

UrbanWell: Benchmarking Multimodal Large Language Models for Spatio-Temporal Urban Wellbeing Analytics

Yanxin Xi
Department of Computer Science,
University of Helsinki
Helsinki, Finland
yanxin.xi@helsinki.fi

Xiang Su*
Department of Computer Science,
Department of Agricultural Sciences,
University of Helsinki
Helsinki, Finland
xiang.su@helsinki.fi

Jie Feng*
Zhongguancun Academy
Beijing, China
fengjie@bza.edu.cn

Yu Liu
Institute of Biomedical Engineering,
Department of Engineering Science,
University of Oxford
Oxford, United Kingdom
yu.liu@eng.ox.ac.uk

Sasu Tarkoma
Department of Computer Science,
University of Helsinki
Helsinki, Finland
sasutarkoma@helsinki.fi

Pan Hui*
Computational Media and Arts
Thrust, Hong Kong University of
Science and Technology (Guangzhou)
Guangzhou, China
Department of Computer Science,
University of Helsinki
Helsinki, Finland
panhui@ust.hk

Abstract

Understanding urban wellbeing from multimodal data requires integrating heterogeneous spatial and temporal signals, posing significant challenges for current multimodal large language models (MLLMs). We introduce UrbanWell, a large-scale benchmark designed to systematically evaluate the spatio-temporal reasoning capabilities of MLLMs for urban wellbeing analytics through joint modeling of satellite and street view imagery. UrbanWell spans 38 cities across multiple years and includes diverse indicators covering (1) environmental conditions (CO₂, NO₂, PM_{2.5}, and Normalized Difference Vegetation Index), (2) spatial accessibility (minimum distance to supermarkets and restaurants), (3) urban form (road length, road density, and land use), (4) urban vitality (population, economic activity diversity, and land use diversity), and (5) subjective perception attributes (e.g., safety, beauty, liveliness, wealth, and quietness). All indicators are aligned at grid level to enable standardized evaluation. Beyond static prediction, UrbanWell defines temporal reasoning tasks, including future value forecasting from historical observations and temporal trend classification. We benchmark 15 state-of-the-art representative MLLMs in a zero-shot setting, providing a comprehensive comparative evaluation across spatial and temporal dimensions. Experimental results indicate that while MLLMs capture salient spatial and perceptual cues, their performance varies substantially across heterogeneous urban indicators spanning environment and subjective perception. UrbanWell serves as a unified benchmark for evaluating multimodal spatial and temporal reasoning in urban wellbeing analytics, offering a

standardized testbed for systematic assessment and future research on multimodal urban intelligence. Our codes and datasets are accessible via <https://github.com/axin1301/UrbanWell-Benchmark>.

CCS Concepts

• **Information systems** → **Data mining**; **Spatial-temporal systems**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Urban Wellbeing; Multimodal Large Language Models; Spatio-Temporal Analytics; Benchmarking

ACM Reference Format:

Yanxin Xi, Xiang Su, Jie Feng, Yu Liu, Sasu Tarkoma, and Pan Hui. 2026. UrbanWell: Benchmarking Multimodal Large Language Models for Spatio-Temporal Urban Wellbeing Analytics. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26)*, August 09–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3770855.3817584>

1 Introduction

Urban areas accommodate over 57% of the global population [55], making cities a major source of large-scale heterogeneous data that reflect environmental, spatial, and socioeconomic dynamics. Understanding multidimensional urban wellbeing requires the integration of diverse signals across space and time, posing substantial challenges for data-driven modeling. Urban wellbeing analytics aims to systematically quantify environmental [6], spatial [58], socioeconomic [17], and perceptual indicators [63] across neighborhoods, providing a computational perspective on how urban systems function and evolve.

Recent advances in computer vision and urban sensing have shown that deep learning models can infer urban indicators from satellite and street view imagery [17, 34, 37, 57]. These approaches

*Corresponding authors.



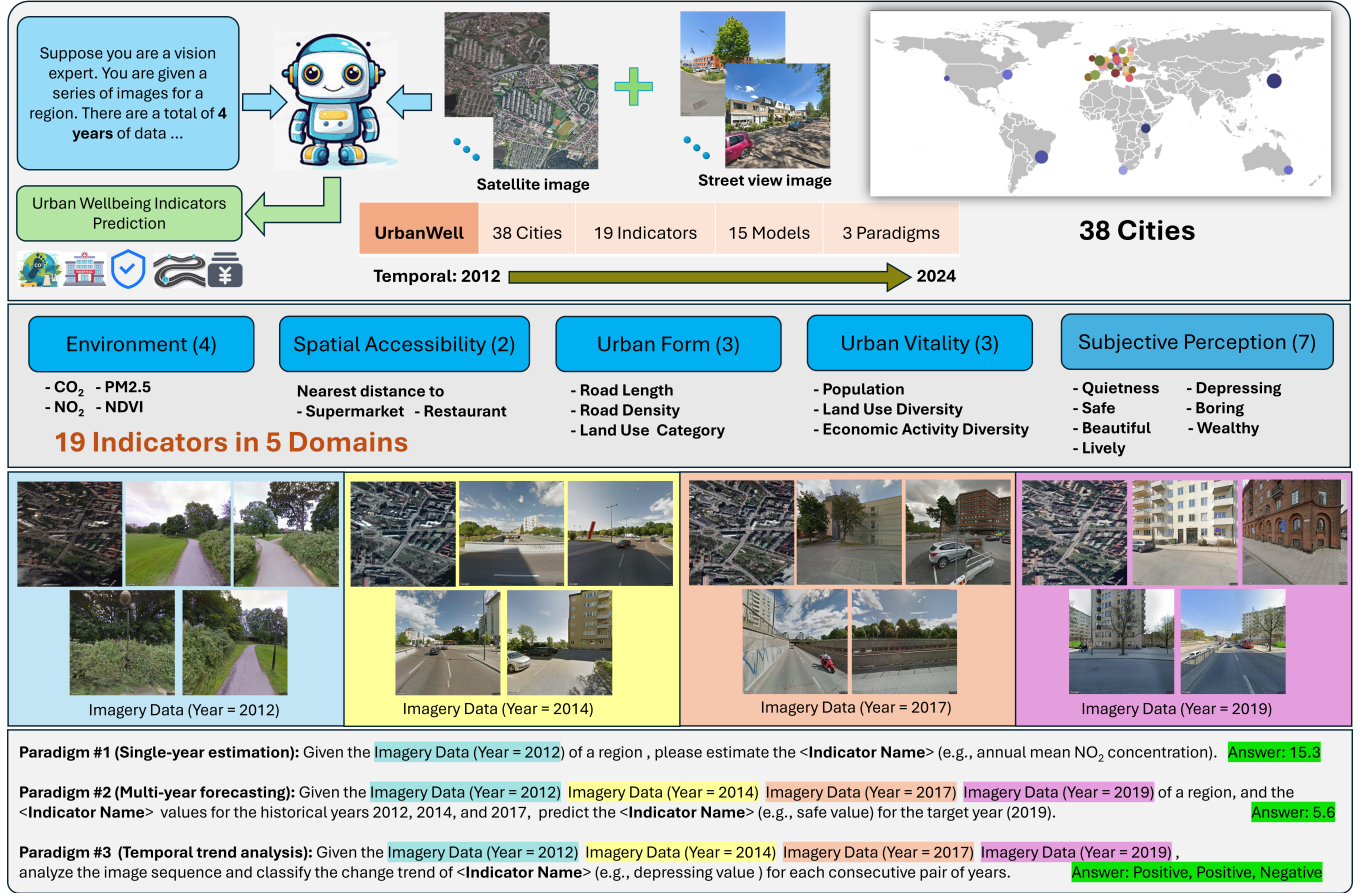


Figure 1: Overview of UrbanWell. UrbanWell integrates satellite and street view imagery with 19 ground-truth urban wellbeing indicators across 38 cities spanning 2012–2024, defining three task paradigms for evaluating the urban intelligence capabilities of 15 MLLMs.

estimate attributes such as land use [4], inequality [65], and socioeconomic conditions [31, 32, 49, 54]. However, existing studies typically focus on isolated tasks, region-specific datasets, or single-year prediction settings, limiting systematic comparison across methods and restricting reproducibility. Moreover, temporal dynamics are rarely incorporated, despite being essential for modeling long-term urban change. The absence of standardized, multimodal benchmarks makes it difficult to comprehensively and quantitatively evaluate model capabilities in spatio-temporal urban reasoning and to assess generalization across cities.

