

UrbanLLaVA: A Multi-modal Large Language Model for Urban Intelligence

Jie Feng[†], Shengyuan Wang[‡], Tianhui Liu[§], Yanxin Xi[¶], Yong Li[†]

[†]Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China

[‡]Department of Computer Science and Technology, Tsinghua University, Beijing, China

[§]School of Electronic and Information Engineering, Beijing Jiaotong University, China

[¶]University of Helsinki, Finland

{fengjie, liyong07}@tsinghua.edu.cn

Abstract

Urban research involves a wide range of scenarios and tasks that require the understanding of multi-modal data. Current methods often focus on specific data types and lack a unified framework in urban field for processing them comprehensively. The recent success of multi-modal large language models (MLLMs) presents a promising opportunity to overcome this limitation. In this paper, we introduce UrbanLLaVA, a multi-modal large language model designed to process these four types of data simultaneously and achieve strong performance across diverse urban tasks compared with general MLLMs. In UrbanLLaVA, we first curate a diverse urban instruction dataset encompassing both single-modal and cross-modal urban data, spanning from location view to global view of urban environment. Additionally, we propose a multi-stage training framework that decouples spatial reasoning enhancement from domain knowledge learning, thereby improving the compatibility and downstream performance of UrbanLLaVA across diverse urban tasks. Finally, we also extend existing benchmark for urban research to assess the performance of MLLMs across a wide range of urban tasks. Experimental results from three cities demonstrate that UrbanLLaVA outperforms open-source and proprietary MLLMs in both single-modal tasks and complex cross-modal tasks and shows robust generalization abilities across cities. Source codes and data are openly accessible to the research community via <https://github.com/tsinghua-fib-lab/UrbanLLaVA>.

1. Introduction

Urban science [53, 57] and geographic science research [35] highlight that urban data spans multiple modalities, including urban visual data [14], geo-text [45], structured geospatial data [1, 2], and spatiotemporal series data [20, 28]. To-

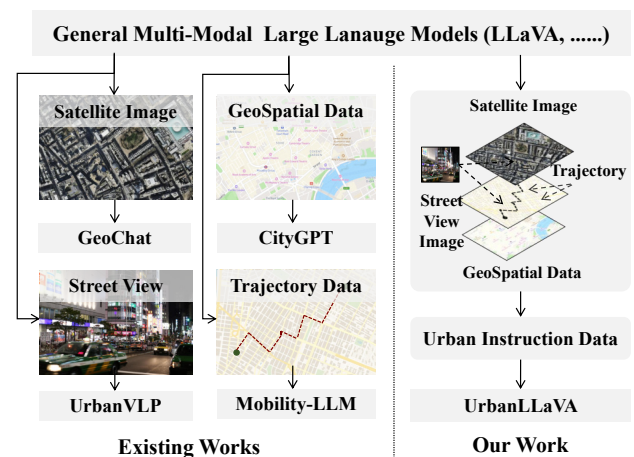


Figure 1. Existing works vs. our *UrbanLLaVA* in urban research.

gether, these data types capture the multi-faceted nature of urban environments, representing a wide range of spatial information and urban knowledge [35, 46, 57]. Integrating these multi-modal data into a cohesive framework is essential for developing a systematic understanding of urban spaces and advancing complex modeling architectures in urban research. However, the inherent heterogeneity of these diverse urban data presents substantial challenges for the integration. While numerous deep learning based methods have been proposed to fuse various cross-domain urban data [57], they are often designed for specific urban tasks, limiting their ability to achieve a comprehensive understanding of urban environment and advanced reasoning for real-world urban applications [46, 53].

Recently, multi-modal large language models (MLLMs) [49] have made notable advancements by leveraging large language models (LLMs) [38] with build-in common sense and reasoning abilities as a central component for unifying the processing data across various modalities, such as images [31], speech [19], and

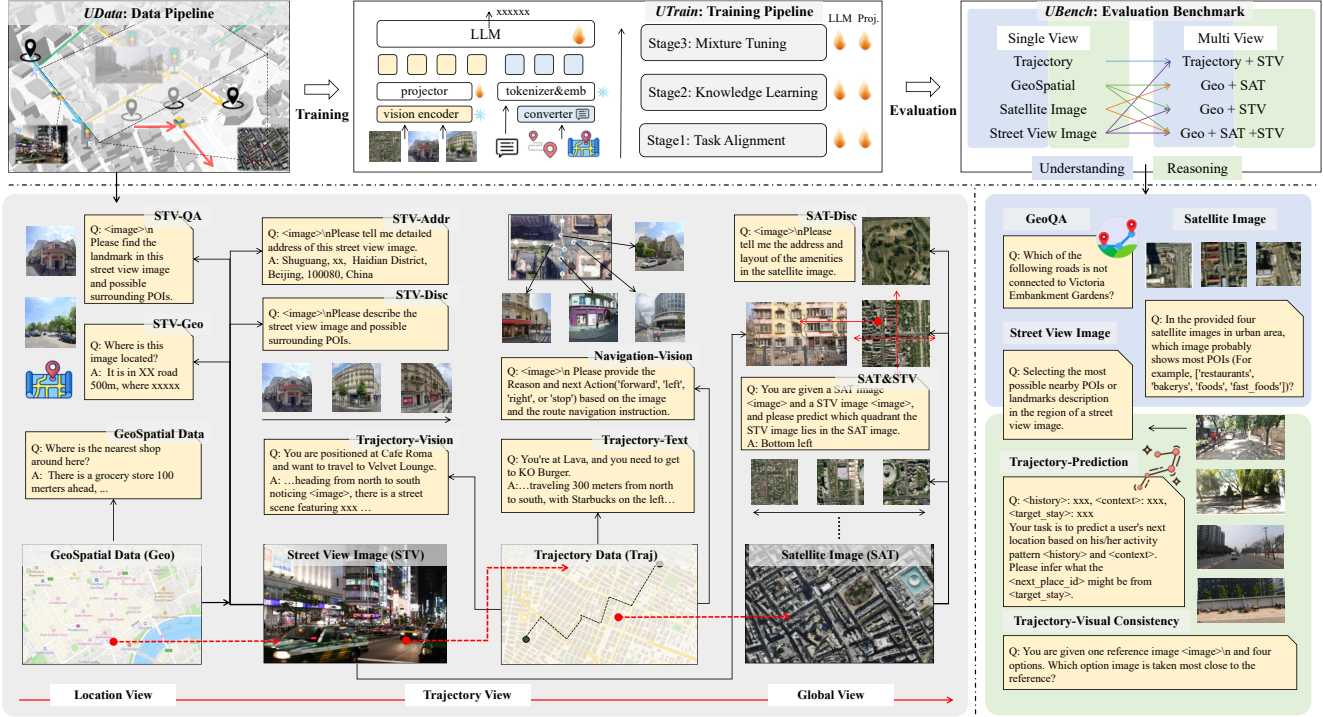


Figure 2. The framework of *UrbanLLaVA*, including *UData*, *UTrain* and *UBench*.

