CITYREP: A UNIFIED BENCHMARK FOR URBAN REPRESENTATIONS ACROSS CITIES, TASKS, AND MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Urban representation learning encodes complex urban environments into general-purpose embeddings for diverse downstream tasks and emerging urban foundation models. However, current evaluations are limited, typically focusing on one or two cities and tasks and relying on random splits that introduce spatial leakage. This can inflate performance, provide weak evidence of cross-location generalization, and hinder fair comparison. To address this, we propose **CityRep**, a unified benchmark that evaluates urban representations across data modalities, cities, and tasks using spatially structured splits. CityRep consists of three key components: (1) a spatial unit-agnostic evaluation framework that supports heterogeneous urban representations through a standardized alignment module; (2) a unified evaluation protocol using block-based spatial splits to mitigate spatial leakage and enable rigorous model comparison; and (3) an extensible multi-city, multi-task benchmark suite spanning 8 cities and 8 tasks across regression, classification, and distribution prediction. We evaluate 11 representative urban representation models. Results show that performance is highly sensitive to the split protocol, with random splits inflating scores and altering model rankings. We also observe substantial variability across cities and tasks, underscoring the need for generalization-aware evaluation. CityRep is released as a reproducible benchmark with datasets, evaluation pipelines, and diagnostic tools to facilitate fair comparison and support future research in urban representation learning.
Data: `https://www.kaggle.com/datasets/cityrep/cityrep`;
Code: `https://anonymous.4open.science/r/CityRep-A2F0`.

1 INTRODUCTION

Urban representation learning seeks to turn heterogeneous observations of cities into reusable spatial embeddings. Recent models draw on remote sensing (Klemmer et al., 2025; Feng et al., 2026; Brown et al., 2025), street-view imagery (Wang et al., 2020; Yong & Zhou, 2024), and points of interest (Yan et al., 2017; Huang et al., 2022; Wang et al., 2025; Balsebre et al., 2024; Liu et al., 2025b) to encode geographic entities, regions, or locations. This motivation parallels the broader shift toward foundation models: representations learned from broad urban data are expected to transfer across tasks, locations, and domains. Yet the evaluation of these representations remains much less unified. Reported results are often tied to a particular model interface or a particular format of downstream task, making it difficult to assess whether an embedding is broadly useful or only effective under a narrow evaluation setup.

Existing evaluations are also too narrow to support claims about general-purpose urban representations. Many studies evaluate on one or two cities, a small number of tasks, or a single label type. Such experiments are valuable for demonstrating a specific application, but they do not reveal whether a representation transfers across urban contexts or across qualitatively different prediction problems. This limitation is especially important for the emerging urban foundation models, whose value depends on broad reuse. A benchmark should therefore cover multiple cities, multiple urban domains, and multiple task types, while remaining extensible to incorporate new cities, tasks, and models.

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.

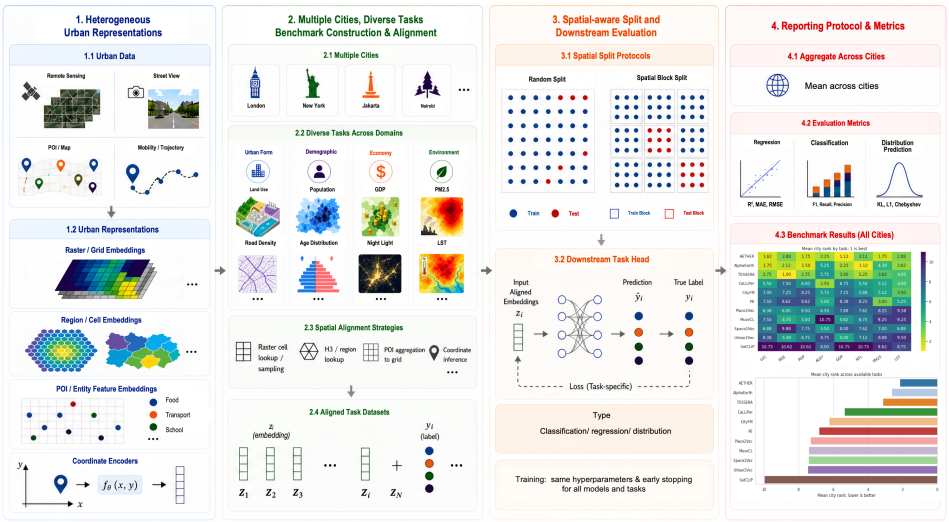


Figure 1: **CityRep benchmark framework.** CityRep standardizes the evaluation of heterogeneous urban representations by aligning different spatial supports to common downstream task units, evaluating them across eight cities and eight tasks, and using spatial block splits.

