AGENTICCITY-BENCH: LONGITUDINAL OBSERVATIONS OF URBAN DYNAMICS FOR MULTIMODAL AGENTIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Urban systems are shaped by interactions among infrastructure, human activity, and environmental conditions. Existing earth observation and urban benchmarks are limited in unified modeling and in capturing their joint dynamics. We introduce AgenticCity-Bench, a longitudinal multimodal benchmark for multi-layer unified state modeling and agentic city task reasoning. The benchmark covers 16 cities in 12 countries across six continents from 2000 to 2025, and includes 18 modalities spanning urban function, activity, and environmental state. It organizes heterogeneous data sources around a common spatial reference and calendar index while preserving spatial and temporal coverage, provenance and tile-level availability. On this foundation, AgenticCity-Bench supports urban cross-modal translation together with diverse agentic city tasks involving forecasting, planning and resource analysis. This study provides a unified benchmark for studying the transition from multimodal urban and earth observation to joint understanding and agentic city decision-making. The dataset is available at https://anonymous.4open.science/r/AgenticCity-Bench.

1 INTRODUCTION

Urban and earth data are essential for understanding how cities function, evolve, and respond to planning and the environment. Cities are dynamic multi-layer systems shaped by population (Zimmer et al., 2026), infrastructure (Bettencourt & Marchio, 2025), economic activity (Li et al., 2026b), resource use (Li et al., 2026a), and environmental conditions (Liao et al., 2025). Sustainable urban development requires coordinated planning that accounts for interactions among these components (Kaufmann et al., 2024).

Recent earth observation (EO) (Su et al., 2026) and urban benchmarks have broadened the data and tasks available for studying cities, covering both physical landscapes (Li et al., 2026c) and urban social activity (Xu et al., 2025; Sun et al., 2026). Artificial intelligence (e.g., GenAI and agentic AI) has further accelerated these developments. This progress has increasingly moved from single-task and single-source data toward broader, multimodal, and temporal observations(Yap et al., 2025). EO benchmarks (Nedungadi et al., 2024; Marsocci et al., 2025; Van Etten et al., 2021) have expanded the use of multimodal and longitudinal observations for understanding physical environments. On the other hand, urban benchmarks (Feng et al., 2025c; Zhou et al., 2025; Liu et al., 2026) have broadened the modeling of social activity, mobility, and dynamic urban processes. These advances motivate linking social and physical knowledge and data to social activity and urban dynamics. Fine-grained alignment requires considering differences in spatial coverage (O'Brien, 2025), periods, and missing data across urban records.

As urban datasets and benchmarks have evolved, they have increasingly enabled the study of agent reasoning for urban applications. In parallel, agentic AI approaches in urban intelligence have extended large language models from interpreting observations toward scientific investigation and interactive planning (Feng et al., 2026a; Lai et al., 2026; Akinboyewa et al., 2025). This interplay also exposes a growing need for urban data that allows agents to connect task-relevant observations, analytical models, domain knowledge, and planning procedures. A multi-layer city should encompass not only environmental observations but also infrastructure, socioeconomic activity, and resource use,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reflecting the interconnected nature of urban systems (McPhearson et al., 2016). However, as Table 1 shows, EO benchmarks tend to focus on imagery and derived physical information, while urban benchmarks emphasize transportation, buildings, economy and other records. This specialization provides valuable but fragmented views of urban systems, with limited cross-domain coverage. (Nedungadi et al., 2024; Wang et al., 2024; Lai et al., 2026).

Integrating these data remains challenging, particularly when records span different spatial and temporal scales. Long-term static physical urban morphology and visual appearance are different from dynamic urban activities (Wang et al., 2024; Xuan et al., 2025). Current agent research has begun to explore tool use and interactive planning (Feng et al., 2026a; Lai et al., 2026). These considerations highlight several limitations in existing EO/urban data and agents: (1) A multi-layer urban system needs multiple sources of data covering human activities, environmental conditions, and urban functions, unlike individual EO data or urban records alone (Mühlematter et al., 2025). (2) Urban data should provide stable long-term coverage across heterogeneous sources, enabling joint temporal modeling. (3) Current EO agents or urban agents only support limited individual data modalities (Zhang et al., 2024) to complete agentic city tasks, with most efforts still centered on conventional EO or urban datasets (Feng et al., 2025a; Xiao et al., 2026). (4) Existing EO/urban agents suffer from mismatched external tool usage (Feng et al., 2026b), a lack of city task-specific reasoning, understanding to action gaps, and inappropriate spatio-temporal scales (Hao et al., 2026).

We introduce AgenticCity-Bench, a longitudinal multimodal dataset and benchmark for multi-layer system cities. It comprises 16 cities in 12 countries across six inhabited continents. Its 18 primary modalities are divided into function, activity, and environmental information, connecting built form, infrastructure, population, property, mobility, energy, and environmental conditions through imagery, numerical and categorical grids, vector-derived data source, official records, and geo-temporal text, in Fig 1. The 2000-2025 calendar provides a common long-term index for investigating cross-domain urban variation, while availability remains explicitly specific to each city, year, and modality. Building on this dataset, the task suite connects cross-modal state generation and analysis with urban planning and forecasting. Our contributions are as follows:

1. We propose AgenticCity-Bench, integrating 18 modalities across urban function, activity, and environmental information for 16 cities from 2000 to 2025. The dataset integrates remote sensing, infrastructure, socioeconomic activity, resource use, and environmental observations into a unified urban data framework.
2. An any-to-any multimodal urban translation benchmark covering 306 directed source-to-target modality pairs is established. It reveals translation relationships among all modalities, directional asymmetry, and joint training imbalance across modalities.
3. We construct 1,000 agentic city cases spanning forecasting, urban planning, and resource analysis tasks. We evaluate diverse frontier general and domain-specific agents on multi-modal urban data, analytical tools usage, and task performance. Beyond the good performance agents have achieved, experiments reveal research gaps between understanding and action, a lack of urban domain-specific knowledge, and mismatched tool usage.

Paper organization. Section 2 describes the data collection, processing, and spatiotemporal alignment, and defines the benchmark tasks. Section 3 presents the experimental setup and results for any-to-any urban-state translation and agentic tasks. Due to space constraints, we defer the discussion of related work to Appendix A, with additional dataset and task details provided in the subsequent Appendix C and Appendix D.

2 AGENTICCITY-BENCH ESTABLISHMENT

AgenticCity-Bench consists of a multimodal data collection and a task framework for urban-state translation and agentic decision-making, in Fig.2. This section describes how source data are organized into spatially and temporally indexed samples.

2.1 DATA COLLECTION.

The modality design aims to capture complementary aspects of urban systems across space and time, supporting diverse tasks rather than a single application. We assemble 18 semantic modalities from satellite imagery, geographic databases, and official city records. The collection includes direct

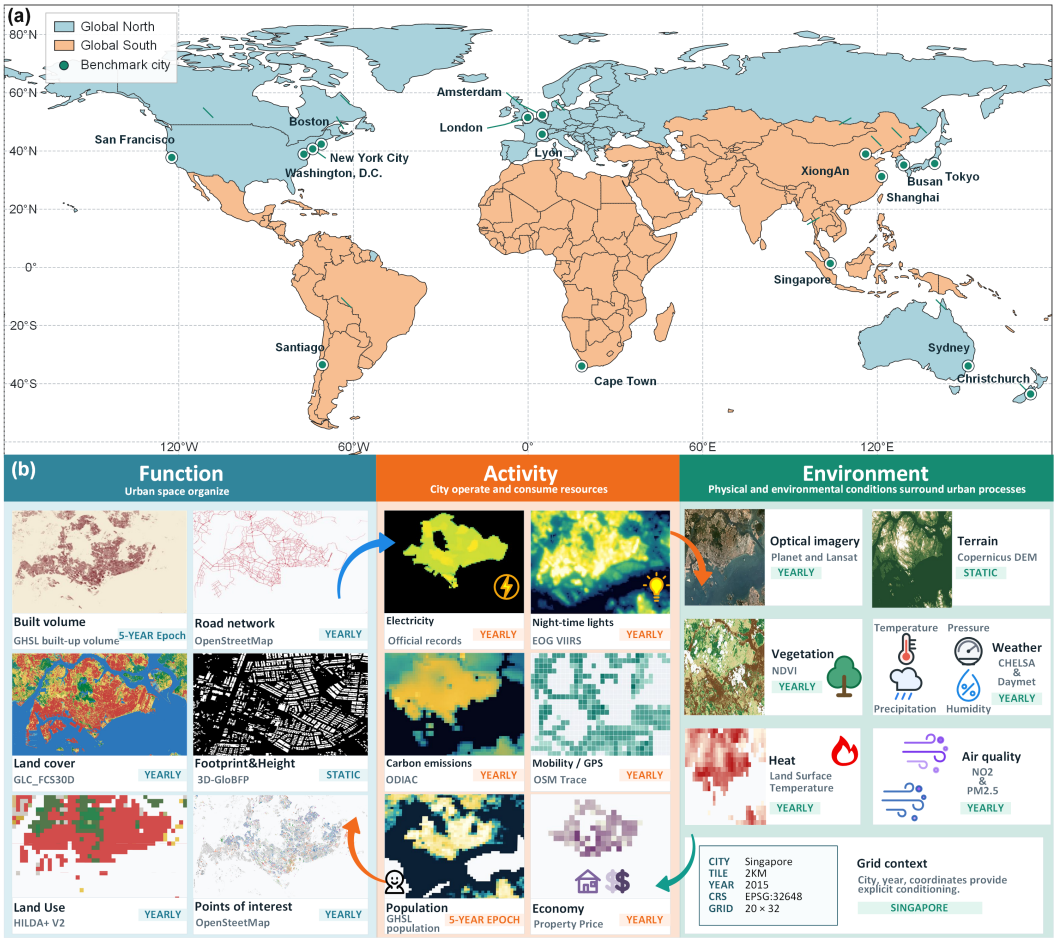


Figure 1: Overview of the proposed AgenticCity-Bench. (a) Geographic distribution of the 16 cities across six continents. (b) Representative examples of the three modality groups: urban function, activity, and environment.

observations, model-derived estimates, and static conditions. Sources are selected based on their complementary information, measurement definitions, and geographic and temporal coverage.

The collection spans 16 cities in 12 countries across six continents from 2000 to 2025, with availability varying across data sources, cities, and years. In Table 1, the 18 modalities are organized into three complementary families: (1) *Function*, covering built volume, road network, land cover, land use, and urban facilities; (2) *Activity*, covering population, economic activity, mobility, electricity consumption, and emissions; and (3) *Environment*, covering remote sensing, terrain, vegetation, weather, air quality, and land surface temperature. Each sample is accompanied by grid context specifying city, country, year, spatial scale, grid location, and geographic bounds. It is used only for conditioning and is not treated as a translation modality. Detailed source descriptions are provided in Appendix B.

2.2 Tile Construction and Processing

Projected city tiles. Each grid divides the study area into regular, non-overlapping tiles at either 2×2 km or 5×5 km, and each tile is identified by its city, scale, row, and column. Samples are indexed by the corresponding tile and year, providing a common spatial footprint for aligning heterogeneous urban observations. Each source is then linked to the corresponding tile and year according to its spatial and temporal coverage. Annual data are aligned with the corresponding year. Five-year epoch data use the first year of each epoch for all years within that epoch (Tesselaar et al., 2023; Yu et al., 2022). Static data are reused across all years (Helfenstein et al., 2024). For

Table 1: Comparison with existing EO and urban datasets. (✓) denotes a released direct semantic data counterpart; (△) denotes a related semantic counterpart; and – denotes no explicitly released or documented counterpart.

Family	Semantic modality	SN7	MME	PAN	CB	URB	CL	**Ours**
Function	Built volume (BV)	△	–	△	–	–	–	✓
	Road network (RD)	–	–	–	✓	–	–	✓
	Land cover (LC)	–	✓	✓	–	△	–	✓
	Land use (LU)	–	–	–	–	✓	–	✓
	Building height & footprint (H&F)	✓	–	✓	–	△	✓	✓
	Points of interest (POI)	–	–	–	✓	–	–	✓
Activity	Electricity (ELEC)	–	–	–	–	–	–	✓
	Nighttime lights (NTL)	–	–	–	–	–	–	✓
	Carbon emissions (CO_2)	–	–	–	–	–	–	✓
	Mobility (GPS)	–	–	–	✓	–	△	✓
	Population (POP)	–	–	–	✓	–	✓	✓
	Property value (PROP)	–	–	–	–	–	✓	✓
Environment	Optical imagery (OPT)	✓	✓	✓	✓	✓	✓	✓
	Terrain (DEM)	–	✓	–	–	–	–	✓
	Vegetation (NDVI)	–	△	△	–	–	–	✓
	Weather (WTH)	–	△	–	–	–	–	✓
	Surface temperature (HEAT)	–	–	–	–	–	–	✓
	Air quality (AQ: NO_2; $PM_{2.5}$)	–	–	–	–	–	–	✓
Coverage	Temporal extent	2017–2020	2017–2020	–	–	2022–2023	–	2000–2025
Matched categories (direct + related)		**2+1**	**3+2**	**3+2**	**5+0**	**2+2**	**4+1**	**18+0**

Notes. SN7: SpaceNet 7 (Van Etten et al., 2021); MME: MMEarth (Nedungadi et al., 2024); PAN: PANGAEA (Marsocci et al., 2025); CB: CityBench (Feng et al., 2025c); URB: UrBench (Zhou et al., 2025); CL: CityLens (Liu et al., 2026).
Grid context is provided as background information during training and is not included in the 18 translation modalities.

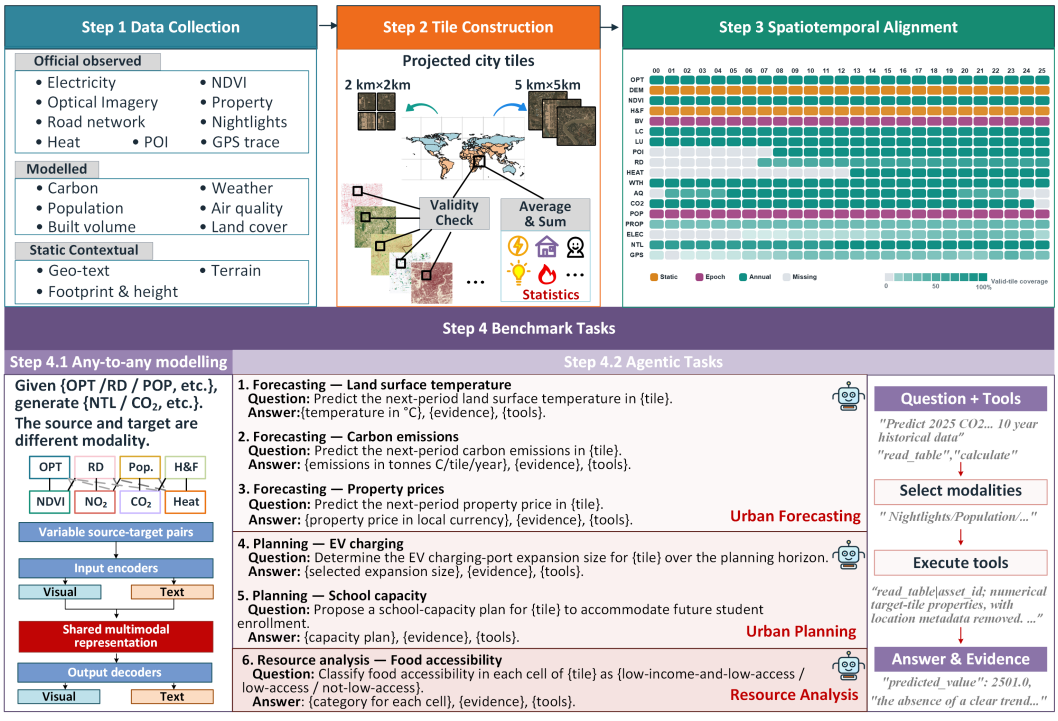


Figure 2: Overview of AgenticCity-Bench. The framework comprises four steps: (1) Data Collection, which gathers multimodal urban data; (2) Tile Construction, which organizes data into standardized spatial tiles; (3) Alignment, which aligns spatial and temporal data; and (4) Benchmark Tasks, which evaluate any-to-any multimodal translation and agentic urban tasks.

explicitly missing intermediate years outside an epoch, no interpolation, reuse, or synthetic values are introduced; the corresponding observations are marked as missing.

Validity check. Data availability is determined separately for each modality and tile. Across all city-year-tile instances and modalities, approximately 84.255% of observations are available. Missing or invalid observations are retained as unavailable rather than imputed. Details about data availability and quality control are provided in Appendix B.4.

Visual and statistical data. For visual data, optical images use RGB bands without per-tile contrast adjustment. Roads, land cover, and other categorical maps use fixed colour schemes. Category labels remain discrete and are not averaged. Numerical inputs summarize valid cells within each tile. Radiance, temperature, and population density use area-weighted means. Property values are converted to 2017 international dollars for consistency. For datasets that report cell totals, values are summed and divided by tile area to obtain densities. Data-specific bands, units and spatial and temporal aggregation rules are detailed in the Appendix B.4.