time series [25]. For example, Ma et al. [34] develop a vision-language model as a conversational assistant for autonomous driving, Brohan et al. [4] introduce RT-2, a vision language model based end-to-end model for flexible robotics control, Li et al. [27] train LLaVA-Med for answering open-ended questions related to biomedicine images. Building on this trend, researchers have begun to explore the potential of MLLMs in urban studies [53]. As shown in the left part of Figure 1, notable examples include GeoChat [26], an early effort in creating MLLMs for remote sensing tasks; Mobility-LLM [20], which extends LLM with capabilities for trajectory modeling; and CityGPT [16], designed to process structured geospatial data with LLMs. In contrast to earlier urban data fusion methods developing in the deep learning era [53, 57], these recent studies incorporate various unimodal urban data into LLMs to create obtain MLLMs that maintain the powerful reasoning abilities and address diverse urban tasks within a single modality.

However, these recent works focus solely on processing unimodal urban data and fall short of achieving a comprehensive understanding and modeling of urban system across diverse tasks involving multi-modal urban data. Unified modeling of multi-modal urban data poses significant challenges. The first challenge is the scarcity of high-quality data for cross-modality alignment. While previous works [16, 26] propose various methods for constructing

instruction tuning dataset for different types of unimodal urban data integrated with language, these efforts are insufficient for unified modeling across multiple modalities. A second challenge lies in the potential conflicts among diverse urban tasks across different modalities, which can lead to unstable training and inconsistent performance.

In this paper, we introduce *UrbanLLaVA*, a multi-modal large language model designed to build comprehensive urban cognition and addressing a wide range of urban tasks, which is shown in the right part of Figure 1. In *UrbanLLaVA*, we first design *UData*, a systematic urban instruction data pipeline that enables the generation of high-quality synthetic data. In *UData*, data generation is meticulously structured to span multiple perspectives—from a localized view for single modality data to trajectory and global view for cross-modality data—capturing the inherently multi-faceted nature of urban system. To improve the training stability and model performance, we conduct extensive experiments to identify key factors impacting the training process and develop an effective three-stage training pipeline *UTrain*, based on these insights. In fact, the proposed multi-stage training framework can be viewed as a promising practice that explicitly decouples the learning of reasoning capabilities from domain-specific knowledge in MLLMs. Finally, we extend existing urban benchmarks to build a systematic urban benchmark *UBench* for evaluating the capabilities of MLLMs in tackling diverse urban

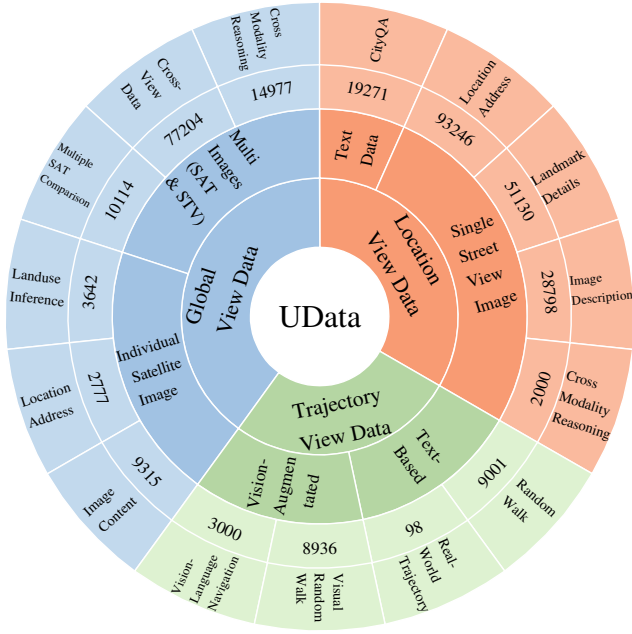


Figure 3. The thorough composition of *UData* in Beijing.

tasks. In summary, our contributions are as follows,

- *UrbanLLaVA* is the first MLLM designed for the unified modeling of four major types of urban data, with the goal of fostering comprehensive understanding and effective task-solving for urban environments, to the best of our knowledge.
- We conduct extensive experiments to identify the key factors influencing training and propose a three-stage training pipeline that ensures stable performance of *UrbanLLaVA* across a wide range of urban tasks involving multiple data modalities.
- *UrbanLLaVA* demonstrates effective integration of multi-modal data, establishing comprehensive spatial cognition and outperforming general MLLMs across various urban tasks based on results from an enhanced urban task benchmark.

2. Methods

As illustrated in Figure 2, *UrbanLLaVA* comprises three key components: 1) the data pipeline, *UData*, designed for generating diverse and high-quality urban instruction data across various urban scenarios; 2) the training pipeline, *UTrain*, which facilitates efficient and stable training across a wide range of urban tasks; 3) the evaluation benchmark, *UBench*, for evaluating the capabilities of MLLMs in multi-modal urban tasks.

2.1. *UData*: Constructing Urban Instruction Data from a Multi-View Perspective of Urban Space

Over the past decade, effectively integrating multi-modal urban data has emerged as a key research question in urban studies [57]. Building on the successes of MLLMs in various fields [49, 53], we extend the modelling of four types of urban data into a unified model, *UrbanLLaVA*, by constructing a diverse urban instruction data from a systematic view on the urban environment. Specifically, we organize the urban instruction data in a sequence that move from location view to a trajectory view, and finally to a global view. This approach ensures both broad spatial coverage and the integrity of relationships between different modalities in the final data. *UData* builds upon four kinds of original urban data: 1) the structured geospatial data from OpenStreetMap¹; 2) public trajectory data, e.g., Foursquare-checkins [48] and OpenStreetMap traces²; 3) satellite images from GoogleEarth³; 4) street view images from GoogleMap⁴ and BaiduMap⁵. Before experiments, we collect original data from above platforms and using the following data pipeline to build instruction data. We follow the license of these platforms and ensure that the data is used only for academic research.

