UrbanCLIP: Learning LLM-Enhanced Urban Region Profiling from the Web

Anonymous Author(s)

ABSTRACT

Urban region profiling from web-sourced data is of utmost importance for urban planning and sustainable development. We are witnessing a rising trend of LLMs for various fields, especially dealing with multi-modal data research such as vision-language learning, where the text modality serves as a supplement information for the image. Since textual modality has never been introduced into modality combinations in urban region profiling, we aim to answer two fundamental questions in this paper: i) *Can textual modality enhance urban region profiling? ii) and if so, in what ways and with regard to which aspects?* To answer the questions, we leverage the power of Large Language Models (LLMs) and introduce the first-ever LLM-enhanced framework that integrates the knowledge of textual modality into urban imagery profiling, named LLM-enhanced <u>U</u>rban Region Profiling with <u>C</u>ontrastive <u>L</u>anguage-<u>I</u>mage <u>P</u>retraining (**UrbanCLIP**). Specifically, it first generates a detailed textual description for each satellite image by an open-source Image-to-Text LLM. Then, the model is trained on the image-text pairs, seamlessly unifying natural language supervision for urban visual representation learning, jointly with contrastive loss and language modeling loss. Results on predicting three urban indicators in four major Chinese metropolises demonstrate its superior performance, with an average improvement of 6.1% on R^2 compared to the state-of-the-art methods. The source code is available at https://anonymous.4open.science/r/UrbanCLIP. The image-language dataset will be released upon paper notification.

1 INTRODUCTION

The rapid pace of urbanization has led to more than half of the global population, totaling 4.4 billion inhabitants [7, 65]. *Urban region profiling*, a pervasive and enduring theme within the domains of web mining and knowledge discovery, is the process of representing and summarizing key features and attributes of urban areas in a lower-dimensional space. By harnessing diverse web-sourced data, such as satellite [13, 22, 23, 28, 79, 86] and street-view imagery [48, 53, 79], this process delivers a comprehensive understanding of urban spaces, spanning the realms of social, economic, and environmental aspects. In this way, urban region profiling empowers decision-makers with critical insights and related web systems into urban planning, sustainable development, and policy formulation.

Scholars and policymakers traditionally rely on manual surveys to gather urban statistics. However, such methods inherently face
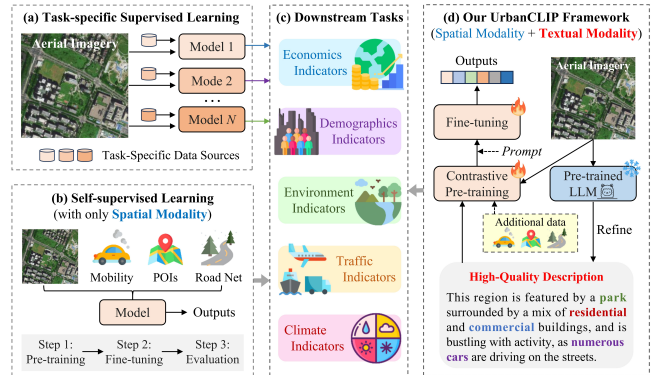
Figure 1: Illustration of different frameworks. Compared with previous works, we present the first attempt to leverage the power of LLMs for urban imagery profiling.

limitations in balancing high spatial resolution and real-time updates due to their prohibitive costs [57, 75, 86]. In contrast, data originating from web platforms boasts consistent updates and easy accessibility, especially high-resolution urban surfaces extracted from Baidu Map or Google Map [48, 53, 82], serving as the foundation for machine learning models to achieve a cost-friendly, high-quality, and timely understanding of urban indicators [9, 48, 78]. Upon revisiting the existing literature, we classify *web-based* urban region profiling into two categories, as shown in Figure 1:

a) **Task-specific supervised learning** acquires urban region representations through supervised training using data sources (e.g., satellite imagery) specific to particular tasks, including poverty levels [4, 5, 23, 30, 63, 86], crop yields [54, 56, 66, 77, 85, 87], population, land cover [26, 76] and commercial activeness [25, 53]. However, the task-specific nature of supervised learning, which requires considerable labeled data, may impede the model's *generalizability*, potentially compromising its overall robustness and efficacy.

b) **Self-supervised learning**, extending beyond satellite imagery, integrates diverse auxiliary spatial modalities to generate comprehensive feature representations. These representations boast wide applicability, readily generalizing across numerous urban indicator tasks, as delineated in Figure 1(c). Typically, [6, 32, 53, 82] integrate the information of Point-of-Interests (POIs) to capture human-inhabited areas and associated activities. Similarly, a series of studies consider aspects like mobility [32] and human trajectory data to enhance urban region profiling [48, 84]. Nevertheless, these approaches often lack sufficient *explanatory significance*, such as explaining in language that can easily be understood by humans.

During the past year, there has been a notable upsurge in the use of LLMs across various fields [1, 10, 34, 72]. The success is attributed to their remarkable proficiency in language understanding and the extensive knowledge they acquire during pre-training. Particularly, LLMs play a pivotal role in advancing multimodal learning, where textual data complements other modalities. As an example, the

integration of rich textual information has proven beneficial in tasks like image captioning [64, 71, 89] and video question-answering [61, 68, 83]. However, the incorporation of the textual modality in conjunction with urban imagery is a relatively unexplored area. Inspired by the significant achievements of LLMs in general fields, we embark on the exploration of two fundamental questions – **Q1:** *Can the inclusion of textual data serve as a powerful complement to satellite imagery for more effective urban region profiling?* **Q2:** *and if so, in what ways and with regard to which specific aspects?*

To answer the aforementioned questions, we integrate the textual modality into urban imagery profiling for the first time, leading to a novel framework, named LLM-enhanced Urban Region Profiling with Contrastive Language-Image Pretraining, termed as UrbanCLIP. At first, we generate a detailed description by a well-trained LLM (LLaMA-Adapter V2 [17]) for each satellite image. Then, the high-quality image-text pairs are fed into UrbanCLIP with an encoder-decoder architecture. It encodes satellite images to latent representation by a visual encoder (vision transformer [16]) and decodes texts with a causal masking transformer decoder. We further design a decoupled decoder mechanism, where unimodal textual representations from the first half of decoder layers would cascade the rest of decoder layers, cross-attending to the image encoder for multimodal representations. Moreover, a contrastive loss is applied between unimodal image and text embeddings, while language modeling loss on the multimodal decoder outputs is utilized for natural language profiling of urban regions with detailed granularity. The text-incorporated visual representations can support the prediction of various urban indicators from different urban regions. Overall, the main contributions of our work are summarized as:

- Powered by LLM, UrbanCLIP is the first-ever framework that integrates the knowledge of text modality into urban region profiling. We show that such comprehensive textual data generated by pre-trained image-to-text LLM is a critical supplement to urban region representations.
- UrbanCLIP infuses textual knowledge into visual representations through deep modality interaction jointly with contrastive loss and language modeling loss, via a contrastive learning-based encoder-decoder architecture, which subsumes model capacities from both contrastive models and generative models.
- Extensive experiments on four cities and three urban indicators demonstrate the effectiveness of UranCLIP. Further analyses are conducted to show the transferability and interpretability of the proposed model. We further develop a novel web-based application enabled by the proposed model to offer insights about urban planning, with an interactive and dynamic experience.

2 PRELIMINARIES

2.1 Formulation

Definition 1 (Urban Region) We follow prior studies [53, 82] to partition an area of interest (*e.g.*, a city) evenly into L urban regions.
Definition 2 (Satellite Image) Based on the real-time monitoring of the Earth's surface by satellites, satellite imagery offers a comprehensive view of the structural characteristics of a given region. Each input satellite image w.r.t. an urban region g can be denoted as $I_g \in \mathbb{R}^{H \times W \times 3}$, where H and W are length and width.