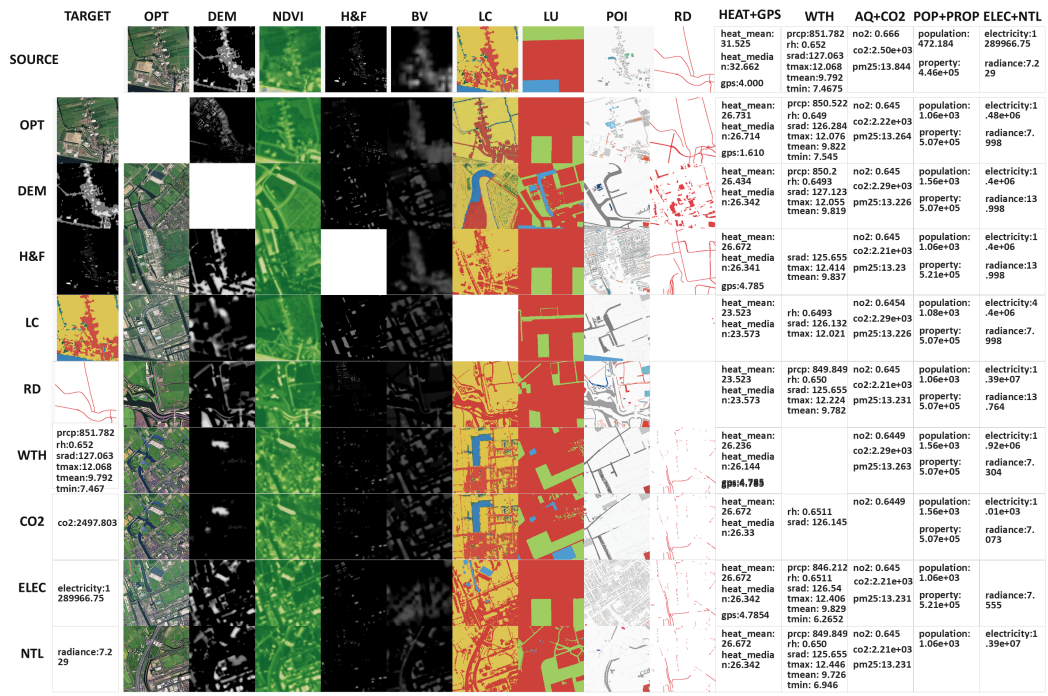


Figure 3: Selected qualitative results for any-to-any cross-modal translation. Rows denote source modalities and columns denote target modalities. Same-modality pairs are omitted. The complete 18×18 any-2-any translation matrix is provided in Appendix C.2.

2.3 Benchmark Tasks

To systematically exploit the heterogeneous observations in AgenticCity-Bench, we construct two benchmark tracks. The first focuses on any-to-any generation across modalities. The second focuses on city agentic tasks from longitudinal records.

2.3.1 Any-to-Any Urban-State Translation

AgenticCity-Bench is a unified benchmark to incorporate imagery, maps, and numerical records spanning urban function, activity, and environment, enabling bidirectional evaluation across modalities. Cross-modal translation generates a requested target modality from an available source modality. Each instance is defined within the same city, spatial scale, tile, and year, with source and target modalities $m_s, m_t \in \mathcal{M}$ and $m_s \neq m_t$, where $\mathcal{M}$ is the set of 18 modalities. The model generates the target from the source and its geographic and temporal context: $\hat{x}^{m_t} c = f_\theta\left(x_c^{m_s}, c\right)$, where c denotes the city, spatial scale, tile, geographic coordinates, and year, and $x_c^{m_s}$ is the source record.

Training pairs and supervision. Source-target pairs are matched by city, tile scale, grid location, and year. Within each tile, any two valid modalities can serve as source and target in either direction.

Modalities are excluded from pairing when: (1) the modality is missing or contains invalid numerical values; (2) Property, GPS and Electricity records must contain positive values, while valid zeros in other modalities, such as road and categorical data, are retained. In Fig.2, for training stage, model-specific input adapters encode images or text-formatted numerical records into a shared multimodal representation. Output adapters generate the requested image or numerical record, using the corresponding target as supervision. This pairwise training setup supports different cross-modal translation directions within a unified interface.

2.3.2 AGENTIC TASKS

The agentic benchmark comprises three task categories: forecasting, urban planning, and resource analysis. Each task requires the agent to plan its own workflow, select modalities, and use the retrieved information to make a prediction or spatial decision. Forecasting tasks target next land surface temperature, carbon emissions, and property prices, using historical observations indexed by relative periods $-10, \ldots, -1$ to predict period 0. Planning tasks include EV charging port expansion and school capacity planning. The resource analysis task contains food-access analysis, which assigns each of four $1\,\mathrm{km} \times 1\,\mathrm{km}$ cells within a target tile to one of three categories: low-income-and-low-access, low-access, or not-low-access. Detailed task settings are given in D.1 and D.3.1.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Any-to-any baseline We fine-tune and evaluate four multimodal models on AgenticCity-Bench: MODUS (7B) Ye et al. (2026), UniGen (Qwen2.5-1.5B backbone)Tian et al. (2026), 4M-21-XL (3B) Mizrahi et al. (2023), Show-o2 (1.5B) Xie et al. (2025). All baselines are fine-tuned end-to-end from their released checkpoints. We update the full backbone, and no additional urban-specific adapter is introduced. As shown in Table 14, the dataset contains more than 28 million possible samples. We select 269,472 samples and split them into training, validation, and test sets with 172,032, 48,720, and 48,720 samples, respectively.

Agentic Baselines We evaluate frontier general agents OpenCode (OpenCode Contributors, 2026) and Codex CLI (OpenAI, 2026a), alongside domain-specific Earth Agent (Feng et al., 2026b). The task model configurations use GPT-5.6 Luna and GPT-5.6 Terra (OpenAI, 2026c;d), GPT-5.5 (OpenAI, 2026b), DeepSeek-V4.1-Flash (DeepSeek-AI, 2026; Guo et al., 2025), and Gemini 3.1 Flash-Lite (Google, 2026); within each task, configurations share the selected cases and allowed observation scope. The EV charger, school capacity planning, and food access experiments expose the same task-specific tools across their respective agent configurations, and the forecasting experiments compare Earth Agent's analytical tools with the native tools of OpenCode and Codex.

Agentic task cases. Our agentic task comprises 1,000 cases spanning forecasting, urban planning, and resource analysis. Forecasting includes 220 land surface temperature, 220 carbon emission, and 160 property-price cases across 15, 16, and 8 cities, respectively. These cover 120 tile–target pairs across 106 distinct $2\,\mathrm{km} \times 2\,\mathrm{km}$ tiles, with annual forecast origins in 2020–2024 and rolling ten-year observation windows spanning 2010–2023. Planning includes 150 EV charger location cases and 100 school capacity cases. EV cases cover 50 tiles each in Boston, New York City, and San Francisco, predicting net charging port growth from September 2024 to September 2026. School cases cover 100 London schools in distinct tiles (60 primary and 40 secondary), using enrollment and capacity histories from academic years 2021/22–2023/24 to plan additional seats for 2024/25. The sample contains 50 balanced and 50 shortage cases under target-year utilization without expansion. Resource analysis comprises 150 food access cases using 2019 reference labels: 38 each in Boston and Washington, D.C., and 37 each in New York City and San Francisco. Each tile contains four $1\,\mathrm{km} \times 1\,\mathrm{km}$ cells, yielding 600 cell-level assessments.

3.2 ANY-TO-ANY CROSS-MODAL URBAN TRANSLATION

Tables 2 and 3 evaluate translation between any two modalities and provide representative cross-model comparisons. Three main findings follow. (1) Any-to-any translation reveals cross-modal complementarity and translation relationships. This task provides a way to examine how information

Table 2: MODUS any-to-any translation performance across target modalities. NMSE is normalized by the training-set 1st–99th percentile range. NMAE and NMSE values are reported as percentages.

(a) Visual target modalities

	OPT		DEM	NDVI	H&F	BV	LC	LU	RD
Input	MSE↓	LPIPS↓	NMAE (%)↓	NMAE (%)↓	NMAE (%)↓	log1p NMAE (%)↓	macro mIoU↑	macro mIoU↑	clDice↑
OPT	–	–	8.147	13.009	24.976	33.548	0.213	0.580	0.288
DEM	0.115	0.614	–	8.127	27.111	36.542	0.187	0.451	0.257
NDVI	0.115	0.643	23.084	–	28.336	41.056	0.212	0.507	0.258
H&F	0.120	0.477	12.681	12.947	–	31.675	0.200	0.508	0.266
BV	0.113	0.685	9.891	5.900	26.777	–	0.205	0.452	0.257
LC	0.113	0.671	26.987	8.217	21.734	63.800	–	0.451	0.256
LU	0.114	0.630	11.060	13.108	26.782	45.226	0.188	–	0.257
POI	0.117	0.559	22.600	8.026	30.799	43.833	0.199	0.508	0.260
RD	0.109	0.700	5.982	9.974	31.358	54.169	0.196	0.407	–
HEAT	0.124	0.714	17.672	14.370	26.045	41.641	0.190	0.292	0.064
WTH	0.126	0.705	25.116	8.783	32.715	69.074	0.111	0.318	0.096
AQ	0.124	0.762	17.395	13.017	26.908	61.996	0.113	0.260	0.069
CO_2	0.125	0.750	18.781	5.943	32.996	72.625	0.125	0.253	0.094
POP	0.125	0.707	27.716	22.428	26.457	42.583	0.123	0.262	0.089
PROP	0.125	0.733	24.732	9.206	28.929	35.753	0.182	0.368	0.100
ELEC	0.124	0.716	26.461	8.962	27.102	38.565	0.169	0.488	0.077
NTL	0.125	0.741	35.103	15.408	24.198	90.330	0.111	0.334	0.074
GPS	0.126	0.738	29.576	39.560	26.378	37.917	0.106	0.267	0.079

(b) Numerical target modalities

Input	HEAT macro NMSE (%)↓	WTH macro NMSE (%)↓	AQ macro NMSE (%)↓	CO_2 log1p NMSE (%)↓	POP log1p NMSE (%)↓	PROP log1p NMSE (%)↓	ELEC log1p NMSE (%)↓	NTL log1p NMSE (%)↓	GPS log1p NMSE (%)↓
OPT	13.756	9.333	23.656	13.450	28.571	10.287	4.093	19.847	28.800
DEM	7.537	0.756	19.861	9.347	25.000	14.294	24.822	42.301	20.015
NDVI	8.208	12.397	32.573	3.196	33.056	13.134	28.571	14.764	28.255
H&F	13.694	16.015	25.851	8.046	20.229	14.286	14.286	12.200	15.504
BV	14.340	2.356	28.102	19.763	16.262	6.370	10.994	12.605	15.366
LC	15.693	1.442	10.153	1.520	29.078	11.966	12.500	22.715	30.037
LU	20.715	9.327	40.779	7.006	28.571	6.018	56.890	33.880	13.584
POI	6.484	1.847	14.737	30.605	24.785	11.150	37.500	19.628	14.605
RD	9.270	1.024	30.054	17.489	28.207	19.842	37.500	30.562	13.720
HEAT	–	12.538	14.153	23.190	14.885	10.476	30.947	27.971	16.174
WTH	29.719	–	49.125	30.252	12.699	13.815	21.316	11.088	21.540
AQ	18.009	2.128	–	33.890	12.500	14.041	12.789	25.761	13.064
CO_2	6.315	0.667	26.212	–	28.945	17.242	6.425	14.821	29.400
POP	24.948	4.944	27.970	1.078	–	0.855	1.318	10.608	12.716
PROP	15.385	19.973	34.070	4.751	48.012	–	12.500	34.104	12.716
ELEC	14.185	12.063	37.791	9.740	2.947	3.953	–	2.450	8.118
NTL	25.261	12.889	15.405	0.845	4.400	9.019	6.796	–	2.488
GPS	5.503	10.093	41.834	20.613	0.219	4.796	5.947	7.708	–

in one modality can support the generation of another. For example, Table 2 shows that activity modalities such as population, electricity, nighttime lights, GPS density, and carbon emissions perform well on several numerical-to-numerical translations. Population yields better generation results for electricity than continuous visual features such as the road network. (2) Cross-modal translation is directionally asymmetric. Translation performance varies by direction. Visual inputs can predict numerical targets, but numerical inputs are less effective at recovering spatial structure. In Fig. 5, numerical inputs can generate plausible images and land-cover maps, but their spatial layouts may differ from the ground truth. (3) Joint any-to-any training is uneven across modalities. Although H&F and DEM are both encoded as height-related fields, their spatial statistics differ substantially: H&F contains sparse, sharp building boundaries, whereas terrain elevation varies more smoothly. The jointly trained model preserves the boundaries of H&F, but generated DEMs sometimes inherit similarly sharp edges and lose the smooth terrain patterns. The OPT→DEM example in Fig. 3 illustrates this effect. In Table 3, different models perform better on different tasks rather than showing consistent performance across modalities.

In summary, these results highlight that cross-modal translation offers a way to further explore relationships among modalities, while pointing to the need for stronger modality-aware optimization and joint-training strategies.

Table 3: Performance on five representative source-target urban translation routes. – indicates that the model does not natively support image-to-image generation.

Models	NDVI → WTH macro NMSE (%)↓	POI → ELEC NMSE (%)↓	H&F → POP NMSE (%)↓	WTH → OPT MSE↓	OPT → RD clDice ↑
ML-4M-3B (Mizrahi et al., 2023)	77.800	32.300	106.100	0.063	0.139
Show-O2-1.5B (Xie et al., 2025)	53.500	**3.324**	164.200	0.299	0.024
UniGen-1.5B (Tian et al., 2026)	19.000	96.900	87.9	**0.036**	–
MODUS-7B (Ye et al., 2026)	**12.397**	37.500	**20.229**	0.126	**0.288**

3.3 Agentic Tasks

3.3.1 Forecasting

Table 4 compares agent forecasts and historical baselines across 600 cases. Three main findings follow. (1) Agents improve carbon forecasts while consulting multimodal data. Across the five agent configurations, carbon emission MAE decreases by a median of 112.08 tonnes C/tile/year relative to the last-year baseline, a 17.334% reduction. Tool execution records show that agents use statistical machine learning tools to interpret target history trends and, in some cases, incorporate other urban modalities and indicators (e.g., nighttime lights and satellite imagery) to refine the final estimate (Appendix case 18 and 17). (2) Historical averaging remains effective for LST. The ten-year mean outperforms all evaluated agents, suggesting that multi-year averaging captures persistent local thermal conditions while smoothing annual fluctuations. Agents do not consistently adapt their forecasting strategies to these temporal differences: carbon emissions may require greater emphasis on recent activity-related changes, whereas LST may benefit more from long-term historical patterns. (3) Multimodal reasoning does not consistently translate into forecast adjustments in complex tasks. According to the error analysis, we find that agents can identify relevant urban context, but often use it to check or justify a historical estimate rather than quantitatively revise their answer. The inspected cases reveal a gap between interpreting additional modalities and using them to calibrate forecasts (Appendix D.3.2).

Table 4: Forecasting performance on 600 cases: 220 LST, 220 carbon, and 160 property cases

Method	LST MAE (°C) ↓	Carbon MAE (tonnes C/tile/year) ↓	Property MAPE (%) ↓
Earth Agent (GPT Luna)	1.436	548.58	9.47
Earth Agent (DeepSeek)	1.400	**511.62**	8.64
Codex (GPT-5.5)	1.341	524.21	**8.47**
OpenCode (GPT Luna)	1.354	534.51	9.43
OpenCode (Gemini Flash Lite)	1.470	584.56	9.59
Last-year value ($t-1$)	1.677	646.59	8.67
Ten-year mean	**1.313**	2,169.03	22.41

3.3.2 Planning

Table 5 summarizes both planning tasks. For EV charger expansion planning, agents outperform the constant-class baseline by up to 16.7 percentage points in accuracy. Successful cases combine multimodal data such as population and POI with charging histories to support decisions (Table 32). However, agents often underestimate the real requirement due to focusing too much on the historical trend instead of considering the increasing need for EV facilities as cities develop. For school capacity planning, agents improve balance rate over the three-year average baseline by up to 4 percentage points in balance rate, showing the agent's ability to translate recent enrollment trends into capacity adjustments. Incorporating more domain-specific knowledge (e.g., age structure) has the potential to improve agents' task-specific ability (Department for Education, 2026).

Table 5: Agentic planning results for 150 EV charging port cases and 100 school capacity cases. The school balanced rate uses a 95–105% enrollment to capacity range. Higher is better for all metrics.

	EV charging		School capacity	
Method	Accuracy (%)	Macro-F1	Balance rate (%)	Mean balance
Constant-class baseline	25.3	0.101	–	–
Three-year-mean baseline	–	–	84.0	96.60
Earth Agent (GPT Terra)	35.3	0.339	81.0	96.91
Earth Agent (DeepSeek)	**42.0**	0.382	**88.0**	**97.50**
Codex (GPT-5.5)	40.7	**0.387**	84.0	97.20
OpenCode (GPT Terra)	36.0	0.333	80.0	96.74
OpenCode (Gemini)	33.3	0.282	84.0	97.45

3.3.3 Resource Analysis

Table 6 reports performance on 150 food access analysis cases, each containing four spatial cells. Identifying joint low-income and low-access areas is challenging; for example, Gemini identifies only 2 of 62 such cells correctly. Inspected agent reasons reveal that agents sometimes misestimate supermarket accessibility from commercial activity, which shows limitations in agent understanding of modality semantics and spatial scale when translating observations into task-specific judgments. Agents could use multimodal information more effectively by distinguishing direct modalities from indirect clues and checking whether the data supports their decisions.

Table 6: Food access classification on 150 cases. Exact-case accuracy requires all four cells on one tile to be correct.

Method	Cell acc. (%)	Macro-F1	Exact-case acc. (%)
Majority-class baseline	49.67	0.221	32.0
Earth Agent (GPT Terra)	50.00	0.385	33.3
Earth Agent (DeepSeek)	51.33	0.385	14.7
Codex (GPT-5.5)	53.83	**0.475**	37.3
OpenCode (GPT Terra)	53.67	0.429	35.3
OpenCode (Gemini)	**57.00**	0.412	**38.7**

Tool-use traces show that Earth Agent relies more on specialized statistical and machine-learning tools for forecasting, while Codex and OpenCode tend to analyze through code (Appendix D.4.1); more tool calls do not consistently correspond to better performance. For example, in food-access analysis, Earth Agent–DeepSeek averages 28.520 calls per case, compared with 12.907 for Codex, but achieves lower macro-F1 (0.385 versus 0.475) and exact-case accuracy (14.7% versus 37.3%). Across the inspected cases, differences in historical window selection, capacity margins, and interpretation of spatial data are more directly reflected in decisions than the number of tool calls, and additional modalities don't consistently lead to numerical forecast adjustments.

4 Conclusion

We present AgenticCity-Bench, a dataset that integrates 18 urban modalities across 16 cities from 2000 to 2025. It provides a foundation for research on urban understanding and agentic analysis. Our experiments demonstrate the potential for cross-modal translation of urban data, while highlighting the need to better balance different modalities. We further evaluate six agentic city tasks across three categories with 1,000 cases and identify both promising capabilities and persistent limitations. The gap between understanding and action, insufficient domain knowledge, and improper tool use can be improved in the future. Researchers can build on our dataset and benchmark to develop more effective urban foundation models and urban agents to provide suggestions to policymakers and advance the development of urban science and the Earth domain.