2.1.1. Location View Data

In the location view data construction stage, we focus on structured geospatial data and single street view images. Following the recent practices [1, 16] for structured geospatial knowledge learning, we create geospatial instruction data by designing question templates that transform basic geospatial data into natural language question and answers. For single street view image, we synthesize three types of questions: 1) two types based on predefined templates populated with information from structured geospatial data, such as, location addresses and landmark details; 2) one general MLLM generated detailed description of the image content, following the common practice for image captioning [6]. Throughout the data construction, we maintain a core principle of integrating street view image content with structured geographical knowledge, such as consistency in location addresses and landmark descriptions.

2.1.2. Trajectory View Data

Here, we construct the trajectory view data, which includes the geospatial data, trajectory data, and street view images. We start by creating two types of text-based trajectory data. The first type is generated by randomly sampling origin and destination points for routing, while the second type uses the real-world trajectory data collected from the public web

¹<https://www.openstreetmap.org>

²<https://www.openstreetmap.org/traces>

³<https://earth.google.com/>

⁴<https://www.google.com/maps>

⁵<https://map.baidu.com/>

source, including Foursquare-checkins and OpenStreetMap traces. To enhance geospatial context of trajectory data, we align the GPS coordinates from these original data sources with the structured geospatial data, using the textual addresses to represent locations within the trajectory. Additionally, we integrate street view images to enrich trajectory data, resulting two types of vision-augmented trajectory data. The first data extends the text-based trajectory data by incorporating street view images captured along the route (excluding intersections). We organize this data with the similar interleaved image-text format in VILA [29]. The second data builds on the navigation instruction format akin to the classic vision-language navigation task [5]. In this data, multiple street view images are presented at intersection during the trajectory, and the correct image is selected to guide the continuation of the journey.

2.1.3. Global View Data

Here, we present the construction of global view data designed to capture relationships among diverse data types over long distances, with street view images and satellite images as primary components and geospatial data serving as auxiliary support. Initially, we create a basic form of global view data by generating captions for single satellite image data enriched with structured geospatial data. Specifically, we define three types of data: 1) prompting general MLLM to produce detailed content description for individual satellite image; 2) sampling location address within satellite image and using a general LLM to summarize the spatial coverage of it based on these location address; 3) prompting general MLLM with land use ground-truth label to generate land use inference results with reason.

Furthermore, we introduce the multiple satellite images for more complex instruction data. The first task is to compare the building densities across multiple satellite images. The second task focuses on identifying functional point of interest within these images. For these tasks, we provide manually crafted reasoning steps in a chain-of-thoughts format, supported by structured geospatial data, to improve the alignment between satellite images and geospatial data. Finally, we design two tasks to strengthen the alignment between the street view images and satellite images. The first task is to select the correct satellite image from a set when given a street view image, requiring the model to understand and match content or address across both image types. The second, more challenging task involves pinpointing the location of the street view image within a specific satellite image, such as identifying it as located in the top-left region of satellite image.

Based on the data generation steps described before, we perform data quality checks and filtering on the synthesized data to ensure its quality.

2.2. *UTrain*: A Multi-Stage Training Pipeline for Decoupling Reasoning and Knowledge Learning

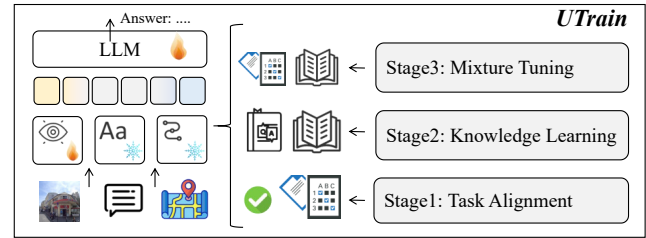


Figure 4. *UTrain*: three-stage training pipeline.

Training *UrbanLLaVA* presents significant challenges due to the heterogeneity of multi-modal urban instruction data and the diversity of urban tasks. Achieving stable training and balancing performance across various tasks is notably difficult. We chose VILA[29] as the base model for our experiments and conduct extensive studies to identify key factors affecting training. We examine the impact of the training order of multi-modal data and trained components, observing minimal effects. However, we find that learning rate has significant effects on training stability and performance. Detailed results of them are provided in the section 3.3. Additionally, inspired by Dong et al. [12], we explore and propose an effective multi-stage training pipeline which is shown in the Figure 4.

We first introduce three kinds of learning procedures: *knowledge learning*, *task alignment* and *mixture learning*. The *knowledge learning* procedure refers to the training process which *UrbanLLaVA* acquires foundational urban knowledge from various urban data, such as the information of geospatial data, pure textual trajectory, and detailed description of street view and satellite images. The *task alignment* learning focuses on equipping *UrbanLLaVA* with task-specific skills for urban applications, including vision-language navigation, trajectory prediction, chain-of-thoughts reasoning across multiple satellite and street view images. Finally, *mixture learning* represents the standard training method used by most MLLMs, which training by directly mixing all types of instruction data.

During our experiments, we observe that different combination of various learning procedures significantly impact the training. Based on the observations, we propose a three-stage tuning pipeline to improve the training stability and performance on diverse urban tasks. This pipeline consists of three sequential stages: *task alignment*, *knowledge learning*, and finally *mixture learning*. Starting with a well-trained general MLLM as our base model, we first introduce the *task alignment* learning procedure, fine-tuning the model with diverse urban task related instructions to prepare it for various urban tasks. Through this phase, the model become familiar with a variety of urban tasks, lever-

aging its pre-existing general knowledge to complete them. However, familiarity with general knowledge alone is insufficient for effectively addressing diverse urban tasks, so we incorporate the second stage, *knowledge learning* procedure. This stage imparts specialized urban knowledge from multi-modal urban data that is essential for task resolution. Finally, we introduce the mixture learning stage to enhance the model’s awareness of combining knowledge and skills for solving diverse urban tasks. Here, we resample 1/3 domain specific data from the first two stages and 1/3 general textual instruction data, e.g. ShareGPT⁶ and UltraChat [11], for final tuning.

