data. Basically, it first performs a unified masking of both POI and satellite image input, and then asks ReFound to recover them based on the joint understanding of remaining textual and visual content, as well as their geospatial relationships. As shown in Figure 2(d), for POI side, we randomly mask 15% of POI name tokens with a special token [M], and ReFound learns how to complete these POI names. While for the satellite image, we follow BEiT [5] to replace a portion of patches (40%) with a mask embedding, and predict the discrete visual tokens at these positions, which are obtained by a publicly available image tokenizer [42].

Formally, denoting positions of masked POI name tokens and satellite image patches as $\mathcal{M}_P$ and $\mathcal{M}_S$ respectively, the input sequence, corrupted at these positions, is encoded by the model described in Section 4.1 into contextualized representation vectors $\boldsymbol{H}$. Then, the training objective is to minimize the negative log-likelihood of the original POI name tokens at positions $\mathcal{M}_P$, as well as the correct visual tokens at $\mathcal{M}_S$:

$$\mathcal{L}_{MGDM} = -\sum_{i\in\mathcal{M}_P} log\, p(y_i^P|\boldsymbol{h}_i) - \sum_{i\in\mathcal{M}_S} log\, p(y_i^S|\boldsymbol{h}_i) \tag{6}$$

where $p(y_i^P|\boldsymbol{h}_i)$ in the first term represents the predicted probability to the correct POI name token y_i^P, based on the encoded vectors at masked positions $\{\boldsymbol{h}_i : i \in \mathcal{M}_P\}$. This prediction is made by a 2-layer Multi-Layer Perceptron (MLP) classifier with softmax function. Similarly, $p(y_i^S|\boldsymbol{h}_i)$ in the second term denotes the probability of predicting the correct image tokens y_i^S at positions $\mathcal{M}_S$. In this process, we only mask POI names and satellite image patches, while keeping the geo-aware position embedding unchanged. It facilitates the model to consider spatial contexts when representing these data.

Cross-Modal Spatial Alignment (CMSA). Though POIs and the satellite image describe a region from very different views, their geospatial relationships serve as a connection between these two modalities, because each POI corresponds to a venue in the satellite image. A previous study [36] also suggests the possibility to bridge the gap between multi-modal geographical data via the spatial relationship. In view of this, the goal of CMSA task is to make the model aware of which POIs correspond to which parts of visual content in the satellite image, thereby further facilitating the alignment of semantic information between two modalities.

Inspired by the alignment task in [23], this objective asks the model to determine whether a POI is located in an area that is masked in the satellite image. As shown in Figure 2(d) we perform a binary classification on representation vectors at the POI side, and optimize the model with binary cross-entropy loss. Note that positions of masked POI name tokens $\mathcal{M}^P$ are not included in the loss calculation, to avoid the trivial solution that simply maps [M] token to the positive class. It can be expressed by:

$$\mathcal{L}_{CMSA} = \sum_{1 \leq i < L_P \wedge i \notin \mathcal{M}_P} -y_i log\, p(y_i|\boldsymbol{h}_i) - (1-y_i) log(1-p(y_i|\boldsymbol{h}_i)) \quad (7)$$

where binary label y_i indicates whether this position is included in the POI that locates at masked image patches ($y_i = 1$) or not ($y_i = 0$), and $p(y_i|\boldsymbol{h}_i)$ is the output probability of a sigmoid classifier.

4.3 Usage of ReFound

Based on POI and satellite image data of urban regions collected from multiple cities, ReFound is jointly pre-trained with three knowledge distillation objectives in Section 4.2.1 and two self-supervised learning objectives in Section 4.2.2. The overall loss function is: $\mathcal{L} = \mathcal{L}_{DLFM} + \mathcal{L}_{DVFM} + \mathcal{L}_{DVLFM} + \mathcal{L}_{MGDM} + \mathcal{L}_{CMSA}$.

After pre-training ReFound, we propose to obtain the final region representation by merging POI and satellite image representations $\boldsymbol{h}_{[\mathrm{P}]}$ and $\boldsymbol{h}_{[\mathrm{S}]}$, through averaging or attentional fusion [58]. Then, the pre-trained ReFound can be transferred to solve downstream urban region understanding tasks in the following two ways: **(1) Fine-tuning** introduces minimal task-specific parameters (e.g., a linear regression layer) following the pre-trained ReFound backbone, and the whole model is optimized together in the downstream tasks. **(2) Feature-based Prediction** only trained the task-specific layers, which take ReFound's region representations as inputs.

5 EXPERIMENTS

In this section, we conduct extensive experiments on six real-world datasets of three downstream urban region understanding tasks in two cities, to evaluate the effectiveness of ReFound. We provide an implementation of ReFound at: https://github.com/PaddlePaddle/PaddleSpatial/tree/main/research/ReFound.

5.1 Experimental Settings

We first briefly introduce the settings including data collection for pre-training, as well as downstream tasks, baselines and metrics for evaluation. Detailed setup is provided in Appendix A.2.

5.1.1 Pre-training Corpora. We collect POI and satellite image data of urban regions from five cities in China to pre-train ReFound, which are Beijing, Guangzhou, Shenzhen, Shanghai, and Suzhou. Firstly, following many previous studies that divide cities into region grids for urban region understanding tasks [33, 59], we create the region set by partitioning these five cities into $256m \times 256m$ grids, which results in approximately 171K regions in total. Then, for each region, we collect POI and satellite image data updated in June 2020, from Baidu Maps. The POI data comprises a POI's textual name, a category-id from 128 categories and coordinates. The satellite image data are 3-channel 256×256 RGB images with the spatial resolution of 1.0 m.