The recent emergence of Multimodal Large Language Models (MLLMs) has introduced a new paradigm for integrating visual and textual information [24, 46, 50]. By jointly modeling images and language, MLLMs enable flexible multimodal reasoning and have shown strong performance across diverse perception and understanding tasks [38, 39]. However, the capabilities of MLLMs in urban analytics remain underexplored. Existing studies typically evaluate models on limited tasks or single-modality datasets, without systematically assessing cross-city generalization, multi-indicator prediction, or temporal reasoning. Moreover, current urban visual benchmarks are constrained in modality coverage and temporal

depth. As summarized in Table 1, most of the datasets lack integration of environmental and air quality measures, perceptual attributes, and longitudinal observations, limiting a comprehensive evaluation across spatial and temporal dimensions. These limitations highlight the need for a standardized benchmark that enables systematic assessment of MLLMs for urban wellbeing analytics.

Table 1: Comparison with existing urban visual dataset. ‘SAT’: satellite imagery, ‘STV’: street view imagery, ‘TMP’: temporal, ‘EAQ’: environment air quality, ‘PER’: perception, ‘SC’: spatial coverage.

dataset	SAT	STV	TMP	EAQ	PER	SC (#cities)
CityBench [20]	✓	✓				13
CityLens [36]	✓	✓				17
UBench [19]	✓	✓				3
UrBench [68]	✓	✓				-
UrbanFeel [27]		✓	✓		✓	11
UrbanWell	✓	✓	✓	✓	✓	38

To address these limitations, we introduce UrbanWell, a large-scale multimodal benchmark for evaluating the spatio-temporal reasoning capabilities of MLLMs in urban wellbeing analytics. As illustrated in Figure 1, UrbanWell integrates environmental, spatial accessibility, urban form, urban vitality, and subjective perception indicators across 38 cities over multiple years. The dataset aligns heterogeneous signals at grid level and provides a standardized evaluation framework. Beyond conventional single-year prediction tasks adopted in prior benchmarks [19, 20, 36, 68], UrbanWell defines two additional temporal reasoning tasks, including (1) forecasting future indicator values from historical observations and (2) classifying temporal trends. These tasks explicitly probe the ability of models to perform longitudinal reasoning from multimodal inputs, extending evaluation beyond static estimation settings. Through systematic benchmarking of 15 state-of-the-art MLLMs under a unified protocol, UrbanWell provides a comprehensive assessment of their capacity to model multidimensional spatio-temporal urban signals. UrbanWell is constructed from publicly accessible data sources and released with standardized annotations and task definitions to support comparison across models. Our contributions are threefold:

- We present UrbanWell, a multimodal benchmark covering 38 cities spanning 2012–2024 and integrating 19 indicators across five domains. It provides a standardized evaluation framework for spatio-temporal understanding of urban wellbeing.
- We formalize single-year estimation, multi-year forecasting, and temporal trend classification tasks to systematically evaluate MLLMs’ spatio-temporal reasoning ability, and benchmark 15 state-of-the-art MLLMs under a unified zero-shot setting.
- We provide a comprehensive empirical analysis of modality contributions and task design, identifying strengths and limitations of current MLLMs in modeling heterogeneous urban indicators.

2 Dataset Construction

UrbanWell is designed as a standardized benchmark to evaluate the spatio-temporal reasoning capabilities of MLLMs for multimodal urban analytics. As illustrated in Figure 1, the benchmark integrates 19 heterogeneous indicators across five domains, spanning 38 cities over the period 2012–2024. All indicators are spatially aligned at grid level and temporally synchronized to enable controlled evaluation across static estimation, temporal forecasting, and trend classification settings. The following sections describe the data construction process in detail.

2.1 Data Collection

The indicator design in UrbanWell follows prior urban analytics literature, with the goal of constructing a heterogeneous and multimodal benchmark rather than focusing on a single predictive task. UrbanWell includes 19 indicators across five domains.

Environment (Carbon Dioxide (CO₂), Nitrogen Dioxide (NO₂), Particulate Matter 2.5 (PM_{2.5}), Normalized Difference Vegetation Index (NDVI) [51]) provides atmospheric and vegetation-related signals commonly used in environmental urban studies [6]. **Spatial accessibility** (minimum distance to restaurants (RTR) and supermarkets (SPM)) encodes service proximity derived from geospatial data, following accessibility modeling approaches [21]. **Urban**

form (road length (RL), road density (RD), land use category (LUC)) captures structural morphology extracted from OpenStreetMap (OSM) [25] and land use datasets, consistent with interpretable urban analysis [3]. **Urban vitality** (population (Popu), land use diversity (LUD), economic activity diversity (EcD)) measures functional intensity and diversity across urban grids [47]. **Subjective perception** (safe, beautiful (BEA), lively (LIV), boring (BOR), depressing (DEP), and wealthy (WEA) [14], quietness suitability index (QSI) [15]) represents perceptual attributes inferred from urban imagery [27, 29].

We construct UrbanWell using data from 38 cities, primarily located in Europe and complemented by representative metropolitan areas from Asia, North America, South America, Africa, and Oceania. The emphasis on European cities is due to the availability of long-term, standardized, and high-coverage open datasets across environmental, geospatial, and socioeconomic domains, which enable consistent temporal alignment from 2012 to 2024. Cities from other continents are included to introduce distributional diversity and support evaluation under heterogeneous urban conditions.

The list of cities is provided in Appendix A.1. Satellite imagery is obtained from Google Earth [22], and street view imagery is collected via the Google Street View API [23]. Environmental indicators include CO₂ from ODIAC [43, 44], NO₂ and PM_{2.5} from the European Environment Agency (EEA) [16], and NDVI from Copernicus [13]. Urban form and accessibility data include urban core boundaries and land use from Urban Atlas [11, 12], as well as road networks, facility distances, and economic POIs from OSM [25]. Population data are obtained from WorldPop [56]. Subjective perception indicators (excluding QSI) are computed using pretrained models applied to large-scale street view imagery, while QSI is obtained from EEA [16]. All data sources are temporally aligned to the same yearly granularity to ensure consistency across tasks.

2.2 Data Processing

Figure 2 illustrates the processing pipeline for harmonizing multi-source data. For each city, we first define the urban core region to ensure spatial consistency across cities. Within this region, satellite imagery, street view imagery, and indicator data are collected for the period 2012–2024. Each satellite image is treated as a basic spatial unit. Given a spatial resolution of approximately 5 meters per pixel and an image size of 256×256 , each spatial unit covers roughly 1.28×1.28 km, corresponding to an area of approximately 1.6 km^2 . This neighborhood-scale granularity balances spatial detail and computational tractability for large-scale benchmarking. All indicators are aggregated to and aligned with this grid to ensure consistent spatial granularity across modalities. For each spatial unit, we extract the geographic boundaries and perform spatial matching with street view imagery using coordinate-based queries. This procedure links aerial and ground-level observations within the same grid cell, enabling multimodal alignment at a unified spatial scale.

To ensure image quality, we apply an automatic filtering strategy to street view imagery. A pretrained semantic segmentation model [66, 69] estimates the proportion of key outdoor categories (e.g., sky, vegetation, building, road) in each image. Images with low outdoor coverage or excessive indoor content are identified

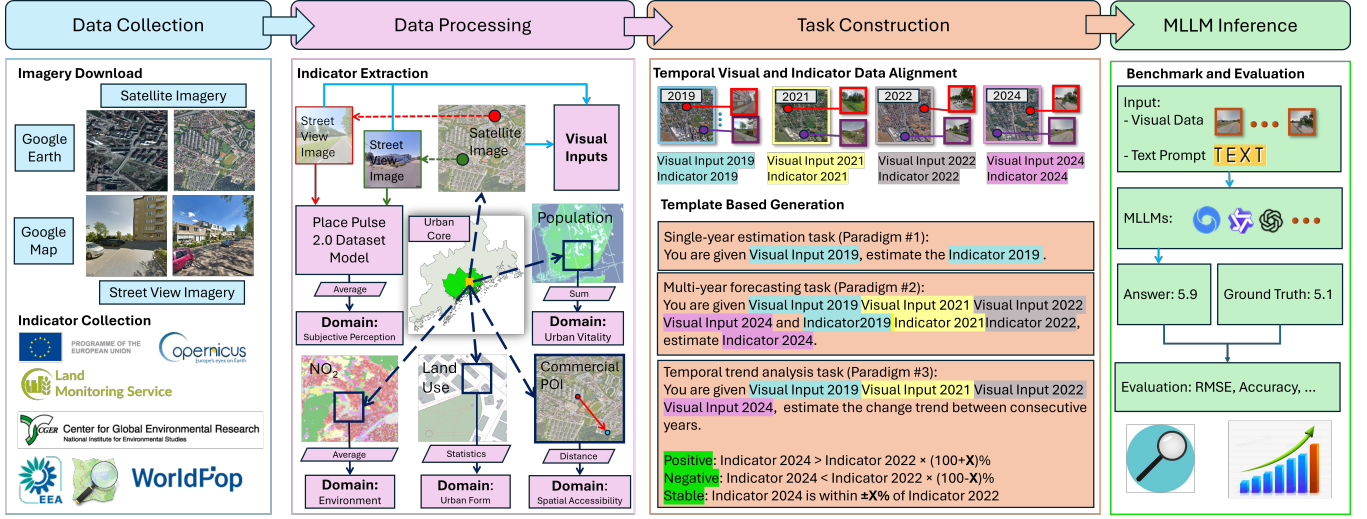


Figure 2: Overview of the UrbanWell benchmark construction pipeline, including data collection, data processing, task construction, and MLLM evaluation.

through threshold-based filtering and subsequently verified manually. Indoor or severely occluded scenes are removed to retain representative outdoor samples. The entire processing pipeline is applied uniformly across all cities and years to guarantee consistency. Next, indicators from different domains are aligned to the common spatial grid.