Definition 3 (Location Description) The description text T_g for an urban region g contains several individual sentences. Such text can be generated manually or using image captioning tools. E.g., by leveraging the well-trained LLM's profound understanding of general-purpose knowledge [27, 47, 80, 92], we can generate the summary text of a given region, especially including its spatial context (e.g., POIs) that significantly reflects its land function [17].
Definition 4 (Urban Indicator) Urban indicators gauge the urban region's standing on the socioeconomic spectrum and the environmental perspective. The K indicators on a set of L urban regions are denoted as $\mathbf{Y} \in \mathbb{R}^{L \times K}$. In this paper, we use *population* (#citizens), *GDP* (million Chinese Yuan), and *carbon emission* (ton) as social, economic, and environmental ground-truth indicators, respectively.
Problem Statement (LLM-Enhanced Urban Region Profiling) Given the above setting, we aim to learn a function $\mathcal{F}$ to map the *satellite imagery*, its *text description*, and other available data (e.g., POIs, road networks) to a representation vector $e_g = \mathcal{F}(I_g, T_g)$. The representations can be further utilized to infer urban indicators $\mathbf{Y} \in \mathbb{R}^{L \times K}$ for an arbitrary set of regions.

2.2 Related Work

2.2.1 Urban Imagery Profiling.
Learning urban region profiling from the web data has been a long-standing research topic in web mining. Current efforts can be broadly classified into two types:

- *Task-specific supervised learning.* This line of research learns prediction models from task-specific data sources. For example, using light intensity as supervision data, Yeh et al. [86] employ a pre-trained CNN model to predict asset levels in Africa. Similar methodologies have been applied in forecasting economic indicators in studies like [25, 28, 62]. Additionally, certain investigations estimate house prices by leveraging learned visual features from both satellite and street-view images, as seen in [40].

- *Self-supervised learning (SSL) with spatial modality.* This research strand mostly focuses on combining urban imagery and spatial modality for urban region profiling. They typically resort to Tobler's First Law of Geography [58], known as "Everything is related to everything else, but near things are more related than distant things", to distill the visual representations of urban imagery, via different designs of similarity metrics [31, 79] or loss forms [8, 36, 82]. Some studies, such as [82], incorporate POI data in a contrastive-learning framework, aiming to ensure that satellite images associated with similar POI distributions exhibit a closer relationship in visual latent space. Furthermore, [53] introduces an urban knowledge graph and infuses such semantic embedding into visual representation learning of satellite imagery via contrastive learning. Technically speaking, SSL-based methods outperform task-specific supervised learning in terms of generalization.

Compared with SSL with spatial modality, UrbanCLIP introduces the textual modality as complementary information for urban region profiling with the first shot. leading to a more comprehensive, generalizable, and interpretable urban region representation.

2.2.2 Large Language Model.
LLM are renowned for their ability to attain comprehensive language understanding the generation, which stem from their training on massive datasets and billions of parameters. Inspired by their impressive performance, there is a rising trend of incorporating LLMs in various fields, such as ChatBot

[2, 67, 72], coding [10, 19, 50], and even time series forecasting [35, 93]. However, the potential of LLMs remains largely untapped in the field of urban computing, including our region profiling task.

2.2.3 Vision-Language Pre-Training (VLP). VLP aims for effective vision-language alignment with frozen unimodal models from the vision and natural language communities. CLIP [64] is proposed as a vision foundation model based on image-text contrastive learning rationale. In certain approaches, the image encoder is frozen to extract visual features, as exemplified by the frozen object detector [90] or the pre-trained image encoder for CLIP used in LiT [91]. In contrast, some methods freeze the language model to leverage the knowledge from LLMs for vision-to-language generation tasks. To align visual features with the fixed text space, Frozen [74] fine-tunes an image encoder whose outputs are fed as soft prompts for LLMs. Flamingo [14] pre-trains a cross-attention layer added into the LLM to inject visual features. Moreover, BLIP-2 [44] takes full advantage of both frozen image encoders and frozen LLMs for various vision-language tasks. Leveraging the VLP scheme, we align generated text with satellite images to produce interpretable and admirable representations for urban regions.

3 METHODOLOGY

As illustrated in Figure 2, the overall framework of UrbanCLIP is composed of two key phases with two optional settings.

- **Phase 1**: In this language-image pre-training phase, we first generate a detailed location description via LLaMA-Adaptor V2 (an image-to-text foundation model) for the satellite imagery crawled from Baidu Map, thus forming a set of high-quality image-text pairs. The image and text are then fed into two unimodal encoders separately. Lastly, a multimodal interaction module is designed to align the representation of the two modalities in the latent space, with an elaborately designed cross-attention mechanism and contrastive learning objective.
- **Phase 2**: In the urban indicator prediction phase, we utilize a frozen unimodal image encoder for downstream tasks, by simply fine-tuning outermost multi-layer perceptrons (MLPs) with a few trainable parameters. Furthermore, we offer two optional choices, which are a flexible infusion of other spatial modalities and prompt-guided urban indicator prediction.
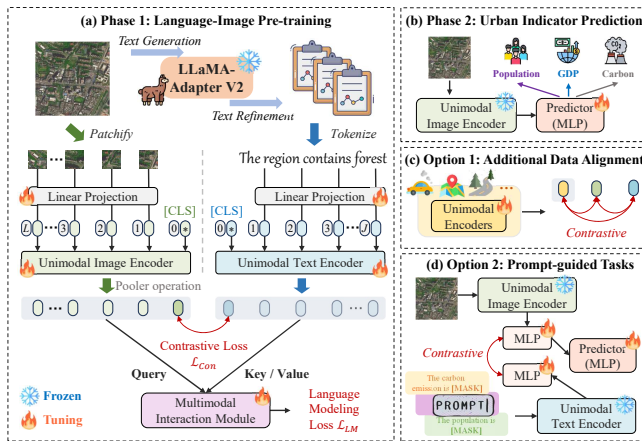

Figure 2: Overall framework of the proposed UrbanCLIP.

3.1 Text Generation and Refinement

Text Generation. For each satellite image, we adopt LLaMA-Adaptor V2, an image-to-text foundation model, to generate a detailed location description as illustrated in Figure 3(a). It takes a satellite image and an elaborately designed instruction as input and outputs a detailed text that describes the spatial information of the image. Through empirical experiments based on different language instructions, *we find that a more detailed prompt, especially including a specific focus such as urban infrastructure, can trigger a more powerful capability of LLM to generate a high-quality summary.*
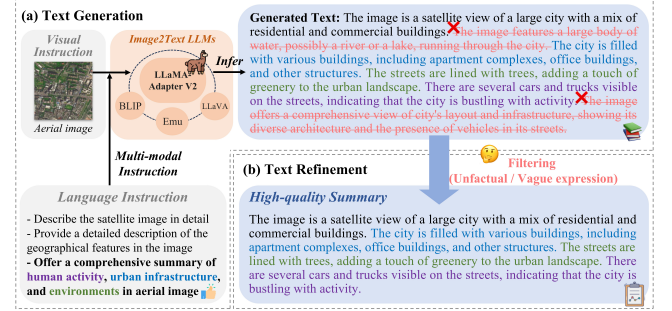

Figure 3: Text generation and refinement.

Text Refinement. As shown in the example of Figure 3(b), the generated description contains unfactual or vague information, and a thorough refinement, particularly the rule-based removal or rewriting, is conducted. As a result, a concise and high-quality summary retains the essential details about the satellite image, including its infrastructure, greenery, activity, etc.