5.1.2 Downstream Tasks. Our model is evaluated on three urban region understanding tasks in Beijing and Shenzhen. We briefly introduce how to build real-world datasets for these tasks.

Urban Village Detection (UVD). This is a binary classification task aimed at identifying whether a region is contained by or overlaps with an urban village (UV) area. The ground-truth UV area data for dataset construction are obtained by crowdsourcing in June 2023. Firstly, we source news reports and official documents from the Internet, to collect potential UVs for verification. These candidate areas are uploaded to an online platform embedded with a map service, where the geographic coordinates, satellite images and street views of these areas can be accessed. Then, we enlist professional participants to select ground-truth UV areas on the platform. To ensure data reliability, each potential area is assigned to three participants, and will be labeled as UV only if all three participants reach a consensus. Following [59], regions overlapping with ground-truth UV areas by more than 20% of their area are labeled as positive samples, while we randomly select five times amount of regions from remaining areas of the city as negative samples. As a result, we construct two datasets with 882 and 552 samples in Shenzhen and Beijing.

Commercial Activeness Prediction (CAP). In this regression task, we follow previous studies to count the number of map users' comments to all POIs in a region, as an indicator of this region's commercial activeness. We also collect the number of comments per POI in Beijing and Shenzhen city from June 2019 to April 2020, with the same map service platform. Then, these counts are aggregated by regions according to POI locations, to obtain the regional commercial activeness data. The Shenzhen and Beijing datasets contain 4196 and 8789 samples, respectively.

Population Prediction (POP). It's also a regression task which predicts the population of regions. The real-world datasets are built based on WorldPop statistics (www.worldpop.org/) for 2020, at a resolution of approximately 100 m. For each region, its population value is contributed by values from several statistical units it overlaps with, according to its overlapping areas to each unit. We randomly sample 10,000 regions in each city for evaluation.

To select the best hyper-parameters for all comparing methods in downstream tasks, we randomly split each dataset into three parts with equal sizes for training, validation and test.

5.1.3 Baselines. We compare our model with two categories of state-of-the-art (SOTA) baselines under different settings. (1) *Foundation Model (FM) + Fine-tuning.* We compare fine-tuning performance between ReFound and three representative general-purpose

Table 1: Performance comparison in three downstream tasks on Shenzhen dataset.

		Urban Village Detection		Commercial Activeness Prediction			Population Prediction		
Usage	Methods	AUC ↑	F1-score ↑	RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑
Fine-tuning	BERT	0.73 ± 0.01	0.40 ± 0.09	17.31 ± 0.34	8.64 ± 0.24	0.44 ± 0.02	361.60 ± 2.11	266.99 ± 2.92	0.60 ± 0.00
	ViT	0.71 ± 0.01	0.39 ± 0.01	21.77 ± 0.19	10.95 ± 0.39	0.12 ± 0.02	338.23 ± 2.94	246.92 ± 2.90	0.65 ± 0.01
	CN-CLIP	0.74 ± 0.01	0.41 ± 0.03	18.39 ± 0.30	8.70 ± 0.11	0.37 ± 0.02	303.61 ± 5.35	220.79 ± 4.87	0.72 ± 0.01
	CN-CLIP-I	0.73 ± 0.02	0.38 ± 0.03	22.22 ± 0.19	11.63 ± 0.53	0.08 ± 0.02	337.68 ± 12.01	244.92 ± 8.20	0.65 ± 0.03
	SpaBERT	0.65 ± 0.02	0.31 ± 0.02	19.45 ± 0.35	10.26 ± 0.32	0.30 ± 0.03	389.93 ± 4.24	296.28 ± 1.45	0.53 ± 0.01
	GFM	0.76 ± 0.01	0.44 ± 0.03	21.43 ± 0.31	11.38 ± 0.45	0.15 ± 0.02	325.36 ± 4.81	237.47 ± 4.66	0.67 ± 0.01
	ReFound	**0.82 ± 0.02**	**0.44 ± 0.03**	**14.85 ± 0.16**	**7.57 ± 0.15**	**0.59 ± 0.01**	**286.10 ± 4.37**	**203.42 ± 3.39**	**0.75 ± 0.01**
Feature-based Prediction	HGI	0.57 ± 0.00	0.28 ± 0.01	20.18 ± 0.01	11.52 ± 0.03	0.24 ± 0.00	347.47 ± 2.09	263.69 ± 1.88	0.63 ± 0.00
	MMGR	0.70 ± 0.00	0.37 ± 0.02	21.86 ± 0.06	12.22 ± 0.22	0.11 ± 0.00	370.79 ± 0.38	279.34 ± 0.92	0.58 ± 0.00
	PG-SimCLR	0.68 ± 0.01	0.35 ± 0.03	21.70 ± 0.07	11.61 ± 0.21	0.13 ± 0.01	403.02 ± 0.99	303.82 ± 0.90	0.50 ± 0.00
	ReFound	**0.77 ± 0.00**	**0.44 ± 0.01**	**17.28 ± 0.20**	**9.96 ± 0.23**	**0.45 ± 0.01**	**308.45 ± 1.21**	**224.97 ± 0.87**	**0.71 ± 0.00**

Table 2: Performance comparison in three downstream tasks in Beijing dataset.