Environment. For each spatial unit and year, raster-based environmental indicators (CO_2 , NO_2 , $\text{PM}_{2.5}$, NDVI) are spatially clipped using the grid boundary coordinates, after which the mean pixel value within the clipped region is computed to produce standardized grid-level environmental measurements.

Spatial Accessibility. For essential facility accessibility, we consider two POI categories: supermarket and restaurant. For each category, we compute the Euclidean distance between the centroid of each spatial unit and the nearest corresponding POI. The minimum distance to the nearest POI is used as the indicator value.

Urban Form. Annual OSM road network data are clipped to each spatial unit to calculate total RL. RD is computed by normalizing RL by the grid area. LUC within each grid is assigned as the land use indicator for that year.

Urban Vitality. LUD is computed using an entropy-based measure [47] over land use categories within each spatial unit, including (i) residential areas, (ii) commercial, industrial, institutional, and governmental areas, and (iii) recreational spaces, parks, and water bodies. EcD is calculated using Shannon entropy [48] over commercial POI categories (e.g., shops, restaurants, hotels) extracted from OSM within each spatial unit. The formulations of LUD and EcD are provided in Appendix A.2. Popu is obtained by mapping the population to each grid and summing the values within the region to derive total population per spatial unit and year.

Subjective Perception. Pretrained models on the Place Pulse 2.0 dataset [14, 29] are applied to street view images located in each spatial unit. For each image, the models output perception scores including safe, BEA, LIV, BOR, DEP, and WEA. Grid-level

perception indicators are computed as the mean score across all matched street view images within the spatial unit. QSI is obtained directly from the source dataset and matched to the corresponding spatial units.

2.3 Task Construction

The construction of benchmark tasks from the processed imagery and indicator data is illustrated in Figure 2. To systematically evaluate spatial-temporal reasoning in MLLMs, we define three task paradigms, i.e., *Single-year Estimation*, *Multi-year Forecasting*, and *Temporal Trend Analysis*.

Single-year Estimation Task. Given satellite and street view images of a spatial unit from the same year, the model predicts predefined urban indicators for that year. This task evaluates cross-modal spatial reasoning in a temporally static setting.

Multi-year Forecasting Task. Given satellite and street view images from multiple consecutive years for a spatial unit, together with historical indicator values, the model predicts the indicator value in the final year of the sequence. This task evaluates temporal reasoning and longitudinal pattern modeling from multimodal observations.

Temporal Trend Analysis Task. Given sequential imagery of the same spatial unit across consecutive years, the model classifies whether the indicator exhibits a positive trend, a negative trend, or a stable trend (i.e., no significant change). This task evaluates the model’s ability to detect temporal dynamics rather than estimate absolute values.

Due to computational constraints, we adopt a stratified sampling strategy to construct the benchmark dataset while preserving geographic and temporal coverage. For each indicator, samples are drawn approximately evenly across all 38 cities. Within each city, grid-year instances are sampled from the available years to maintain temporal diversity. This design ensures broad coverage across geographic regions and time periods.

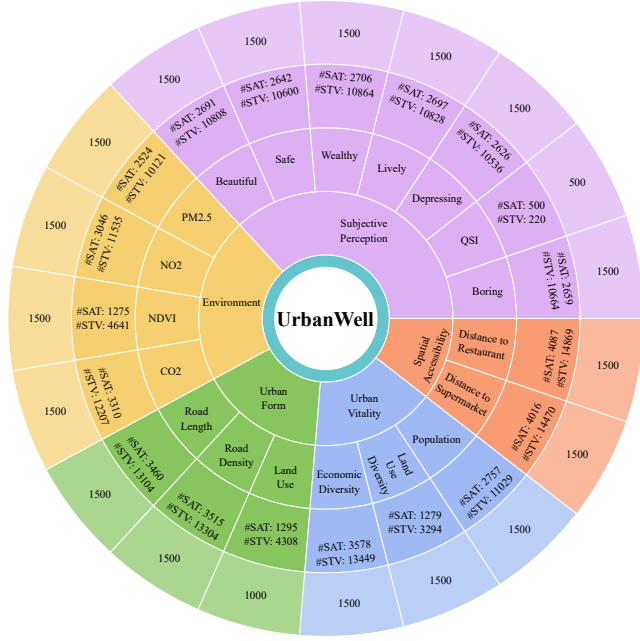


Figure 3: Summary statistics of UrbanWell, showing the numbers of samples, satellite images, and street view images across the 19 urban indicators. Statistics are aggregated over all task paradigms.

Dataset Statistics. Figure 3 summarizes the number of samples, satellite images, and street view images across the 19 indicators in 5 domains. These statistics provide an overview of benchmark scale and modality coverage across indicators.

2.4 MLLM Inference

We evaluate all models under a unified zero-shot setting using task-specific prompt templates. Prompts are constructed according to the three task paradigms and provide the corresponding multimodal inputs, historical observations, and task instructions. For temporal trend analysis, prompts additionally define explicit percentage thresholds for classifying trends as positive, negative, or stable. All prompts follow fixed templates with placeholders for years, image sequences, indicator names, units, and output formats. This standardized schema ensures consistent inputs and outputs across model evaluation. Example prompts are provided in Appendix A.3.

Land use prediction is formulated as a multiple-choice classification task. Temporal trend analysis is formulated as a three-class classification problem. All other indicators in the single- and multi-year tasks are treated as regression problems, where models output scalar values. This formulation enables standardized evaluation across heterogeneous urban indicators while maintaining task-specific prediction objectives.

3 Benchmark and Experiments

3.1 Experimental Setup

Model Deployment. We evaluate 15 representative MLLMs, covering both open-source and proprietary systems across different model families and parameter scales. Following [36], the evaluated

models include Gemma3-4B/12B/27B [50], Qwen2.5VL-3B/7B/32B [7], InternVL3-9B [70], InternVL3.5-14B [53], Phi-3.5-vision [1], Phi-4-multimodal [2], Mistral-small-3.1-24B [41], Llama4-Scout [40], Amazon-Nova-Lite [5], Gemini-2.0-Flash [24], and GPT-5-nano [46]. All models are evaluated using identical prompt templates and inference configurations. No task-specific hyperparameter tuning is performed.

Evaluation Metrics. Evaluation metrics are determined according to task type. LUC is evaluated using classification accuracy. Temporal trend analysis is evaluated using three-class accuracy, with accuracy computed over all predicted trend labels. All remaining indicator estimation tasks are evaluated using Root Mean Squared Error (RMSE). RMSE is reported in the original measurement units, and results are presented separately for each indicator due to differences in scale and semantics across tasks.

3.2 Performance on Single-year Estimation Task

Single-year estimation remains challenging and exhibits clear indicator stratification. Table 2 reveals substantial instability in quantitative prediction from static multimodal inputs. Environmental indicators are particularly challenging. Although RMSE values are not directly comparable across indicators due to differences in measurement scales (e.g., CO_2 is reported in monthly carbon tonnes, whereas NO_2 is measured in $\mu\text{g}/\text{m}^3$), CO_2 exhibits extreme cross-model variation, with RMSE spanning multiple orders of magnitude. This suggests substantial difficulty in inferring emission-related quantities from visual observations and points to calibration challenges in mapping visual cues to physically meaningful emission scales. NO_2 and $\text{PM}_{2.5}$ also exhibit noticeable performance variation across models, indicating that air-quality estimation remains challenging even when environmental signals are partially observable. Urban form metrics (RL and RD) yield relatively consistent performance across models, indicating partial alignment between visible spatial structures and numerical measurements. Popu consistently exhibits large prediction errors, highlighting the limited direct correspondence between imagery and aggregated demographics. In contrast, several urban vitality and perception-related attributes (e.g., economic diversity, BEA, and LIV) exhibit comparatively lower cross-model variance, suggesting that some higher-level urban characteristics can be inferred more consistently from visual cues.