3.2 Single-modality Representation Learning

Visual Representation Learning. For an urban region g with its satellite imagery I_g, we first split it into a sequence of patches I_p (the default patch size is 16×16), which are then linearly embedded into a dense vector: $e_p^I = \mathbf{W}_p I_p^\top + b_p$, where $\mathbf{W}_p$ and b_p are learnable parameters. The learnable positional embeddings $\mathbf{E}$ are further added to provide information about the relative position of each patch: $e_E^I = e_p^I + \mathbf{E}$. Then, e_{PE}^I is sent to the layers of the self-attention module to integrate the sequence information:

$$\left(\mathbf{Q}^I, \mathbf{K}^I, \mathbf{V}^I\right)^\top = e_{PE}^I \left(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\right)^\top \qquad (1)$$

where $\mathbf{W}_Q, \mathbf{W}_K$, and $\mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable matrices. The single-head and multi-head self attention (MSA) are defined as:

$$e_{(i)}^I = \text{Softmax}\left(\mathbf{Q}^I \mathbf{K}^{I^\top} / \sqrt{d}\right) \mathbf{V}^I,$$
$$e_{MSA}^I = \text{Concat}(e_{(1)}^I, e_{(2)}^I, ..., e_{(\#head)}^I) \mathbf{W}_O, \qquad (2)$$

where $\mathbf{W}_O$ is a learnable weight matrix, and Concat($\cdot$) denotes the concatenation function. After residual connection and layer normalization, the latent visual representation can be obtained as:

$$e^I = \text{LayerNorm}\left(e_E^I + \text{MSA}\left(e_E^I\right)\right). \qquad (3)$$

It is noteworthy that the input satellite image patch sequence incorporates a learnable image [CLS] token at first to obtain dynamic interaction representation between patches. However, unlike traditional paradigms, this token does not guide learning through

task label information. Inspired by [41], we have implemented task-specific temporary pooling (abbreviated as pooler) to customize visual representation for distinct pre-training tasks while sharing the previous backbone encoder. The pooler serves as a task-specific self-attention layer, which acts as a natural task adapter. Specifically, we employ the overall sequence as the query attention pooler for the subsequent fine-grained cross-modality interaction task, and the [CLS] token as the query attention pooler for the subsequent coarse-grained cross-modality alignment task.

Textual Representation Learning. For an urban region g, a high-quality text summary T_g is generated from LLMs through text generation and refinement. Similar to the prior visual representation, it is desirable to encode this summary into a latent textual representation. Normally, BERT-style [37] models with encoder-only architecture can be generalized to capture latent textual representations through denoising with token masks. However, such traditional bidirectional attention may encounter low-rank issues [15], potentially weakening the model's expressive capacity and yielding limited generative capabilities. As this is incompatible with the intended pre-training task, we choose a decoder-only architecture for the text encoding module. The primary distinction in this approach is that the foundational textual representation is acquired via causally masked multi-head self-attention:

$$e^T = \text{LayerNorm}\left(e_E^T + \text{M-MSA}\left(e_E^T\right)\right). \quad (4)$$

where M-MSA means masked multi-head self attention operation, and e_E^T is the token representation of added location information.

Noted that we also add a learnable text [CLS] token to obtain the overall dynamic interaction information of the summary.

3.3 Cross-modality Representation Learning

Modality Alignment Task. While more visual tokens can help multi-modal understanding tasks, visual embeddings of image [CLS] tokens as global representations are beneficial for visual recognition and alignment tasks [88]. We therefore first focus on the alignment task of visual and textual modalities. Specifically, for the underlying visual representation learning sequence, we obtain a new sequence representation through self-attention and pooling operation:

$$e_{pool}^I = \text{Pooling}(\text{Softmax}(e_q^I E_k^{I^\top}) \cdot E_v^I) \quad (5)$$

where E_k^I equals E_v^I represents the sequence of visual representations before transform, and e_q^I represents global visual embedding of [CLS] image token to be queried. For Pooling, we select the Mean-Pooling to detect the global information.

Contrastive learning has demonstrated its superiority. Inspired by that, we propose an Image-text contrastive loss $\mathcal{L}_{\text{Con}}$, which is inspired by the fact that both LLM-enhanced semantic representation (i.e., text embedding) and visual representation (i.e., satellite imagery representation) of the same urban region should be as close to one another as possible. It can maximize the agreement of representations learned across different modalities while capturing different relationships. Thus, the two unimodal encoders are jointly optimized by contrasting the image-text pairs against others in the

sampled batch of m samples:

$$\mathcal{L}_{Con} = \mathcal{L}_{Con}^{\text{Image} \to \text{Text}} + \mathcal{L}_{Con}^{\text{Text} \to \text{Image}}$$

$$= -\log \frac{\exp\left(\text{sim}\left(e_{pool}^I, e^T\right)\right)}{\sum_{i=1}^m \exp\left(\text{sim}\left(e_{pool}^I, e_i^T\right)\right)} - \log \frac{\exp\left(\text{sim}\left(e^T, e_{pool}^I\right)\right)}{\sum_{i=1}^m \exp\left(\text{sim}\left(e^T, e_{pool_i}^I\right)\right)},$$

where $\text{sim}(\cdot)$ is inner product; $\mathcal{L}_{con}^{\text{Image} \to \text{Text}}$ and $\mathcal{L}_{con}^{\text{Text} \to \text{Image}}$ are image-to-text and text-to-image contrastive losses, repsectively.

Modality Interaction Task. Unlike the previous studies where cross-modality interaction is shallow (e.g., via dot product-based similarity) [48, 53], UrbanCLIP emphasizes the deep inter-modal interaction learning through layers for a contextualized feature sequence. Motivated by [38, 88], Transformer-based decoder architecture are then leveraged to fuse unimodal visual and textual representations together as multimodal representations. Specifically, UrbanCLIP employs multimodal decoder layers to effectively learn joint image-text representations, by leveraging unimodal textual encoder outputs and employing cross-attention mechanisms towards image encoder outputs. The key difference between multimodal cross-attention here and unimodal MSA is that cross-attention uses the visual modality as a query and the textual modality as key and value. Here, visual representations are obtained via pooler operations.

Besides, to generate a natural language-based description for comprehensive urban region profiling, we introduce language modeling loss $\mathcal{L}_{\text{LM}}$ that enables the model to predict the next tokenized texts autoregressively with detailed granularity. Hence, the multimodal decoder would learn to maximize the conditional likelihood of the paired text T via the autoregressive factorization mechanism:

$$\mathcal{L}_{\text{LM}} = -\sum_{l=1}^L \log P_\theta\left(T_l \mid T_{<l}, I\right), \quad (6)$$

3.4 Urban Indicator Prediction

Pre-training Stage. UrbanCLIP enables both unimodal text and multimodal representations to be generated simultaneously. To achieve this, both image-text contrastive loss $\mathcal{L}_{\text{Con}}$ and language modeling loss $\mathcal{L}_{\text{LM}}$ are applied, we minimize the following objective function for model learning during pre-training stage:

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{Con}} \cdot \mathcal{L}_{\text{Con}} + \lambda_{\text{LM}} \cdot \mathcal{L}_{\text{LM}}, \quad (7)$$

where λ_{Con} and λ_{LM} are loss weighting hyperparameters.

An additional notable benefit of the loss design lies in its training efficiency [88]. The decoupled autoregressive decoder enables high-efficiency computation of two training losses. Unidirectional language models, trained with causal masking on complete texts, allow the decoder to generate outputs for both contrastive and generative training objectives in a single forward propagation. In contrast, the bidirectional approach requires two passes [46], which is more time-consuming. As for UrbanCLIP, most computation is shared between the two losses. We provide a detailed complexity analysis in Appendix 6.2.