		Urban Village Detection		Commercial Activeness Prediction			Population Prediction		
Usage	Methods	AUC ↑	F1-score ↑	RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑
Fine-tuning	BERT	0.79 ± 0.02	0.48 ± 0.04	16.80 ± 0.28	7.73 ± 0.10	0.48 ± 0.02	193.64 ± 9.12	140.73 ± 10.82	0.57 ± 0.04
	ViT	0.88 ± 0.02	0.71 ± 0.03	21.16 ± 0.15	10.76 ± 0.10	0.17 ± 0.01	149.39 ± 1.69	102.90 ± 1.08	0.74 ± 0.01
	CN-CLIP	0.92 ± 0.01	0.74 ± 0.02	16.52 ± 0.19	7.78 ± 0.18	0.50 ± 0.01	**138.55 ± 2.14**	98.35 ± 2.14	**0.78 ± 0.01**
	CN-CLIP-I	0.94 ± 0.01	0.73 ± 0.05	22.10 ± 0.14	11.04 ± 0.08	0.10 ± 0.01	141.62 ± 2.00	99.58 ± 1.97	0.77 ± 0.01
	SpaBERT	0.77 ± 0.03	0.46 ± 0.07	18.18 ± 0.07	8.82 ± 0.15	0.39 ± 0.00	212.47 ± 0.72	160.01 ± 1.39	0.48 ± 0.00
	GFM	0.94 ± 0.01	0.69 ± 0.02	20.32 ± 0.07	10.29 ± 0.13	0.24 ± 0.01	139.20 ± 0.39	**97.13 ± 0.60**	0.78 ± 0.00
	ReFound	**0.97 ± 0.00**	**0.80 ± 0.02**	**13.56 ± 0.34**	**6.61 ± 0.05**	**0.66 ± 0.02**	140.62 ± 1.32	97.26 ± 0.69	0.77 ± 0.00
Feature-based Prediction	HGI	0.86 ± 0.01	0.54 ± 0.02	19.89 ± 0.02	10.30 ± 0.01	0.27 ± 0.00	181.36 ± 1.88	136.24 ± 1.55	0.62 ± 0.01
	MMGR	0.90 ± 0.00	0.66 ± 0.02	20.75 ± 0.13	10.99 ± 0.31	0.21 ± 0.01	185.37 ± 0.31	140.49 ± 0.70	0.61 ± 0.00
	PG-SimCLR	0.84 ± 0.00	0.56 ± 0.01	21.84 ± 0.04	11.35 ± 0.19	0.12 ± 0.00	226.49 ± 0.92	172.20 ± 1.33	0.41 ± 0.00
	ReFound	**0.94 ± 0.00**	**0.67 ± 0.05**	**15.09 ± 0.17**	**8.17 ± 0.06**	**0.58 ± 0.01**	**143.89 ± 0.24**	**104.02 ± 0.46**	**0.76 ± 0.00**

FMs (**BERT**[12], **ViT** [15] and **CN-CLIP** [61]), and two recent FMs in geospatial domain (**SpaBERT** [32] and **GFM** [37]). Among these models, text-based BERT and SpaBERT use POI data as inputs, while ViT and GFM work with satellite images. For CN-CLIP, we implement it in two ways: CN-CLIP makes use of both POI and satellite data, while **CN-CLIP-I** only encodes satellite images.
(2) *Region Embedding Model + Feature-based Prediction.* To evaluate ReFound's performance in extracting region representations for feature-based prediction, we compare it with three SOTA region embedding methods (**HGI** [22], **MMGR** [3] and **PG-SimCLR** [58]) based on POI and satellite image data. Detailed descriptions of these two categories of baselines are introduced in Appendix A.1.

5.1.4 Evaluation Metrics. Evaluation metrics for two regression tasks include Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R^2). For the binary classification task, we use Area Under Curve (AUC) and F1-score.

5.2 Performance Evaluation

5.2.1 ***Overall Performance.*** The performance comparison across three downstream tasks in two cities is presented in Table 1 and Table 2, with mean and standard deviation of all metrics derived from five random runs. As we can see, ReFound achieves outstanding performance in all three tasks. In the fine-tuning setting, it brings average performance gains of 5.5% on AUC in urban village detection (UVD), as well as 16.1% and 2.1% on RMSE in commercial activeness prediction (CAP) and population prediction (POP), respectively, over the most competitive baseline of each task. When performing feature-based prediction, ReFound can achieve 7.2%, 19.3% and 15.9% average improvements UVD, CAP and POP tasks respectively. Moreover, we have the following observations:

- A model's performance is greatly affected by the geospatial information it considers. To be specific, the image-based ViT, CN-CLIP-I and GFM are inferior in CAP task, because they cannot leverage the information of region functionality and human activities reflected in POI data. While BERT and SpaBERT get better results in CAP task based on POI data, they perform worse in UVD and POP tasks, due to the inability to capture the spatial distribution of buildings from satellite images, such as building density and height. Compared with CN-CLIP-I, CN-CLIP which incorporates both POI and satellite image data evidently performs better in three tasks, highlighting the importance of integrating multi-modal data for a variety of downstream tasks.
- Region embedding baselines that make feature-based prediction generally have lower performance than fine-tuned models, as only a small amount of parameters are optimized for specific tasks. In contrast, when ReFound also performs feature-based prediction as a region embedding model without fine-tuning, it still achieves promising performance, even surpassing the majority of fine-tuned baselines. This indicates that ReFound is capable of more effectively capturing and integrating the unique properties of POI and satellite image data, thereby improving its understanding of regions.