5 AI DISCLOSURE

The authors declare that artificial intelligence (AI) tools, specifically ChatGPT, were used solely to assist with language polishing and grammar checking of the manuscript text. All intellectual content, data analysis, interpretations, and conclusions were conceived, written, and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Temitope Akinboyewa, Zhenlong Li, Huan Ning, and M. Naser Lessani. GIS Copilot: Towards an autonomous gis agent for spatial analysis. *International Journal of Digital Earth*, 18(1):2497489, 2025. doi: 10.1080/17538947.2025.2497489.

Luís MA Bettencourt and Nicholas Marchio. Infrastructure deficits and informal settlements in sub-saharan africa. *Nature*, 645(8080):399–406, 2025.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026. URL https://arxiv.org/abs/2609.19969. Model card: https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash. Accessed September 26, 2026.

Department for Education. School Capacity: Academic Year 2024/25, 2026. URL https://explore-education-statistics.service.gov.uk/find-statistics/school-capacity/2024-25. Published March 19, 2026; updated March 20, 2026. Accessed September 25, 2026.

Department for Education. Get Information about Schools, n.d. URL https://www.get-information-schools.service.gov.uk/. Continuously updated establishment register. Accessed September 25, 2026.

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 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 591–602, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736878. URL https://doi.org/10.1145/3711896.3736878.

Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. UrbanLLaVA: A multi-modal large language model for urban intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6209–6219, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Feng_UrbanLLaVA_A_Multi-modal_Large_Language_Model_for_Urban_Intelligence_ICCV_2025_paper.html.

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 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5413–5424, 2025c. doi: 10.1145/3711896.3737375.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, Lei Bai, Conghui He, and Weijia Li. Earth-Agent: Unlocking the full landscape of earth observation with agents. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=dkIXAbWuxO.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, Lei Bai, Conghui He, et al. Earth-agent: Unlocking the full landscape of earth observation with agents. In *International Conference on Learning Representations*, volume 2026, pp. 55700–55764, 2026b.

Google. Gemini 3.1 Flash-Lite. Model documentation, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite. Accessed September 23, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Xixuan Hao, Yutian Jiang, Jiabo Liu, Yihang Yang, Guangyin Jin, Song Gao, and Yuxuan Liang. Multi-agent collaborative reasoning with tool-augmented evidence for urban region profiling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 1590–1601, 2026.

Anatol Helfenstein, Vera L Mulder, Gerard BM Heuvelink, and Mirjam JD Hack-ten Broeke. Three-dimensional space and time mapping reveals soil organic matter decreases across anthropogenic landscapes in the netherlands. *Communications Earth & Environment*, 5(1):130, 2024.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. TerraMind: Large-scale generative multimodality for earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jakubik_TerraMind_Large-Scale_Generative_Multimodality_for_Earth_Observation_ICCV_2025_paper.html.

David Kaufmann, Michael Wicki, Stefan Wittwer, and Jake Stephan. Democratic discrepancies in urban sustainable development. *Nature Sustainability*, 7(11):1409–1418, 2024.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiaoxiang Zhu. GEO-Bench: Toward foundation models for earth monitoring. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a0644215d9cff6646fa334dfa5d29c5a-Abstract-Datasets_and_Benchmarks.html.

Siqi Lai, Yansong Ning, Zirui Yuan, Zhixi Chen, and Hao Liu. USTBench: Benchmarking and dissecting spatiotemporal reasoning capabilities of LLMs as urban agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ETzBStUFJy.

Kaisan Li, Shengyu Tao, Zuxun Xiong, Ruojun Ren, Yuhang Zhang, Rui Miao, He Qi, Borong Lin, Jian Kang, Yi Jiang, et al. Facilitating megacity electricity decarbonization via grid-interactive demand-side resource management. *Nature Communications*, 2026a.

Tian Li, Zhuosen Wang, Christopher Kyba, Miguel O Román, Karen C Seto, Yun Yang, Shi Qiu, Theres Kuester, Michail Fragkias, Xiang Chen, et al. Satellite imagery reveals increasing volatility in human night-time activity. *Nature*, 652(8109):379–386, 2026b.

Zhuohong Li, Linxin Li, Ting Hu, Mofan Cheng, Wei He, Tong Qiu, Liangpei Zhang, and Hongyan Zhang. Satellite mapping of every building's function in urban china reveals deep built environment disparities. *Nature Communications*, 17(1):2827, 2026c.

Shujie Liao, Wei Pan, Li Wen, Rongkai Chen, Dongyang Pan, Renjie Wang, Cheng Hu, Hongbo Duan, Hong Weng, Chenxiao Tian, et al. Temperature-related hospitalization burden under climate change. *Nature*, 644 (8078):960–968, 2025.

Tianhui Liu, Hetian Pang, Xin Zhang, Tianjian Ouyang, Zhiyuan Zhang, Jie Feng, Yong Li, and Pan Hui. CityLens: Evaluating large vision-language models for urban socioeconomic sensing. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kswX9NfAlo.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020. doi: 10.1109/TGRS.2020.2988782.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. Pangaea: Assessing geospatial foundation models capabilities through a global and inclusive benchmark. *IEEE geoscience and remote sensing magazine*, 2025.

Timon McPhearson, Steward T. A. Pickett, Nancy B. Grimm, Jari Niemelä, Marina Alberti, Thomas Elmqvist, Christiane Weber, Dagmar Haase, Juergen Breuste, and Salman Qureshi. Advancing urban ecology toward a science of cities. *BioScience*, 66(3):198–212, 2016. doi: 10.1093/biosci/biw002.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36: 58363–58408, 2023.

Dominik J Mühlematter, Lin Che, Ye Hong, Martin Raubal, and Nina Wiedemann. Urbanfusion: Stochastic multimodal fusion for contrastive learning of robust spatial representations. In *Forty-third International Conference on Machine Learning*, 2025.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/8085_ECCV_2024_paper.php.

ODIAC Project. Open-Data Inventory for Anthropogenic Carbon Dioxide (ODIAC): Data Product, n.d. URL https://www.odiac.org/data-product.html. Data-product documentation. Accessed September 25, 2026.

OpenAI. Codex CLI. Software documentation, 2026a. URL https://learn.chatgpt.com/docs/cli. Version 0.155.1 used in the experiments. Accessed September 23, 2026.

OpenAI. GPT-5.5. Model documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.5. Accessed September 23, 2026.

OpenAI. GPT-5.6 Luna. Model documentation, 2026c. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed September 23, 2026.

OpenAI. GPT-5.6 Terra. Model documentation, 2026d. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra. Accessed September 23, 2026.

OpenCode Contributors. OpenCode: The open source coding agent. Software, 2026. URL https://github.com/anomalyco/opencode. Version 1.18.31 used in the experiments. Accessed September 23, 2026.

Daniel T O'Brien. The pointillistic city and geographic scale in urban science. *Nature Cities*, 2(5):379–386, 2025.

Yang Su, Yidi Xu, Xianglin Zhang, David Makowski, Agnes Pellissier-Tanon, Saverio Francini, Tianqi Shi, Ke Yu, Siyu Liu, Haotian Chen, et al. Disturbance characteristics and forest properties regulate surface warming and recovery across europe. *Nature Geoscience*, pp. 1–9, 2026.

Kailai Sun, Mingyi He, Heye Huang, Can Rong, Alok Prakash, Baoshen Guo, Shenhao Wang, and Jinhua Zhao. Sense: Satellite-based energy synthesis for sustainable environment. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 12030–12041, 2026.

Max Tesselaar, WJ Wouter Botzen, Timothy Tiggeloven, and Jeroen CJH Aerts. Flood insurance is a driver of population growth in european floodplains. *Nature Communications*, 14(1):7483, 2023.

Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *Advances in Neural Information Processing Systems*, 38:152386–152415, 2026.

U.S. Department of Agriculture, Economic Research Service. Food Access Research Atlas: 2019 Data, 2021. URL https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data. 2019 reference data, updated April 27, 2021; subsequently listed as the Large Retailer Access Map. Accessed September 25, 2026.

U.S. Department of Energy. Alternative Fuels Data Center: Alternative Fueling Station Locator, n.d. URL https://afdc.energy.gov/stations. Station database and historical station-data exports. Accessed September 25, 2026.

U.S. Geological Survey. Landsat Collection 2 Surface Temperature, n.d. URL https://www.usgs.gov/landsat-missions/landsat-collection-2-surface-temperature. Data-product documentation. Accessed September 25, 2026.

Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development SpaceNet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6398–6407, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Van_Etten_The_Multi-Temporal_Urban_Development_SpaceNet_Dataset_CVPR_2021_paper.html.

Yiheng Wang, Tianyu Wang, Yuying Zhang, Hongji Zhang, Haoyu Zheng, Guanjie Zheng, and Linghe Kong. UrbanDataLayer: A unified data pipeline for urban science. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0db7f135f6991e8cec5e516ecc66bfba-Abstract-Datasets_and_Benchmarks_Track.html.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Xia_DOTA_A_Large-Scale_CVPR_2018_paper.html.

Aoran Xiao, Shihao Cheng, Yonghao Xu, Yexian Ren, Hongruixuan Chen, and Naoto Yokoya. Geommbench and geommagent: Toward expert-level multimodal intelligence in geoscience and remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 47490–47518. Curran Associates, Inc., 2025. doi: 10.52202/085713-1584. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/43e8fd8b9581faa71a6a61602bc28435-Paper-Conference.pdf.

Yichen Xu, Song Gao, Qunying Huang, Aslıgül Göçmen, Qiang Zhu, and Feng Zhang. Predicting human mobility flows in cities using deep learning on satellite imagery. *Nature Communications*, 16(1):10372, 2025.

Weihao Xuan, Junjue Wang, Heli Qi, Zihang Chen, Zhuo Zheng, Yanfei Zhong, Junshi Xia, and Naoto YOKOYA. Dynamicvl: Benchmarking multimodal large language models for dynamic city understanding. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-1114. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2fb80f455dc6056422c8af87dca50489-Paper-Datasets_and_Benchmarks_Track.pdf.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Winston Yap, Abraham Noah Wu, Clayton Miller, and Filip Biljecki. Revealing building operating carbon dynamics for multiple cities. *Nature Sustainability*, 8(10):1199–1210, 2025.

Mingqiao Ye, Zhaochong An, Zhitong Gao, Xian Liu, François Fleuret, Chuan Li, Amir Zadeh, Serge Belongie, Afshin Dehghan, Jesse Allardice, David Mizrahi, Oğuzhan Fatih Kar, Roman Bachmann, and Amir Zamir. Modus: Decoder-only any-to-any modeling of diverse modalities. *arXiv preprint arXiv:2607.25948*, 2026.

Zhen Yu, Philippe Ciais, Shilong Piao, Richard A Houghton, Chaoqun Lu, Hanqin Tian, Evgenios Agathokleous, Giri Raj Kattel, Stephen Sitch, Daniel Goll, et al. Forest expansion dominates china's land carbon sink since 1980. *Nature Communications*, 13(1):5374, 2022.

Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. CityFlow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *The World Wide Web Conference*, pp. 3620–3624, 2019. doi: 10.1145/3308558.3314139.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

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*, volume 39, pp. 10707–10715, 2025. doi: 10.1609/aaai.v39i10.33163.

Andrew Zimmer, Nina Brooks, Andrea E Gaughan, and Cascade Tuholske. Global divergence in urban demographic change and migration patterns. *Nature Cities*, pp. 1–9, 2026.

A Related Work

Multimodal and Temporal Urban and EO Benchmarks Research on EO and urban benchmarks can be organized around three complementary goals: learning from annotated scenes, transferring across heterogeneous observations, and understanding urban conditions across modalities and time (Xia et al., 2018; Nedungadi et al., 2024; Xuan et al., 2025; Feng et al., 2025b). Early task-focused resources established reusable supervision: DOTA provided oriented object annotations for aerial detection, while RSVQA paired remotely sensed imagery with natural-language questions and answers (Xia et al., 2018; Lobry et al., 2020). Subsequent resources broadened both training and evaluation, with geographically aligned multimodal pretraining data and curated task suites for assessing transferable representations (Nedungadi et al., 2024; Lacoste et al., 2023). Temporal benchmarks introduced a different axis of coverage, progressing from monthly building tracking to multi-year analyses of urban transformation (Van Etten et al., 2021; Xuan et al., 2025). In parallel, urban data pipelines and instruction benchmarks have connected imagery with structured geospatial and mobility data, extending evaluation beyond surface appearance to cross-modal urban understanding (Wang et al., 2024; Feng et al., 2025b).

These advances highlight the challenge of jointly modeling heterogeneous urban data across modalities and time (Wang et al., 2024; Feng et al., 2025b; Xuan et al., 2025). Generative EO models can synthesize missing modalities from available observations, but integrating environmental observations with urban activity records over time remains challenging (Jakubik et al., 2025).

Geospatial Agents for Urban Analysis and Planning Agentic geospatial research brings together simulation-based decision environments, language-guided analytical workflows, and increasingly process-aware evaluation (Zhang et al., 2019; Akinboyewa et al., 2025; Feng et al., 2026a; Lai et al., 2026). Before the rise of LLM agents, CityFlow already provided a large-scale traffic simulation environment and interfaces for multi-agent reinforcement learning, establishing an action-and-feedback setting for urban control (Zhang et al., 2019). In the broader language-agent literature, ReAct subsequently articulated how interleaving reasoning with actions allows models to obtain external information and update their plans (Yao et al., 2023). Domain-specific agentic systems have adapted these capabilities to GIS workflow generation, multimodal scientific analysis, and the coordination of retrieval, perception, and reasoning tools (Akinboyewa et al., 2025; Feng et al., 2026a; Xiao et al., 2026). Urban evaluation has developed alongside this analytical strand: CityBench combines heterogeneous city data with interactive simulation, while USTBench diagnoses understanding, forecasting, planning, and reflection within urban tasks (Feng et al., 2025c; Lai et al., 2026).

These developments make it possible to examine how agents reach their results (Feng et al., 2026a; Lai et al., 2026). However, failures in tool use, domain-specific reasoning, and long-horizon adaptation show that successful perception does not guarantee reliable analysis or decision-making (Feng et al., 2026a; Xiao et al., 2026; Lai et al., 2026). Agents must therefore select relevant information, identify missing prerequisites, and construct spatial plans under explicit constraints.

B Benchmark Illustration

B.1 Study Regions and Grid Geometry

The benchmark covers 16 city-centered regions selected to span six inhabited continents and different urban forms, climates, and data environments. Each region is projected to a local metric coordinate reference system before tiling. Separate, non-overlapping grids are constructed at 2-km and 5-km scales; the 5-km grid is not produced by aggregating the 2-km grid. A tile is identified by city, scale, row, and column, which provides a stable spatial key for joining heterogeneous sources without implying that their native resolutions are identical. Table 7 reports the projected grid geometry and the number of tiles retained at each scale.

B.2 Modality Sources

The 18 semantic modalities combine satellite observations, derived EO products, geographic databases, model-based estimates, and city-specific official records. Table 8 records the primary data source and access for each modality.

Table 7: Multi-scale city-grid inventory across 16 cities. Grid dimensions are rows×columns and parenthetical values give cell counts. Each scale uses its own projected grid and is evaluated separately.

City	Continent	CRS	2 km grid (cells)	5 km grid (cells)
Amsterdam	Europe	EPSG:32631	10 × 14 (140)	3 × 4 (12)
Boston	North America	EPSG:32619	13 × 13 (169)	4 × 4 (16)
Busan	Asia	EPSG:32652	16 × 13 (208)	5 × 5 (25)
Cape Town	Africa	EPSG:32734	14 × 13 (182)	4 × 5 (20)
Christchurch	Oceania	EPSG:32759	30 × 32 (960)	11 × 12 (132)
Washington, D.C.	North America	EPSG:32618	13 × 11 (143)	5 × 4 (20)
London	Europe	EPSG:32630	24 × 30 (720)	9 × 11 (99)
Lyon	Europe	EPSG:32631	11 × 13 (143)	3 × 4 (12)
New York City	North America	EPSG:32618	26 × 24 (624)	10 × 9 (90)
San Francisco	North America	EPSG:32610	17 × 19 (323)	6 × 6 (36)
Santiago	South America	EPSG:32719	13 × 14 (182)	4 × 5 (20)
Shanghai	Asia	EPSG:32651	13 × 13 (169)	5 × 4 (20)
Singapore	Asia	EPSG:32648	20 × 32 (640)	7 × 12 (84)
Sydney	Oceania	EPSG:32756	13 × 13 (169)	5 × 4 (20)
Tokyo	Asia	EPSG:32654	17 × 17 (289)	6 × 6 (36)
XiongAn	Asia	EPSG:32650	25 × 31 (775)	9 × 12 (108)
Total	Six continents	-	5,836	750

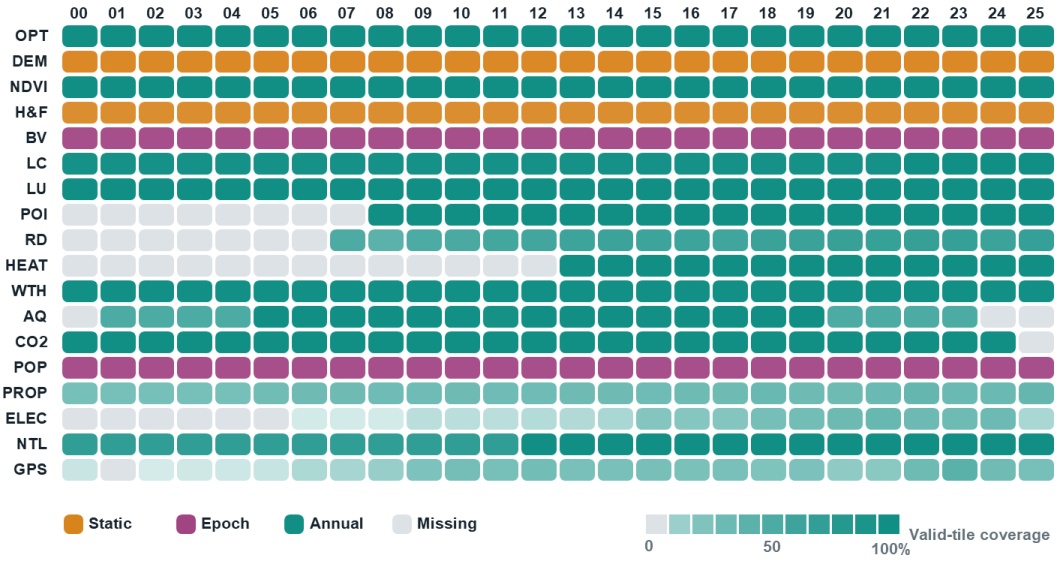


Figure 4: **Temporal availability of the 18 benchmark modalities from 2000 to 2025.** Columns denote calendar years and rows denote modalities. Color intensity indicates the proportion of valid tiles across the 16 cities.