Model performance is shaped by both parameter scale and architectural design. Performance differences across MLLMs are substantial. Within several model families (e.g., Gemma and Qwen), larger variants generally achieve lower errors, although improvements are not uniformly observed across all indicators. This suggests that increased model capacity can benefit quantitative estimation. However, cross-family comparisons indicate that architecture and alignment strategy remain equally important. Gemini-2.0-Flash achieves the strongest performance on several environmental indicators and LUC, while GPT-5-nano performs particularly well on urban form indicators, attaining the lowest errors for both RL and RD. Other models exhibit strengths on specific perception and vitality-related indicators. The absence of a universally dominant model suggests that multimodal numerical reasoning

Table 2: Results of single-year estimation tasks. Accuracy is reported for the LUC task, while RMSE is reported for all other indicators. The best result for each indicator is highlighted in bold, and the second-best result is underlined.

Model	Environment				Accessibility		Urban Form			Urban Vitality			Subjective Perception						
	CO ₂	NO ₂	PM2.5	NDVI	RTR	SPM	RD	RL	LUC	Popu	LUD	EcD	QSI	Safe	LIV	WEA	BEA	BOR	DEP
Gemma3-4B	6.2e3	19.34	15.45	0.75	6.94	6.54	17.69	16.76	0.60	1.2e5	0.31	1.15	4.5e2	4.84	3.45	2.80	2.35	8.30	7.32
Gemma3-12B	1.5e4	18.67	3.93	0.87	7.17	7.66	16.63	17.67	0.55	7.4e4	0.43	0.98	4.8e2	1.83	1.66	1.69	2.16	4.16	4.49
Gemma3-27B	2.5e3	12.75	3.68	0.86	6.27	4.27	16.19	13.82	0.59	1.7e4	0.47	0.81	3.7e2	1.99	2.86	2.15	2.64	6.65	7.81
Qwen2.5VL-3B	1.0e14	6.2e2	93.21	0.80	3.6e2	77.71	2.2e2	70.51	0.40	9.6e6	0.52	1.82	4.3e2	7.27	6.45	5.47	6.73	9.64	9.19
Qwen2.5VL-7B	1.1e8	18.83	10.28	0.86	10.87	8.07	16.43	36.40	0.51	7.2e4	0.53	1.48	5.7e2	3.99	4.17	4.38	3.35	7.34	7.48
Qwen2.5VL-32B	5.4e4	9.32	4.24	0.80	47.63	8.47	17.07	16.69	0.48	7.8e4	0.41	1.42	6.2e2	3.13	4.64	5.03	4.46	8.95	7.61
InternVL3-9B	3.8e3	19.78	15.89	0.87	7.64	7.59	16.86	15.11	0.44	7.6e4	0.45	1.41	5.8e2	3.82	5.08	4.25	3.21	7.09	9.44
InternVL3.5-14B	2.4e4	16.65	5.15	0.72	7.30	6.34	15.10	15.41	0.41	2.4e4	0.44	1.23	4.2e2	1.82	2.28	2.98	2.40	3.62	3.90
Phi-3.5	1.4e3	29.96	19.85	0.90	4.77	6.94	15.61	27.36	0.40	1.5e5	0.54	1.43	4.7e2	4.72	4.43	3.86	3.48	8.72	10.89
Phi-4	2.8e3	69.07	63.68	0.91	4.92	5.40	14.51	1.4e2	0.25	2.4e7	0.54	1.02	5.4e2	9.31	7.54	8.19	8.86	10.96	11.16
Mistral-Small-3.1-24B	6.3e7	7.60	4.57	0.71	7.50	6.05	17.65	16.69	0.52	1.8e4	0.42	1.30	5.6e2	2.65	4.55	5.10	3.91	4.42	6.06
Llama4-Scout	6.3e4	8.78	3.92	0.69	2.57	2.33	16.84	17.86	0.07	7.4e4	0.46	0.96	5.1e2	2.62	3.54	4.64	3.98	6.21	6.41
Nova-lite-v1	5.7e6	15.29	6.66	0.78	6.40	5.48	17.07	16.99	0.49	6.9e4	0.47	0.87	6.1e2	2.43	2.09	2.34	2.06	3.86	6.95
Gemini-2.0-Flash	2.2e2	5.50	2.98	0.69	2.61	2.66	16.15	17.27	0.66	2.3e4	0.39	1.19	4.7e2	3.67	3.60	4.29	3.50	7.18	6.74
GPT-5-nano	1.3e4	7.30	4.10	0.80	2.52	3.42	12.79	13.62	0.58	3.1e4	0.57	1.21	5.4e2	2.19	5.44	3.86	2.83	5.12	7.05

Table 3: Results of multi-year forecasting tasks. Accuracy is reported for the LUC task, while RMSE is reported for all other indicators. For each indicator, the best value across all models is highlighted in bold, and the second-best value is underlined.

Model	Environment				Accessibility		Urban Form			Urban Vitality			Subjective Perception					
	CO ₂	NO ₂	PM2.5	NDVI	RTR	SPM	RD	RL	LUC	Popu	LUD	EcD	Safe	LIV	WEA	BEA	BOR	DEP
Gemma3-4B	11.51	3.82	2.71	0.50	0.38	0.35	5.37	5.02	0.11	8.3e2	0.13	0.33	1.30	1.23	1.25	1.49	1.04	1.25
Gemma3-12B	13.33	2.93	2.23	0.58	0.32	0.30	4.91	5.08	0.44	2.5e2	0.07	<u>0.29</u>	1.44	1.42	1.37	1.62	1.03	1.33
Gemma3-27B	13.99	3.06	2.24	0.39	0.32	<u>0.27</u>	4.46	4.80	0.61	2.0e2	0.07	0.28	1.45	1.41	1.37	1.56	1.05	1.34
Qwen2.5VL-3B	27.60	3.49	2.59	0.48	0.40	0.43	5.94	5.42	0.21	3.6e2	0.18	0.42	1.33	1.34	1.26	1.47	0.98	1.23
Qwen2.5VL-7B	14.20	3.67	2.71	0.43	0.34	0.30	4.81	4.98	0.59	3.7e2	0.11	0.30	1.34	<u>1.26</u>	1.27	1.42	1.03	1.27
Qwen2.5VL-32B	12.17	2.80	2.11	0.52	<u>0.30</u>	0.30	4.54	4.79	0.66	2.5e2	0.09	0.30	1.42	1.36	1.33	1.52	1.04	1.31
InternVL3-9B	14.27	2.71	2.09	0.47	0.31	0.28	4.76	5.07	0.45	2.5e2	0.10	0.32	1.48	1.49	1.45	1.65	1.16	1.41
InternVL3.5-14B	13.93	2.79	1.97	0.59	0.40	0.29	4.78	5.38	0.33	2.6e2	0.11	<u>0.29</u>	1.46	1.46	1.43	1.59	1.09	1.35
Phi-3.5	13.12	3.50	2.73	0.60	0.33	0.33	5.00	5.28	0.18	2.8e2	0.08	0.32	1.41	1.34	1.31	1.50	1.03	1.23
Phi-4	14.24	3.40	2.72	0.51	0.34	0.24	5.33	5.24	0.10	3.0e2	0.05	0.30	<u>1.32</u>	1.30	1.23	1.49	1.04	1.26
Mistral-Small-3.1-24B	12.66	<u>2.48</u>	<u>1.96</u>	0.60	0.20	0.24	5.09	<u>4.48</u>	0.58	4.1e2	0.12	0.31	1.34	1.39	1.40	1.55	1.01	1.23
Llama4-Scout	12.10	2.42	1.92	0.35	0.33	0.29	2.56	4.09	0.08	3.1e2	0.21	<u>0.29</u>	1.33	1.30	1.16	1.35	<u>0.99</u>	1.14
Nova-lite-v1	14.75	2.62	1.99	0.46	0.33	0.30	4.61	4.92	0.48	2.5e2	<u>0.06</u>	0.30	1.38	1.33	1.32	1.46	1.06	1.27
Gemini-2.0-Flash	13.93	2.73	2.16	0.34	0.32	0.41	4.90	5.01	<u>0.63</u>	2.4e2	<u>0.06</u>	0.28	1.39	1.36	1.32	1.47	<u>0.99</u>	1.22
GPT-5-nano	12.54	2.97	2.12	0.59	0.45	0.32	<u>4.42</u>	5.04	0.52	<u>2.2e2</u>	0.23	0.36	1.46	1.38	1.44	1.54	1.08	1.33

remains highly task-dependent rather than being determined solely by model scale.

3.3 Performance on Multi-year Forecasting Task

Temporal context substantially improves numerical stability. Table 3 presents results on the multi-year forecasting task. Compared to single-year estimation, error magnitudes decrease dramatically and become tightly bounded across models, eliminating the extreme outliers previously observed for CO₂ and Popu. This indicates that incorporating temporal context significantly improves calibration for emission-related and demographic quantities. Moreover, performance differences across model families are noticeably reduced, suggesting that temporal context provides a strong signal that partially compensates for architectural differences.