5.2.2 ***Ablation study.*** To verify the effectiveness of each design in this work, we further compare ReFound with its seven variants:

- **w/o MoGE** replaces the MoGE Transformer with the vanilla one where a shared feed-forward network is used in each layer.
- **w/o DLFM**, **w/o VLFM** and **w/o DVLFM** each remove the knowledge distillation from language, visual and visual-language foundation model, while **w/o Dist** removes all of them.
- **w/o MGDM** removes masked geospatial data modeling. Meanwhile, cross-modal spatial alignment is also disused, as it relies

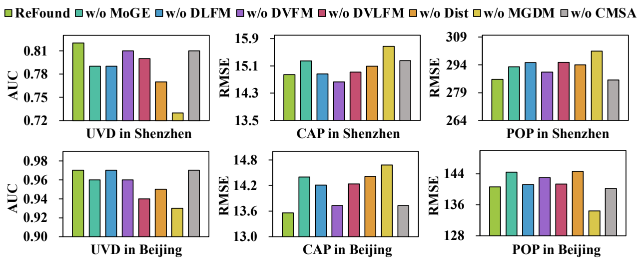
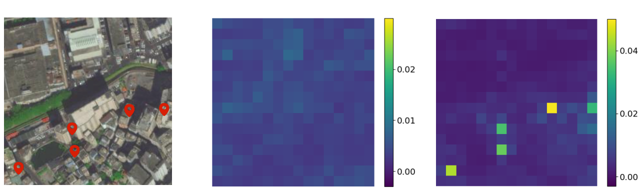


Figure 3: Ablation study.

on masked satellite image patches. In contrast, **w/o CMSA** only removes the cross-modal spatial alignment task.

As shown in Figure 3, our designs can generally improve ReFound's performance in various downstream tasks. Specifically, removing MoGE Transformer (w/o MoGE) worsens performance. It indicates the importance of such an architecture that can not only adjust to unique characteristics of POI and satellite image data, but also deeply fuse them for comprehensive urban region understanding. Additionally, the notable performance decline of w/o Dist highlights the advantage of harnessing knowledge from multiple FMs to improve ReFound's versatility. The three distilling objectives contribute to the improvement in different downstream tasks in varying degrees (w/o DLFM, w/o DVFM and, w/o DVLFM). Furthermore, performance evidently degrades without the self-supervised MGDM task (w/o MGDM) in most cases, which verifies the importance to learn in-domain feature from geospatial data for region understanding. CMSA task also benefits ReFound a lot, as it facilitates the semantic alignment of two modalities (w/o CMSA).

5.2.3 ***Visualization of Cross-Modal Spatial Alignment.*** To further explain the advantage of cross-modal semantic alignment brought by CMSA objective, we compare the attention maps between POIs and satellite image patches, produced by ReFound and its variants w/o CMSA. Specifically, Figure 4(a) shows the distribution of convenience stores in a region. For each token from these five stores, we use ReFound and w/o CMSA variant to respectively calculate its normalized attention scores in the last Transformer layer to different satellite image patches. The obtained score vectors are averaged among these tokens and visualized in Figure 4(b)-(c), where each square tile represents a patch. As observed, when using ReFound trained with CMSA objectives, these POIs (convenience stores) attend more to patches they locates in. It suggests that our model is able to align the semantics between two modalities, which facilitates the more accurate characterization of the region.



(a) Locations of stores (b) w/o CMSA (c) ReFound

Figure 4: Visualization of cross-modal spatial alignment.

5.2.4 ***Efficiency Evaluation.*** We also conduct an experiment on the population prediction dataset in Shenzhen to evaluate the efficiency of our framework. To be specific, we compare our ReFound and the foundation model baselines, in terms of the average time to fine-tune for one epoch on more than 3300 training samples and the average inference time per instance. For a fair comparison, the batch size is uniformly set to 1 for all models in this experiment. As shown in Table 3, the time costs of ReFound to fine-tune one epoch and to infer one instance are less than the time costs of the most competitive baseline (CN-CLIP). It indicates that our framework can achieve superior performance while keeping good efficiency, and is practical for real-world scenarios.

Table 3: Training (s/epoch) and inference (s/instance) time.

	BERT	ViT	CN-CLIP	CN-CLIP-I	SpaBERT	GFM	ReFound
Training	270.1	265.7	451.6	236.0	276.0	520.6	340.8
Inference	0.013	0.015	0.023	0.012	0.013	0.027	0.016

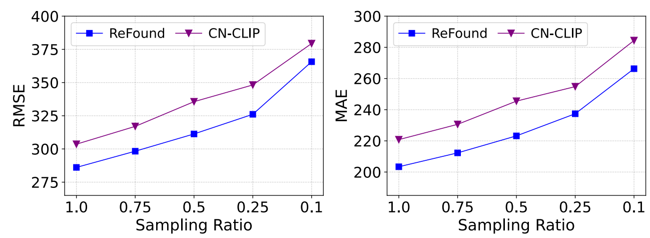


Figure 5: Population prediction performance in Shenzhen with sampling different ratios of training data.