2.3. *UBench*: An Enhanced Multimodal Benchmark for Urban Intelligence Tasks

To assess the potential of MLLMs in urban studies, CityBench [18] and Urbench [56] have been recently introduced. Drawing from the diverse evaluation tasks in these two benchmarks, we reorganize and expand them to create the urban evaluation benchmark *UBench*, which includes 12 tasks for our experiments. All the evaluation tasks are presented in Table 1. We select 6 of these tasks based on the utility of the evaluation data and their relevance to urban scenarios involving *UrbanLLaVA*’s urban data. For structured geospatial data and trajectory modelling, we incorporate the GeoQA, trajectory prediction and navigation task from CityBench. For cross-view urban tasks involving both street view and satellite images, we adopt the image retrieval, camera localization, and scene comparison task from UrBench. In addition, we introduce 6 new tasks in *UBench*. Four of these tasks are designed for single street view and satellite images, including address inference for both image types, landmark recognition for street view images, and land use inference for satellite images. These single-image tasks are aligned with the urban instruction data, and we partition the original dataset into training and validation sets to prevent potential data leakage. Moreover, we build 2 additional tasks involving multiple images: 1) STV-Outlier, is a spatial consistency task for street view image, where multiple images from a single trajectory are compared to identify an outlier image not part of the trajectory; 2) SceneFunc, extends the scene comparison task from UrBench, challenging model to select the correct satellite image that fulfills specific functional requirements.

3. Experiments

3.1. Settings

We select Beijing, London and New York to conduct experiments. Due to the large volume of data, we select a region

⁶<https://huggingface.co/datasets/shareAI/ShareGPT-Chinese-English-90k>

Table 1. Detailed information about *UBench* for Beijing, ‘STV’ refers to street view image, and ‘SAT’ refers to satellite image.

| Tasks | Data | Category | Metrics | Samples | Source |
|---------------------|-----------------|----------|---------------|---------|---------------|
| GeoQA | Geospatial Data | GeoQA | Avg. Accuracy | 1450 | CityBench |
| TrajPredict | Trajectory Data | Geo+Traj | Top-1 | 500 | CityBench |
| Navigation | Single STV | Geo+Traj | Success Rate | 50 | CityBench |
| SceneComp | Multi SAT | Geo+SAT | Accuracy | 200 | UrBench |
| ImgRetrieval | Multi STV & SAT | Geo+SS | Accuracy | 200 | UrBench |
| CameraLoc | Multi STV & SAT | Geo+SS | Accuracy | 200 | UrBench |
| STV-Address | Single STV | Geo+STV | Accuracy | 200 | <i>UBench</i> |
| STV-Landmark | Single STV | Geo+STV | Accuracy | 200 | <i>UBench</i> |
| SAT-Address | Single SAT | Geo+SAT | Accuracy | 200 | <i>UBench</i> |
| SAT-Landuse | Single SAT | Geo+SAT | Accuracy | 200 | <i>UBench</i> |
| STV-Outlier | Multi STV | Geo+STV | Accuracy | 200 | <i>UBench</i> |
| SceneFunc | Multi SAT | Geo+SAT | Accuracy | 200 | <i>UBench</i> |

from each cities to conduct experiments. The spatial coverage of each region is shown in supplementary material.

MLLMs We consider the following MLLMs as baselines: Qwen2VL-7B/72B [41], InternVL2-8B/26B [7, 8], VILA1.5-3B/8B/13B [29], LLama3.2-11B/90B [36], and GPT4o and GPT4o-mini [40]. For open source MLLMs, we deploy them through VLMEvalKit [13]. The max output tokens are set to 1000, and the temperature is set as 0.

Metrics Table 1 contains all the metrics for *UBench*. For general evaluation tasks including LLaVA-Bench(In-the-Wild) [30], RealWorldQA [44], and MM-Vet [50], RealWorldQA uses accuracy as the metric, while LLaVA-Bench(In-the-Wild) and MM-Vet use rating score form GPT4o as the judgement.

Implementation We use codes from official repository⁷ of VILA [29] for fine-tuning on a single 8xA100 node. The training parameters are set as follows: a learning rate of 1e-5, a maximum sequence length of 2048, a batch size of 8 per GPU, and one training epoch. Training UrbanLLaVA for Beijing on 4xA100 took a total of 10.7 hours.

3.2. Main Results

The main results of *UrbanLLaVA* on three cities are presented in Table 2. We use VILA1.5-8B as the default base model in most experiments and use *UData* with *UTrain* methods to fine-tune it to obtain the final model *UrbanLLaVA*.

We analyze the results in Beijing first. One point to note is that, since LLama3.2 series models currently do not support multi-image input, the results for evaluation tasks involving multiple images in the *UBench* are left blank. For models within the same series, the general trend is that larger parameter models tend to perform better, e.g., VILA1.5-13b significantly outperforms VILA1.5-3b on 5 out of 6 tasks, including both single modal and cross modal tasks. Additionally, we observe that the latest released Qwen2VL series models outperform the GPT4o se-

⁷<https://github.com/NVlabs/VILA>

Table 2. Results on *UBench* at Beijing, London, and New York. *UrbanLLaVA* significantly outperforms other baselines in most task across cities. Here, ‘STV’ denotes street view images related tasks, ‘Geo’ denotes geospatial data, ‘Traj’ denotes trajectory related task, ‘SAT’ denotes satellite images related tasks, and ‘SS’ denotes street view + satellite images. Detailed subtask and metrics can refer to Table 1.