B.3 Data Availability

Coverage is summarized at two distinct levels. Figure 4 and Tables 10,11 describe coverage for each city, year, and modality. A combination is considered available when the release contains at least one valid tile for the specified city, calendar year, and semantic modality.

Across the 16 cities and 26 years from 2000 to 2025, the dataset includes 18 modalities for the any-to-any translation tasks, indexed by city, year, and modality. In addition, a grid-context modality is provided only during training to supply spatial background information and is not used as a source or target in the translation tasks. The grid context is available for all cities and is therefore excluded

Table 8: Primary source registry for the AgenticCity-Bench modalities.

Modality	Dataset name	Source website
Optical Imagery	Landsat Collection 2 Level-2	https://www.usgs.gov/landsat-missions/landsat-collection-2-level-2-science-products
Optical Imagery	Planet RapidEye(2009–2015) and Mosaics (2016–2025)	https://docs.planet.com/data/imagery/basemaps/
Roads	Open Street Map	https://planet.openstreetmap.org/planet/full-history/
POI	Open Street Map	https://www.openstreetmap.org/
GPS	Open Street Map	https://api.openstreetmap.org/api/0.6/trackpoints
Height & Footprint	3D-GloBFP global building-footprint height	https://essd.copernicus.org/articles/16/5357/2024/
DEM	Copernicus DEM GLO-30	https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM
Built Volume	GHSL GHS-BUILT-V R2023A	https://data.jrc.ec.europa.eu/dataset/ab2f107a-03cd-47a3-85e5-139d8ec63283
Population	GHSL GHS-POP R2023A	https://data.europa.eu/data/datasets/2ff68a52-5b5b-4a22-8f40-c41da8332cfe?locale=en
NDVI	Landsat Collection 2 Level-2	https://www.usgs.gov/landsat-missions/landsat-collection-2-level-2-science-products
Heat	Landsat Collection 2 Level-2 Surface Temperature	https://www.usgs.gov/landsat-missions/landsat-collection-2-surface-temperature
Land Cover	GLC_FCS30D	https://zenodo.org/records/8239305
Land Use	HILDA+ v2	https://doi.org/10.1594/PANGAEA.974335
NO_2	Surface-no2	https://sites.wustl.edu/acag/datasets/surface-no2/
$PM_{2.5}$	SatPM2.5	https://sites.wustl.edu/acag/surface-pm2-5/
Weather	CHELSA v2.1	https://www.chelsa-climate.org/datasets
Weather	Daymet	https://daymet.ornl.gov/getdata
Carbon Emissions	ODIAC	https://www.odiac.org/data-product.html
Nighttime Lights	VIIRS nighttime lights (VNL)	https://eogdata.mines.edu/products/vnl/
Property	City-specific property valuation records, harmonized to 2017 PPP international dollars	https://api.worldbank.org/v2/country/all/indicator/FP.CPI.TOTL?format=json https://api.worldbank.org/v2/country/all/indicator/PA.NUS.PPP?format=json https://gisportal.boston.gov/arcgis/rest/services/Assessing https://www.data.go.kr/data/15126469/openapi.do https://www.capetown.gov.za/ https://services.arcgis.com/neT9SoYxizqTHZPH/arcgis/rest/services/OCFO_ITSPE_view_05212026/FeatureServer/53 https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads https://files.data.gouv.fr/geo-dvf/latest/ https://data.cityofnewyork.us/Housing-Development/Annualized-Rolling-Sales-Update/w2pb-icbu https://data.sfgov.org/Housing-and-Buildings/Assessor-Historical-Secured-Property-Tax-Rolls/wv5m-vpq2 https://data.gov.sg/collections/189 https://www.ura.gov.sg/property-data/private-residential-properties https://www.nsw.gov.au/housing-and-construction/land-values-nsw/resource-library/property-sales-data-guide https://www.reinfolib.mlit.go.jp/
Electricity	City-specific official records.	https://data.boston.gov/dataset/building-emissions-reduction-and-disclosure-ordinance https://citymaps.capetown.gov.za/agsext/rest/services/Theme_Based/ODP_SPLIT_10/MapServer/12 https://www.emi.ea.govt.nz/Retail/Reports/0YUCE0 https://www.gov.uk/government/collections/sub-national-electricity-consumption-data https://data.london.gov.uk/dataset/london-energy-and-greenhouse-gas-inventory-leggi-2ko63 https://www.data.gouv.fr/datasets/donnees-locales-de-consommation-delectricite-de-gaz-naturel-et-de-chaleur-et-de-froid-iris-a-partir-de-2018 https://data.cityofnewyork.us/Environment/NYC-Building-Energy-and-Water-Data-Disclosure-for-/5zyy-y8am https://data.sfgov.org/Energy-and-Environment/Existing-Buildings-Benchmark-Reports/4ua7-5sfx/about_data https://www.ine.gob.cl/herramientas/portal-de-mapas/siedu https://www.ema.gov.sg/resources/singapore-energy-statistics/chapter3 https://data.cityofsydney.nsw.gov.au/datasets/electricity-consumption-by-suburb https://www.kankyo.metro.tokyo.lg.jp/documents/d/kankyo/co2-reduction-dashboard_source_202508

Table 9: Spatial granularity and temporal coverage of the 18 semantic modalities in AgenticCity-Bench.

Semantic modality	Spatial granularity	Temporal coverage
Built volume	∼100 m	2000, 2005, 2010, 2015, 2020, 2025 (5-year epochs)
Road network	Vector geometry / rasterized grid	Source-specific historical snapshots
Land cover	30 m	2000–2022
Land use	Source raster grid	2000–2025
Building height & footprint	Building / raster grid	Static
Points of interest	Point / building-level raster	Source-specific historical snapshots
Electricity	Building / point / administrative unit	City-specific years
Nighttime lights	∼500 m	2012–2025; source-specific historical coverage
Carbon emissions	1 km	2000–2024 (annual)
Mobility (GPS)	GPS sampling points / tile-level density	Source-specific years
Population	∼1 km	2000, 2005, 2010, 2015, 2020, 2025 (5-year epochs)
Property value	Parcel / point / postcode / neighborhood	City-specific years
Optical imagery	∼30 m (Landsat); ∼5 m (Planet)	2000–2025
Terrain	∼30 m	Static
Vegetation	∼30 m	2000–2025 (annual)
Weather	∼1 km	2000–2025 (annual aggregation)
Surface temperature	Landsat-derived spatial support	2000–2025; source-dependent annual/composite observations
Air quality	Source-specific coarse grid	Source-specific annual coverage

*Note.*Epoch-based and static data may be reused across calendar years, but such reuse does not constitute independent annual observations.

Table 10: Data availability by modality in the current 2-km release. Each modality has $16 \times 26 = 416$ possible city-year combinations. Static and epoch-based data are counted as available when explicitly reused as model inputs.

Modality	**Available combinations.**	**Availability**
Optical imagery	416 / 416	100.000%
Terrain	416 / 416	100.000%
Vegetation	416 / 416	100.000%
Footprint & height	364 / 416	87.500%
Built volume	416 / 416	100.000%
Land cover	416 / 416	100.000%
Land use	416 / 416	100.000%
POI	288 / 416	69.231%
Road network	300 / 416	72.115%
Heat exposure	358 / 416	86.058%
Weather	416 / 416	100.000%
Air quality	240 / 416	57.692%
Carbon emissions	400 / 416	96.154%
Population	416 / 416	100.000%
Property economy	232 / 416	55.769%
Electricity	111 / 416	26.683%
Night-time lights	404 / 416	97.115%
GPS density	284 / 416	68.269%
Overall	**6,309 / 7,488**	**84.255%**

from the modality coverage analysis. Static and epoch-based data are reused across the corresponding years when applicable, rather than being treated as independent annual observations. For the 18 translation modalities, there are $16 \times 26 \times 18 = 7{,}488$ possible city, year, and modality combinations in Table 10 . Among these, 6,309 combinations are available, corresponding to an overall availability rate of 84.3%. The main gaps arise from electricity, property, air quality, GPS, POI, and historical road-network records in Fig 4.

Table 11: Data availability by city in the current 2-km release. Each city has $26 \times 18 = 468$ possible year-modality combinations.

City	Availability	City	Availability
Amsterdam	392 / 468 (83.761%)	Boston	413 / 468 (88.248%)
Busan	403 / 468 (86.111%)	Cape Town	395 / 468 (84.402%)
Christchurch	353 / 468 (75.427%)	Washington, D.C.	415 / 468 (88.675%)
London	420 / 468 (89.744%)	Lyon	390 / 468 (83.333%)
New York City	411 / 468 (87.821%)	San Francisco	411 / 468 (87.821%)
Santiago	379 / 468 (80.983%)	Shanghai	316 / 468 (67.521%)
Singapore	418 / 468 (89.316%)	Sydney	420 / 468 (89.744%)
Tokyo	401 / 468 (85.684%)	Xiong'an	372 / 468 (79.487%)
Overall			**6,309 / 7,488 (84.255%)**

Table 12: Visual spatial modalities used by any2any models. Each source GeoTIFF is converted to a tile-aligned PNG with a modality-specific fixed rendering rule.

Modality	Source field(s)	MODUS representation and processing	Unit
Landsat optical imagery	B1=red, B2=green, B3=blue	Native RGB	
Planet optical imagery	B1=red, B2=green, B3=blue	Native RGB	-
Road network	B1=major-road raster	Fixed white-background/red-road image	-
Height& footprint	B1=`height_m`	Fixed-range grayscale PNG: black is zero/background; brighter gray indicates taller buildings. Zero is a valid height.	m
DEM	B1=`elevation_m`	Train-global linear grayscale `dem.png`. Zero is valid elevation.	m
NDVI	B1=`NDVI`	Fixed-range grayscale PNG: darker values indicate lower NDVI and brighter values indicate higher NDVI	1
Land cover	B1=raw land-cover class ID	Fixed categorical-palette PNG of B1 land-cover classes	Class ID
Land use	B1=raw HILDA class ID	Fixed categorical-palette PNG of B1 land use classes; class 99 is no-data.	Class ID
POI	B1= class ID	Fixed categorical-palette PNG of B1 POI classes.	Class ID
Built volume	B1= total built-volume density	`log1p`-scaled grayscale PNG of total built-volume density.	m^3/m^2

Note. B1, B2, and B3 denote the first, second, and third bands of the source GeoTIFF, respectively; band meanings are modality-specific.

B.4 Preprocessing and Alignment

Data availability is also determined using finite-value checks, data-specific no-data values, and coverage masks when available. Image modalities require at least 98% valid pixels, while numerical modalities require valid source support above a predefined threshold. If a modality does not meet these criteria, it is marked as unavailable for the corresponding city, year, tile combination.

Missing observations are represented using explicit availability masks and are neither zero-filled nor replaced with synthetic observations. For example, the electricity data for Washington DC do not cover the surrounding areas and suburbs. Tiles outside the reported coverage are therefore marked as unavailable rather than assigned zero consumption. Conversely, zero is retained when it represents a meaningful observation, including no road in `road network`, ocean in Land Use, and physical zeros in continuous fields.

Tables 12 and 13 summarize the fields used for the 18 urban modalities, their processing, and the corresponding units.

B.4.1 Function modalities Preprocessing

Road network (OSM). In Table 12, we use band 1 of the supplied OSM raster as the road layer. Nonzero pixels are rendered in a fixed red RGB value $(224, 64, 69)$ on a white background. Zero pixels indicate no mapped road and remain valid. Therefore, a tile without mapped roads can still meet the 98% validity threshold. The release converter does not add road buffers, dilate road lines, or

Table 13: Statistical and geo-temporal text modalities used by MODUS. Numerical fields are serialized as typed JSON within the prompt or target text. All raster summaries are computed over valid source support only.

Modality	JSON field(s)	Tile-level processing	Unit
Population_GHSL	`population_density`	Area-weighted mean of the population-density display field.	persons/km^2
Air quality NO2	`annual_mean_no2`	Area-weighted mean of annual surface NO_2.	ppb
Air quality PM25	`annual_mean_pm25`	Area-weighted mean of annual $PM_{2.5}$.	μg/m^3
Temperature_LandsatST	`mean_celsius`; `median_celsius`	Area-weighted means of source B1 (mean) and B2 (median), converted to Celsius. Source p90, p95, and standard-deviation bands are excluded in V1.	$^\circ$C
Weather	`tmin_c`; `tmax_c`; `tmean_c`; `prcp_mm`; `srad_w_m2`; `rh_fraction`	The six shared fields are area-weighted over valid source cells.	`tmin_c`, `tmax_c`, `tmean_c`: $^\circ$C; `prcp_mm`: mm; `srad_w_m2`: W/m^2; `rh_fraction`: %
Property value	`property_mean_2017ppp`	Area-weighted mean property value, expressed in 2017 PPP international dollars	2017 PPP international dollars per property observation
Electricity	`electricity`	Valid source-cell electricity values are summed and divided by tile area.	kWh/km^2
CO_2	`carbon`	Valid source-cell carbon emissions are summed and divided by tile area.	tonne C/km^2/year
Nighttime lights	`radiance`	Area-weighted mean annual nighttime radiance.	nW/cm^2/sr
OSM_GPS_Trace	`gps_trace_density_log1p`;	Aggregated to the tile from GPS-trace point records. The primary value is $\log(1 + \text{points}/\text{km}^2)$.	log1p(points/km^2)
grid	City, country, year, scale, row, column, and WGS84 bounds	Automatic geo-temporal text context. It is always included in the prompt and never selected as a generation target.	–

create new road pixels. Each raster keeps the year or snapshot provided with the original OSM data. A blank road image indicates that no roads are mapped in the supplied data, not that the area has no roads in reality.

Building height & footprint (H&F). In Table 12, band 1 of BuildingHeight_3DGloBFP provides building height in metres. We treat it as static built-environment information, even when the same data appear in multiple year directories. Valid zero values are retained, and an image is included only when at least 98% of its pixels are valid. Heights are mapped linearly to a fixed grayscale range and replicated across the three RGB channels, then saved as a PNG image. The raster encodes height values, including zero-height pixels, rather than providing a separate binary building-footprint layer. The spatial pattern of nonzero pixels may therefore show where buildings are present, but these pixels are not extracted or converted into a separate footprint layer. The PNG preserves the relative height values within the display range, but may lose the precision and extreme values of the original floating-point data.

Built volume (BV). Built volume (GHSL) is provided as 5-year epoch data. We use band 1, which in Table 12 represents total built-volume density in m^3/m^2, and exclude the non-residential band. Zero values are valid and indicate no built volume, while the explicit -9999 value is treated as invalid. After applying the 98% validity filter, we apply a log transform and map the values to a fixed percentile-based grayscale range to produce a PNG image. We use the spatial grid provided in the original GHSL data without additional spatial reconstruction or resampling. According to previous works (Tesselaar et al., 2023; Yu et al., 2022), each 5-year epoch value is reused for the years covered by that epoch and does not represent separate yearly observations.

Land use (LU). In Table 12, band 1 of LU (HILDA) provides categorical land-use IDs. The fixed legend includes ocean, urban land, crop systems, pasture, forest types, unmanaged grass or shrubland, sparse vegetation, and inland water. ID 0 is valid and represents ocean, while ID 99 represents nodata. IDs outside the defined classes are treated as invalid. Images are included only when at least 98% of the pixels are valid and are saved as a PNG image. The class IDs represent discrete categories rather than continuous values, so we use a fixed RGB lookup table instead of continuous interpolation. During generation, each predicted pixel is mapped to the class with the nearest predefined class color.

Building-function POI. POI refers to OSM-derived points of interest and building-function categories rasterized to the study grid. In Table 12, band 1 of POI is a 13-class building-function raster. ID 0 denotes outside/no building, while ID 1 denotes an unlabeled building. IDs 2–12 represent residential, commercial/retail/office, education, healthcare, food service, civic/public, transport/parking, tourism/leisure, industrial/warehouse, religious, and other functions. Both 0 and 1 are valid classes, not missing data. The converter uses the supplied rasterized footprint/function map, applies the 98% valid-pixel threshold, and writes a PNG image using a fixed class palette. It does not reclip polygons or infer labels for unlabeled buildings. Temporal meaning follows the supplied OSM snapshot, and the absence of a class is distinct from missing map data. During generation, each predicted pixel is mapped to the class with the nearest predefined class color.

Land cover and green-space composition (LC). In Table 12, land cover uses the original band-1 IDs. A fixed legend retains crop, forest, shrub, grass, wetland, impervious, bare, water, and snow/ice distinctions and maps them to a PNG image. Thus the image also supports analysis of green-space composition. The current legend accepts ID 0 as non-green or unclassified; this must not be interpreted as the road-background class or automatically as verified non-green land. Accepted images require 98% valid pixels. Temporal assignments are inherited; class IDs are not continuously normalized or interpolated. During generation, each predicted pixel is mapped to the class with the nearest predefined class color.

B.4.2 Activity modalities Preprocessing

Population (POP). Population (GHSL) provides population density in persons/km^2. In Table 13, it provides population density rather than population counts for individual pixels. We use the spatial mean of valid pixels within each tile to obtain the tile-level population density. At least 50% of the cells must be valid, and zero population is retained as a valid value. Population data are provided for specific source epochs, and the same value is reused for the corresponding years in the dataset without additional temporal interpolation (Yu et al., 2022; Tesselaar et al., 2023). A log transform is applied before standardization.