5.2.5 ***Analysis on Downstream Training Data Scale.*** Additionally, we further investigate ReFound's performance in the case when the labeled data available for adapting the pre-trained model to the downstream task is limited. This evaluation is also conducted on the population prediction dataset in Shenzhen. To achieve this, we randomly sample 75%, 50%, 25% and 10% of training data, to gradually reduce the data scale for fine-tuning the pre-trained model. Figure 5 presents the prediction error of our framework and the most competitive baseline (CN-CLIP). As we can see, ReFound consistently outperforms CN-CLIP under different sampling ratios, which demonstrates that our model has potential to better solve the downstream urban tasks with limited training data.

6 CONCLUSION

In this paper, we propose a novel framework to pre-train a foundation model for urban region understanding, that harnesses the strength of well-established language and visual FMs to enhance its generality. An important advantage of this framework is the ability to sustainably leverage advancements in language and visual FMs. As these general domains continues to evolve and release more powerful FMs, we believe that our framework can produce stronger FMs for better urban region understanding in the future.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (Grant No.2023YFF0725001), National Natural Science Foundation of China (Grant No.92370204), Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2023A03J0008), Education Bureau of Guangzhou Municipality.

REFERENCES

[1] Jacob Levy Abitbol and Marton Karsai. 2020. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *NMI* 2, 11 (2020), 684–692.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Lubin Bai, 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 P&RS* 201 (2023), 193–208.

[4] Pasquale Balsebre, Weiming Huang, Gao Cong, and Yi Li. 2023. City Foundation Models for Learning General Purpose Representations from OpenStreetMap. *arXiv preprint arXiv:2310.00583* (2023).

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[6] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *NeurIPS* 35 (2022), 32897–32912.

[7] Tom Brown, 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. *NeurIPS* 33 (2020), 1877–1901.

[8] Rui Cao, 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 (2018), 1553.

[9] Longbiao Chen, Chenhui Lu, Fangxu Yuan, Zhihan Jiang, Leye Wang, Daqing Zhang, Ruixiang Luo, Xiaoliang Fan, and Cheng Wang. 2021. UVLens: Urban village boundary identification and population estimation leveraging open government data. *IMWUT* 5, 2 (2021), 1–26.

[10] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *NeurIPS* 35 (2022), 197–211.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 248–255.

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

[13] Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang Zhang, and Yao Xu. 2023. Mgeo: Multi-modal geographic language model pre-training. In *SIGIR*. 185–194.

[14] Lei Dong, Carlo Ratti, and Siqi Zheng. 2019. Predicting neighborhoods' socioeconomic attributes using restaurant data. *PNAS* 116, 31 (2019), 15447–15452.

[15] Alexey Dosovitskiy, 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* (2020).

[16] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *ACL*. 320–335.

[17] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, and Xiaolin Li. 2019. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. In *AAAI*, Vol. 33. 906–913.

[18] Yunfan Gao, Yun Xiong, Siqi Wang, and Haofen Wang. 2022. GeoBERT: Pre-Training Geospatial Representation Learning on Point-of-Interest. *Applied Sciences* 12, 24 (2022), 12942.

[19] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *IJCV* 129 (2021), 1789–1819.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*. 16000–16009.

[21] Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. 2022. ERNIE-GeoL: A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps. In *SIGKDD*. 3029–3039.

[22] Weiming Huang, Daokun Zhang, Gengchen Mai, Xu Guo, and Lizhen Cui. 2023. Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS P&RS* 196 (2023), 134–145.

[23] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *MM*. 4083–4091.

[24] Yingjing Huang, Fan Zhang, Yong Gao, Wei Tu, Fabio Duarte, Carlo Ratti, Diansheng Guo, and Yu Liu. 2023. Comprehensive urban space representation with varying numbers of street-level images. *CEUS* 106 (2023), 102043.

[25] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. 2019. Unsupervised representation learning of spatial data via multimodal embedding. In *CIKM*. 1993–2002.

[26] Carsten Keßler and Grant McKenzie. 2018. A geoprivacy manifesto. *Transactions in GIS* 22, 1 (2018), 3–19.

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 12888–12900.

[28] Ruiyuan Li, Huajun He, Rubin Wang, Yuchuan Huang, Junwen Liu, Sijie Ruan, Tianfu He, Jie Bao, and Yu Zheng. 2020. Just: Jd urban spatio-temporal data engine. In *ICDE*. IEEE, 1558–1569.

[29] Tong Li, Yanxin Xi, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. 2023. Learning Representations of Satellite Imagery by Leveraging Point-of-Interests. *TIST* 14, 4 (2023), 1–32.

[30] Tong Li, Shiduo Xin, Yanxin Xi, Sasu Tarkoma, Pan Hui, and Yong Li. 2022. Predicting multi-level socioeconomic indicators from structural urban imagery. In *CIKM*. 3282–3291.

[31] Yi Li, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. 2023. Urban Region Representation Learning with OpenStreetMap Building Footprints. In *SIGKDD*. 1363–1373.

[32] Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. 2022. SpaBERT: A Pretrained Language Model from Geographic Data for Geo-Entity Representation. *arXiv preprint arXiv:2210.12213* (2022).

[33] Yuxuan Liang, Kun Ouyang, Junkai Sun, Yiwei Wang, Junbo Zhang, Yu Zheng, David Rosenblum, and Roger Zimmermann. 2021. Fine-grained urban flow prediction. In *WWW*. 1833–1845.

[34] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. 2023. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *WWW*. 4150–4160.