| City Task Group | Beijing | | | | London | | | | New York | | | | | | |
|-------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | GeoQA | Geo+Traj | Geo+STV | Geo+SAT | Geo+SS | GeoQA | Geo+Traj | Geo+STV | Geo+SAT | Geo+SS | GeoQA | Geo+Traj | Geo+STV | Geo+SAT | Geo+SS |
| VILA1.5-3B | 0.3873 | 0.0200 | 0.3967 | 0.3200 | 0.2575 | 0.4362 | 0.0400 | 0.2557 | 0.2850 | 0.2725 | 0.3954 | 0.0400 | 0.4400 | 0.2713 | 0.2425 |
| VILA1.5-8B | 0.4322 | 0.0589 | 0.4300 | 0.3488 | 0.2425 | 0.4841 | 0.0884 | 0.4495 | 0.4575 | 0.2575 | 0.4575 | 0.1200 | 0.4983 | 0.3763 | 0.2525 |
| VILA1.5-13B | 0.4410 | 0.1156 | 0.5167 | 0.3638 | 0.2400 | 0.4592 | 0.1298 | 0.4991 | 0.4538 | 0.2625 | 0.4501 | 0.2350 | 0.5583 | 0.4025 | 0.2825 |
| InternVL2-8B | 0.4709 | 0.1578 | 0.4667 | 0.3313 | 0.2325 | 0.4973 | 0.1347 | 0.4477 | 0.4763 | 0.2400 | 0.4632 | 0.1830 | 0.4917 | 0.4175 | 0.2400 |
| InternVL2-26B | 0.4877 | 0.1478 | 0.4550 | 0.3825 | 0.2275 | 0.5168 | 0.1288 | 0.4923 | 0.5138 | 0.2425 | 0.4766 | 0.2240 | 0.5217 | 0.4738 | 0.2375 |
| Qwen2VL-7B | 0.4950 | 0.1389 | 0.4383 | 0.3638 | 0.2675 | 0.4991 | 0.1560 | 0.4381 | 0.4863 | 0.2775 | 0.4567 | 0.1700 | 0.5117 | 0.5100 | 0.2950 |
| Qwen2VL-72B | 0.5491 | 0.1611 | 0.5817 | 0.3588 | 0.2975 | 0.5802 | 0.2322 | 0.6375 | 0.4375 | 0.3250 | 0.5273 | 0.2540 | 0.6333 | 0.3788 | 0.3275 |
| LLaMA3.2-11B | 0.4229 | 0.0756 | 0.4375 | 0.3075 | / | 0.4804 | 0.1180 | 0.4000 | 0.3800 | / | 0.4127 | 0.1100 | 0.5200 | 0.2225 | / |
| LLaMA3.2-90B | 0.4502 | 0.1056 | 0.5325 | 0.2925 | / | 0.5659 | 0.2010 | 0.5450 | 0.4700 | / | 0.5234 | 0.1570 | 0.6825 | 0.3400 | / |
| GPT4o-mini | 0.4542 | 0.1622 | 0.4350 | 0.3800 | 0.2475 | 0.5357 | 0.1278 | 0.4752 | 0.5388 | 0.2675 | 0.5075 | 0.2320 | 0.5633 | 0.4775 | 0.2350 |
| GPT4o | 0.5479 | 0.1522 | 0.4300 | 0.4125 | 0.3025 | 0.6446 | 0.1300 | 0.5469 | 0.6050 | 0.2850 | 0.6232 | 0.2340 | 0.5767 | 0.5400 | 0.2900 |
| <i>UrbanLLaVA</i> -VILA1.5-8B | 0.5682 | 0.2800 | 0.8650 | 0.6663 | 0.7025 | 0.6399 | 0.2680 | 0.7500 | 0.7100 | 0.4325 | 0.5773 | 0.3060 | 0.8500 | 0.7725 | 0.5825 |
| vs. VILA1.5-8B | +31.47% | +375.38% | +101.16% | +91.03% | +189.69% | +32.18% | +203.17% | +66.85% | +55.19% | +67.96% | +26.19% | +155.00% | +70.57% | +105.32% | +130.69% |
| vs. Best Baseline | +3.48% | +72.63% | +48.70% | +61.53% | +132.23% | -0.73% | +15.42% | +17.65% | +17.36% | +33.08% | -7.37% | +20.47% | +24.54% | +43.06% | +77.86% |

ries models on 2 tasks. These results demonstrate the validity and usability of our *UBench*. Our *UrbanLLaVA* shows marked improvements over all baselines across all tasks in *UBench*. Against the best baselines, *UrbanLLaVA* achieves performance gains ranging from 3.48% to 132.23% for each task. When compared to the base model VILA1.5-8B, the minimum increase is 31.47% on the GeoQA task, while the maximum reaches an impressive 375.38% on the Geo+Traj task. These results highlight the effectiveness of the proposed multi modal dataset, *UData*, which successfully equips smaller MLLMs with a variety of capabilities within urban space, achieving superior performance over all advanced general MLLMs.

The results in New York and London are similar to those in Beijing. Out of 5 tasks, *UrbanLLaVA*@London and *UrbanLLaVA*@NewYork both perform best in 4 tasks. However, in GeoQA task, their performance is slightly inferior to GPT4o, with reductions of -0.73% and -7.37%, respectively. For *UrbanLLaVA*’s performance falling short of expectations on certain task, we speculate two possible reasons: first, the quality of relevant data in the two cities may be lower than that in Beijing, preventing the model from acquiring urban capabilities through learning in the training stage; second, the base model VILA1.5-8B may have comparatively weaker capabilities than commercial API GPT4o, e.g., for the GeoQA task, *UrbanLLaVA*@London outperforms VILA1.5-8B by 32.18% but falls short of GPT4o by 0.73%.

Overall, the proposed *UrbanLLaVA* successfully enhance the performance of small MLLMs on diverse urban tasks.

3.3. Effects of Training Strategies

Since *UrbanLLaVA* is trained with multi-modal urban instruction data, we conduct various experiments to explore a

stable and well-performing training strategies. Due to the limitation of space, we only report the multi-stage results here, more results on learning rate, modality and trained components can refer the supplementary material.

We divide the training dataset into two categories: basic knowledge data and task format aligned data, aiming to develop a training pipeline that enables the model to perform stably and effectively on diverse urban tasks. As Figure 5a shows, ‘Three stage: TA→K→Mix’ performs best in most tasks and maintains reliable performance, surpassing the default tuning method for MLLMs. We also probe the effects of the order between *knowledge learning* and *task alignment* in Figure 5b and Figure 5c. We find that in two-stage training, K→TA slightly outperforms TA→K. However, when the third *mixture learning* is added in a two-stage training model, having *task alignment* first achieves better results, surpassing the models in two-stage training. We hypothesize that this is because in the two-stage training, the model first learns the foundational knowledge and then learns how to solve specific tasks. In the three-stage training, the two-stage model that *knowledge learning* first and then *task alignment* already possesses considerable capabilities, so the impact of mixture training is not significant. However, for two-stage model that completes *task alignment* before *knowledge learning*, *mixture learning* can enhance its abilities, allowing it to recall how to solve urban tasks learned previously.