Finally, evaluation must explicitly account for spatial dependence. In urban representation learning, we expect models to leverage information from observed regions to make predictions in unseen areas. While this challenge has been well recognized in spatial validation studies (Roberts et al., 2017; Valavi et al., 2019; Meyer et al., 2019), most existing urban representation models are still evaluated using random splits that ignore spatial structure. We address this gap by establishing a unified spatially structured evaluation protocol. Our benchmark provides empirical evidence that random splits can substantially inflate performance and lead to over-optimistic conclusions about model generalization. Moreover, our cross-city diagnostic shows that globally pretrained geospatial representations retain more transferable downstream structure when target-city labels are withheld, whereas traditional city-specific urban representations exhibit substantially weaker transfer.

We introduce **CityRep**, a unified benchmark for urban representations across modalities, tasks, and cities, as shown in Figure 1. The central goal of CityRep is to move urban representation evaluation beyond narrow, single-setting comparisons. Rather than assessing an embedding on one city, one task, or one type of urban label, CityRep asks whether a representation remains useful across a sufficiently broad set of urban phenomena and geographic contexts. To operationalize this goal, we construct downstream evaluation data from four key dimensions of urban systems: morphology, demographics, economy, and environment. These dimensions are instantiated as eight tasks across eight cities, covering classification, regression, and distribution prediction. CityRep pairs this broad task suite with a common evaluation protocol for heterogeneous representations. Models based on rasters, regions, entities, or coordinates can be evaluated under the same downstream interface, while spatially structured splits are used to reduce leakage between nearby training and test samples. This provides a basis for assessing whether urban representations are genuinely general-purpose, rather than effective only under narrow evaluation settings.

In summary, CityRep makes the following contributions. 1) **Unified benchmark.** CityRep provides a common evaluation framework for heterogeneous urban representations, covering 11 representative models, 8 cities, and 8 downstream tasks with different modalities and spatial supports. 2) **Spatial evaluation.** We standardize representation alignment and use shared block-based splits for spatially controlled comparison, reducing the leakage associated with commonly used random splits. 3) **Empirical findings.** Our evaluation reveals substantial variation across cities and tasks: random splits can inflate performance and alter comparative conclusions, while globally pretrained geospatial representations show stronger cross-city transfer of downstream predictors than earlier city-specific representations. 4) **Open resources.** We release the processed datasets, model interfaces, evaluation pipeline, and diagnostic tools to support reproducible and extensible evaluation of future urban representations.

2 Related Work

Urban Representation Learning Existing urban representation learning methods are highly heterogeneous, differing in the data they use, the spatial units they operate upon, and the urban signals they encode. Following geographic information systems (GIS) taxonomy, this heterogeneity is largely shaped by whether models ingest vector data, such as points, polylines, and polygons, or raster data, such as satellite and street-view imagery. Vector-based methods often rely on POI data but produce different outputs: category embedding methods, including Place2Vec (Yan et al., 2017), POI2Vec (Feng et al., 2017), and SPPE (Huang et al., 2022), capture spatial co-occurrence patterns of POI types and require aggregation to represent urban spaces, whereas entity embedding models such as Urban2Vec (Wang et al., 2020), HGI (Huang et al., 2023), and CityFM (Balsebre et al., 2024) directly encode regions, buildings, or roads. Raster-based methods, by contrast, naturally produce grid-cell embeddings, with earth-observation foundation models such as AlphaEarth Foundation (Brown et al., 2025) and TESSERA (Feng et al., 2026) enabling dense large-scale representations, and AETHER further incorporating POI semantics into raster foundations (Liu et al., 2025a). A related line of coordinate-based encoders, including Space2Vec (Mai et al., 2020), SatCLIP (Klemmer et al., 2025), and CaLLiPer (Wang et al., 2025), learns representations for continuous locations from POIs, imagery, or language supervision. Consequently, the resulting embeddings operate over disparate spatial supports (e.g., regions, H3 cells (Uber Technologies, Inc., 2026), raster grids, and coordinates) and capture varying urban information. This creates a central evaluation challenge: heterogeneous methods cannot be fairly compared without spatial alignment to standard task units, and evaluation on only a few downstream tasks is insufficient for representations capturing diverse urban signals.