Property economy (PROP). Property values are obtained from two sources. For most cities, we use the existing property-value datasets listed in Table 8. For Shanghai and Cape Town, where comparable gridded datasets are unavailable, records are collected from city-specific property websites.

For existing raster data, we use the mean value. A pixel is valid when it is finite and is not excluded by the source nodata value or raster mask. Let x_i be the stored value of pixel i and $m_i \in \{0, 1\}$ its validity indicator. The valid-support ratio is $r_{\text{valid}} = \frac{\sum_i m_i}{N}$, where N is the number of pixels in the tile. A tile is retained when $r_{\text{valid}} \geq 0.5$, and its property value is computed as $x_{\text{PROP}} = \frac{\sum_i m_i x_i}{\sum_i m_i}$.

All monetary values are converted from the reported local currency and year to 2017 purchasing-power-parity (PPP) international dollars. This harmonization reduces differences caused by currencies, price levels, and purchasing power across cities and years. The released field represents the mean property value within a tile in 2017 PPP international dollars. The original property values are also retained for reference and reproducibility.

Missing values remain unavailable rather than being replaced with zero. The released value remains in physical units; the `log1p` transformation and train split standardization are applied only in the model input pipeline.

Electricity (ELEC). For electricity, we aggregate the valid cell values within each tile by summing them and dividing by the nominal tile area. This produces the tile-level electricity density in

kWh/km^2/year. For partially supported tiles, we do not estimate the contribution from the missing area. The released value remains in its physical unit, while the model input pipeline applies log transformation and standardization. No additional resampling is applied to the original cell values.

Carbon emissions (ODIAC, CO_2). For ODIAC, we aggregate the valid cell values within each tile by summing them and dividing by the nominal tile area. This produces the tile-level fossil-fuel CO_2 emission density in tonne C/km^2/year. Valid zero values are retained. For partially supported tiles, we do not fill or scale up the missing area. The released value remains in its physical unit, while the model input pipeline applies log transformation and standardization. Each value corresponds to the supplied annual ODIAC product, with no additional temporal interpolation.

Nighttime lights (NTL). For Nighttime Lights, we compute the masked spatial mean of band 1 radiance within each tile and serialize the result with units of nW/cm^2/sr. Radiance is averaged rather than summed because it represents an intensity field. Zero radiance is retained as a valid observation. The released value remains in its physical unit, while the model input pipeline applies log transformation and standardization. No additional sensor harmonization or reconstruction of missing historical years is performed.

GPS trace activity (GPS). For GPS traces, we use tile-level attributes from the OSM GPS Trace GeoJSON rather than a rasterized trajectory image. Records are matched to benchmark tiles using their row and column indices or a parseable tile identifier. A tile is considered unavailable when no corresponding GPS feature or valid point-density value is present. The resulting modality represents the density of GPS sampling points captured in the public OSM trace data. It reflects observed GPS trace activity in the available sample and should not be interpreted as the number of unique travelers or trips, or as a direct measure of overall mobility.

B.4.3 Environment modalities Preprocessing

Optical imagery (OPT). In Table 12, optical imagery is treated as RGB images and retain bands 1-3 in the declared RGB order, while Planet band 4 is excluded. The supplied uint8 RGB values are preserved without per-tile contrast stretching and saved as PNG images.

Terrain elevation (DEM). In Table 12, band 1 of DEM provides elevation in metres. We retain valid zero and negative elevations, exclude pixels marked as missing, and require at least 98% valid pixels. A fixed global linear grayscale mapping is used to produce DEM PNG image. As a result, a genuinely flat tile may have low visual contrast. DEM is treated as static terrain information and reused across years rather than as an annual measurement, according to previous work (Helfenstein et al., 2024).

Normalized Difference Vegetation Index(NDVI). In Table 12, band 1 of NDVI (Landsat) is rendered as NDVI PNG image, using a fixed palette over the range $[-1, 1]$. Zero and negative NDVI values are valid observations rather than nodata. Pixel validity is determined before rendering, and at least 98% of pixels must be valid. Finite values outside the $[-1, 1]$ range are clipped during rendering rather than treated as invalid. We use the annual or composite NDVI provided by the source without recomputing it from the red and NIR bands. No logarithmic transformation or tile-specific contrast stretching is applied.

Air quality (AQ). In Table 13, NO_2 and $PM_{2.5}$ are independently aggregated from AirQualityNO2 and AirQualityPM25, respectively. Their coarse source grids are preserved without additional resampling or artificial fine-scale detail. For each product, the tile-level value is computed as the masked spatial mean of valid cells. Valid zero values are retained. If one product is unavailable, it is excluded rather than imputed from the other. The values retain their physical units of ppb for NO_2 and μg/m^3 for $PM_{2.5}$, while separate normalizers apply log transformation and standardization. The annual values follow the supplied annual concentration products.

Land-surface temperature (HEAT). In Table 13, LandsatST uses only bands 1 and 2, which provide the supplied mean and median temperature fields, respectively. We compute the spatial mean of each band. Valid zero degrees Celsius is retained. This annual or composite surface-temperature

product is distinct from weather air temperature and is represented as numerical data rather than an image.

Weather (WTH). In Table 13, the weather representation contains TMIN, TMAX, TMEAN, PRCP, SRAD, and RH. For each tile, we first compute an area-weighted spatial mean over all valid source grid cells. Thus, TMIN, TMAX, and TMEAN are reported in degrees Celsius, SRAD as mean shortwave radiation, and RH as mean relative humidity. PRCP is not spatially summed across grid cells.

C Any-to-Any Translation

C.1 Experiment Details

C.1.1 Any-to-Any Evaluation Metrics

In Table 2, for optical imagery (OPT), we report RGB mean squared error (MSE) and LPIPS to measure pixel-level fidelity and perceptual similarity, respectively. For continuous spatial fields (DEM, NDVI, and H&F), we use normalized mean absolute error (NMAE), while built volume (BV) is evaluated with log1p NMAE to account for its skewed dynamic range. For categorical land-cover (LC) and land-use (LU) maps, we use macro mIoU to measure class-wise spatial overlap while giving equal weight to each class. Road networks (ROAD) are evaluated with clDice, which is better suited to thin and connectivity-sensitive structures.

For reference values y_i and predictions $\hat{y}_i$, we define MSE, NMAE, and normalized mean squared error (NMSE) as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2, \qquad \mathrm{NMAE} = \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|}{\sum_{i=1}^{N}|y_i| + \epsilon},$$

$$\tilde{y}_i = \frac{T(y_i) - p_{01}}{p_{99} - p_{01}}, \qquad \tilde{\hat{y}}_i = \frac{T(\hat{y}_i) - p_{01}}{p_{99} - p_{01}}, \qquad \mathrm{NMSE} = \frac{1}{N}\sum_{i=1}^{N}(\tilde{\hat{y}}_i - \tilde{y}_i)^2.$$

For multi-field numerical targets, including surface temperature (HEAT), weather (WTH), and air quality (AQ), we first compute NMSE independently for each target field using its field-specific percentile range, and then macro-average across the K fields:

$$\text{macro NMSE} = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{N_k}\sum_{i=1}^{N_k}\left(\tilde{\hat{y}}_{ik} - \tilde{y}_{ik}\right)^2\right],$$

where

$$\tilde{y}_{ik} = \frac{T_k(y_{ik}) - p_{01,k}}{p_{99,k} - p_{01,k}}, \qquad \tilde{\hat{y}}_{ik} = \frac{T_k(\hat{y}_{ik}) - p_{01,k}}{p_{99,k} - p_{01,k}}.$$

Here, $p_{01,k}$ and $p_{99,k}$ are the 1st and 99th percentiles of the training distribution for field k. This field-wise normalization prevents variables with larger numerical ranges from dominating the metric, while macro-averaging gives equal weight to each target field.

For strongly skewed non-negative targets, we evaluate in the log-transformed domain by replacing the reference and prediction with

$$y'_i = \log(1 + y_i), \qquad \hat{y}'_i = \log(1 + \hat{y}_i). \tag{1}$$

We refer to the resulting metrics as log1p NMAE and log1p NMSE. We use log1p NMAE for BV and log1p NMSE for ODIAC, population (POP), property value (PROP), electricity (ELEC), nighttime lights (NTL), and GPS activity to reduce the disproportionate influence of extreme values.

For categorical targets, macro mIoU averages IoU equally across semantic classes, while clDice evaluates centerline agreement and is therefore used for road-network structure. LPIPS complements MSE for OPT by measuring similarity in learned perceptual feature space rather than only pixel-wise differences. For every source-to-target route, the metric is determined solely by the target modality and is identical across all source modalities. NMAE and NMSE variants are multiplied by 100 and reported as percentages; MSE, LPIPS, macro mIoU, and clDice are reported on their native scales.

Table 14: Exhaustive one-to-one candidate pairs for 2-km and 5-km scales (28,249,584 directed pairs). Rows indicate sources and columns indicate targets. Panel (a) lists visual target modalities, and panel (b) lists numerical target modalities.

(a)

Source / Target	RS	DEM	NDVI	Height	BV	GLC	LU	POI	Road
RS	–	193,379	203,199	178,798	207,654	167,984	119,315	166,579	159,147
DEM	193,379	–	120,028	104,078	123,578	102,304	72,408	85,554	88,844
NDVI	203,199	120,028	–	111,913	133,144	110,430	77,667	92,491	86,312
Height	178,798	104,078	111,913	–	114,894	94,346	70,077	80,046	74,822
BV	207,654	123,578	133,144	114,894	–	113,390	79,632	94,950	89,072
GLC	167,984	102,304	110,430	94,346	113,390	–	72,492	74,370	69,897
LU	119,315	72,408	77,667	70,077	79,632	72,492	–	49,361	46,921
POI	166,579	85,554	92,491	80,046	94,950	74,370	49,361	–	85,117
Road	159,147	88,844	86,312	74,822	89,072	69,897	46,921	85,117	–
Temp	124,092	61,789	66,897	57,811	68,575	49,580	30,376	68,939	61,945
Weather	196,112	116,784	126,187	108,757	129,457	111,090	79,737	88,757	83,029
AQ	306,563	179,484	194,021	167,699	199,163	182,159	131,631	147,200	138,905
ODIAC	198,038	118,801	127,885	110,451	131,851	113,367	79,632	89,659	84,291
Pop	181,718	114,530	111,106	98,774	114,530	94,300	72,387	79,290	82,232
Prop	66,712	37,008	40,986	42,245	42,063	30,137	27,809	30,544	28,347
Elec	19,027	9,112	9,694	10,353	10,353	6,357	5,095	10,345	9,125
NTL	189,232	114,002	117,604	100,326	121,310	98,282	68,700	89,670	85,179
GPS	163,824	87,403	94,518	81,158	96,918	75,519	53,498	88,996	80,701

(b)

Source / Target	Temp	Weather	AQ	ODIAC	Pop	Prop	Elec	NTL	GPS
RS	124,092	196,112	306,563	198,038	181,718	66,712	19,027	189,232	163,824
DEM	61,789	116,784	179,484	118,801	114,530	37,008	9,112	114,002	87,403
NDVI	66,897	126,187	194,021	127,885	111,106	40,986	9,694	117,604	94,518
Height	57,811	108,757	167,699	110,451	98,774	42,245	10,353	100,326	81,158
BV	68,575	129,457	199,163	131,851	114,530	42,063	10,353	121,310	96,918
GLC	49,580	111,090	182,159	113,367	94,300	30,137	6,357	98,282	75,519
LU	30,376	79,737	131,631	79,632	72,387	27,809	5,095	68,700	53,498
POI	68,939	88,757	147,200	89,659	79,290	30,544	10,345	89,670	88,996
Road	61,945	83,029	138,905	84,291	82,232	28,347	9,125	85,179	80,701
Temp	–	62,882	94,793	63,289	57,265	22,450	9,965	68,575	63,630
Weather	62,882	–	195,363	128,651	109,384	40,421	10,204	113,833	90,912
AQ	94,793	195,363	–	199,125	166,260	61,494	13,680	175,470	148,981
ODIAC	63,289	128,651	199,125	–	110,101	40,116	10,204	116,023	91,815
Pop	57,265	109,384	166,260	110,101	–	37,008	9,112	105,362	81,185
Prop	22,450	40,421	61,494	40,116	37,008	–	9,013	37,634	34,596
Elec	9,965	10,204	13,680	10,204	9,112	9,013	–	10,320	10,067
NTL	68,575	113,833	175,470	116,023	105,362	37,634	10,320	–	89,683
GPS	63,630	90,912	148,981	91,815	81,185	34,596	10,067	89,683	–

C.1.2 Benchmark Instance Construction and Data Splits

To prevent spatial leakage across resolutions, training, validation, and test partitions are defined using a common $10\,\mathrm{km} \times 10\,\mathrm{km}$ parent grid. All $2\,\mathrm{km}$ and $5\,\mathrm{km}$ tiles within the same parent region are assigned to the same partition. All years and modalities associated with these tiles are kept together. Thus, spatially overlapping observations cannot enter different partitions through a change of scale, modality, or year. Candidate pairs are then sampled independently within each partition. For categorical modalities, content diversity is considered during sampling. For example, the coarse spatial support of land use data can produce tiles containing only one class; more diverse tiles are prioritized to provide useful supervision, while single-class tiles are retained at a lower sampling rate rather than removed entirely. Directed source–target routes are assigned approximately balanced sampling quotas, subject to their available pair counts, so that frequent routes do not dominate training and rare routes remain represented. Table 14 shows the number of samples provided by the benchmark. The sample selection procedure produces 269,472 directed translation instances: 172,032 for training, 48,720 for validation, and 48,720 for testing.

The splits are saved for all models. For each target modality, all source modalities are evaluated on the same target-anchored test instances, identified by identical city, spatial scale, tile, and year keys. The target references are therefore fixed within each column of Table 2, and only the source modality changes. This design makes source-wise performance directly comparable for a fixed target.

C.1.3 ANY-TO-ANY EXPERIMENTAL SETTING.

All models were fine-tuned on NVIDIA H200 GPUs with 140 GB of memory. No LoRA modules or parameter-efficient adapters were introduced. All four baselines are fine-tuned from their released checkpoints without additional urban-specific adapters. For MODUS-7B, we freeze the VAE and fine-tune the LLM and unified diffusion transformer with a learning rate of 1×10^{-5}, cosine decay to 5×10^{-7}, and a 500-step warm-up on $3\times$ NVIDIA H200 GPUs. For Show-O2-1.5B, we freeze the Wan2.1 VAE and image and positional embeddings, while fine-tuning the remaining trainable components with learning rates of 5×10^{-6} for the visual encoder and 1×10^{-5} for the projection and generation components, using cosine decay with a 5% warm-up on one H200 GPU. For UniGen-1.5B, we freeze the MagViT-v2 visual tokenizer and fine-tune the Qwen2.5-1.5B backbone and generation layers with a learning rate of 2×10^{-5}, cosine decay to 1×10^{-6}, and a 500-step warm-up on one H200 GPU. For ML-4M-3B, we keep the pretrained modality tokenizers frozen and fully fine-tune the 4M transformer with a learning rate of 3.125×10^{-6} and cosine decay on one H200 GPU.

C.2 DETAILED ANALYSIS OF CROSS-MODAL TRANSLATION

Any-to-any translation reveals cross-modal complementarity and translation relationship. This task provides a way to examine how information in one modality can support the generation of another. Table 2 provides an empirical map of which modalities contain information useful for reconstructing others. Because different targets have different value distributions and prediction characteristics, performance is not directly comparable across different targets. Comparisons are therefore most informative when the target modality is held fixed. Table 2 shows that activity modalities such as population, electricity, nighttime lights, GPS density, and carbon emissions exhibit clear relationships with each other. For example, GPS→POP, POP→ELEC, ELEC→NTL, and NTL→GPS achieve NMSEs of 0.219%, 1.318%, 2.450%, and 2.488%, respectively. For the same target, these activity-related sources also provide useful information compared with other modality types. In particular, POP predicts ELEC with an NMSE of 1.318%, compared with 37.500% for RD→ELEC, indicating that population intensity is considerably more informative about electricity distribution than road geometry alone.

Complementarity also extends across broad modality families. DEM→WTH obtains 0.756% NMSE and ranks second among all inputs for WTH; LC→AQ obtains the best AQ result at 10.153% NMSE; and OPT→ELEC achieves the second-best ELEC result at 4.093% NMSE. These relationships suggest that the model captures a structured dependency graph connecting terrain, land cover, environmental conditions, and human activity, rather than relying only on similarity between modalities of the same type. Nevertheless, these translation relationships should be interpreted as predictive associations rather than causal dependencies.

Cross-modal translation is directionally asymmetric. The information that modality A provides about modality B is not generally recoverable in the reverse direction. Because the two directions are evaluated using different target-specific metrics, their raw scores are not directly comparable; nevertheless, their rankings within the corresponding target columns reveal clear asymmetries. For example, DEM→WTH ranks second among 17 available non-self inputs, with a macro NMSE of 0.756%, whereas WTH→DEM ranks twelfth, with an NMAE of 25.116%. Similarly, LC→AQ is the best route to AQ, with a macro NMSE of 10.153%, whereas AQ→LC ranks fourteenth, with an mIoU of 0.113. Directionality is also evident within the numerical modalities: POP→PROP is the best route to PROP, with a log1p NMSE of 0.855%, whereas PROP→POP is the worst route to POP, with 48.012%.

The asymmetry is especially pronounced for targets that require accurate semantic boundaries or topology. Averaged over the available visual and structural inputs, with unavailable self-translations excluded, the LC, LU, and RD targets obtain scores of 0.200 mIoU, 0.483 mIoU, and 0.260 clDice, respectively. The corresponding averages for numerical inputs are 0.137, 0.316, and 0.082. These descriptive gaps suggest that numerical modalities may constrain the broad appearance, semantic composition, or overall intensity of an output without uniquely determining its fine-scale boundaries and topology. This interpretation is also consistent with Fig. 5, where numerical inputs produce globally plausible images and land-cover maps but do not always recover the ground-truth spatial layout. Conversely, structure-rich inputs preserve local cues that can be aggregated to predict coarser numerical fields.