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*. 10012–10022.

[36] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. 2023. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798* (2023).

[37] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. 2023. Towards geospatial foundation models via continual pretraining. In *ICCV*. 16806–16816.

[38] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *NeurIPS* 35 (2022), 9564–9576.

[39] Jack A Orenstein and Tim H Merrett. 1984. A class of data structures for associative searching. In *SIGACT-SIGMOD*. 181–190.

[40] Alec Radford, 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 *ICML*. PMLR, 8748–8763.

[41] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[42] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*. PMLR, 8821–8831.

[43] Wonyoung Shin, Jonghun Park, Taekang Woo, Yongwoo Cho, Kwangjin Oh, and Hwanjun Song. 2022. e-clip: Large-scale vision-language representation learning in e-commerce. In *CIKM*. 3484–3494.

[44] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*. 843–852.

[45] Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. 2023. DIME-FM: DIstilling Multimodal and Efficient Foundation Models. *arXiv preprint arXiv:2303.18232* (2023).

[46] Waldo R Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, sup1 (1970), 234–240.

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* 30 (2017).

[49] Dongjie Wang, Kunpeng Liu, David Mohaisen, Pengyang Wang, Chang-Tien Lu, and Yanjie Fu. 2021. Automated feature-topic pairing: Aligning semantic and embedding spaces in spatial representation learning. In *SIGSPATIAL*. 450–453.

[50] Dongjie Wang, Kunpeng Liu, David Mohaisen, Pengyang Wang, Chang-Tien Lu, and Yanjie Fu. 2021. Towards semantically-rich spatial network representation learning via automated feature topic pairing. *Frontiers in big Data* 4 (2021), 762899.

[51] Pengyang Wang, Kunpeng Liu, Dongjie Wang, and Yanjie Fu. 2021. Measuring urban vibrancy of residential communities using big crowdsourced geotagged data. *Frontiers in big Data* 4 (2021), 690970.

[52] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al.

2023. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *CVPR*. 19175–19186.
[53] Wenshan Wang, Su Yang, Zhiyuan He, Minjie Wang, Jiulong Zhang, and Weishan Zhang. 2018. Urban perception of commercial activeness from satellite images and streetscapes. In *Companion Proceedings of the The Web Conference 2018*. 647–654.
[54] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *AAAI*, Vol. 34. 1013–1020.
[55] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163* (2022).
[56] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2023. Llmrec: Large language models with graph augmentation for recommendation. *arXiv preprint arXiv:2311.00423* (2023).
[57] Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. 2022. Multi-graph fusion networks for urban region embedding. *arXiv preprint arXiv:2201.09760* (2022).
[58] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *WWW*. 3308–3316.
[59] Congxi Xiao, Jingbo Zhou, Jizhou Huang, Hengshu Zhu, Tong Xu, Dejing Dou, and Hui Xiong. 2023. A contextual master-slave framework on urban region graph for urban village detection. In *ICDE*. IEEE, 736–748.
[60] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. UrbanCLIP: Learning Text-enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web. In *WWW*. 4006–4017.
[61] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335* (2022).
[62] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).
[63] Liang Zhang, Cheng Long, and Gao Cong. 2022. Region Embedding With Intra and Inter-View Contrastive Learning. *TKDE* (2022).
[64] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2021. Multi-view joint graph representation learning for urban region embedding. In *IJCAI*. 4431–4437.
[65] Weijia Zhang, Jindong Han, Zhao Xu, Hang Ni, Hao Liu, and Hui Xiong. 2024. Towards Urban General Intelligence: A Review and Outlook of Urban Foundation Models. *arXiv preprint arXiv:2402.01749* (2024).
[66] Zhilun Zhou, Yu Liu, Jingtao Ding, Depeng Jin, and Yong Li. 2023. Hierarchical knowledge graph learning enabled socioeconomic indicator prediction in location-based social network. In *WWW*. 122–132.
[67] Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. 2024. Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook. *arXiv preprint arXiv:2402.19348* (2024).

A APPENDIX

A.1 Baseline Descriptions

Following *Foundation Model + Fine-tuning* paradigm, we compare ReFound with five SOTA FMs.

- **BERT** [12] is a pre-trained language FM. Similar to our model, we use BERT to encode the POI name sequence to produce the region representation. Since POI names in this work are in Chinese, we utilize "bert-base-chinese" pre-trained on Chinese corpora.
- **ViT** [15] is a visual FM pre-trained on ImageNet-21k [11], which can address region understanding tasks based on satellite images. We select "ViT-Base" model for comparison, and perform 2D position embedding interpolation to fit 256×256 satellite images during the fine-tuning stage, as suggested by [15].
- **CN-CLIP** [61] is a contrastive language-image pre-training model for Chinese image-text pairs, excelling in joint understanding of Chinese text and image data. We use CN-CLIP to encode both POI names and satellite images, then average them to obtain the region representation for downstream tasks. In our experiments, the selected version uses "ViT-Base" as the image encoder and "RoBERTa-wwm-Base" as the text encoder. The position embedding interpolation is also applied during fine-tuning.
- **CN-CLIP-I** [61] is a variant of CN-CLIP that makes prediction solely based on satellite images with the image encoder.
- **SpaBERT** [32] is a language FM pre-trained to represent POIs, which can jointly consider POIs' name and spatial distances. We use "SpaBERT-Base" to encode the POI sequence organized by Z-order for region representation. Specifically, we select the first POI in the sequence as the "pivot", and derive its contextualized representation that aggregates information from all POIs in the region. Since this model is pre-trained on English texts, we translate the Chinese POI names into English using an open translation service: https://api.fanyi.baidu.com/.
- **GFM** [37] is a recent visual FM that achieves SOTA performance in geospatial applications. It is pre-trained by masked image modeling on satellite images, with an auxiliary distilling objective from Swin-B model [35] pre-trained on ImageNet-22k [11].