Geospatial Benchmarks and Spatial Evaluation Urban representation learning can be viewed as a fine-grained, city-focused branch of geospatial representation learning. Existing geospatial benchmarks have largely started from image- or raster-centered settings. TorchGeo provides reusable infrastructure for geospatial data loading, sampling, and model development (Stewart et al., 2025), while GEO-Bench, SatlasPretrain, and PANGAEA standardize Earth-observation evaluation and pretraining across tasks, sensors, resolutions, regions, and temporal settings (Lacoste et al., 2023; Bastani et al., 2023; Marsocci et al., 2026). Recent benchmarks move closer to spatial representation learning. TorchSpatial evaluates general-purpose location encoders (Wu et al., 2024), OBSR evaluates geospatial embedders on regional and trajectory tasks (Moska et al., 2025), and MoRA introduces human-centric social and economic prediction tasks based on mobility-centered representations (Wen et al., 2026). However, they still provide limited evidence on whether representations capture fine-grained intra-urban structure and functions. Spatial evaluation is also critical. Prior work shows that random splits can overestimate performance under spatial dependence and recommends spatially structured validation for assessing transfer to unseen areas (Roberts et al., 2017; Valavi et al., 2019; Meyer et al., 2019; Ploton et al., 2020). These gaps motivate a unified multi-city benchmark for heterogeneous urban representations under spatially robust evaluation protocols.

3 CityRep Framework and Benchmark

3.1 Problem Definition

Urban representation learning aims to compress heterogeneous observations of a city into reusable spatial embeddings. CityRep evaluates whether the resulting representation captures transferable urban information that is useful for downstream prediction across multiple task domains and spatial units. For representation model m and city c, let $E_{m,c}$ denote the resulting urban representation. For task t in city c, let $\mathcal{U}_{c,t} = \{u_i\}_{i=1}^{n_{c,t}}$ denote the task units with corresponding labels $\mathbf{y}_{c,t}$. Different representations may use spatial supports ranging from raster grids and regions to spatial entities and continuous coordinates. To evaluate these representations on the same task units, CityRep maps each representation to the corresponding downstream support:

$$\mathbf{X}_{m,c,t} = A(E_{m,c}, \mathcal{U}_{c,t}) \in \mathbb{R}^{n_{c,t} \times d_m}, \tag{1}$$

where A denotes the spatial alignment procedure and d_m is the representation dimension. Each row of $\mathbf{X}_{m,c,t}$ is the feature vector associated with one task unit.

After alignment, all models for a given city–task pair use the same spatial train, validation, and test splits and are evaluated with the same downstream prediction protocol. This separates differences in representation quality from differences in spatial support or evaluation setup, allowing CityRep to compare heterogeneous representations across models, cities, and tasks.

3.2 Spatial Alignment

CityRep uses each downstream task's native spatial units as the evaluation anchor. The task support remains fixed, and each representation adapts to it. Accordingly, the alignment operator A maps representations with heterogeneous spatial supports to the task units while preserving those units. Rather than resampling all representations and labels onto an arbitrary universal grid, CityRep selects an alignment strategy according to the spatial relationship between the representation support and the task support.

For raster- and region-based representations, CityRep uses spatial overlap between representation units and task units. When multiple representation units fall within a task unit, their embeddings are mean-aggregated; when a representation unit is coarser than the task units, its embedding is assigned to the task units it covers. Directly compatible supports are handled through raster sampling, cell lookup, or spatial join. For sparse entity representations, such as POI- or map-entity embeddings, CityRep first aggregates entity embeddings to an intermediate H3 support and then aligns the resulting cells to the downstream task units, avoiding the low feature coverage that can arise from direct aggregation of unevenly distributed entities. Appendix C.3 compares the two strategies in terms of feature coverage and downstream performance. For coordinate encoders, A directly queries the encoder at a representative coordinate of each task unit, using the original coordinate for point tasks and the cell center or representative point for areal units. Task units with no matched representation are retained and assigned zero vectors, so all models use the same task samples and split assignments; feature-coverage diagnostics are reported in Appendix C.3. Together, these strategies instantiate the common interface in Equation 1. Model-specific spatial supports and alignment details are provided in Appendix C.2.