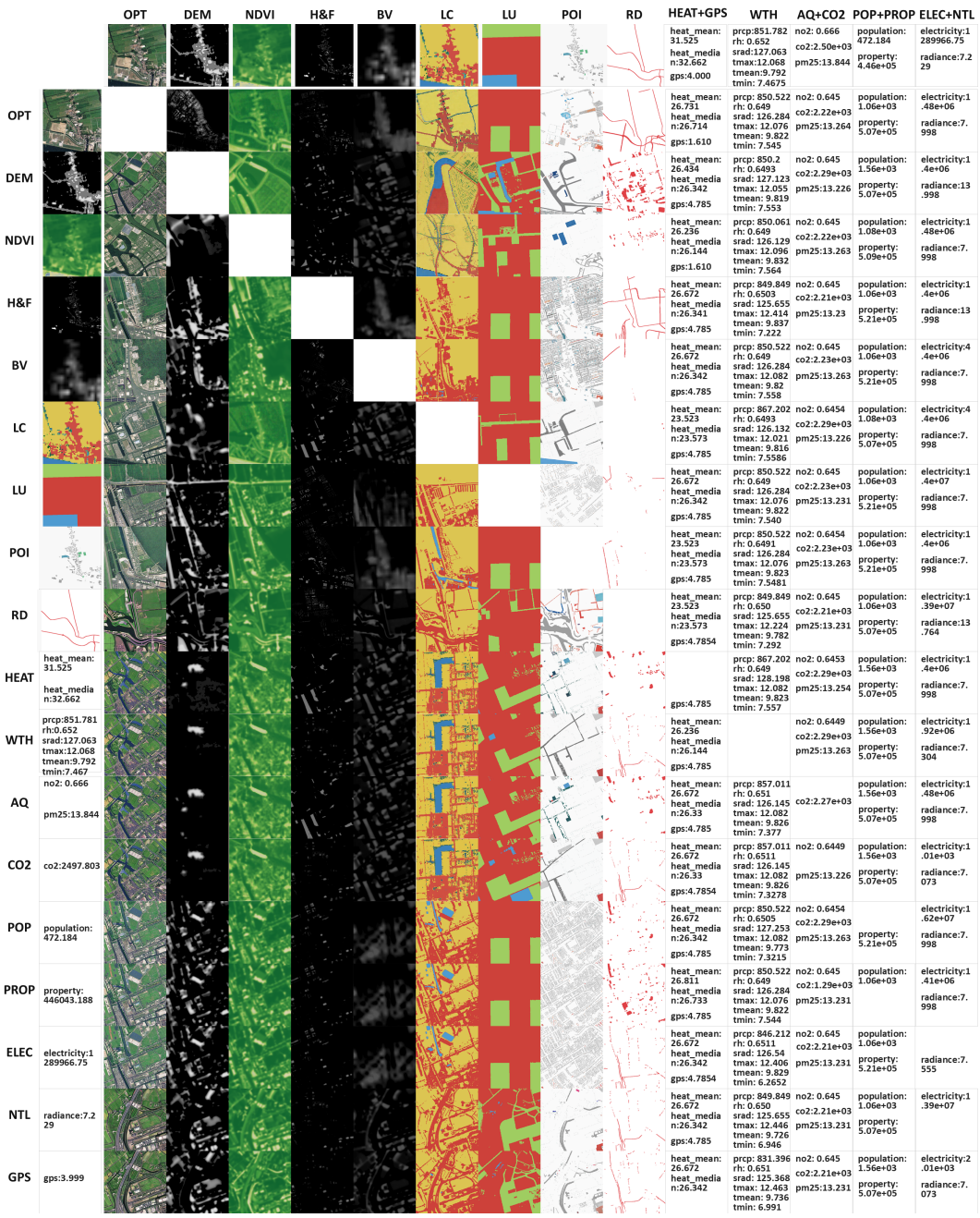


Figure 5: Qualitative any-to-any cross-modal translation across 18 urban modalities. Rows denote source modalities and columns denote target modalities. Same-modality source-target pairs are omitted.

Joint any-to-any training is uneven across modalities. The imbalance in joint any-to-any training does not arise from uneven sampling. The model must jointly optimize targets that differ not only in semantics, but also in spatial frequency, sparsity, output distribution, and evaluation objective. For example, H&F and DEM illustrate this conflict particularly clearly. Although both modalities encode height-related information, H&F is dominated by sparse discontinuities and sharp building contours, whereas DEM primarily contains smooth, low-frequency terrain variation. Shared features and gradients that improve boundary preservation for H&F can therefore bias DEM generation toward high-frequency structures. In the OPT→DEM example in Fig. 5, the generated DEM inherits unnaturally sharp edges and consequently loses part of the smooth terrain pattern.

More generally, joint training must reconcile pixel-wise regression and perceptual reconstruction for continuous images, semantic objectives for LC and LU, and a topology-sensitive objective for RD. Improvements induced by one type of supervision therefore need not transfer uniformly to the others. The results in Table 3, where different models perform best on different translation routes rather than one model consistently dominating all targets, are consistent with such cross-task interference. Joint training consequently provides modality-dependent benefits rather than a uniform improvement across the translation matrix.

These observations motivate a combination of shared and modality-specific optimization. Possible strategies include route-balanced sampling, target-dependent loss normalization or gradient balancing, modality-specific normalization layers, and lightweight output adapters. Frequency-aware objectives may be particularly useful for the height modalities: edge-sensitive supervision can preserve building boundaries in H&F, while smoothness or low-frequency consistency constraints can prevent the same edge prior from leaking into DEM generation.

D AGENTIC TASKS

D.1 CASE CONSTRUCTION AND SELECTION

The case counts and geographic coverage are reported in the main text. Here we describe the eligibility criteria, sampling procedures, and reference outcomes used to construct the evaluation set. Selection operates on task-specific records linked to the benchmark grid: a tile–target pair for forecasting, a tile for EV expansion and food access, and an individual school for capacity planning. Reference outcomes are stored separately from agent-visible observations.

Forecasting. The forecasting targets use the benchmark's annual Landsat surface temperature products (U.S. Geological Survey, n.d.), ODIAC carbon-emission records (ODIAC Project, n.d.), and city-specific property records documented in Appendix B.2. For each city and target, we retain tiles with valid annual target observations throughout 2010–2024 and a consistent measurement unit across years. This complete-panel requirement supports all five forecast origins on the same spatial sample. The initial cohort samples two eligible tiles per city–target combination using a fixed pseudorandom seed of 0. The expanded cohort adds 14 LST, 12 carbon-emission, and 16 property-price tile–target pairs, excluding those already selected for the same target. Additional candidates are ordered by hash with seed `prediction600-seed42` and selected in round-robin order across eligible cities. Selection depends on data availability and city coverage, without ranking candidates by target magnitude. Each selected pair generates five cases; the reference answer is the observed target value at the corresponding forecast origin. This yields 44 LST, 44 carbon-emission, and 32 property-price spatial series. Different targets may share a tile, while each tile–target–year combination remains a distinct case.

EV charging expansion. We use the archived alternative-fueling-station records underlying the EV case package, associated with the Alternative Fuels Data Center station database (U.S. Department of Energy, n.d.). We construct reference outcomes by aggregating public charging-port counts to tiles at the baseline and target dates and taking their difference. These observed changes define four expansion categories using city-specific intervals (Table 15). Sampling is stratified jointly by city and category. Within each stratum, candidates are ordered by a deterministic hash of the seed, city, category, and tile identifier. Pilot tiles are excluded, and the additional cohort also excludes previously selected tiles. The final sample contains 38 cases in each of categories A and B and 37 in each of C and D. The category boundaries are derived from the full observed outcome data; the resulting sample provides approximately balanced coverage of expansion levels rather than their population frequencies.

School capacity planning. We use the UK Department for Education's *School capacity: Academic year 2024/25* release, which provides school-level places and pupil enrollment records (Department for Education, 2026). School locations are joined using coordinates derived from Get Information about Schools (Department for Education, n.d.). We link longitudinal records by the school Unique Reference Number (URN) and retain primary and secondary schools with exactly one record in each required academic year, a consistent education phase, valid nonnegative enrollment, positive

Table 15: City-specific strata for EV case selection, defined by the observed net increase in public charging ports over the evaluation horizon.

City	A	B	C	D
Boston	≤ 0	1–5	6–10	≥ 11
New York City	≤ 0	1–4	5–18	≥ 19
San Francisco	≤ 0	1–3	4–8	≥ 9

capacity, and a coordinate that maps to the benchmark grid. Schools are assigned to development or test pools by a deterministic tile-level hash, keeping colocated schools in the same pool. Evaluation cases are selected from the test pool with at most one school per tile. Let E_{+1} denote observed target-year enrollment and C_0 the last historical capacity. Sampling distinguishes balanced cases with $0.95 \leq E_{+1}/C_0 \leq 1.05$ from shortage cases with $E_{+1}/C_0 > 1.05$; surplus cases are excluded. The final cohort retains the 20 pilot cases and fills the two strata to 50 cases each using a fixed hash ordering. Target-year enrollment supplies the reference demand for evaluating the proposed capacity, while the surrounding tile supplies contextual information for the individual school.

Food-access analysis. We derive reference labels from the 2019 Food Access Research Atlas of the USDA Economic Research Service (U.S. Department of Agriculture, Economic Research Service, 2021). The source provides census-tract-level low-income and food-access indicators under the urban 1-mile and rural 10-mile distance convention. These tract indicators are spatially aligned to the four assessment cells within each tile. The resulting cell labels distinguish low-income-and-low-access, low-access without low-income status, and not-low-access areas. Source coverage and assignment scores are retained alongside each label. Eligible tiles must contain all four assessment cells, each with at least 95% source coverage and a valid label. Tiles are stratified by the number of cells marked as low access, including those also classified as low income: zero, one to three, or four. Within each city and stratum, selection prioritizes the number of low-income-and-low-access cells, then spatial separation from already selected tiles measured by Manhattan distance on the grid, then the minimum source-assignment score across the four cells; a deterministic hash resolves ties. Pilot tiles are excluded, and the expansion excludes the original cohort. The final set contains 48 tiles with no low-access cells, 54 with mixed access, and 48 with all four cells marked as low access. This procedure preserves varied within-tile patterns and emphasizes underserved areas; these strata describe sampling composition rather than citywide prevalence.

D.2 Experimental Setup

Agent tools. Tool access is defined separately for each task. For forecasting, Earth Agent uses specialized statistical tools, including trend estimation, summary statistics, and time-series diagnostics, together with tools for retrieving target histories, discovering assets, and inspecting tables and rasters. Its adapter retains 90 of the 104 tools in the original implementation and exposes 20 additional helper functions. The excluded tools require unavailable model assets or GDAL-specific processing. Codex and OpenCode retain their native shell and file-operation tools within case-scoped workspaces, allowing them to implement analyses through code. Forecasting comparisons therefore evaluate complete agent configurations with different analytical interfaces.

For planning and resource analysis, all configurations share the task-specific tool interface within each task. EV charging provides nine tools for retrieving charging histories, discovering and inspecting urban data, calculation, and submission. School capacity planning provides eight tools for retrieving school histories, inspecting urban assets, calculation, and plan submission. Food-access analysis provides seven tools for asset discovery, table and raster inspection, cell-level summaries, calculation, and submission of the four cell labels. Tools expose source metadata and observation periods; reference answers and target-period outcomes are unavailable through these interfaces. A successful task-specific submission terminates the case.

Agents choose which available tools to invoke and which data to inspect. Tool-call counts describe interactions at each framework's interface rather than equivalent amounts of computation: a single shell invocation may perform multiple reads and calculations. The counting protocol and observed tool-use statistics are reported in Appendix D.4.1.

Agentic evaluation metrics. We use task-specific metrics to evaluate forecasting, urban planning, and resource analysis. **(1) Forecasting:** Land-surface-temperature and carbon-emission forecasts are evaluated using mean absolute error (MAE), measured in °C and tonnes C/tile/year, respectively. Property-price forecasts are evaluated using mean absolute percentage error (MAPE) to accommodate different local currencies:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \qquad \mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{|y_i|},$$

where y_i and $\hat{y}_i$ denote observed and predicted values. We report results separately for each forecasting target.

(2) EV charging: We evaluate predictions of the four charging-port expansion categories using accuracy, macro-F1, and balanced accuracy. Macro-F1 averages class-specific F1 scores, while balanced accuracy averages class-specific recall.

(3) School capacity: We evaluate next-year enrollment forecasts using MAE and capacity plans using the balanced-case rate and mean capacity-balance score. Let E denote observed target-year enrollment and $C = C_0 + \Delta C$ denote existing capacity plus proposed additional seats. A plan is balanced when $0.95 \le E/C \le 1.05$, and its capacity-balance score is

$$S_{\mathrm{balance}} = 100\left(1 - \frac{|C - E|}{C}\right).$$

Higher scores indicate closer alignment between capacity and observed enrollment. We additionally report total added seats, shortfall seats $\max(0, E - C)$, and surplus seats $\max(0, C - E)$, separating pre-existing surplus from surplus introduced by expansion.

(4) Food accessibility: We report cell-level accuracy and macro-F1 across the three accessibility categories over 600 assessment cells, together with the number of tiles for which all four cells are classified correctly. These metrics measure agreement with the benchmark's tract-derived reference labels.

D.3 Task Setting

D.3.1 Task Instruction

We report the case-independent task instructions used in the six evaluated tasks, formatted for the appendix. Case IDs, locations, historical observations, available assets, and answer ranges are supplied by the corresponding case or tools. The three forecasting targets share one instruction template and differ in the target fields of the case message. EV charging, school capacity, and food access use separate task prompts.

Forecasting: land surface temperature, carbon emissions, and property values. The shared forecasting instruction reads:

> Predict the requested annual value for one $2\,\mathrm{km} \times 2\,\mathrm{km}$ tile at period 0, using the supplied ten-year history (periods -10 to -1). The case message specifies the target, unit, and spatial aggregation. Calendar years are hidden; no target-period observations are supplied. Available modalities include satellite imagery, surface temperature, vegetation, land cover, terrain, buildings, weather, population, nighttime lights, carbon emissions, and property prices; availability varies by tile and period. You may explore multiple relevant modalities, including imagery, to inform your prediction. Use `get_history` for the target history and `list_assets` to discover available modalities. Use `read_table` for numerical properties, `describe_raster` for bands and statistics, `view_raster` for image previews, and `materialize_asset` when an analysis tool needs a local asset path. Return one JSON object with `prediction` (number), `unit`, `reason`, `modalities_used`, and `modules_used`; list only sources and tools actually used.

The case message supplies the city and tile, `prediction_period=0`, the target, its unit and aggregation, and the history window $[-10, -1]$. The three task-specific instantiations are:

Task	Target field	Requested output
LST	`lst`	Period-0 tile-mean LST in °C
Carbon emissions	`carbon`	Period-0 tile total in tonnes C/tile/year
Property values	`property`	Period-0 median of valid 500-m cell medians

In each case, the ten historical values are retrieved through `get_history`; the target-period value is withheld. Property values use the city's local currency.

EV charging-site planning. The task prompt is instantiated with the case ID, city, tile, and four city-specific port-growth ranges:

> Forecast the net change in publicly available EV charging ports in this $2\,\mathrm{km} \times 2\,\mathrm{km}$ tile from $t-2$ to t (two years). Choose one of the supplied ranges A–D for $\mathrm{ports}(t) - \mathrm{ports}(t-2)$. Use `inspect_case` for charging history at $t-4$, $t-3$, and $t-2$. Use `list_modalities` and `list_assets` to discover historical data through $t-2$; inspect relevant tables, rasters, and baseline records, and use `calculate` for arithmetic. Each time unit is one year; calendar years are hidden. Ports count vehicles that can charge simultaneously (Level 1, Level 2, and DC fast), whereas stations are separate counts. Option A includes zero or negative growth. Follow each source's units and observation dates; missing values are unknown, not zero. Use only this case's supplied tools and evidence. Do not access future data, answer labels, other cases, or the web. Cite only tool-call IDs marked citable and distinguish observations from inference. Call `submit_forecast` with the selected range, confidence, a concise explanation, a reason for each A–D range, evidence-call IDs, and limitations. A successful submission ends the task.

School capacity planning. The school prompt is instantiated with a school case ID and city:

> Forecast next-year enrollment and plan period-$+1$ capacity for the specified school. Read enrollment and capacity for periods -2, -1, and 0 with `read_school_history`. Choose an integer `added_seats`≥ 0; planned capacity equals period-0 capacity plus added seats. Reductions are unavailable, but zero additions are allowed. Use only supplied case tools and evidence through period 0. Urban evidence is optional: check its age, units, and spatial coverage. A tile is not a verified school catchment, and a single-period proxy cannot establish growth. Missing data are unknown; do not access future outcomes or infer student counts from display pixels. Aim for next-year enrollment to be 95%–105% of planned capacity. Submit `predicted_enrollment`, `added_seats`, evidence, and a concise reason through `submit_plan`. Explain the forecast calculation and assumptions, compare utilization before and after expansion, and justify additions or zero. Cite each factual basis using a successfully read `school-history` or asset reference, separating observations from inference. A successful submission ends the task.

Food access. The food-access prompt is instantiated with a tile case ID and city:

> Predict FoodAccess for four $1\,\mathrm{km} \times 1\,\mathrm{km}$ cells in this $2\,\mathrm{km} \times 2\,\mathrm{km}$ tile: `s00` (NW), `s01` (NE), `s10` (SW), and `s11` (SE). Assign each cell one of `not_low_access_1and10`, `low_access_1and10`, or `lila_1and10`. Low access means that at least 500 people or 33% of a census tract live beyond one mile (urban) or ten miles (rural) from a qualifying supermarket; these tract-derived labels are transferred to the grid rather than measured independently per cell. Discover assets with `list_reference_assets`; inspect relevant tables, raster descriptions, images, and cell-level summaries. Read units, class legends, source ages, and spatial support. Worksite earnings and property values are indirect income evidence; generic retail or food-service classes are not supermarkets. Missing evidence is unknown, not absence. Assess access and income separately for each cell, distinguishing observations from inference. Use only this case's tools and evidence, without labels, other cases, web, filesystem, or future data. Call `submit_predictions` with one label, confidence, reason, and evidence for each cell, plus limitations. Each cited observation must refer to a successfully inspected result in the form `evidence-NNNN/asset_ID: observation`. A successful submission ends the task.