On the whole, the proposed three-stage training pipeline *UTrain* integrates cross-modal data to achieve stable training and balanced performance across various urban tasks.

3.4. Model Generalization Study

Here, we show that *UrbanLLaVA* can be generalized to different data distributions and tasks, which are crucial for

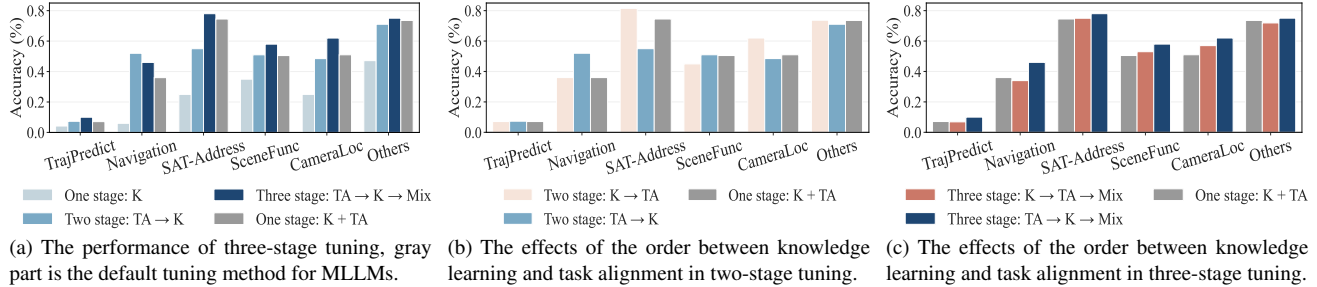


Figure 5. Performance of different training strategies. ‘K’ refers to *knowledge learning*, ‘TA’ refers to *task alignment*, and ‘Mix’ refers to *mixture learning*. ‘One stage: K + TA’ means *knowledge learning* and *task alignment* are merged in the same stage. ‘Two stage: TA→K’ means *task alignment* first then *knowledge learning* in the second stage. ‘Three stage: TA→K→Mix’ adds a step in the third stage: *mixture learning*. The tasks detailed in the table are those with significant differences across different training strategies, while ‘Others’ refers to other tasks in *UBench* with smaller differences.

general urban intelligence. As Table 3 shows, while our *UrbanLLaVA* performs well in diverse urban tasks, it also maintains the original stability in general scenarios, including LLaVA-Bench [30], RealWorldQA [44], and MM-Vet [50]. The results demonstrate that *UrbanLLaVA* is competitive in the dimension of various daily-life visual tasks, real-world spatial understanding and integrated capabilities which is the base for general urban intelligence.

Table 3. General benchmark results. Rating Score refers to result from the LLM-as-a-judge method with GPT4o. For LLaVA-Bench, scores range from 0 to 100, for MM-Vet, scores range from 0.0 to 1.0. Higher scores indicate better performance.

| Test@General | LLaVA-Bench (In-the-Wild) | RealWorldQA | MM-Vet |
|--------------|---------------------------|-------------|--------------|
| Metric | Rating Score | ACC | Rating Score |
| VILA1.5-8B | 60.75 | 0.3765 | 0.3518 |
| Ours-8B | 58.95 | 0.4052 | 0.3239 |

Different cities exhibit various natural and artificial features. Thus, the transferability of urban model is important for its application. As Figure 6 shows, apart from in-domain capabilities empowering and performance improvement after learning, *UrbanLLaVA* can generalize to out-of-domain tasks in various cities. Here, we examined our model trained in the Beijing training set and it exhibits competitive capabilities when tested on London and New York benchmarks. We can see from Figure 6, performance improvements are observed across all tasks in London and New York. Notably, for challenging aspects such as trajectory and regional tasks, the enhancements are significant, indicating the presence of similarity structures across cities that go beyond simple, naive differences.

3.5. Data Ablation Study

Here, we investigate the influences of different data compositions, with results shown in Table 4. As outlined in Sec-

tion 2.1, the urban instruction data is divided into three different subsets: *local view*, *trajectory view*, and *global view*. We remove each subset individually and observe the resulting performance differences. **Local view:** It consists of textual urban geography denoted as CityQA and street view related data denoted as STV. *Local view* data is important for different tasks requiring intelligence about a local part of cities. Noticeable deterioration is observed in both single-modal and multi-modal tasks, indicating the importance of locality knowledge for overall urban understanding. **Trajectory view:** It describes the knowledge about continuous spaces in urban areas. It contains text-trajectory (random walk routing and real-world trajectories) and visual-trajectory (visual-language navigation instructions and random walk with visual input). Both text and multi-modal trajectory view datasets are essential for navigation task. It is also shown that trajectory view data is helpful for different tasks like SceneFunc and GeoQA. **Global view:** It includes a subset of single satellite images that focus on urban knowledge from a specific area, as well as a subset of multiple satellite images that highlight the correlations between different regions and cross alignment between satellite and street view images. Results show that they are essential to empower MLLM to handle urban tasks from a global view, e.g. ImgRetrieval and CameraLoc, while local capabilities are already competitive.

4. Related Work

4.1. Multi-modal Large Language Model

Since the success of GPT4-V [39], MLLM [49] have become a major area of focus in research community, exemplified by the development of models like the LLaVA [30, 31], VILA [29], QwenVL [41] and InternVL [7, 8]. One of the most promising solution to develop advanced MLLM is constructing diverse and high-quality instruction dataset. For example, LLaVA [31] use GPT4-V to create visual in-

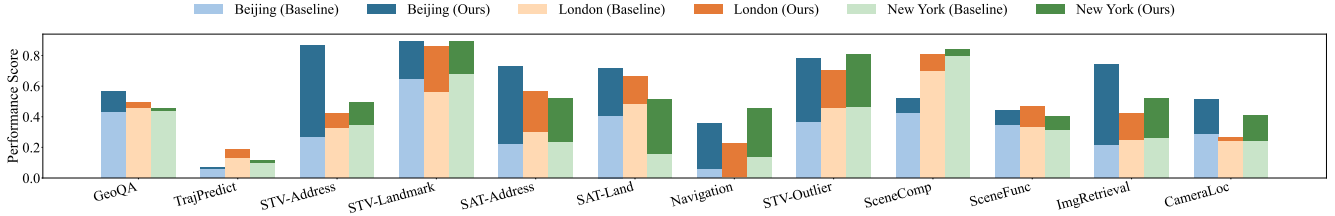


Figure 6. Learning from one city (Beijing) can be directly generalized to other cities (London and New York). In this figure, Baseline is VILA1.5-8b, and our *UrbanLLaVA* is only trained with the urban instruction data from Beijing.