Predict Stage. Through optimizing the loss function in Eq. 7, we can obtain the final text-enhanced visual representations e_g based on the frozen image encoder. Consequently, given any satellite image I_g, we can use a simple yet effective Predictor MLP to predict the urban indicators as $\mathbf{Y}_g = \text{MLP}\left(\mathbf{I}_g\right)$.

3.5 Discussion

3.5.1 Additional Data Alignment and Integration. In reality, other spatial modalities such as POIs [6, 32, 52, 82] and trajectories [48, 84] may be available which can contribute to urban region profiling. Considering this, we improve the flexibility of UrbanCLIP from the following two aspects: i) better alignment among diverse modalities. As illustrated in Figure 2(c), multimodality contrastive learning shows great capability in learning joint representations, by maximizing the agreement between semantically aligned examples (i.e., positive sample) across modalities while minimizing the agreement between non-aligned ones. For more modalities, an example of a positive sample could be the combination of a satellite image, a text description, the majority of POI categories as parks, and the road network of a given area. ii) better interaction with existing modalities. An intuitive way is adopting cross-attention mechanisms in UrbanCLIP. For instance, each modality engages in attention with every other modality, creating pairwise interactions. In summary, UrbanCLIP supports a flexible infusion with other modalities as a plug-and-play integration for better urban region profiling.

3.5.2 Prompt-guided Downstream Tasks. Prompting was proposed initially in natural language processing domain, and it refers to the generation of task-relevant instructions to obtain the desired output from a pre-trained model [33, 52]. Hence, a simple, task-specific prompt can be designed manually as one option to boost the downstream prediction performance of UrbanCLIP. As illustrated in Figure 2(d), for the carbon emission prediction task, a simple prompt can be designed during fine-tuning as "The carbon emission is *[MASK]*", guiding the model to concentrate on the environment-related spatial information for visual representation learning. *[MASK]* token serves as an indication that our textual encoder should predict the next token in an autoregressive generation manner. Furthermore, motivated by recent prompt learning-based studies, language instructions could be learned by training discrete [18, 33] or continuous [42, 49] vectors, consequently steering the performance of downstream urban indicators prediction.

4 EXPERIMENTS

In this section, we conduct extensive experiments to investigate the following Research Questions (RQ):

- **RQ1**: Can UrbanCLIP outperform prior approaches and generalize well to various urban indicator tasks?
- **RQ2**: How does each component (e.g., textual modality, text refinement, training objectives) contribute to UrbanCLIP?
- **RQ3**: How is the transferability of UrbanCLIP across cities?
- **RQ4**: How do we envision the practicality of UrbanCLIP?

4.1 Experimental Setup

4.1.1 Datasets. The datasets used in this paper include satellite imagery, textual description, and three urban indicators for four representative cities in China: Beijing, Shanghai, Guangzhou, and Shenzhen. The satellite images obtained from Baidu Map API have a fixed size of 256×256 with a spatial resolution of around 13 meters per pixel, which leads to an area of approximately 1 km^2. The textual information for each satellite image is generated from LLaMA-Adapter V2 [17], which has the most detailed and high-quality text generation compared with other up-to-date open-source Image-to-Text foundation models [3, 20, 43–45, 51, 69, 70] via empirical

Table 1: Dataset statistics.

Dataset	Coverage		#Satellite Image	#Location Description
	Bottom-left	Top-right		
Beijing	39.75°N, 116.03°E	40.15°N, 116.79°E	4,592	20,642
Shanghai	30.98°N, 121.10°E	31.51°N, 121.80°E	5,244	23,455
Guangzhou	22.94°N, 113.10°E	23.40°N, 113.68°E	3,402	15,539
Shenzhen	22.45°N, 113.75°E	22.84°N, 114.62°E	4,324	18,113

experiment. There exists a one-to-many relationship between images and associated texts. We filter out low-quality descriptions and then adopt a random selection to choose one high-quality summary text that matches each satellite image. The overall statistics of satellite imagery and textual description can be seen in Table 1. As for urban indicator data, we collect population from WorldPop [81] as a social indicator, GDP from [59] as an economic indicator and carbon emission from ODIAC [60] as the environmental indicator. All urban indicators per grid cell are aligned with corresponding satellite imagery and converted into a logarithmic scale. In this paper, we randomly partition the dataset into 60% for training, 20% for validation, and 20% for test.

4.1.2 Baselines. We compare UrbanCLIP with the following baselines in the field of urban imagery-based socioeconomic prediction:

- **Autoencoder** [39]. A neural network architecture that acquires representations from unlabeled satellite images as input, with the training objective of minimizing the reconstruction error.
- **PCA** [73]. Principal Component Analysis (PCA) is utilized to transform original satellite imagery into extended vectors and compute the first 10 principal components for each image.
- **ResNet-18** [24]. It is a well-established deep learning model pre-trained on ImageNet. It directly transfers a model trained on natural imagery to satellite imagery.
- **Tile2Vec** [31]. An unsupervised model that employs a triplet loss to learn the visual representations, with the goal of minimizing the similarity of proximate satellite image pairs, while maximizing the dissimilarity of distant pairs.
- **READ** [22]. Representation Extraction over an Arbitrary District (READ) is a semi-supervised model that leverages limited labeled data and transfer learning methods on a partially-labeled dataset to extract robust and lightweight satellite image representations, utilizing a teacher-student network with pre-trained models.
- **PGSimCLR** [82]. A satellite image representation method for its competitive performance in socioeconomic prediction, leveraging SimCLR [11] to encourage similar representations for grids with analogous facility distribution and geo-adjacency.

4.1.3 Metrics and Implementation. To assess the prediction performance, we adopt three commonly used evaluation metrics: coefficient of determination (R^2), rooted mean squared error (RMSE), and mean absolute error (MAE) [30, 82]. Higher R^2, and lower RMSE, MAE means better performance. As for the default implementation of UrbanCLIP, Vision Transformer (ViT) [16] and the first half of transformer decoder are applied to convert the satellite image and location description into their unimodal representations, respectively; and the rest of transformer decoder can be used for multimodal interaction to generate image-text representations. The parameter initialization follows the setting from [12, 29]. Adam optimizer is chosen to minimize the training loss during parameter learning. A grid search on hyperparameters is conducted, where

Table 2: Urban indicators prediction results in four datasets. The best results are in bold, and the second-best results are underlined. The last row indicates the relative improvement in percentage.

Dataset	Beijing									Shanghai								
Model	Carbon			Population			GDP			Carbon			Population			GDP		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Autoencoder	0.099	0.936	0.621	0.094	0.988	0.712	0.115	1.603	0.858	0.119	0.968	0.617	0.101	0.967	0.800	0.077	1.782	0.900
PCA	0.124	0.921	0.598	0.109	0.968	0.700	0.102	1.696	0.882	0.123	0.952	0.588	0.131	0.958	0.802	0.103	1.702	0.890
ResNet-18	0.393	0.599	0.411	0.202	0.858	0.680	0.203	1.280	0.758	0.451	0.512	0.460	0.233	0.852	0.692	0.217	1.297	0.777
Tile2Vec	0.599	0.512	0.468	0.204	0.813	0.635	0.182	1.356	0.792	0.572	0.462	0.390	0.249	0.801	0.620	0.169	1.380	0.806
READ	0.284	0.678	0.545	0.301	0.813	0.632	0.208	1.281	0.759	0.399	0.588	0.527	0.322	0.801	0.600	0.229	1.296	0.773
PG-SimCLR	<u>0.613</u>	<u>0.489</u>	<u>0.360</u>	<u>0.362</u>	<u>0.799</u>	<u>0.599</u>	<u>0.317</u>	<u>1.114</u>	<u>0.688</u>	<u>0.597</u>	<u>0.442</u>	<u>0.356</u>	<u>0.410</u>	<u>0.790</u>	<u>0.584</u>	<u>0.319</u>	<u>1.181</u>	<u>0.725</u>
UrbanCLIP	**0.662**	**0.327**	**0.302**	**0.407**	**0.788**	**0.589**	**0.319**	**1.102**	**0.684**	**0.652**	**0.331**	**0.300**	**0.429**	**0.778**	**0.578**	**0.320**	**1.119**	**0.702**
Improvement	8.11%	33.22%	16.00%	12.35%	1.39%	1.69%	0.73%	1.04%	0.62%	9.28%	25.12%	15.73%	4.59%	1.54%	1.06%	0.38%	5.28%	3.06%