Forecasting difficulty remains indicator-dependent. Despite improved stability, substantial differences persist across indicator types. Accessibility indicators and EcD generally achieve more consistent forecasting performance, suggesting relatively regular temporal dynamics. In contrast, Popu remains among the most challenging indicators, while environmental forecasting performance

varies considerably across variables. These patterns indicate that indicators influenced by complex external factors, such as demographic change and emissions, remain difficult to predict even when historical multimodal observations are available. By comparison, indicators more directly tied to observable urban structure exhibit greater temporal regularity and are therefore more amenable to forecasting.

Temporal context narrows performance differences across models. Compared with the single-year estimation task, inter-model performance gaps become narrower across most indicators. While larger variants within the same model family sometimes achieve lower forecasting errors, scaling benefits are less consistently observed than in single-year estimation. Notably, forecasting performance becomes noticeably more homogeneous across models, suggesting that temporal context provides a strong predictive signal that narrows performance differences across architectures. As a result, scaling benefits become less pronounced in the forecasting setting.

Table 4: Results of temporal trend analysis tasks. Accuracy is reported for all indicators. For each indicator, the best value across all models is highlighted in bold, and the second-best value is underlined.

Model	Environment				Accessibility		Urban Form		Urban Vitality			Subjective Perception					
	CO ₂	NO ₂	PM2.5	NDVI	RTR	SPM	RD	RL	Popu	LUD	EcD	Safe	LIV	WEA	BEA	BOR	DEP
Gemma3-4B	0.32	<u>0.36</u>	0.41	<u>0.53</u>	0.45	0.42	0.23	0.23	0.36	0.12	0.35	0.34	0.32	0.30	<u>0.37</u>	0.33	0.37
Gemma3-12B	0.26	0.25	0.21	0.06	0.30	0.33	0.30	0.25	0.38	<u>0.80</u>	0.27	0.15	0.13	0.13	<u>0.15</u>	0.20	0.19
Gemma3-27B	0.23	0.13	0.14	0.09	<u>0.37</u>	0.33	0.36	0.33	0.39	0.77	0.28	0.09	0.10	0.13	0.11	0.15	0.21
Qwen2.5VL-3B	<u>0.37</u>	0.38	0.35	0.44	0.36	0.35	0.32	0.30	0.36	0.17	0.34	<u>0.37</u>	<u>0.36</u>	<u>0.35</u>	0.36	0.35	0.37
Qwen2.5VL-7B	0.24	0.21	0.18	0.12	0.31	0.33	0.33	0.32	0.37	0.64	0.28	0.13	0.15	0.13	0.15	0.16	0.16
Qwen2.5VL-32B	0.26	0.25	0.20	0.08	0.31	0.32	0.31	0.31	<u>0.40</u>	0.77	0.27	0.12	0.10	0.09	0.12	0.17	0.16
InternVL3-9B	0.23	0.17	0.17	0.10	0.33	0.33	0.32	0.34	0.38	0.67	0.27	0.17	0.19	0.20	0.22	0.24	0.19
InternVL3.5-14B	0.30	0.28	0.31	0.30	0.31	0.32	0.38	0.39	0.39	0.63	0.34	0.31	0.30	0.30	0.32	0.30	0.29
Phi-3.5	0.23	0.23	0.20	<u>0.27</u>	0.32	0.31	0.33	0.35	0.37	0.68	0.30	0.25	0.30	0.26	0.25	0.27	0.28
Phi-4	0.30	0.27	0.27	0.54	0.32	0.32	0.43	0.43	0.37	0.26	0.40	0.33	0.35	0.32	0.31	0.27	0.30
Mistral-Small-3.1-24B	0.00	0.31	0.20	0.14	0.00	0.00	0.56	0.67	0.36	0.89	0.00	0.21	0.21	0.19	0.19	0.21	0.19
Llama4-Scout	0.67	0.00	0.04	0.00	0.31	0.31	0.37	0.36	0.52	0.75	0.36	0.21	0.19	0.16	0.20	0.21	0.16
Nova-lite-v1	0.33	0.28	0.30	0.31	0.29	0.29	0.45	<u>0.46</u>	0.39	0.45	0.37	0.31	0.34	0.34	0.35	0.29	0.28
Gemini-2.0-Flash	0.21	0.11	0.10	0.05	0.30	0.31	0.32	0.32	<u>0.40</u>	0.79	0.26	0.08	0.08	0.08	0.08	0.11	0.07
GPT-5-nano	0.31	0.32	0.23	0.33	<u>0.37</u>	0.34	<u>0.48</u>	0.42	0.35	0.37	<u>0.39</u>	0.38	0.38	0.40	0.39	<u>0.34</u>	<u>0.36</u>

3.4 Performance on Temporal Trend Analysis

Temporal trend analysis remains challenging. Table 4 presents element-wise accuracy for multi-year trend analysis. For many indicators, accuracy remains modest, with a large proportion of results concentrated around 0.2–0.4. Unlike value forecasting, trend analysis requires models to identify the direction of change from image sequences and capture subtle inter-annual variations through explicit temporal comparison. These results suggest that current MLLMs struggle to reliably distinguish fine-grained temporal changes across many urban indicators.

Trend predictability varies substantially across indicators. While several indicators remain close to chance-level performance, others are considerably more predictable. LUD achieves the highest accuracy across multiple models, with several models exceeding 0.75 accuracy. Urban form indicators such as RL and RD also achieve relatively strong performance. In contrast, many environmental and subjective perception-related indicators remain clustered around moderate accuracy levels. These patterns suggest that indicators with more direct visual manifestations are more predictable from temporal image sequences, whereas subtle environmental and perceptual changes remain difficult to identify reliably.

Model scale provides limited advantage in temporal trend reasoning. Performance differences across model families and parameter scales are inconsistent. Larger variants do not systematically outperform smaller counterparts within the same family, and improvements on one indicator are often offset by declines on others. This contrasts with the partial scaling benefits observed in single-year estimation. The relatively homogeneous performance across models suggests that temporal trend detection remains challenging regardless of model scale, highlighting a common limitation in capturing subtle temporal changes from image sequences.

To better understand these failures, we inspect model predictions and observe a consistent tendency toward temporal over-smoothing. When year-to-year differences are subtle, MLLMs frequently predict *stable* even when the ground-truth trend is positive or negative. This behavior suggests that current MLLMs rely primarily on static

visual cues and struggle to perform explicit temporal comparison across image sequences.

3.5 Detailed Analysis

In this section, we analyze performance variation across cities and investigate how different visual, temporal, and adaptation factors affect model performance. We further compare model performance against human baselines.

City-Level Performance Variations. Using Gemma3-27B as a representative model, we analyze city-level prediction error for the BEA indicator (Figure 4(a)). RMSE varies noticeably across cities in both single-year and multi-year settings, indicating heterogeneous prediction difficulty under different urban environments. Compared with single-year estimation, multi-year forecasting generally produces lower RMSE with reduced cross-city variance, suggesting that temporal context improves prediction stability across urban environments. Nevertheless, certain cities remain consistently challenging across both settings, implying that city-specific visual characteristics and urban heterogeneity continue to affect model performance even when temporal information is available.

Effect of Satellite Imagery. We evaluate the contribution of satellite imagery by comparing two configurations: (i) satellite + street view imagery (with SAT) and (ii) street view imagery only (w/o SAT). Figure 4(b) reports results for PM2.5, BEA, and RL using Gemma3-27B. The effect of satellite imagery is indicator- and task-dependent. Environmental indicators exhibit limited sensitivity to aerial observations, whereas perceptual and structural indicators generally benefit more from additional spatial context. The gains are particularly evident in the single-year setting, while the contribution of satellite imagery becomes less consistent for multi-year forecasting. Overall, these results suggest that aerial context is most useful when the target indicator is visually grounded and the task relies on static spatial structure.

Effect of Street View Image Quantity. We analyze the impact of the number of street view images per spatial unit (1–4 images) on PM2.5, BEA, and RL (Figure 4(c)). Increasing the number of street view images does not consistently improve performance across

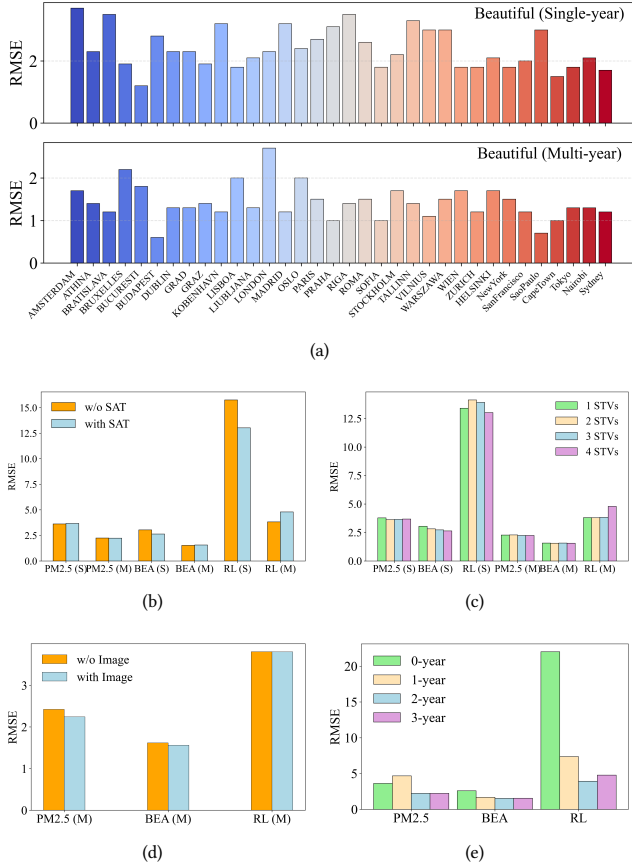


Figure 4: RMSE results for (a) city-level *beautiful* estimation and analyses of the effects of (b) satellite imagery, (c) street view image quantity, (d) image modality, and (e) temporal length. ‘S’ denotes the single-year estimation task, and ‘M’ denotes the multi-year forecasting task.

indicators. Environmental indicators remain largely insensitive to additional street-level coverage, while perceptual attributes benefit from more diverse visual perspectives. Structural indicators exhibit non-monotonic behavior, where adding more views may improve performance in some cases but also introduce redundancy or noisy visual signals. These results suggest that additional street view observations provide complementary information to varying degrees across different indicator types.