Table 4. Ablation results on different urban instruction data compositions. The arrows indicate corresponding comparison with ours. Only significant differences are denoted. For TrajPredict task, the threshold is 1%, for other tasks, the threshold is 5%. All models are trained using the one-stage strategy to optimize experimental efficiency.

| Task | Data View | GeoQA | TrajPredict | Address | Landmark | Address | LandUse | Navigation | STV-Outlier | SceneComp | SceneFunc | ImgRetrieval | CameraLoc |
|-------------------|------------|---------------|-----------------|---------------|---------------|---------------|---------------|--------------|---------------|---------------|---------------|--------------|-----------|
| Metric | | Avg. Acc | Acc@1 | Acc | Acc | Acc | Acc | Success Rate | Acc | Acc | Acc | Acc | Acc |
| Ours | - | 0.5741 | 0.0711 | 0.8550 | 0.8750 | 0.7450 | 0.7850 | 0.3600 | 0.7800 | 0.5500 | 0.5050 | 0.7300 | 0.5100 |
| w/o CityQA | Local | 0.5409 | 0.0822 ↑ | 0.8700 | 0.8900 | 0.7150 | 0.6950 ↓ | 0.4000 | 0.8050 | 0.5400 | 0.5200 | 0.7750 | 0.5200 |
| w/o STV | Local | 0.5192 ↓ | 0.0622 | 0.4300 ↓ | 0.7300 ↓ | 0.4700 ↓ | 0.7200 ↓ | 0.4200 ↑ | 0.6700 ↓ | 0.4900 ↓ | 0.4550 ↓ | 0.6250 ↓ | 0.4250 ↓ |
| w/o Traj-Text&Nav | Trajectory | 0.4769 ↓ | 0.0644 | 0.8100 | 0.8800 | 0.6350 ↓ | 0.7050 ↓ | 0.0000 ↓ | 0.7600 | 0.4950 ↓ | 0.4300 ↓ | 0.6800 ↓ | 0.4600 ↓ |
| w/o Traj-Vision | Trajectory | 0.5590 | 0.0690 | 0.8350 | 0.9050 | 0.7300 | 0.7100 ↓ | 0.3000 ↓ | 0.8000 | 0.5150 | 0.4650 | 0.7150 | 0.4950 |
| w/o SAT-Single | Global | 0.5345 | 0.0778 | 0.8600 | 0.9100 | 0.5550 ↓ | 0.4550 ↓ | 0.3800 | 0.7800 | 0.5150 | 0.4100 ↓ | 0.7200 | 0.4800 |
| w/o SAT-Multi | Global | 0.5420 | 0.0778 | 0.8500 | 0.8700 | 0.6200 ↓ | 0.6800 ↓ | 0.3400 | 0.6450 ↓ | 0.3500 ↓ | 0.3400 ↓ | 0.3950 ↓ | 0.2600 ↓ |

instruction tuning data, leading to the training of the first open source MLLM. Following LLaVA, VILA [29] explore the effects of training pipelines and data formats during the pre-training stage. ShareGPT4v [6] further expand data scale by developing a superb caption model trained on high-quality caption data from GPT4-V. While general MLLM demonstrate strong visual understanding and reasoning capabilities [9, 21, 43, 47] in common scenarios, they often face challenges in many specialized fields such as medical applications and remote sensing tasks. Thus, domain-specific multi-modal large language models [42] are proposed, such as, Dolphins [34] for autonomous driving, GeoChat [26] for remote sensing tasks, and various models for medical application [23]. In this paper, we propose the first MLLM for urban intelligence which can handle various data and diverse tasks in urban field.

4.2. Multi-modal Model for Urban Study

Urban research is an interdisciplinary field that exists multi-modal data sources [10, 17, 35, 53, 57], including structured geospatial data [2], spatiotemporal series data [57], remote sensing data [35, 55] and street view data [3, 14, 54]. Inspired by the recent advances of MLLMs, researchers explore their potential in urban studies. For structured geospatial data, Balsebre et al. [1] and Feng et al. [16] propose various methods to convert structured geospatial data into a language-compatible format to enhance the geospatial knowledge of large language models. For remote sensing data [24, 33, 37, 51], Kuckreja et al. [26] and Zhang et al. [52] design various remote sensing instruction data to

fine-tune general MLLMs for various downstream remote sensing tasks. For street view data, Hao et al. [22] fine-tune CLIP model for improved urban indicator prediction by integrating street view data and remote sensing data. Liu et al. [32] evaluate the potential of multi-modal language model for urban socioeconomic sensing. For spatiotemporal series data, Li et al. [28] and Gong et al. [20] introduce domain-specific encoders to enhance LLM capabilities for spatiotemporal series modeling. Feng et al. [15] propose agentic framework to unleash the power of LLM for zero-shot mobility prediction. Unlike these works that focus on limited data types and specific tasks, our method is designed to process all these data types and address a wide range of urban tasks.

5. Conclusion

In this paper, we propose *UrbanLLaVA*, a MLLM with enhanced urban spatial cognition by integrating four types of urban data and supporting a wide range of urban tasks. Our approach begin with the development of diverse and high-quality urban instruction data, spanning from local view to global view of urban environment. We then design a three-stage training pipeline to ensure the stable training and improved performance of model on diverse urban tasks. Experimental results from three cities on an extended urban benchmark highlight the effectiveness of *UrbanLLaVA* for integrating multi-modal urban data and solving urban tasks. In summary, *UrbanLLaVA* sheds lights for building the unified foundation model with powerful perception and reasoning abilities for general urban intelligence.

6. Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under grant 2024YFC3307603, in part by the National Natural Science Foundation of China under grant U23B2030, in part by the China Postdoctoral Science Foundation under grant 2024M761670 and GZB20240384, in part by the Tsinghua University Shuimu Scholar Program under grant 2023SM235. This work was also sponsored by Tsinghua University-Toyota Research Center.

References

- [1] Pasquale Balsebre, Weiming Huang, and Gao Cong. Lamp: A language model on the map. [arXiv preprint arXiv:2403.09059](#), 2024. 1, 3, 8
- [2] Pasquale Balsebre, Weiming Huang, Gao Cong, and Yi Li. City foundation models for learning general purpose representations from openstreetmap. In [Proceedings of the 33rd ACM International Conference on Information and Knowledge Management](#), pages 87–97, 2024. 1, 8
- [3] Filip Biljecki and Koichi Ito. Street view imagery in urban analytics and gis: A review. [Landscape and Urban Planning](#), 215:104217, 2021. 8
- [4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. [arXiv preprint arXiv:2307.15818](#), 2023. 2
- [5] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snaveley, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 12538–12547, 2019. 4
- [6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. [arXiv preprint arXiv:2311.12793](#), 2023. 3, 8
- [7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. [arXiv preprint arXiv:2404.16821](#), 2024. 5, 7
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 24185–24198, 2024. 5, 7
- [9] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. [arXiv preprint arXiv:2406.01584](#), 2024. 8
- [10] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. [ACM Computing Surveys](#), 2024. 8
- [11] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. [arXiv preprint arXiv:2305.14233](#), 2023. 5
- [12] Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. [arXiv preprint arXiv:2310.05492](#), 2023. 4
- [13] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024. 5
- [14] Zhuangyuan Fan, Fan Zhang, Becky PY Loo, and Carlo Ratti. Urban visual intelligence: Uncovering hidden city profiles with street view images. [Proceedings of the National Academy of Sciences](#), 120(27):e2220417120, 2023. 1, 8
- [15] Jie Feng, Yuwei Du, Jie Zhao, and Yong Li. Agentmove: A large language model based agentic framework for zero-shot next location prediction. In [Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies \(Volume 1: Long Papers\)](#), pages 1322–1338, 2025. 8
- [16] Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. Citygpt: Empowering urban spatial cognition of large language models. In [Proceedings of the 31th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining](#), 2025. 2, 3, 8
- [17] Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, et al. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. [arXiv preprint arXiv:2504.09848](#), 2025. 8
- [18] Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. Citybench: Evaluating the capabilities of large language models for urban tasks. In [Proceedings of the 31th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining](#), 2025. 5
- [19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 15180–15190, 2023. 1
- [20] Letian Gong, Yan Lin, Xinyue Zhang, Yiwen Lu, Xuedi Han, Yichen Liu, Shengnan Guo, Youfang Lin, and Huaiyu Wan. Mobility-llm: Learning visiting intentions and travel preferences from human mobility data with large language models. [arXiv preprint arXiv:2411.00823](#), 2024. 1, 2, 8

- [21] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024. 8
- [22] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. Urbanvlp: A multi-granularity vision-language pre-trained foundation model for urban indicator prediction. *arXiv preprint arXiv:2403.16831*, 2024. 8
- [23] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review, 2024. 8
- [24] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023. 8
- [25] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuanfang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023. 2
- [26] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 2, 8
- [27] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [28] Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5351–5362, 2024. 1, 8
- [29] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 4, 5, 7, 8
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 5, 7
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 7
- [32] Tianhui Liu, Jie Feng, Hetian Pang, Xin Zhang, Tianjian Ouyang, Zhiyuan Zhang, and Yong Li. Citylens: Benchmarking large language-vision models for urban socio-economic sensing. *arXiv preprint arXiv:2506.00530*, 2025. 8
- [33] Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024. 8
- [34] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. *arXiv preprint arXiv:2312.00438*, 2023. 2, 8
- [35] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial (vision paper). *ACM Transactions on Spatial Algorithms and Systems*, 2024. 1, 8
- [36] Meta AI. LLaMA 3.2: Advancing Vision, Edge, and Mobile Devices. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, 2024. Accessed: 2024-11-01. 5
- [37] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. *arXiv preprint arXiv:2402.02544*, 2024. 8
- [38] OpenAI. Introducing chatgpt. <https://openai.com/blog/chatgpt/>, 2022. 1
- [39] OpenAI. Gpt-4v(ision) system card. 2023. 7
- [40] OpenAI. Hello GPT-4. <https://openai.com/index/hello-gpt-4o/>, 2024. 5
- [41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5, 7
- [42] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Wenhai Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint arXiv:2406.08394*, 2024. 8
- [43] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 8
- [44] XAI Organization. RealworldQA Dataset. <https://huggingface.co/datasets/xai-org/RealworldQA>, 2024. Accessed: 2024-10-01. 5, 7
- [45] Zhaomin Xiao, Eduardo Blanco, and Yan Huang. Analyzing large language models’ capability in location prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 951–958, 2024. 1
- [46] Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*, 2023. 1
- [47] Fengli Xu, Qian Yue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025. 8

- [48] Dingqi Yang, Daqing Zhang, and Bingqing Qu. Participatory cultural mapping based on collective behavior data in location-based social networks. ACM Transactions on Intelligent Systems and Technology (TIST), 7(3):1–23, 2016. 3
- [49] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549, 2023. 1, 3, 7
- [50] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In International conference on machine learning. PMLR, 2024. 5, 7
- [51] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeeyegt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. arXiv preprint arXiv:2401.09712, 2024. 8
- [52] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing, 2024. 8
- [53] Weijia Zhang, Jindong Han, Zhao Xu, Hang Ni, Hao Liu, and Hui Xiong. Urban foundation models: A survey. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 6633–6643, 2024. 1, 2, 3, 8
- [54] Xin Zhang, Tianjian Ouyang, Yu Shang, Qingmin Liao, and Yong Li. UrbanMLLM: Joint learning of cross-view imagery for urban understanding, 2025. 8
- [55] Yunke Zhang, Ruolong Ma, Xin Zhang, and Yong Li. Perceiving urban inequality from imagery using visual language models with chain-of-thought reasoning. In Proceedings of the ACM on Web Conference 2025, pages 5342–5351, 2025. 8
- [56] Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 10707–10715, 2025. 5
- [57] Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. Information Fusion, 113:102606, 2025. 1, 2, 3, 8