Dataset	Guangzhou									Shenzhen								
Model	Carbon			Population			GDP			Carbon			Population			GDP		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Autoencoder	0.068	0.992	0.736	0.163	0.991	0.833	0.122	1.753	0.887	0.099	0.970	0.704	0.122	0.989	0.817	0.093	1.901	0.899
PCA	0.087	0.989	0.688	0.179	0.989	0.812	0.134	1.693	0.862	0.133	0.956	0.677	0.134	0.977	0.810	0.087	1.902	0.899
ResNet-18	0.388	0.500	0.513	0.244	0.883	0.711	0.215	1.290	0.791	0.409	0.556	0.503	0.250	0.880	0.701	0.165	1.398	0.844
Tile2Vec	0.482	0.499	0.501	0.269	0.855	0.683	0.173	1.346	0.799	0.466	0.501	0.486	0.289	0.841	0.649	0.123	1.500	0.881
READ	0.353	0.589	0.589	0.301	0.849	0.633	0.200	1.289	0.766	0.378	0.600	0.551	0.301	0.811	0.631	0.186	1.356	0.823
PG-SimCLR	<u>0.503</u>	<u>0.401</u>	<u>0.401</u>	<u>0.370</u>	<u>0.823</u>	<u>0.603</u>	<u>0.309</u>	<u>1.109</u>	<u>0.702</u>	<u>0.523</u>	<u>0.412</u>	<u>0.417</u>	<u>0.386</u>	<u>0.791</u>	<u>0.610</u>	<u>0.290</u>	<u>1.172</u>	<u>0.741</u>
UrbanCLIP	**0.587**	**0.390**	**0.389**	**0.388**	**0.801**	**0.602**	**0.309**	**1.109**	**0.700**	**0.597**	**0.373**	**0.387**	**0.391**	**0.791**	**0.602**	**0.293**	**1.153**	**0.734**
Improvement	16.77%	2.65%	3.02%	4.89%	2.70%	0.10%	0.10%	0.04%	0.37%	14.12%	9.58%	7.27%	1.48%	0.04%	1.39%	0.86%	1.65%	0.96%

search ranges for learning rate and batch size are set as $\{2e^{-6}, 2e^{-5}, 2e^{-4}, 2e^{-3}, 2e^{-2}\}$ and $\{4, 8, 16, 32, 64\}$, respectively.

4.2 RQ1: Performance Comparison

We empirically evaluate the performance of different models on the four datasets. The experimental results are shown in Table 2, from which we can obtain the following findings:

i) UrbanCLIP consistently achieves the best performance across all the datasets. It outperforms the best baseline, PG-SimCLR, by 7.06%, 4.75%, 7.25% and 5.49% in terms of R^2 for Beijing, Shanghai, Guangzhou and Shenzhen, respectively. Besides, the average performance gain of UrbanCLIP on RMSE and MAE are 7.02% and 4.27%, respectively. The results further prove the effectiveness of introducing the text modality into the urban region profiling.

ii) UrbanCLIP achieves promising results across all three urban indicators, with carbon emission being the best, followed by population, and GDP ranking last. The average R^2 improvement percentages for carbon emission, population and GDP prediction are 12.07%, 5.83% and 0.52%, respectively. A better performance in environmental indicators may come from the text-enhanced nature of UrbanCLIP, since the location summary containing key POIs such as parks can help indicate whether the corresponding region is environmentally friendly but cannot deduce the wealth class around that area. This insight inspires future work to leverage non-spatial information (such as economic-related time series) to enhance economic indicators' prediction performance.

iii) Existing satellite imagery-based prediction approaches still lack the capability to profile urban regions comprehensively. Taking the spatial correlations of regions into account, PG-SimCLR [82] (the best baseline model) and Tile2Vec [31] achieve competitive results among most prediction tasks compared to other baselines, which indicates that extra knowledge is beneficial for visual representation learning. Nevertheless, these methods may not capture crucial semantics in satellite imagery, such as significant POIs, where textual information can enhance understanding.

4.3 RQ2: Ablation Studies

Next, we conduct ablation studies to investigate the effectiveness of different components in UrbanCLIP, including the generation and refinement of textual information, cross-modality interaction, and training objectives. The results on R^2 are depicted in Figure 4.

4.3.1 Effectiveness of Textual Modality. The core idea of UrbanCLIP is the introduction of textual modality for urban region profiling. Thus, it is natural to ask for the effectiveness of textual information. To this end, we compare UrbanCLIP with a standard ViT-based model [16], termed as UrbanViT, which has the same setting as the unimodal visual encoder of UrbanCLIP. The extracted visual representations without textual enhancement would be used to predict three urban indicators.
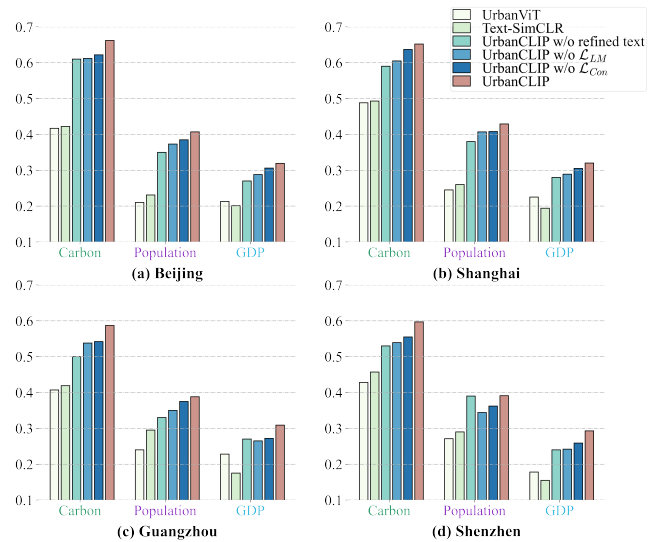
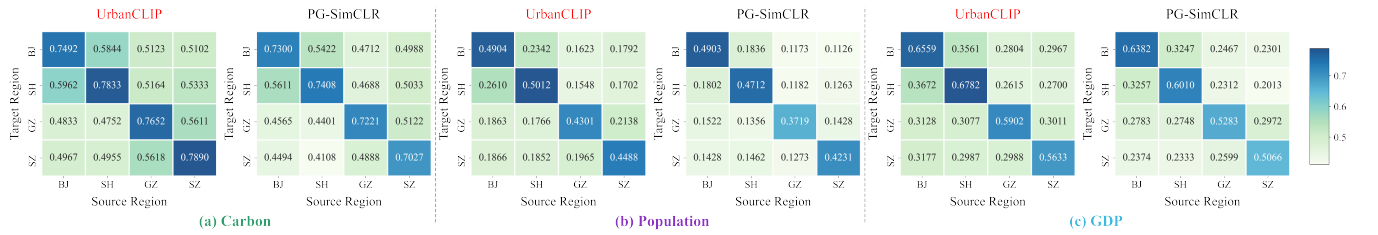

Figure 4: Results of Ablation Study on R^2 Metric.