D.3.2 EXAMPLE TASK CASES

The six examples below pair one held-out reference outcome with an agent's submitted answer and a condensed account of its recorded reasoning. They illustrate individual decisions; aggregate performance is reported separately. Reference answers were withheld from the agents.

Table 16: Example land-surface-temperature forecasting case using historical and urban evidence.

Task Description

Predict the period-0 mean land-surface temperature (LST) of a $2\,\mathrm{km} \times 2\,\mathrm{km}$ New York City tile from ten historical periods and available urban modalities. The target-period observation is withheld.

Case Information

Case ID: `New_York_City_2KM_r3_c4__lst__2021` **City:** New York City.
Reference period: 0; calendar year withheld from the agent.
Reference answer: $25.595\,^{\circ}\mathrm{C}$. **Last year:** $24.475\,^{\circ}\mathrm{C}$. **Ten-year mean:** $24.813\,^{\circ}\mathrm{C}$.

Target and Output

One numerical prediction of tile-mean LST in $^{\circ}\mathrm{C}$, with an evidence-grounded reason.

Answer

Earth Agent–GPT Luna: $24.800\,^{\circ}\mathrm{C}$.

Reasoning Requirements

Submitted rationale, condensed: The ten-year history has a mean near $24.813\,^{\circ}\mathrm{C}$ and recent values near 25–$26\,^{\circ}\mathrm{C}$. The tile is predominantly built up with little green cover. The agent nevertheless uses a historical-level estimate near the long-term mean; the contextual observation does not produce a numerical adjustment.

Table 17: Example carbon emission forecasting case with a recent upward trend.

Task Description

Predict the period-0 annual carbon-emission total of a $2\,\mathrm{km} \times 2\,\mathrm{km}$ Shanghai tile from ten historical periods and available urban modalities.

Case Information

Reference period: 0; calendar year withheld from the agent.
Reference answer: 131,991.969 tonnes C/tile/year. **Last year:** 127,921.686. **Ten-year mean:** 116,367.980.

Target and Output

One numerical prediction of annual tile-total carbon emissions in tonnes C/tile/year, with an evidence-grounded reason.

Answer

Earth Agent–DeepSeek: 131,689 tonnes C/tile/year.

Reasoning Requirements

Submitted rationale, condensed: The agent verifies that the target equals the sum of four ODIAC raster cells in the observed history. It identifies a rising series, with a steeper increase in recent periods, and compares recent linear extrapolations before selecting 131,689. Growth in nighttime lights and GHSL population is used as corroborating context rather than as a measured emissions increment.

Table 18: Example carbon-emission forecasting case using historical trends and partial target-year observations.

Task Description

Estimate the target-year annual carbon-emission total of an unnamed $2\,\mathrm{km} \times 2\,\mathrm{km}$ urban tile using ten years of historical multimodal observations. The location and calendar year are withheld.

Case Information

Historical observations: `history/t-10/` through `history/t-01/`, covering the ten years preceding the target year.
Target-year observations: `current/`, containing partial observations from the target year.
Data documentation: Layer meanings and units are provided in `../data_dictionary.md`.

Target and Output

Predict the period-0 annual carbon-emission total of a $2\,\mathrm{km} \times 2\,\mathrm{km}$ tile from ten historical periods and available urban modalities.

Answer

Reported estimate: 126,000 tonnes C/tile/year.

Reasoning Requirements

Submitted rationale, condensed: The historical series shows an upward trend with year-to-year fluctuations. A linear fit over all ten years predicts approximately 123,303 tonnes C/tile/year. Adding the mean historical increment to the latest value gives approximately 126,548, while using the mean of the last three increments gives approximately 125,776. The agent also compares structural observations, reporting an increase of approximately 3.96% in the OSM building-layer sum and 8.56% in the optical-imagery mean, alongside slightly lower POI/amenity sums. It interprets the structural evidence as broadly stable conditions with a small building-growth signal and selects 126,000 as a rounded estimate near the recent-increment forecasts and above the full-history linear prediction.

Table 19: Example property value forecasting case following a change in historical level.

Task Description

Predict the period-0 property value of a $2\,\mathrm{km} \times 2\,\mathrm{km}$ Singapore tile from ten historical periods and available urban modalities.

Case Information

Case ID: `Singapore_2KM_r7_c16__property__2023` **City:** Singapore.
Reference period: 0; calendar year withheld from the agent.
Reference answer: SGD 996,250. **Last year:** SGD 870,000. **Ten-year mean:** SGD 560,147.200.

Target and Output

One numerical prediction of the median of valid $500\,\mathrm{m}$ cell-level property-value medians, in SGD.

Answer

Earth Agent–DeepSeek: SGD 966,433.

Reasoning Requirements

Submitted rationale, condensed: The target series remains near SGD 500,000 through period -3, then rises to SGD 751,750 at -2 and SGD 870,000 at -1. The agent compares the recent raw-series increase with a smoother PPP/GDP-adjusted property series and checks nighttime-light and GDP evidence for a reversal. It projects further growth with damping, between the two recent-increment estimates.

Table 20: Example EV charging-port planning case with an underestimated expansion.

Task Description

For a $2\,\mathrm{km} \times 2\,\mathrm{km}$ Boston tile, plan the number of new EV charger ports based on your estimated needs, then choose the range of net public charging-port growth from $t-2$ to t using charging history and available urban evidence through $t-2$.

Case Information

Case ID: `evport_67af07e6ec128cbd` **Tile:** `Boston_2KM_r8_c8`.

Observed history: 2, 2, and 13 known ports at $t-4$, $t-3$, and $t-2$; the latest increase includes an 11-port university garage.

Reference answer: D; the held-out net increase is 21 ports.

Target Labels

A: ≤ 0 ports; B: 1–5; C: 6–10; D: ≥ 11 (Boston-specific ranges).

Answer

Earth Agent–GPT Terra: B (1–5 additional ports); submitted confidence 0.420.

Reasoning Requirements

Submitted rationale, condensed: The agent recognizes the recent 11-port garage and two existing sites, along with dense urban activity. It expects some further deployment but considers another garage-scale project less likely without direct evidence of a committed installation. Its recorded limitations note that historical port sums can be incomplete and that station openings are lumpy.

Table 21: Example school-capacity plan that brings observed utilization into the target band.

Task Description

Plan the next period enrollment capacity for a London secondary school and choose nonnegative added seats so projected utilization is within 95–105% of planned capacity.

Case Information

Case ID: `school_plan_c4cac5a4a2cc` **Tile:** `London_2KM_r12_c11`.

Observed history: enrollment 1,358, 1,379, 1,395 at periods -2, -1, 0; period-0 capacity 1,333.

Reference answer: next-period enrollment 1,402.

Planning Constraint

Predict enrollment and an integer number of added seats ≥ 0; planned capacity equals period-0 capacity plus added seats.

Answer

Earth Agent–GPT Terra: predicted enrollment 1,414; add 14 seats, yielding planned capacity 1,347. With the observed 1,402 pupils, utilization is 104.083%.

Reasoning Requirements

Submitted rationale, condensed: Enrollment increased by 21 and then 16 pupils. Extending their mean increase of 18.5 gives a forecast of 1,414. Without expansion, forecast utilization is 106.077%, above the target band. Adding 14 seats brings forecast utilization to 104.974%. The agent does not use tile-level urban proxies as verified catchment or enrollment-growth measures.

Table 22: Example FoodAccess prediction task using multimodal urban evidence under a masked reference period.

Task Description

Given a $2\,\mathrm{km} \times 2\,\mathrm{km}$ urban tile in Washington, DC, predict the food-access category of each of its four $1\,\mathrm{km} \times 1\,\mathrm{km}$ cells. Assess supermarket access and low-income status separately, integrate the available evidence, and provide an evidence-grounded prediction for every cell.

Case Information

Case ID: `fa_full_0105ec127c44` **City:** Washington, DC.
Reference period: 0; calendar year withheld.
Target cells: `s00` (NW), `s01` (NE), `s10` (SW), and `s11` (SE).
Reference answer: all four cells are `not_low_access_1and10`.

Target Labels

`not_low_access_1and10`: Not classified as low access.
`low_access_1and10`: Low access without joint low-income classification.
`lila_1and10`: Both low income and low access.

A census tract is considered low access when at least 500 people or 33% of its population live more than 1 mile (urban) or 10 miles (rural) from a qualifying supermarket. Reference labels are transferred from census tracts to the grid; they are not independent measurements at the $1\,\mathrm{km}$ cell level.

Answer

Earth Agent–GPT Terra: `s00`, `s01`, `s10`, and `s11` $\rightarrow$ `not_low_access_1and10`.

Reasoning Requirements

Submitted rationale, condensed: For each cell, the agent cites mapped building and food-service footprints and a typical property-value median of USD 692,874.940 as indirect evidence against joint low income and low access. For example, it reports 209 food-service pixels in `s00` and 1,219 in `s01`. Its stated limitations acknowledge that these classes do not identify qualifying supermarkets, while tract income and supermarket distance are unavailable.

D.4 TASK RESULT

Additional forecasting metrics Tables 23–28 supplement the main MAE/MAPE results with RMSE and R^2 for the same final predictions. For an evaluation group of N cases, we compute

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \qquad R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{2}$$

where $\bar{y}$ is the mean of the reference values within that evaluation group. LST and carbon metrics pool cases within each task; property metrics are computed separately for each city in its local currency. Negative R^2 values are retained. Pooled R^2 reflects both spatial and temporal variation and does not separately measure the accuracy of year-to-year changes. All reported values use three decimal places.

Block-bootstrap confidence intervals. Brackets report pointwise 95% percentile intervals from 10,000 block-bootstrap replicates. Each block contains all five forecast years (2020–2024) for one spatial tile. Within each city, we sample tiles with replacement, retaining the original number of tiles and all five observations from each sampled tile. We recompute RMSE and R^2, including the reference mean in the denominator, on each replicate and take its 2.5th and 97.5th percentiles. All agents and baselines use identical resampled blocks. LST and carbon each contain 44 blocks; each property city contains four. These intervals describe variation across sampled tiles conditional on the observed cities and years, preserve within-tile dependence, and do not model dependence between different tiles. Property intervals have limited resolution because each city contains only four blocks.

Interpretation and sample-size limitations. Each property city contributes four tile-level series, with five forecast years per tile. The resulting 20 cases should not be treated as 20 independent spatial observations. With so few blocks, bootstrap intervals can be unstable, particularly for R^2, whose denominator also changes with the resampled tiles. For example, OpenCode–Luna in Sydney has a point estimate of $R^2 = 0.831$ and an interval of $[-0.502, 0.884]$, indicating substantial uncertainty across tile resamples. We therefore treat the reported values and rankings as descriptive results on the evaluated cohort, rather than evidence of statistically established method superiority. The tables report intervals for individual methods, not paired performance differences; interval overlap alone is not used to establish either a difference or equivalence. More spatially diverse evaluation samples and repeated agent runs would be needed for more reliable comparisons and broader generalization.

Table 23: Land-surface temperature: point estimates and 95% tile-block bootstrap intervals on 220 cases (44 tiles). RMSE is in °C.

Method	RMSE ↓ [95% CI]	R^2 ↑ [95% CI]
Earth Agent (GPT Luna)	1.778 [1.578, 1.978]	0.899 [0.865, 0.924]
Earth Agent (DeepSeek)	1.761 [1.508, 2.028]	0.901 [0.860, 0.930]
Codex (GPT-5.5)	1.671 [1.498, 1.850]	0.910 [0.883, 0.931]
OpenCode (GPT Luna)	1.677 [1.482, 1.871]	0.910 [0.881, 0.932]
OpenCode (Gemini)	1.834 [1.595, 2.093]	0.892 [0.852, 0.921]
Last-year value	2.116 [1.818, 2.414]	0.856 [0.802, 0.898]
Ten-year mean	1.666 [1.491, 1.852]	0.911 [0.883, 0.932]

Prediction diagnostics. Tables 29–31 report city- and year-level error for the completed prediction configurations. Calendar years are restored only for this post-hoc breakdown; they were hidden from agents. LST and carbon use MAE in their target units, while property uses MAPE because local currencies differ. Compare carbon errors within a city because emission totals vary greatly in scale. The five agent columns refer to complete 600-case evaluations. The ten-year mean has the lowest LST MAE in 8 of 15 city groups; an agent achieves the lowest property MAPE in six of eight city groups, while the last-year baseline performs best in the remaining two.

Planning diagnostics. Table 32 shows where EV growth-range decisions concentrate. The highest-growth class D has 37 reference cases, yet only 6–9 are recovered by each configuration; class C

Table 24: Carbon emissions: point estimates and 95% tile-block bootstrap intervals on 220 cases (44 tiles). RMSE is in tonnes C/tile/year.

Method	RMSE ↓ [95% CI]	R^2 ↑ [95% CI]
Earth Agent (GPT Luna)	1,043.065 [818.575, 1,218.253]	0.999 [0.997, 0.999]
Earth Agent (DeepSeek)	989.874 [776.753, 1,161.460]	0.999 [0.997, 0.999]
Codex (GPT-5.5)	1,013.169 [819.737, 1,163.004]	0.999 [0.997, 0.999]
OpenCode (GPT Luna)	1,052.534 [782.250, 1,268.355]	0.999 [0.998, 0.999]
OpenCode (Gemini)	1,215.050 [866.028, 1,494.668]	0.998 [0.997, 0.999]
Last-year value	1,355.533 [1,000.532, 1,595.534]	0.998 [0.996, 0.998]
Ten-year mean	4,935.183 [3,370.034, 5,988.697]	0.974 [0.959, 0.981]

Table 25: Property prices: point estimates and 95% tile-block bootstrap intervals. Each city has 20 cases from four tiles. RMSE uses the currency scale stated in each panel; R^2 is dimensionless.

Method	RMSE ↓ [95% CI]	R^2 ↑ [95% CI]
Boston (RMSE: 10^3 USD; 4 tiles)		
EA–Luna	28.005 [23.449, 32.922]	0.959 [0.558, 0.963]
EA–DeepSeek	24.255 [22.348, 26.273]	0.969 [0.657, 0.976]
Codex	22.640 [21.623, 23.764]	0.973 [0.671, 0.980]
OC–Luna	24.800 [21.666, 27.401]	0.968 [0.542, 0.980]
OC–Gemini	32.566 [24.606, 42.727]	0.944 [0.583, 0.948]
Last year	34.915 [29.810, 38.153]	0.936 [0.279, 0.955]
Ten-year mean	189.845 [179.598, 196.409]	-0.901 [-24.491, -0.376]
Busan (RMSE: 10^6 KRW; 4 tiles)		
EA–Luna	101.851 [29.727, 165.006]	0.529 [-1.403, 0.912]
EA–DeepSeek	96.857 [32.743, 151.730]	0.574 [-1.058, 0.895]
Codex	105.079 [41.840, 160.921]	0.499 [-1.549, 0.824]
OC–Luna	108.882 [45.205, 162.141]	0.462 [-1.598, 0.813]
OC–Gemini	88.758 [36.586, 133.882]	0.642 [-0.660, 0.907]
Last year	102.324 [36.132, 159.641]	0.525 [-1.558, 0.875]
Ten-year mean	104.700 [45.839, 144.383]	0.502 [-2.282, 0.770]

EA: Earth Agent; OC: OpenCode. GPT forecasting configurations use Luna. Intervals are based on only four spatial blocks per city.

Table 26: Property prices : point estimates and 95% tile-block bootstrap intervals. Each city has 20 cases from four tiles. RMSE uses the currency scale stated in each panel; R^2 is dimensionless.

Method	RMSE ↓ [95% CI]	R^2 ↑ [95% CI]
DC (RMSE: 10^3 USD; 4 tiles)		
EA–Luna	37.454 [27.194, 45.455]	0.982 [-0.537, 0.988]
EA–DeepSeek	39.700 [28.299, 48.490]	0.980 [-0.892, 0.987]
Codex	33.597 [24.462, 40.732]	0.986 [-0.241, 0.992]
OC–Luna	44.006 [31.672, 51.499]	0.975 [-2.237, 0.987]
OC–Gemini	43.856 [33.954, 54.153]	0.975 [-1.353, 0.981]
Last year	37.626 [23.862, 52.629]	0.982 [-0.205, 0.986]
Ten-year mean	162.937 [130.264, 207.561]	0.660 [-33.469, 0.718]
London (RMSE: 10^3 GBP; 4 tiles)		
EA–Luna	109.601 [45.439, 171.976]	0.398 [-2.978, 0.750]
EA–DeepSeek	107.925 [34.781, 178.928]	0.416 [-0.823, 0.849]
Codex	107.420 [37.759, 175.376]	0.422 [-0.885, 0.775]
OC–Luna	102.560 [37.197, 165.496]	0.473 [-0.886, 0.806]
OC–Gemini	103.042 [31.021, 164.235]	0.468 [-0.693, 0.746]
Last year	99.440 [30.641, 163.781]	0.505 [-0.475, 0.847]
Ten-year mean	115.542 [63.597, 171.475]	0.331 [-4.073, 0.379]

EA: Earth Agent; OC: OpenCode. GPT forecasting configurations use Luna. Intervals are based on only four spatial blocks per city.

Table 27: Property prices: point estimates and 95% tile-block bootstrap intervals. Each city has 20 cases from four tiles. RMSE uses the currency scale stated in each panel; R^2 is dimensionless.