We also compare with three SOTA region embedding models under *Region Embedding Model + Feature-based Prediction* setting.

- **HGI** [22] learns region representations based on POI data and hierarchical spatial information. It constructs a POI-level spatial graph to encoder the relative distance between POIs, and build a region-level adjacency graph to allow spatial interactions among regions. The model is trained with hierarchical Graph Infomax at both levels in a self-supervised manner.
- **MMGR** [3] is a multi-modal region embedding model. It design two encoders to encode POI categories and satellite images into two views of region representations, and adopts a cross-modal contrastive learning strategy to fuse them.
- **PG-SimCLR** [58] trains an image encoder via contrastive learning to generate region representations based on satellite image data. It uses spatial proximity and the POI category distribution as two metrics to measure the similarity between regions, for constructing contrastive samples.

A.2 Experimental Settings

A.2.1 Pre-training Setup. Our model adopts 12-layer Transformer blocks with 768 hidden size, 3,072 intermediate size of feed-forward networks, and 12 attention heads. For POIs in a region, we serialize them into the name sequence by Z-ordering strategy [39], and then tokenize it using BERT-Chinese tokenizer [12] with maximum length $L^P = 512$. For those sequences whose lengths are larger than L^P, we random exclude some POIs to meet the length limit, while sequences shorter than L^P will be appended [PAD] tokens. To obtain grid-level geo-aware position embedding E^{gp} described in Section 4.1.1, we partition the $256m \times 256m$ region into 16×16 grids with size of $16m \times 16m$.

For tasks (DLFM, DVFM and DVLFM) of joint knowledge distillation from multiple pre-trained FMs, we select ChatGLM [16, 62] and CN-CLIP [61] as teacher models. In DLFM task, with the textual prompt derived based on POI names $\mathbb{P}$, we employ "ChatGLM-6B" to generate the description of region functionality $\overline{\mathbb{P}}$, and apply sentence embedding model "M3E-Base" (https://huggingface.co/moka-ai/m3e-base) to encode it into LLM-features $\boldsymbol{u}_P$. For the generated description longer than the maximum input length of sentence

Table 4: Population prediction performance comparison on Guangzhou, Shanghai and Suzhou dataset.

	Guangzhou			Shanghai			Suzhou		
	RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑
BERT	255.10 ± 2.20	179.49 ± 1.39	0.60 ± 0.01	354.77 ± 2.47	239.70 ± 2.49	0.60 ± 0.01	90.14 ± 0.73	64.21 ± 1.10	0.43 ± 0.01
ViT	216.34 ± 4.64	152.07 ± 5.08	0.71 ± 0.01	315.52 ± 6.43	202.24 ± 8.38	0.68 ± 0.01	72.12 ± 0.30	48.43 ± 0.37	0.64 ± 0.00
CN-CLIP-I	215.51 ± 2.55	149.40 ± 1.68	0.72 ± 0.01	323.30 ± 8.14	205.06 ± 4.62	0.67 ± 0.02	71.02 ± 0.45	47.75 ± 0.28	0.65 ± 0.00
CN-CLIP	205.10 ± 2.94	142.83 ± 1.93	0.74 ± 0.01	295.97 ± 9.61	194.30 ± 4.77	0.72 ± 0.02	71.25 ± 0.75	49.31 ± 0.64	0.64 ± 0.01
SpaBERT	273.90 ± 3.40	194.84 ± 1.39	0.54 ± 0.01	416.63 ± 8.05	274.25 ± 5.85	0.45 ± 0.02	92.99 ± 0.41	65.91 ± 0.77	0.39 ± 0.01
GFM	203.53 ± 1.15	141.78 ± 0.98	0.75 ± 0.00	319.94 ± 4.55	199.16 ± 1.87	0.68 ± 0.01	70.92 ± 1.94	48.36 ± 1.02	0.65 ± 0.02
ReFound	**193.31 ± 2.64**	**133.55 ± 1.27**	**0.77 ± 0.01**	**276.77 ± 2.66**	**179.28 ± 2.76**	**0.76 ± 0.00**	**69.95 ± 0.62**	**47.47 ± 0.33**	**0.66 ± 0.01**

embedding model, we divide it into chunks and embed each chunk individually, then combine them with averaging weighted by the size of each chunk. In DVFM task, we adopt "ViT-L/14" image encoder of CN-CLIP as the VFM to extract feature-based knowledge $\boldsymbol{u}_S$ for distillation. For DVLFM task, we also derive the cosine similarity matrix M using this image encoder. M is scaled by a temperature parameter set to 0.07, before being normalized into a probability distribution via the softmax function.

In masked geospatial data modeling (MGDA) task, for the POI side, we randomly mask 15% of name tokens for prediction, where these masked tokens are replaced with [M] 80% of the time, a random token 10% of the time, and an unchanged token 10% of the time. For satellite image patches, we mask 40% of them. When deriving target visual tokens, the satellite image is resized into 224×224 so as to ensure the same number of tokens and patches of an image. We directly use the publicly available image tokenizer [42] whose vocabulary size is 8192.