(a) Carbon

UrbanCLIP

Target \ Source	BJ	SH	GZ	SZ
BJ	0.7492	0.5844	0.5123	0.5102
SH	0.5962	0.7833	0.5164	0.5333
GZ	0.4833	0.4752	0.7652	0.5611
SZ	0.4967	0.4955	0.5618	0.7890

PG-SimCLR

Target \ Source	BJ	SH	GZ	SZ
BJ	0.7300	0.5422	0.4712	0.4988
SH	0.5611	0.7408	0.4688	0.5033
GZ	0.4565	0.4401	0.7221	0.5122
SZ	0.4494	0.4108	0.4888	0.7027

(b) Population

UrbanCLIP

Target \ Source	BJ	SH	GZ	SZ
BJ	0.4904	0.2342	0.1623	0.1792
SH	0.2610	0.5012	0.1548	0.1702
GZ	0.1863	0.1766	0.4301	0.2138
SZ	0.1866	0.1852	0.1965	0.4488

PG-SimCLR

Target \ Source	BJ	SH	GZ	SZ
BJ	0.4903	0.1836	0.1173	0.1126
SH	0.1802	0.4712	0.1182	0.1263
GZ	0.1522	0.1356	0.3719	0.1428
SZ	0.1428	0.1462	0.1273	0.4231

(c) GDP

UrbanCLIP

Target \ Source	BJ	SH	GZ	SZ
BJ	0.6559	0.3561	0.2804	0.2967
SH	0.3672	0.6782	0.2615	0.2700
GZ	0.3128	0.3077	0.5902	0.3011
SZ	0.3177	0.2987	0.2988	0.5633

PG-SimCLR

Target \ Source	BJ	SH	GZ	SZ
BJ	0.6382	0.3247	0.2467	0.2301
SH	0.3257	0.6010	0.2312	0.2013
GZ	0.2783	0.2748	0.5283	0.2972
SZ	0.2374	0.2333	0.2599	0.5066

Figure 5: R^2 heatmap for the transferability test between UrbanCLIP and PG-SimCLR, on 3 urban indicators across 4 datasets.

From Figure 4, the absence of supplementary textual information (i.e., UrbanViT) results in significant performance deterioration, demonstrating the importance of textual modality for achieving a comprehensive visual representation. UrbanViT slightly outperforms ResNet-18 [24], which mainly comes from the powerful capability of ViT to capture global dependencies and contextual understanding in images [16, 21].

4.3.2 Effectiveness of Refined Text. Before feeding the text input into UrbanCLIP, we refine the generated satellite imagery summary for more robust model performance. To validate the effectiveness of this process, we report the performance of using raw generated summary (i.e., UrbanCLIP w/o refined text) for comparison.

Figure 4 clearly shows that UrbanCLIP consistently outperforms this variant across all cities and indicators, though the magnitude of the difference varies. Such result indicates that more relevant and noise-free textual information may align better with image features, leading to a more coherent and meaningful visual representation.

4.3.3 Effectiveness of Knowledge Infusion. UrbanCLIP introduces contrastive learning-based cross-modality interaction coupled with image-text contrastive loss. To validate the efficacy of our approach in infusing textual knowledge, we introduce a direct image-image contrastive loss, denoted as Text-SimCLR, which is similar to PG-SimCLR [82] (the best baseline). In particular, Text-SimCLR calculates textual embedding similarity for positive region pairs, and mandates that the associated satellite images of these pairs be proximate in the visual latent space.

Figure 4 shows the performance comparison between Urban-CLIP and Text-SimCLR over different datasets. The substantial performance gaps observed between these two models suggest that relying solely on the conventional image view-based contrastive loss fails to accomplish effective knowledge infusion. In particular, directly capturing the semantic knowledge inherent in location summaries as a similarity metric, yields a relatively weak self-supervision signal for visual representation learning. In contrast, our proposed cross-modality interaction mechanism, grounded in text-image contrastive learning, more effectively incorporates text-enhanced information within the multimodal representation space. In summary, the results highlight the efficacy of our proposed textual knowledge infusion, with potential applications extending to other research areas involving satellite imagery.

4.3.4 Effectiveness of Loss Design. We further investigate the effects of the two losses, i.e., image-text contrastive loss and language modeling loss. As depicted in Figure 4, we assess the performance of UrbanCLIP in urban indicator prediction concerning contrastive-only and generative-only scenarios (denoted as UrbanCLIP w/o

$\mathcal{L}_{LM}$ and w/o $\mathcal{L}_{Con}$, respectively) across four datasets. The findings reveal that, when compared to UrbanCLIP utilizing both losses, both single-loss variants exhibit relatively inferior R^2 performance. Furthermore, UrbanCLIP exclusively employing language modeling loss outperforms the counterpart with only contrastive loss. This observation implies that the generative objective contributes to refining text representations, thereby augmenting text comprehension for multimodal fusion with visual representations [88]. In essence, combining both losses fosters the acquisition of more semantically rich visual representations of satellite images.

4.4 RQ3: Transferability Study

We then focus on the transferability of UrbanCLIP, by investigating its performance on unseen regions (not included in training).

4.4.1 Performance Across Cities. We conduct experiments of UrbanCLIP and PG-SimCLR on metropolises in China with different geological and demographic characteristics: 1) Beijing, located in the northern part of China as the capital, is densely populated and characterized by a mix of traditional architecture and modern facilities; 2) Shanghai, situated on the eastern coast, serves as a global financial center known for its cosmopolitan atmosphere and iconic skyline; 3) Guangzhou, positioned in southern China, is a major trading and manufacturing center and has an intricate network of waterways; 4) Shenzhen has the almost same location distribution as Guangzhou, but it has transformed into a bustling metropolis characterized by technology parks and industrial zones.

As shown in Figure 5, UrbanCLIP performs better than PG-SimCLR on 36 source-target pairs across three urban indicators. UrbanCLIP achieves an average R^2 of around 0.411, while that of PG-SimCLR is 0.365. Specifically, UrbanCLIP has higher R^2 values for respective urban indicators (carbon emission, population, and GDP) as 0.588, 0.384, and 0.261, but those of PG-SimCLR are only 0.543, 0.338, and 0.215. Such results indicate the stable transferability of our proposed UrbanCLIP in urban regions, although the chosen cities have the aforementioned differences in terms of geological and demographic characteristics.

The good transferability of our proposed UrbanCLIP may be attributed to our cross-modality mutual information maximization paradigm, through effective alignment and information preservation across visual representations and spatial semantics-enhanced textual representations. UrbanCLIP can better extract the inclusive functional semantics hidden behind satellite imagery, especially in urban scenarios involving spatial distribution shifts. Hence, although explicit differences exist among different cities, UrbanCLIP has the potential to address inaccuracies in the unseen satellite imagery of urban regions.
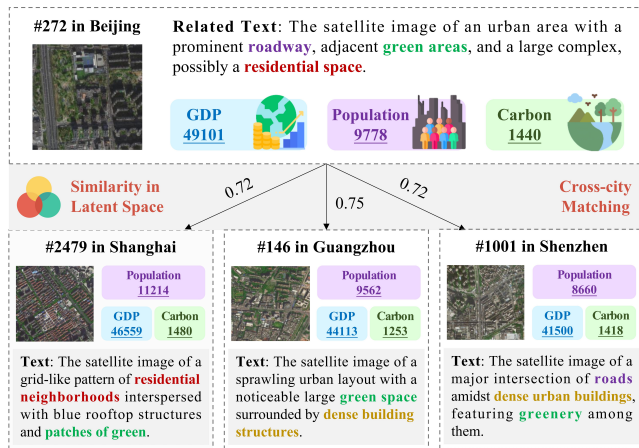
Figure 6: Case study of most similar satellite imagery matching between Beijing and other three cities through text-enhanced visual representations by UrbanCLIP.
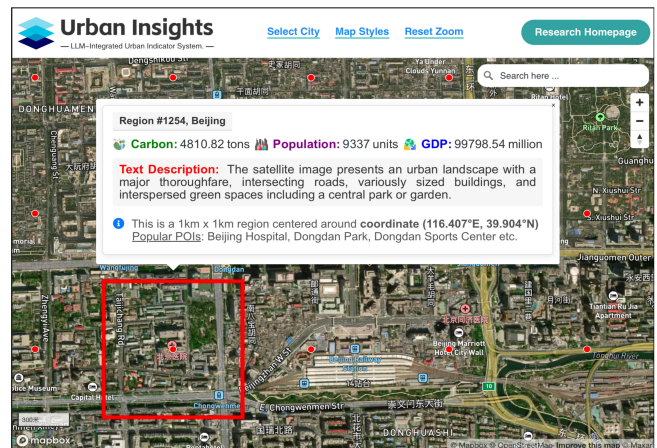