Method	RMSE ↓ [95% CI]	R^2 ↑ [95% CI]
New York City (RMSE: 10^3 USD; 4 tiles)		
EA–Luna	67.079 [40.838, 85.625]	0.910 [-1.170, 0.947]
EA–DeepSeek	64.188 [34.347, 84.027]	0.917 [-0.618, 0.957]
Codex	61.009 [35.084, 80.067]	0.925 [-0.691, 0.966]
OC–Luna	72.192 [32.251, 96.866]	0.895 [-0.857, 0.947]
OC–Gemini	85.311 [50.303, 111.837]	0.854 [-2.292, 0.932]
Last year	51.316 [25.703, 67.868]	0.947 [-0.077, 0.973]
Ten-year mean	170.349 [91.146, 242.217]	0.417 [-19.647, 0.512]
San Francisco (RMSE: 10^3 USD; 4 tiles)		
EA–Luna	137.259 [114.793, 156.533]	-0.918 [-3.798, -0.306]
EA–DeepSeek	95.075 [83.398, 105.467]	0.080 [-1.127, 0.300]
Codex	86.821 [79.080, 92.623]	0.233 [-0.863, 0.484]
OC–Luna	93.658 [74.413, 109.574]	0.107 [-1.330, 0.336]
OC–Gemini	128.069 [92.314, 160.319]	-0.669 [-3.596, -0.073]
Last year	77.496 [70.897, 83.576]	0.389 [-0.492, 0.547]
Ten-year mean	260.979 [195.595, 297.383]	-5.933 [-17.512, -3.266]

EA: Earth Agent; OC: OpenCode. GPT forecasting configurations use Luna. Intervals are based on only four spatial blocks per city.

Table 28: Property prices: point estimates and 95% tile-block bootstrap intervals. Each city has 20 cases from four tiles. RMSE uses the currency scale stated in each panel; R^2 is dimensionless.

Method	RMSE ↓ [95% CI]	R^2 ↑ [95% CI]
Singapore (RMSE: 10^3 SGD; 4 tiles)		
EA–Luna	107.225 [52.255, 162.016]	0.759 [0.355, 0.800]
EA–DeepSeek	92.433 [55.224, 127.521]	0.821 [0.275, 0.847]
Codex	90.460 [43.114, 127.864]	0.828 [0.561, 0.856]
OC–Luna	95.167 [62.199, 127.373]	0.810 [0.026, 0.845]
OC–Gemini	97.088 [65.061, 130.549]	0.802 [0.001, 0.820]
Last year	109.267 [50.398, 164.023]	0.749 [0.400, 0.788]
Ten-year mean	235.602 [112.755, 354.118]	-0.165 [-2.001, 0.047]
Sydney (RMSE: 10^3 AUD; 4 tiles)		
EA–Luna	231.405 [140.671, 298.131]	0.876 [-0.411, 0.911]
EA–DeepSeek	242.987 [134.795, 346.046]	0.863 [-0.048, 0.887]
Codex	255.910 [164.569, 318.882]	0.848 [-0.372, 0.902]
OC–Luna	269.539 [165.134, 346.105]	0.831 [-0.502, 0.884]
OC–Gemini	241.767 [131.780, 334.735]	0.864 [-0.106, 0.894]
Last year	263.068 [162.017, 341.523]	0.839 [-0.487, 0.885]
Ten-year mean	575.934 [301.856, 805.350]	0.229 [-3.373, 0.425]

EA: Earth Agent; OC: OpenCode. GPT forecasting configurations use Luna. Intervals are based on only four spatial blocks per city.

Table 29: Land-surface temperature: MAE by city and forecast year (°C). The two panels regroup the same 220 cases.

Group	n	Earth–Luna	Earth–DeepSeek	OC–Luna	OC–Gemini	Codex	Last year	Ten-year mean
By city								
Amsterdam	15	1.264	1.156	1.403	1.349	1.177	1.705	1.198
Boston	15	1.269	1.403	1.309	1.290	1.223	2.063	1.216
Busan	15	0.723	0.675	0.825	0.759	0.674	0.836	0.561
Cape Town	15	1.264	1.209	1.291	1.193	1.154	1.049	1.602
Christchurch	15	1.387	1.207	1.220	1.428	1.173	1.355	1.270
DC	15	1.204	1.254	1.096	1.193	1.055	1.720	1.037
London	15	2.087	2.308	2.020	2.354	1.952	2.746	1.985
Lyon	15	2.123	1.927	1.996	2.061	2.123	3.047	1.885
New York City	15	1.285	1.153	1.163	1.373	1.195	1.973	1.084
San Francisco	15	1.418	1.393	1.254	1.623	1.383	1.325	1.087
Santiago	15	1.669	1.723	1.386	1.819	1.625	1.472	1.359
Singapore	10	0.904	0.850	0.825	0.931	0.862	0.848	0.640
Sydney	15	1.829	1.611	1.609	1.684	1.478	1.564	1.551
Tokyo	15	1.045	1.275	1.172	1.128	1.150	1.414	1.134
Xiong'an	15	1.886	1.679	1.573	1.691	1.728	1.766	1.863
By forecast year								
2020	44	1.256	1.368	1.231	1.292	1.218	1.688	1.240
2021	44	1.776	1.538	1.587	1.734	1.676	1.949	1.437
2022	44	1.403	1.326	1.192	1.323	1.225	1.608	1.062
2023	44	1.355	1.438	1.384	1.535	1.299	1.478	1.380
2024	44	1.390	1.333	1.378	1.467	1.287	1.664	1.447

Table 30: Carbon emissions: MAE by city and forecast year (tonnes C/tile/year). The two panels regroup the same 220 cases.

Group	n	Earth–Luna	Earth–DeepSeek	OC–Luna	OC–Gemini	Codex	Last year	Ten-year mean
By city								
Amsterdam	15	276.613	226.749	288.957	281.078	295.122	297.730	1,006.523
Boston	10	114.388	132.012	121.332	107.740	93.660	123.812	180.381
Busan	15	514.681	488.062	581.498	517.622	473.480	304.751	403.713
Cape Town	15	331.489	331.532	281.313	288.556	380.377	256.331	385.654
Christchurch	15	0.084	0.073	0.074	0.064	0.068	0.061	0.073
DC	15	260.834	291.934	335.708	295.742	309.908	366.783	504.131
London	15	194.675	151.211	154.035	214.126	149.211	159.943	801.164
Lyon	15	379.862	438.242	385.666	375.469	505.938	529.565	1,149.527
New York City	15	276.473	345.825	335.307	327.485	306.492	352.280	513.236
San Francisco	15	38.262	41.215	48.304	41.091	37.875	49.863	72.645
Santiago	15	1,867.614	1,465.229	1,509.240	1,727.339	1,424.114	1,377.077	824.526
Shanghai	15	1,871.851	1,380.796	1,905.988	2,218.115	1,760.247	3,189.553	13,648.485
Singapore	15	909.223	1,184.343	948.862	932.095	1,032.982	1,562.649	8,494.535
Sydney	10	609.744	511.565	529.353	728.179	480.664	314.393	674.577
Tokyo	10	861.923	999.746	877.195	1,091.602	1,248.750	929.696	4,356.582
Xiong'an	10	100.146	94.542	68.852	104.682	91.841	187.160	800.882
By forecast year								
2020	44	873.673	701.876	737.474	815.289	888.710	967.304	2,113.089
2021	44	469.740	502.044	517.421	435.300	410.509	467.062	1,816.894
2022	44	386.049	373.406	337.557	300.538	385.077	556.525	1,984.669
2023	44	628.651	681.511	539.449	819.490	653.820	944.726	2,524.329
2024	44	384.771	299.272	540.634	552.196	372.937	297.324	2,406.187

Table 31: Property values: MAPE by city and forecast year (%). The two panels regroup the same 160 cases.

Group	n	Earth–Luna	Earth–DeepSeek	OC–Luna	OC–Gemini	Codex	Last year	Ten-year mean
By city								
Boston	20	3.120	2.897	3.137	3.768	2.887	4.460	28.308
Busan	20	15.627	17.512	19.443	15.539	17.520	17.269	18.847
DC	20	5.112	5.548	5.996	6.237	4.577	4.803	27.320
London	20	14.454	11.929	12.367	12.114	12.240	10.916	18.306
New York City	20	6.237	5.645	6.221	8.055	5.286	3.987	16.074
San Francisco	20	9.442	5.780	5.888	8.702	4.990	5.159	18.916
Singapore	20	11.074	9.794	10.809	11.725	9.022	10.963	23.436
Sydney	20	10.709	10.014	11.545	10.597	11.256	11.767	28.059
By forecast year								
2020	32	6.991	6.405	6.674	6.595	5.792	7.037	22.037
2021	32	10.328	9.857	11.311	10.856	9.507	11.427	25.402
2022	32	8.074	8.377	9.119	9.274	8.291	7.976	25.017
2023	32	11.673	9.934	10.761	10.664	10.935	8.895	19.635
2024	32	10.293	8.626	9.263	10.571	7.837	7.993	19.949

is also often assigned to A or B. Table 33 reports the scale of school expansions and the resulting balance outcomes, revealing different seat-addition policies even when agents use the same historical enrollment evidence.

Table 32: EV charging-port growth: reference-by-prediction confusion counts on the recovered 150-case cohort. Recall is the diagonal count divided by the reference-row total. A–D denote the city-specific growth ranges.

Agent	Reference	A	B	C	D	Recall (%)
Earth–Terra	A	19	14	5	0	50.000
	B	13	20	5	0	52.632
	C	12	17	5	3	13.514
	D	4	15	9	9	24.324
Earth–DeepSeek	A	27	11	0	0	71.053
	B	8	23	5	2	60.526
	C	13	17	4	3	10.811
	D	9	12	7	9	24.324
Codex–GPT-5.5	A	20	17	1	0	52.632
	B	3	27	8	0	71.053
	C	7	22	6	2	16.216
	D	3	17	9	8	21.622
OpenCode–Terra	A	28	8	2	0	73.684
	B	20	12	5	1	31.579
	C	16	12	6	3	16.216
	D	9	13	7	8	21.622
OpenCode–Gemini	A	31	7	0	0	81.579
	B	25	10	3	0	26.316
	C	25	7	3	2	8.108
	D	19	9	3	6	16.216

Table 33: School capacity plans across 100 cases. Planned capacity is the last observed capacity plus proposed seats; balance requires observed next-period enrollment to be 95%–105% of planned capacity.

Agent	Added seats: median [IQR]	Zero addition (%)	Balanced (%)	Shortage (%)	Surplus (%)
Earth–Terra	0.000 [0.000, 38.250]	57.000	81.000	16.000	3.000
Earth–DeepSeek	20.000 [0.000, 75.750]	35.000	88.000	7.000	5.000
Codex–GPT-5.5	0.000 [0.000, 46.500]	57.000	84.000	13.000	3.000
OpenCode–Terra	0.000 [0.000, 33.750]	53.000	80.000	17.000	3.000
OpenCode–Gemini	16.500 [0.000, 61.500]	38.000	84.000	10.000	6.000

Resource-analysis diagnostics. Tables 34 and 35 separate food-access class errors from within-tile spatial consistency. Recall for the low-income-and-low-access class ranges from 3.226% to 48.387%.

Exact classification of mixed-label tiles remains rare, even when uniform-label tiles are classified more accurately.

Table 34: Food-access cell-label confusion counts on the recovered 150-tile cohort. NLA: not low access; LA: low access; LILA: low income and low access. Rows are reference classes and columns are predicted classes.

Agent	Reference	NLA	LA	LILA	Recall (%)
Earth–Terra	NLA	242	22	34	81.208
	LA	188	43	9	17.917
	LILA	38	9	15	24.194
Earth–DeepSeek	NLA	224	68	6	75.168
	LA	146	78	16	32.500
	LILA	45	11	6	9.677
Codex–GPT-5.5	NLA	218	29	51	73.154
	LA	143	75	22	31.250
	LILA	13	19	30	48.387
OpenCode–Terra	NLA	259	14	25	86.913
	LA	186	43	11	17.917
	LILA	31	11	20	32.258
OpenCode–Gemini	NLA	219	74	5	73.490
	LA	116	121	3	50.417
	LILA	35	25	2	3.226

Table 35: Food-access performance by within-tile reference-label composition: 93 uniform-label tiles and 57 mixed-label tiles. Cell accuracy counts correct labels across all four cells; exact-tile accuracy requires all four to be correct.

Agent	Uniform: cell (%)	Uniform: exact tile (%)	Mixed: cell (%)	Mixed: exact tile (%)
Earth–Terra	56.183	53.763	39.912	0.000
Earth–DeepSeek	52.957	21.505	48.684	3.509
Codex–GPT-5.5	61.559	58.065	41.228	3.509
OpenCode–Terra	57.527	53.763	47.368	5.263
OpenCode–Gemini	68.548	61.290	38.158	1.754

D.4.1 Tool-use Statistics

Counting protocol. Tables 36 and 37 report tool use across the six tasks. We count each native tool invocation once in the attempt used for the final scored answer, retaining calls that returned errors within that attempt. Final-submission calls and automatic answer processing are excluded from the call totals. Values are mean calls per case, except the error column, which reports the percentage of calls returning an explicit error or a nonzero process exit. Calls are measured at the framework interface: a single shell invocation can perform several reads and calculations, so counts describe interaction patterns rather than equivalent units of computation.

Observed patterns. Forecasting exposes a clear difference in tool use: Earth Agent invokes named statistical functions, whereas Codex and OpenCode use shell execution and file operations to implement their analyses. Earth Agent with DeepSeek makes more statistical calls than its GPT counterpart in all three forecasting tasks. In school planning, only Earth Agent with DeepSeek invokes urban table or raster tools; the other configurations primarily use enrollment history and, where needed, arithmetic. Food-access analysis involves substantially more table and raster inspection across configurations. These statistics describe evidence acquisition and processing; the case studies in Appendix D.3.2 examine whether the retrieved evidence changes the final estimate.

Data: history retrieval, asset discovery, table access, and raster inspection or aggregation. Stats: named trend, summary-statistic, and time-series diagnostic tools. Calc: the arithmetic calculator. Exec: shell or command execution, including any analyses implemented within it. Files: native file discovery, reading, writing, and editing. Other: task-list updates. A zero denotes no calls through that tool category, not absence of the corresponding analytical capability.

Table 36: Tool use in the three forecasting tasks. Entries are mean non-submission calls per case; category columns partition the total. All values are rounded to three decimal places.

Agent	Calls	Data	Stats	Calc	Exec	Files	Other	Error (%)
Land surface temperature (220 cases)								
Earth Agent (GPT Luna)	18.850	16.786	1.668	0.395	0.000	0.000	0.000	0.048
Earth Agent (DeepSeek)	27.491	18.595	4.245	4.650	0.000	0.000	0.000	0.579
Codex (GPT-5.5)	17.068	0.000	0.000	0.000	15.164	1.905	0.000	6.365
OpenCode (GPT Luna)	26.186	0.000	0.000	0.000	3.909	19.014	3.264	2.066
OpenCode (Gemini)	11.673	0.000	0.000	0.000	3.091	8.518	0.064	1.947
Carbon emissions (220 cases)								
Earth Agent (GPT Luna)	11.405	9.382	1.668	0.355	0.000	0.000	0.000	0.199
Earth Agent (DeepSeek)	20.350	12.541	3.536	4.273	0.000	0.000	0.000	0.849
Codex (GPT-5.5)	17.155	0.000	0.000	0.000	15.277	1.877	0.000	6.280
OpenCode (GPT Luna)	24.618	0.000	0.000	0.000	3.509	17.891	3.218	2.234
OpenCode (Gemini)	11.205	0.000	0.000	0.000	2.745	8.405	0.055	2.556
Property prices (160 cases)								
Earth Agent (GPT Luna)	11.738	9.469	2.006	0.263	0.000	0.000	0.000	0.053
Earth Agent (DeepSeek)	17.881	8.425	4.231	5.225	0.000	0.000	0.000	0.245
Codex (GPT-5.5)	16.269	0.000	0.000	0.000	14.481	1.788	0.000	5.647
OpenCode (GPT Luna)	24.556	0.000	0.000	0.000	3.456	17.725	3.375	1.451
OpenCode (Gemini)	11.262	0.000	0.000	0.000	2.694	8.450	0.119	2.608

Data: history retrieval, asset discovery, table access, and raster inspection or aggregation. Stats: named trend, summary-statistic, and time-series diagnostic tools. Calc: the arithmetic calculator. Exec: shell or command execution, including any analyses implemented within it. Files: native file discovery, reading, writing, and editing. Other: task-list updates. A zero denotes no calls through that tool category, not absence of the corresponding analytical capability.

Table 37: Tool use in planning and resource analysis. Configurations share the task-specific evidence interface within each task. Entries are mean non-submission calls per case.

Agent	Calls	History	Discover	Tables	Raster	Calc	Error (%)
EV charger planning (150 cases)							
Earth Agent (GPT Terra)	23.860	3.293	10.547	7.227	2.267	0.527	0.447
Earth Agent (DeepSeek)	20.913	4.013	5.367	7.113	1.700	2.720	4.877
Codex (GPT-5.5)	11.273	3.520	4.700	2.447	0.413	0.193	0.000
OpenCode (GPT Terra)	25.067	3.433	10.807	7.820	2.720	0.287	0.293
OpenCode (Gemini)	4.793	1.753	2.633	0.387	0.000	0.020	2.364
School capacity planning (100 cases)							
Earth Agent (GPT Terra)	1.330	1.000	0.050	0.000	0.000	0.280	0.000
Earth Agent (DeepSeek)	8.020	1.000	1.000	1.050	0.540	4.430	0.249
Codex (GPT-5.5)	1.770	1.000	0.070	0.000	0.000	0.700	1.695
OpenCode (GPT Terra)	2.650	1.000	0.000	0.000	0.000	1.650	1.132
OpenCode (Gemini)	1.160	1.000	0.160	0.000	0.000	0.000	1.724
Food-access analysis (150 cases)							
Earth Agent (GPT Terra)	16.360	0.000	1.013	6.600	8.747	0.000	1.630
Earth Agent (DeepSeek)	28.520	0.000	1.020	13.047	12.087	2.367	16.620
Codex (GPT-5.5)	12.907	0.000	1.000	5.920	5.987	0.000	5.165
OpenCode (GPT Terra)	18.640	0.000	1.007	6.760	10.873	0.000	3.469
OpenCode (Gemini)	10.933	0.000	1.887	3.747	5.260	0.040	17.805

History: charging evidence or school-history retrieval. Discover: case inspection and modality or asset listing. Tables: table reads. Raster: asset or raster descriptions, raster viewing, and cell-level raster summaries. Calc: arithmetic. Food access uses asset-discovery and inspection tools rather than the common history-retrieval tool.