Effect of Image Modality. We assess the contribution of visual modalities by comparing multi-year forecasting performance with and without image inputs (Figure 4(d)). Incorporating satellite and street view imagery improves performance for certain indicators, particularly PM2.5 and BEA, while yielding marginal changes for RL. These results suggest that visual observations provide complementary information beyond historical numerical inputs, but their contribution varies across indicators.

Effect of Temporal Length. We examine the effect of the number of historical years provided as input (Figure 4(e)). Across all three indicators, incorporating temporal context generally reduces

RMSE compared to using no historical input, confirming the importance of historical observations for multi-year forecasting. However, the benefit of longer temporal sequences is not monotonic. PM2.5 and RL exhibit fluctuating performance as more years are included, whereas the BEA indicator improves consistently and then stabilizes. These results suggest that temporal context provides strong predictive signals, but longer sequences do not uniformly translate into better forecasting performance.

Effect of Fine-Tuning. To examine whether the zero-shot setting underestimates model capability, we perform parameter-efficient LoRA fine-tuning [30] on Qwen2.5VL-7B using UrbanWell samples excluded from the benchmark split. Specifically, we keep the pretrained backbone frozen and train only the LoRA adapter parameters, then load the learned adapter on top of the same base model for multimodal inference. Table 5 reports results on five representative indicators covering each benchmark domain. Fine-tuning substantially improves performance in the single-year setting, particularly for Popu and CO₂ estimation, while gains in the multi-year forecasting setting are comparatively smaller. These results suggest that adaptation improves absolute performance but does not eliminate the heterogeneous difficulty across indicators, implying that the observed challenges extend beyond the zero-shot setting.

Table 5: Performance comparison between zero-shot and LoRA fine-tuned Qwen2.5VL-7B across representative UrbanWell indicators. Lower RMSE indicates better performance. ‘S’ and ‘M’ denote single-year estimation and multi-year forecasting, respectively.

	Model	Popu	RD	CO ₂	SPM	BEA
S	Zero-shot	7.20e4	16.43	1.10e8	8.07	3.35
S	LoRA	7942	13.35	1898	2.13	2.13
M	Zero-shot	370	4.81	14.20	0.30	1.42
M	LoRA	319	4.04	15.12	0.27	1.35

Comparison with Human Performance. We additionally evaluate human participants on subsets of four single-year estimation tasks (CO₂, Popu, BEA, and WEA). Humans participants achieve lower RMSE than the best-performing MLLM on CO₂ (50 vs. 175) and Popu (2901 vs. 6400), while remaining competitive on BEA (2.38 vs. 2.04) and WEA (1.84 vs. 1.69). These results reveal varying gaps between human and MLLM performance across different urban indicators.

Additional analyses on validation of pretrained subjective perception labels, prompt sensitivity, temporal forecasting baselines, and failure cases are provided in Appendices A.4–A.7.

4 Related Work

4.1 MLLMs in Urban Applications

Recent advances in MLLMs have enabled joint modeling of visual and textual information, supporting cross-modal reasoning over complex inputs [62]. Representative models such as Qwen [7], GPT [45], and InternVL [70] demonstrate strong performance in image understanding [8], cross-modal alignment [60], and visual reasoning tasks [61]. While these models show robust generalization

across general-purpose benchmarks [33], systematic evaluation in structured urban settings remains limited.

In urban applications, MLLMs have been explored for urban perception, spatial reasoning, and multimodal urban understanding [10, 18, 35, 64]. UrbanCLIP [59] aligns satellite imagery with textual descriptions using contrastive learning, and UrbanVLP [26] leverages the descriptive capabilities of MLLMs to bridge satellite and street view imagery for urban prediction. UrbanLLaVA [19] further integrates satellite and street view imagery with structured urban datasets for improved generalization on urban tasks. However, existing efforts primarily focus on static prediction settings. Standardized benchmarks for evaluating temporal reasoning over longitudinal urban dynamics in MLLMs remain limited.

4.2 Urban Visual Dataset and Benchmark

Urban imagery, such as satellite and street view imagery, has been leveraged to quantify spatial patterns, urban dynamics, and socioeconomic conditions [34, 37]. Satellite imagery captures large-scale spatial structure from a bird’s-eye perspective [57, 59], while street view imagery complements satellite observations with fine-grained ground-level visual information [17, 29]. Building upon these multimodal urban imagery, recent benchmarks have begun to evaluate the capabilities of MLLMs in urban understanding and reasoning tasks. CityBench [20] focuses on urban navigation and identity tasks, UrBench [68] integrates multi-view imagery for spatial reasoning, CityLens [36] incorporates socioeconomic data for urban function modeling and UrbanFeel [27] introduces temporal sequences with perceptual indicators. CityEQA [67] studies embodied urban question answering, while MapDR [9] aligns traffic semantics with structured map representations.

Despite these advances, existing benchmarks primarily focus on perception, navigation, or task-specific urban reasoning, with limited support for comprehensive urban wellbeing analytics. In particular, few benchmarks jointly model multi-year satellite and street view imagery together with heterogeneous indicators under a unified evaluation framework. Moreover, temporal urban reasoning remains underexplored, especially for forecasting future urban conditions and understanding longitudinal urban trends across cities. UrbanWell addresses these limitations through a large-scale multi-year benchmark that systematically evaluates spatio-temporal urban wellbeing reasoning from satellite and street view imagery.

5 Ethical Use of Data

UrbanWell is constructed from satellite imagery, street view imagery, and publicly available geospatial datasets. The spatial resolution of the satellite imagery is insufficient for identifying individuals, while street view images are obtained from officially released platforms where sensitive information has been anonymized prior to publication. All derived indicators are aggregated at grid level and contain no personally identifiable information. Consistent with prior urban computing research using similar data sources [27, 36], we comply with applicable platform policies and data usage regulations. To respect intellectual property and licensing constraints, UrbanWell does not redistribute raw satellite or street view imagery. Instead, we release benchmark annotations, geographic metadata,

timestamps, and a deterministic retrieval pipeline enabling reconstruction of the benchmark inputs from the original providers under their respective terms of service. This design supports reproducible evaluation and standardized comparison across future methods.

6 Conclusion

We introduce UrbanWell, a large-scale multimodal benchmark for evaluating spatio-temporal reasoning capabilities of MLLMs in urban wellbeing analytics. UrbanWell spans 38 cities and 19 heterogeneous urban indicators, integrating multi-year satellite and street view imagery with grid-level socio-environmental measurements. The benchmark provides standardized evaluation framework for single-year estimation, multi-year forecasting, and temporal trend classification. Extensive experiments on 15 state-of-the-art MLLMs reveal substantial performance variation across indicators, task settings, and model architectures. The results further show that, although current models perform relatively well on general-purpose visually grounded tasks, they remain challenged by environmental estimation, demographic inference, and fine-grained temporal reasoning, highlighting limitations in multimodal quantitative and longitudinal urban understanding. These findings underscore the need for more robust multimodal reasoning capabilities in urban-scale AI systems and motivate future research on multimodal spatio-temporal reasoning for urban analytics.

Limitations & Future Work. Subjective perception indicators in UrbanWell are derived from pretrained vision models trained on human-labeled datasets [14], and may therefore inherit biases present in the original annotations. Future extensions may incorporate additional human-validated signals and cross-cultural calibration to improve robustness of perception-related evaluation. UrbanWell currently covers 38 cities, primarily in Europe, and future work will extend the benchmark to more geographically diverse and underrepresented urban regions with richer temporal observations and broader multimodal coverage. A further limitation concerns temporal alignment, as imagery captures specific moments while some indicators are annual aggregates. Although year-level alignment is common in urban computing and urban environments are generally stable at this timescale, some mismatch may still exist. Finally, UrbanWell evaluates MLLMs as deployed, including knowledge potentially acquired during pretraining. While the benchmark is constructed from grid-level imagery and year-specific indicators, making direct memorization unlikely, potential exposure to related geographic information during pretraining cannot be completely excluded. Future work will also investigate more robust multimodal temporal reasoning and calibration strategies for urban-scale quantitative prediction.