Figure 7: User interface of our Urban Insights System. It provides an interactive Mapbox-based platform [81] for urban region query (e.g., showing image captions and POI queries) and profiling (i.e., calculating the urban indicators like carbon emission, population, and GDP.

4.4.2 Similarity Analysis Across Cities. To better validate the transferability and explainability of UrbanCLIP across diverse urban regions, we compute the similarity between text-enhanced visual representations of satellite imagery. In particular, for a given satellite image from a source city, we compute the cosine similarity of visual representations among all others from different target cities. We assess whether there are commonalities in terms of urban indicators and description texts generated by UrbanCLIP. We further investigate the capability of generated text to guide associated image representations to focus on similar spatial information.

As illustrated by Figure 6, a randomly chosen satellite image in Beijing corresponds to three satellite images from other cities (Shanghai, Guangzhou, and Shenzhen) with the highest similarities (0.72, 0.75, and 0.72, respectively) in text-enhanced visual representations. In terms of urban indicators of regions corresponding to these satellite images, we can see that they are very close to each other. This phenomenon suggests that UrbanCLIP can capture similar spatial characteristics and distributions among comparable regions, thereby contributing to effective urban region profiling.

Furthermore, the location summary especially identifies the significant spatial attributes of the urban region. Infusing such key knowledge into the visual representation leads to a more comprehensive representation. For instance, the summary of the Beijing example can pinpoint the presence of roadways, green areas, and residential spaces, aligning with mentions in the summaries of the other three examples with the highest similarity. These findings support the notion that UrbanCLIP exhibits robust transferability and explainability across diverse urban regions.

4.5 RQ4: Practicality

We finally envision and develop a novel web-based application called Urban Insights, which is an LLM-Integrated Urban Indicator System built on the Mapbox platform [55]. It displays urban landscapes in satellite projection, offering an interactive user experience. As shown in Figure 7, users can easily navigate the map by zooming in and out, searching for special locations, and switching between different areas. Overlaid on this imagery are target grid areas, which will furnish users with detailed metrics, including carbon emissions, population, and GDP once clicked. Complementing the visual data, the system also features a descriptive image captioning module, which provides an easy-to-read text for understanding the spatial attributes of the selected grid, making it simpler for users to comprehend the spatial characteristics of the chosen grid. In addition, the system also supports popular POI query features within a region to better understand region functions. In summary, the Urban Insights System has great potential to provide users with a comprehensive and enriched view of varied urban landscapes and their prominent indicators, translating intricate urban data into a more accessible and intuitive visual representation. More details will be released upon paper notification to obey anonymity.

5 CONCLUSION AND FUTURE WORK

Profiling urban areas in terms of social, economic, and environmental metrics is critical for urban planning and sustainable development. This paper investigates whether and how the text modality benefits urban region profiling. To answer the question, we propose UrbanCLIP, the first-ever framework that integrates textual modality into urban imagery profiling. Powered by LLM, UrbanCLIP first generates a high-quality text description for an urban image. Then the text-image pairs are fed into the proposed model that seamlessly unifies natural language supervision for urban visual representation learning. Extensive experiments demonstrate the effectiveness of incorporating the textual modality.

We aspire that this work motivates future research of urban region profiling on the following areas: 1) Investigating efficient and effective methods for integrating urban multimodal data and facilitating prompt-enhanced learning; 2) Exploring the automatic, high-quality text generation and refinement using more up-to-date LLMs; 3) Identifying more potentially beneficial downstream tasks, encouraging other researchers to explore diverse use cases.

REFERENCES

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022).

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390* (2023).

[4] Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. 2020. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612* (2020).

[5] Kumar Ayush, Burak Uzkent, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. 2021. Efficient poverty mapping from high resolution remote sensing images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 12–20.

[6] 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 Journal of Photogrammetry and Remote Sensing* 201 (2023), 193–208.

[7] The World Bank. 2022. *Urban Development*. https://www.worldbank.org/en/topic/urbandevelopment/overview Accessed: 2022-11-01.

[8] Johan Bjorck, Brendan H Rappazzo, Qinru Shi, Carrie Brown-Lima, Jennifer Dean, Angela Fuller, and Carla Gomes. 2021. Accelerating ecological sciences from above: Spatial contrastive learning for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14711–14720.

[9] Marshall Burke, Anne Driscoll, David B. Lobell, and Stefano Ermon. 2021. Using satellite imagery to understand and promote sustainable development. *Science* 371, 6535 (2021), eabe8628. https://doi.org/10.1126/science.abe8628

[10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[12] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2818–2829.

[13] 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. *Advances in Neural Information Processing Systems* 35 (2022), 197–211.

[14] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on CLIP via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386* (2022).

[15] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*. PMLR, 2793–2803.

[16] 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).

[17] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* (2023).

[18] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).

[19] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks Are All You Need. *arXiv preprint arXiv:2306.11644* (2023).

[20] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. ImageBind-LLM: Multi-modality Instruction Tuning. arXiv:2309.03905 [cs.MM]

[21] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 87–110.

[22] Sungwon Han, Donghyun Ahn, Hyunji Cha, Jeasurk Yang, Sungwon Park, and Meeyoung Cha. 2020. Lightweight and robust representation of economic scales from satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 428–436.

[23] Sungwon Han, Donghyun Ahn, Sungwon Park, Jeasurk Yang, Susang Lee, Jihee Kim, Hyunjoo Yang, Sangyoon Park, and Meeyoung Cha. 2020. Learning to score economic development from satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2970–2979.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[25] Zhiyuan He, Su Yang, Weishan Zhang, and Jiulong Zhang. 2018. Perceiving commerial activeness over satellite images. In *Companion Proceedings of the The Web Conference 2018*. 387–394.

[26] Danfeng Hong, Lianru Gao, Jing Yao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. 2020. Graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* 59, 7 (2020), 5966–5978.

[27] Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A Survey of Knowledge Enhanced Pre-Trained Language Models. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[28] Tianyuan Huang, Zhecheng Wang, Hao Sheng, Andrew Y Ng, and Ram Rajagopal. 2021. M3G: Learning urban neighborhood representation from multi-modal multi-graph. In *Proceedings of the DeepSpatial 2021: 2nd ACM KDD Workshop on Deep Learning for Spatio-Temporal Data, Applications and Systems*.

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. https://doi.org/10.5281/zenodo.5143773

[30] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

[31] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3967–3974.

[32] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. 2019. Unsupervised representation learning of spatial data via multimodal embedding. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1993–2002.

[33] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.

[34] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2023. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *arXiv preprint arXiv:2310.01728* (2023).

[35] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2023. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *arXiv preprint arXiv:2310.01728* (2023).