We pre-train ReFound using AdamW optimizer with a batch size of 80 for 300 epochs. The weight decay is set to 0.01 and $(\beta_1, \beta_2) = (0.9, 0.999)$. We set a peak learning rate of 5e-5 with linear warm-up over the first 5 epochs, and then a linear decay strategy is applied.

A.2.2 Fine-tuning Setup. We next introduce the setup for fine-tuning our model. Without specification, the following settings are applied to all three downstream tasks: 2-layer MLP is utilized to make predictions, taking the region representation from ReFound as inputs; we set batch size to 12, and use AdamW optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 0.01 during fine-tuning; the 3-epoch linear warm-up and linear decay scheduler on learning rate (lr) are adopted; following [5, 20], a layer-wise lr decay with a ratio 0.75 is further applied to Transformer model. Other specific settings for different datasets are listed in Table 5, where "Fusion" is the way to merge the POI and satellite image representation, which includes *average* and attentional fusion (*attention*).

Table 5: Fine-tuning setup.

	UVD	CAP	POP
Shenzhen	Epoch: 30 Peak lr: 1e-4 Fusion: *attention*	Epoch: 40 Peak lr: 1e-4 Fusion: *average*	Epoch: 40 Peak lr: 1e-4 Fusion: *attention*
Beijing	Epoch: 30 Peak lr: 1e-5 Fusion: *average*	Epoch: 40 Peak lr: 1e-4 Fusion: *attention*	Epoch: 40 Peak lr: 1e-4 Fusion: *average*

A.2.3 Feature-based Prediction Setup. In addition, we evaluate ReFound 's performance for feature-based prediction. The shared setting for different datasets are as follows: the task-specific predictor is implemented by 2-layer MLP; we set batch size to 12, and use AdamW optimizer, with the weight decay set to 0.01 and $(\beta_1, \beta_2) = (0.9, 0.999)$; the learning rate (lr) is linearly warmed up during the first 3 epochs and then controlled by the linear decay scheduler. Note that only the predictor will be trained in downstream tasks. Other dataset-specific settings are listed in Table 6.

Table 6: Feature-based prediction setup.

	UVD	CAP	POP
Shenzhen	Epoch: 40 Peak lr: 5e-4 Fusion: *average*	Epoch: 40 Peak lr: 5e-4 Fusion: *attention*	Epoch: 40 Peak lr: 5e-4 Fusion: *attention*
Beijing	Epoch: 40 Peak lr: 5e-5 Fusion: *attention*	Epoch: 40 Peak lr: 5e-4 Fusion: *attention*	Epoch: 40 Peak lr: 1e-4 Fusion: *attention*

A.3 Additional Experimental Results

To demonstrate the effectiveness of ReFound across more diverse cities, we further evaluate it in other three cities: Guangzhou, Shanghai and Suzhou. This evaluation is conducted on the population prediction task, as the population data is public available and can be used to construct these datasets. The dataset construction process is consistent with that of Shenzhen and Beijing datasets described in Section 5.1.2. We compare the fine-tuning performance of ReFound and the FM baselines, with the results presented in Table 4. As we can see, our model also consistently outperforms other approaches across these cities, which shows ReFound's good generalizability.

Prompt based on POI data:

区域中有下列设施：雅点联体开关、名科门窗、纱艺纺纱窗、卡诺亚定制家居、…、一零八汽车维修有限公司、有家公寓。请推断这个区域的功能，以及可能的居民活动。

English Translation: There are the following facilities in the region: Yadian Union Switch, Mingke Doors and Windows, Shayi Textile Screen Windows, Canoya Customized Home Furnishings, …, 108 Automotive Maintenance Co., Ltd., Youjia Apartment. Please infer the functionality of this region and potential residential activities.

LLM generation:

这个区域的功能可能是一个家居建材市场，居民可能会有购买家居建材、定制家具、窗帘、衣柜、陶瓷、门窗、监控、净水器等产品的需求。居民可能会有装修房屋、家居装饰的需求，也可能会需要进行维修、保养等服务。这个区域的居民活动可能包括在商场内逛街购物、选购产品、观看家居装饰和装修相关的节目、参加家居装饰和装修相关的讲座、与商家和工作人员交流、购买商品和接受服务等等。此外，居民也可能会在商场内休息、娱乐、用餐等。

English Translation: The functionality of this region is likely to be a home furnishing and building materials market. Residents may have needs for purchasing home building materials, custom furniture, curtains, wardrobes, ceramics, doors and windows, surveillance equipment, water purifiers, and other products. Residents may also have demands for home renovation, interior decoration, as well as services like repair and maintenance. Activities of residents in this area may include shopping, product selection, watching programs related to home decoration and renovation, attending lectures on home decoration and renovation, interacting with merchants and staff, buying goods, and receiving services. Additionally, residents may also engage in activities such as resting, entertainment, and dining within the shopping mall.

Figure 6: An example of prompting LLM based on POI data.

A.4 LLM Generation Example

In DLFM task, we prompt LLM based on POI data, to generate a supplemental description related to region's functionality. An example is shown in Figure 6. As we can see, LLM is able to summarize a region's functionality, and infer potential resident activities within the region. Distilling such LLM knowledge can boost our model capturing the functionality semantics underlying POI data.