7 Acknowledgments

This work was supported in part by NordForsk through Nordic University Cooperation on Edge Intelligence (168043), in part by Finnish Foundation for Technology Promotion Grant (11246), in part by City of Helsinki Research Grant (2025), in part by Research Council of Finland XRise project (371849), and in part by the Zhongguancun Academy (Grant No. XTS0074).

References

- [1] 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv:2404.14219* [cs.CL]
- [2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905* (2024).
- [3] Jacob Levy Abitbol and Marton Karsai. 2020. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence* 2, 11 (2020), 684–692.
- [4] Adrian Albert, Jasleen Kaur, and Marta C Gonzalez. 2017. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1357–1366.
- [5] Amazon. 2025. Amazon Nova. <https://aws.amazon.com/cn/ai/generative-ai/nova/>. Accessed: 2025-11-29.
- [6] Jean Vasile Andrei, Dragoş Sebastian Cristea, Florian Marcel Nuță, Ștefan-Mihai Petrea, Alina Cristina Nuță, Adriana Tiron Tudor, and Luminita Chivu. 2024. Exploring the causal relationship between PM 2.5 air pollution and urban areas economic welfare and social-wellbeing: Evidence from 15 European capitals. *Sustainable Development* 32, 5 (2024), 4567–4592.
- [7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).
- [8] Zhipeng Bian, Jieming Zhu, Qijiong Liu, Wang Lin, Guohao Cai, Zhaocheng Du, Jiacheng Sun, Zhou Zhao, and Zhenhua Dong. 2025. ICG: Improving Cover Image Generation via MLLM-based Prompting and Personalized Preference Alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 12279–12289.
- [9] Xinyuan Chang, Maixuan Xue, Xinran Liu, Zheng Pan, and Xing Wei. 2025. Driving by the rules: A benchmark for integrating traffic sign regulations into vectorized hd map. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 6823–6833.
- [10] Jiabin Chen, Haiping Wang, Jinpeng Li, Yuan Liu, Zhen Dong, and Bisheng Yang. 2025. SpatialLLM: From Multi-modality Data to Urban Spatial Intelligence. *arXiv preprint arXiv:2505.12703* (2025).
- [11] Copernicus Land Monitoring Service. 2012. Urban Atlas 2012. <https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012>. Accessed: November 16, 2025.
- [12] Copernicus Land Monitoring Service. 2018. Urban Atlas 2018. <https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018>. Accessed: November 16, 2025.
- [13] Copernicus Land Monitoring Service. 2025. Copernicus Land Monitoring Service. <https://land.copernicus.eu>. Accessed: November 16, 2025.
- [14] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *European conference on computer vision*. Springer, 196–212.
- [15] European Environment Agency. 2014. *Good practice guide on quiet areas*. Technical Report Technical report No. 4/2014. Publications Office of the European Union, Luxembourg. <https://www.eea.europa.eu/en/analysis/publications/good-practice-guide-on-quiet-areas>
- [16] European Environment Agency. 2025. European Environment Agency. <https://www.eea.europa.eu>. Accessed: November 16, 2025.
- [17] Zhuangyuan Fan, Fan Zhang, Becky PY Loo, and Carlo Ratti. 2023. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences* 120, 27 (2023), e2220417120.
- [18] Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025. Citygpt: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 591–602.
- [19] Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. 2025. UrbanLLaVA: A Multi-modal Large Language Model for Urban Intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6209–6219.
- [20] Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2025. Citybench: Evaluating the capabilities of large language models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5413–5424.
- [21] Manuel Ganter, Malte Toetzke, and Stefan Feuerriegel. 2022. Mining points-of-interest data to predict urban inequality: evidence from Germany and France. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 216–227.
- [22] Google. 2025. Google Earth. <https://earth.google.com>. Accessed: November 16, 2025.
- [23] Google. 2025. Google Street View API. <https://developers.google.com/maps/documentation/streetview>. Accessed: November 16, 2025.
- [24] Google DeepMind. 2025. Gemini 2.0. <https://cloud.google.com/vertex-ai/generative-ai/docs/gemini-v2>. Accessed: 2025-11-29.
- [25] Mordechai Haklay and Patrick Weber. 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive computing* 7, 4 (2008), 12–18.
- [26] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2025. UrbanVLP: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28061–28069.
- [27] Jun He, Yi Lin, Zilong Huang, Jiacong Yin, Junyan Ye, Yuchuan Zhou, Weijia Li, and Xiang Zhang. 2025. Urbanfeel: A comprehensive benchmark for temporal and perceptual understanding of city scenes through human perspective. *arXiv preprint arXiv:2509.22228* (2025).
- [28] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [29] Yujun Hou, Matias Quintana, Maxim Khomiakov, Winston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, Tianhong Zhao, and Filip Biljecki. 2024. Global StreetScapes – A comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS Journal of Photogrammetry and Remote Sensing* 215 (2024), 216–238. doi:10.1016/j.isprsjprs.2024.06.023
- [30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr* 1, 2 (2022), 3.
- [31] Jiahui Jin, Yifan Song, Dong Kan, Haojia Zhu, Xiangguo Sun, Zhicheng Li, Xigang Sun, and Jinghui Zhang. 2025. Urban region pre-training and prompting: A graph-based approach. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 1071–1082.
- [32] 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.
- [33] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895* (2024).
- [34] 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.
- [35] Zhuoheng Li, Yaochen Wang, Zhixue Song, Yuqi Huang, Rui Bao, Guanjie Zheng, and Zhenhui Jessie Li. 2024. What can LLM tell us about cities? *arXiv preprint arXiv:2411.16791* (2024).
- [36] Tianhui Liu, Jie Feng, Hetian Pang, Xin Zhang, Tianjian Ouyang, Zhiyuan Zhang, and Yong Li. 2025. CityLens: Benchmarking Large Language-Vision Models for Urban Socioeconomic Sensing. *arXiv preprint arXiv:2506.00530* (2025).
- [37] 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.
- [38] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680* (2024).
- [39] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213* (2023).
- [40] Meta AI. 2025. Llama 4: Industry Leading, Open-Source AI. <https://www.llama.com/models/llama-4/>. Accessed: November 16, 2025.
- [41] Mistral AI. 2025. Introducing Mistral Small 3.1: Powering Efficient AI Applications. <https://mistral.ai/news/mistral-small-3-1>. Accessed: November 16, 2025.
- [42] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730* (2022).
- [43] Tomohiro Oda and Shamil Maksyutov. 2011. A very high-resolution (1 km×1 km) global fossil fuel CO₂ emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmospheric Chemistry and Physics* 11, 2 (2011), 543–556.
- [44] Tomohiro Oda, Shamil Maksyutov, and Robert J Andres. 2018. The Open-source Data Inventory for Anthropogenic CO₂ version 2016 (ODIAC2016): a global monthly fossil fuel CO₂ gridded emissions data product for tracer transport simulations and surface flux inversions. *Earth System Science Data* 10, 1 (2018), 87–107.
- [45] OpenAI. 2025. GPT-4.1: The Next Generation of AI from OpenAI. <https://openai.com/index/gpt-4-1/>. Accessed: November 16, 2025.
- [46] OpenAI. 2025. GPT-5 System Card. <https://cdn.openai.com/gpt-5-system-card.pdf>. Accessed: November 16, 2025.
- [47] Sanja Scepanovic, Sagar Joglekar, Stephen Law, and Daniele Quercia. 2021. Jane Jacobs in the sky: Predicting urban vitality with open satellite data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.
- [48] Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.
- [49] Fengze Sun, Yanchuan Chang, Egemen Tanin, Shanika Karunasekera, and Jianzhong Qi. 2025. FlexiReg: Flexible Urban Region Representation Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 2702–2713.
- [50] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane

- Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).
- [51] Compton J Tucker. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment* 8, 2 (1979), 127–150.
- [52] Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. 2023. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *The eleventh international conference on learning representations*.
- [53] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265* (2025).
- [54] 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.
- [55] World Bank. 2024. Urban Population (% of Total Population) - World. <https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS> Accessed: 2024-08-07.
- [56] WorldPop. 2025. WorldPop. <https://www.worldpop.org>. Accessed: November 16, 2025.
- [57] 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.
- [58] Yanxin Xi, Yu Liu, Zhicheng Liu, Sasu Tarkoma, Pan Hui, and Yong Li. 2024. From pixels to progress: generating road network from satellite imagery for socioeconomic insights in impoverished areas. *arXiv preprint arXiv:2406.11282* (2024).
- [59] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*. 4006–4017.
- [60] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. 2024. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732* (2024).
- [61] Congzhi Zhang, Jiawei Peng, Zhenglin Wang, Yilong Lai, Haowen Sun, Heng Chang, Fei Ma, and Weijiang Yu. 2025. VReST: Enhancing Reasoning in Large Vision-Language Models through Tree Search and Self-Reward Mechanism. *arXiv preprint arXiv:2506.08691* (2025).
- [62] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024* (2024), 12401–12430.
- [63] Fan Zhang, Bolei Zhou, Liu Liu, Yu Liu, Helene H Fung, Hui Lin, and Carlo Ratti. 2018. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning* 180 (2018), 148–160.
- [64] Jiaxin Zhang, Yunqin Li, Tomohiro Fukuda, and Bowen Wang. 2025. Urban safety perception assessments via integrating multimodal large language models with street view images. *Cities* 165 (2025), 106122.
- [65] Yunke Zhang, Ruolong Ma, Xin Zhang, and Yong Li. 2025. Perceiving urban inequality from imagery using visual language models with chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference 2025*. 5342–5351.
- [66] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *CVPR*.
- [67] Yong Zhao, Kai Xu, Zhengqiu Zhu, Yue Hu, Zhiheng Zheng, Yingfeng Chen, Yatai Ji, Chen Gao, Yong Li, and Jincai Huang. 2025. Cityeqa: A hierarchical llm agent on embodied question answering benchmark in city space. *arXiv preprint arXiv:2502.12532* (2025).
- [68] Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. 2025. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10707–10715.
- [69] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 633–641.
- [70] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).

A Appendix

A.1 List of Covered Cities

UrbanWell covers 38 cities: Amsterdam, Athens, Belgrade, Berlin, Bratislava, Brussels, Bucharest, Budapest, Cape Town, Copenhagen,

Dublin, Graz, Helsinki, Lisbon, Ljubljana, London, Madrid, Nairobi, New York, Oslo, Paris, Prague, Riga, Rome, San Francisco, Sao Paulo, Skopje, Sofia, Stockholm, Sydney, Tallinn, Tirana, Tokyo, Vienna, Vilnius, Warsaw, Zagreb, Zurich.

A.2 Indicator Formulation

Land Use Diversity (LUD). The land use diversity indicator for grid r is computed as [47]:

$$\text{LUD}_r = - \frac{\sum_{j=1}^{n_{LUC}} P_{r,j} \log(P_{r,j})}{\log(n_{LUC})}, \quad (1)$$

where $P_{r,j}$ denotes the proportion of land use category j within grid r , and $n_{LUC} = 3$ is the total number of land use categories: (i) residential areas, (ii) commercial, industrial, institutional, and governmental areas, and (iii) recreational spaces, parks, and water bodies.

Economic Activity Diversity (EcD). The economic activity diversity indicator for grid r is computed using Shannon entropy [48]:

$$\text{EcD}_r = - \sum_{k=1}^{n_{POI}} P_{r,k} \log(P_{r,k}), \quad (2)$$

where $P_{r,k}$ denotes the proportion of POIs belonging to category k within grid r , and n_{POI} is the total number of commercial POI categories. The commercial POI categories include restaurants, fast food outlets, cafes, bars, pubs, supermarkets, convenience stores, malls, clothing stores, shoe stores, electronics stores, jewellers, bakeries, butcher shops, florists, bookshops, hairdressers, beauty shops, laundries, repair shops, photo studios, travel agencies, car rentals, car washes, hotels, motels, guesthouses, hostels, campsites, cinemas, theatres, museums, galleries, nightclubs, casinos, sports centres, and fitness centres.

A.3 Task Prompt Templates

Table 6 summarizes the prompt templates used in UrbanWell for the three task paradigms: Single-year Estimation, Multi-year Forecasting, and Temporal Trend Analysis. Each task adopts a standardized prompt structure specifying the input imagery, temporal context, historical observations, and expected output format.

For single-year estimation, the prompt provides satellite and street view imagery from a specific year and asks the model to estimate a target urban indicator. For multi-year forecasting, the prompt includes sequential imagery together with historical indicator values to predict the target value in the final year. For temporal trend analysis, the prompt presents imagery from consecutive years and requires classification of the indicator trend according to predefined thresholds. All prompts follow unified templates with fixed placeholders to ensure controlled and reproducible evaluation across tasks and models. Additional prompt examples and corresponding MLLM outputs for all task paradigms are provided in the project repository: <https://github.com/axin1301/UrbanWell-Benchmark>.

A.4 Validation of Subjective Perception Labels

Since subjective perception indicators are derived from pretrained models rather than direct human annotations, we compare benchmark labels against a subset of 100 human-annotated samples. For

Table 6: Prompt templates used for single-year estimation, multi-year forecasting, and temporal trend analysis in UrbanWell.

Task Paradigm	Prompt Template
Single-year Estimation	Suppose you are a vision expert specializing in socioeconomic and environmental analysis. Examine {number_of_total_images} images from the same region: 1 satellite image from year {year}, and {number_of_street_view_images} street views from the same year. Analyze the imagery content, and provide your assessment of {indicator} in {indicator_unit} according to the {measurement_definition} provided by its data source. Reply only with the numeric value. Nothing else should be included.
Multi-year Forecasting	Suppose you are a vision expert. You are given a series of images for a region spanning {number_of_total_years} years, from {first_year} to {last_year}. For each historical year, you are provided with 1 satellite image, {number_of_street_view_images} street view image(s), and the corresponding {indicator} in {indicator_unit} based on the {measurement_definition}. For the target year {target_year}, You are provided with only the images. Using the historical years as context, estimate {indicator} in {indicator_unit} for {target_year}. Reply only with the numeric value.
Temporal Trend Analysis	You are given images for a region across {number_of_total_years} years from {first_year} to {last_year} (each year includes 1 satellite image and {number_of_street_view_images} street view image(s)). For each consecutive two-year step, classify the trend of {indicator} in {indicator_unit} according to the {measurement_definition} as positive, negative, or stable based on predefined relative-change thresholds. Reply only with the label.

the Qwen-series models, evaluation against human annotations and pretrained scores yields highly comparable average RMSE (BEA: 4.3 vs. 4.7; WEA: 4.8 vs. 4.9), suggesting that the proxy labels remain closely aligned with human perception for large-scale benchmarking.

A.5 Comparison with Temporal Baselines

To contextualize MLLM performance, we additionally evaluate several temporal forecasting baselines, including ARIMA, Ridge Regression (RR), LSTM [28], PatchTST [42], and MICN [52], using the same historical indicator observations. As shown in Table 7, temporal baselines remain competitive for population forecasting, while MLLMs achieve lower RMSE on CO₂ and RD prediction. For the perception-related indicator (BEA), Ridge Regression achieves the best performance. These results suggest that temporal models remain effective for certain urban indicators.

Table 7: Comparison between MLLMs and temporal forecasting baselines on representative indicators. Lower RMSE indicates better performance.

Model	Popu	CO ₂	BEA	RD
ARIMA	179	16	1.92	3.65
RR	5826	130	1.28	8.85
LSTM	5301	127	1.36	8.66
PatchTST	6872	140	1.80	9.79
MICN	5320	131	1.31	8.34
Best MLLM	200	11.5	1.35	2.56

A.6 Prompt Sensitivity Analysis

To assess prompt sensitivity, we evaluate two alternative prompt variants: (1) a reasoning-enhanced prompt that encourages step-by-step inference, and (2) a structured prompt with explicit formatting

and concise instructions. Table 8 reports results for Gemma3-27B on representative indicators under both single-year estimation and multi-year forecasting settings. Structured prompts generally improve performance for certain indicators, such as CO₂ and Popu in the single-year setting, likely because the default prompts intentionally provide minimal task guidance. However, the gains are inconsistent across tasks and indicators. Reasoning-enhanced prompts often substantially degrade performance, particularly for multi-year forecasting tasks.

Table 8: Prompt sensitivity analysis across representative indicators under single-year estimation (S) and multi-year forecasting (M). Lower RMSE indicates better performance.

Indicator	Setting	Default	Reasoning	Structured
CO ₂	S	371	1531	261
Popu	S	16312	12792	5031
BEA	S	2.7	4.0	4.1
CO ₂	M	10	596	11
Popu	M	229	9965	179
BEA	M	1.5	1.5	1.5

A.7 Failure Case Analysis

We observe several recurring failure patterns across tasks. Indicators that are not directly observable from imagery, such as CO₂ and Popu, often exhibit large estimation errors in the single-year setting due to the lack of strong visual signals. For example, a grid with ground-truth CO₂ value 89 may be predicted as 12. In temporal tasks, models frequently fail to capture subtle urban changes, such as gradual population-related densification, and often predict *stable* when year-to-year differences are visually minor.