[36] Jian Kang, Ruben Fernandez-Beltran, Puhong Duan, Sicong Liu, and Antonio J Plaza. 2020. Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Transactions on Geoscience and Remote Sensing* 59, 3 (2020), 2598–2610.

[37] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[38] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[39] Mark A Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37, 2 (1991), 233–243.

[40] Stephen Law, Brooks Paige, and Chris Russell. 2019. Take a look around: using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 5 (2019), 1–19.

[41] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*. PMLR, 3744–3753.

[42] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[43] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7241–7259.

https://doi.org/10.18653/v1/2022.emnlp-main.488

[44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

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

[46] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[47] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273* (2022).

[48] Tong Li, Shiduo Xin, Yanxin Xi, Sasu Tarkoma, Pan Hui, and Yong Li. 2022. Predicting multi-level socioeconomic indicators from structural urban imagery. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3282–3291.

[49] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[50] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science* 378, 6624 (2022), 1092–1097.

[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.

[52] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[53] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. 2023. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*. 4150–4160.

[54] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. 2019. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 75–82.

[55] Mapbox. [n. d.]. *Mapbox - Location Data & Maps for Developers*. https://www.mapbox.com/

[56] Jorge Andres Chamorro Martinez, Laura Elena Cué La Rosa, Raul Queiroz Feitosa, Ieda Del'Arco Sanches, and Patrick Nigri Happ. 2021. Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing* 171 (2021), 188–201.

[57] MEASURE DHS et al. 2013. *Demographic and Health Surveys*. Measure DHS, Calverton.

[58] Harvey J Miller. 2004. Tobler's first law and spatial analysis. *Annals of the association of American geographers* 94, 2 (2004), 284–289.

[59] Guofeng Cao Eric L. Samson Naizhuo Zhao, Ying Liu and Jingqi Zhang. 2017. Forecasting China's GDP at the pixel level using nighttime lights time series and population images. *GIScience & Remote Sensing* 54, 3 (2017), 407–425. https://doi.org/10.1080/15481603.2016.1276705

[60] Tomohiro Oda, Shamil Maksyutov, and Robert J Andres. 2018. The Open-source Data Inventory for Anthropogenic CO 2, version 2016 (ODIAC2016): a global monthly fossil fuel CO 2 gridded emissions data product for tracer transport simulations and surface flux inversions. *Earth System Science Data* 10, 1 (2018), 87–107.

[61] Junting Pan, Ziyi Lin, Yuying Ge, Xiatian Zhu, Renrui Zhang, Yi Wang, Yu Qiao, and Hongsheng Li. 2023. Retrieving-to-Answer: Zero-Shot Video Question Answering with Frozen Large Language Models. *arXiv preprint arXiv:2306.11732* (2023).

[62] Sungwon Park, Sungwon Han, Donghyun Ahn, Jaeyeon Kim, Jeasurk Yang, Susang Lee, Seunghoon Hong, Jihee Kim, Sangyoon Park, Hyunjoo Yang, et al. 2022. Learning economic indicators by aggregating multi-level geospatial information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12053–12061.

[63] Anthony Perez, Christopher Yeh, George Azzari, Marshall Burke, David Lobell, and Stefano Ermon. 2017. Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654* (2017).

[64] 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 *International conference on machine learning*. PMLR, 8748–8763.

[65] Hannah Ritchie and Max Roser. 2018. Urbanization. *Our World in Data* (2018). https://ourworldindata.org/urbanization.

[66] Marc Rußwurm and Marco Körner. 2020. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing* 169 (2020), 421–435.

[67] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188* (2022).

[68] Hung-Ting Su, Yulei Niu, Xudong Lin, Winston H Hsu, and Shih-Fu Chang. 2023. Language Models are Causal Knowledge Extractors for Zero-shot Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4950–4959.

[69] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355* (2023).

[70] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative Pretraining in Multimodality. arXiv:2307.05222 [cs.CV]

[71] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17918–17928.

[72] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

[73] Michael E Tipping and Christopher M Bishop. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61, 3 (1999), 611–622.

[74] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.

[75] UN Department of Economic and Social Affairs. 2022. *The Sustainable Development Goals Report 2021*. Technical Report. United Nations. https://unstats.un.org/sdgs/report/2021/

[76] Burak Uzkent, Evan Sheehan, Chenlin Meng, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. 2019. Learning to interpret satellite images using wikipedia. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

[77] Anna X Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. 2018. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–5.

[78] 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* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 647–654. https://doi.org/10.1145/3184558.3186581

[79] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1013–1020.

[80] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[81] WorldPop. [n. d.]. *WorldPop - Open Spatial Demographic Data and Research*. https://www.worldpop.org/

[82] 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 *Proceedings of the ACM Web Conference 2022*. 3308–3316.

[83] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems* 35 (2022), 124–141.

[84] Jianzhong Yang, Xiaoqing Ye, Bin Wu, Yanlei Gu, Ziyu Wang, Deguo Xia, and Jizhou Huang. 2022. DuARE: Automatic road extraction with aerial images and trajectory data at Baidu maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4321–4331.

[85] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. 2021. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724* (2021).

[86] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications* 11, 1 (2020), 2583.

[87] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[88] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation

Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=Ee277P3AYC

[89] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23465–23476.

[90] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529* 1, 6 (2021), 8.

[91] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).

[92] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[93] Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. *Advances in Neural Information Processing Systems* (2023).

6 APPENDIX

6.1 Image-to-Text Foundation Models

We provide a brief introduction of the Image-to-Text foundation models that we used for text generation:

- **BLIP**. A VLP framework that leverages noisy web data for bootstrapping captions; it involves a captioner generating synthetic captions and a filter to eliminate noisy ones.
- **Emu**. A multimodal foundation model trained with a unified objective, either classifying the next text token or regressing the next visual embedding in the multimodal sequence.
- **ImageBind-LLM**. A multimodal instruction model that unifies various modalities such as images and video into a single framework by aligning ImageBind's visual encoder with an LLM using a learnable bind network.

- **PandaGPT**. A unified approach that can handle multimodal inputs, allowing natural composition of their semantics by combining multimodal encoders from ImageBind and LLMs from Vicuna.
- **OpenFlamingo**. An open-source multimodal framework that is capable of handling diverse visual language tasks through autoregressive vision-language modeling.
- **mPLUG**. A VLP model with an efficient vision-language architecture, equipped with innovative cross-modal skip-connections.
- **LLaVA**. An instruction tuning-based model that utilizes multimodal language-image data derived from GPT4.

6.2 Complexity Analysis

We use the following notations: m_1 represents the number of visual tokens of ViT, d is the dimension of the representation, L denotes the number of layers in the transformer (assuming uniformity across ViT, textual transformer, and multimodal transformer), and m_2 stands for the sequence length of textual tokens. For the visual encoder, the complexity of ViT is $\mathbf{O}(L(m_1^2 d + m_1 d^2))$ and that of attentional pooling is $\mathbf{O}(m_1^2 d)$. The textual encoder has an embedding lookup complexity of $\mathbf{O}(m_2 d)$ and transformers with $\mathbf{O}(L(m_2^2 d + m_2 d^2))$. The multimodal interaction involves cross attention with a complexity of $\mathbf{O}(Lm_1 m_2 d)$. The final complexity, when summing up, is $\mathbf{O}(L(m_1^2 + m_2^2)d + m_1^2 d + Lm_1 m_2 d)$, which for large values of $m_1 n$ and m_2 approximates to $\mathbf{O}(L(m_1^2 d + m_2^2 d))$. Besides, LLM pre-training is excluded from UrbanCLIP backbone training, and text generation and refinement remain at the preprocessing phase, thus indicating the feasibility of UrbanCLIP in practice.