3.3 Spatial Split

CityRep uses spatial block splitting to evaluate models while reducing leakage from nearby training samples. For each city–task pair, let $\mathcal{B}_{c,t}$ denote a set of non-overlapping spatial blocks and let $b_{c,t}(u_i)$ assign task unit u_i to its block. For seed k, the blocks are partitioned into disjoint sets $\mathcal{B}^s_{c,t,k}$, which induce the task-unit splits

$$\mathcal{U}^s_{c,t,k} = \left\{ u_i \in \mathcal{U}_{c,t} : b_{c,t}(u_i) \in \mathcal{B}^s_{c,t,k} \right\}, \qquad s \in \{\text{train}, \text{val}, \text{test}\}. \tag{2}$$

Thus, task units within the same spatial block always belong to the same split.

The partition is task-specific because tasks differ in spatial extent, valid masks, and label support, but remains model-invariant. All representations evaluated on the same city–task pair use identical split assignments. This prevents nearby samples within the same block from appearing across training and test sets and reduces the extent to which test performance can be driven by local spatial interpolation.

Operationally, CityRep constructs an axis-aligned 10×10 grid over each task's valid spatial extent and partitions its non-empty blocks. For each seed, 20% of the blocks are assigned to test, 10% of the remaining blocks to validation, and the remainder to training. We use five seeds, each controlling both the randomized block assignment and downstream-head training randomness, including parameter initialization and minibatch order. All models within a city–task–seed setting use identical split assignments, enabling paired comparisons. We compare against conventional random splitting as a diagnostic of spatial leakage in Section 4.3. Appendix C.1 provides split visualizations, while Appendix E.3 evaluates sensitivity to a finer 20×20 partition.

3.4 Tasks and Dataset

CityRep evaluates eight tasks spanning four urban domains and three prediction types: morphology includes land-use classification (**LUC**; city-specific sources are listed in Appendix B.1) and road-density regression (**RDE**) based on OpenStreetMap roads (OpenStreetMap contributors, 2026); demographics includes population regression (**POP**) and age-distribution prediction (**AGE**), both

Table 1: Number of prediction units for each downstream task across eight cities.

Task	London	New York	Singapore	Sydney	Mumbai	Nairobi	Jakarta	Cape Town
Morphology								
Land use	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000
Road density	297,314	187,028	83,393	612,027	58,778	81,389	75,724	343,593
Demographics								
Population	266,183	114,276	64,172	416,254	43,522	61,677	74,885	167,771
Age distribution	234,428	87,736	32,990	265,815	43,455	55,521	74,096	136,093
Economy								
Gross domestic product	2,977	1,812	806	6,121	575	817	756	3,420
Nighttime lights	2,977	1,867	839	6,119	583	818	756	3,434
Environment								
$PM_{2.5}$	2,072	1,038	551	4,231	407	569	527	2,380
Land surface temperature	295,364	112,433	70,009	592,033	55,394	81,389	75,117	339,759

derived from WorldPop (Bondarenko et al., 2025b;a); economy includes gross domestic product regression (**GDP**) (Kummu et al., 2025) and nighttime-lights regression (**NTL**) (Earth Observation Group, Payne Institute for Public Policy, Colorado School of Mines, 2024; Elvidge et al., 2021); and environment includes $PM_{2.5}$ regression (**$PM_{2.5}$**) (van Donkelaar et al., 2024) and land-surface-temperature regression (**LST**) (Wan et al., 2021). Appendix B.1 details source versions, licenses, spatial supports, and label construction.

The main benchmark covers London, New York, Singapore, Sydney, Mumbai, Nairobi, Jakarta, and Cape Town. Its task units retain each label's native or processed spatial support, from sampled land-use points to fine and coarse raster cells as in Table 1. AGE aggregate results use London, New York, Singapore, and Sydney because of source-data quality concerns. The globally available raster tasks also support other cities. We provide an exploratory 26-city extension in Appendix F.

3.5 Selected Baselines

We evaluate 11 representations across location, POI/map, multimodal, and raster model groups. The location group contains the deterministic **PE** baseline using **SphereC** (Mai et al., 2023), the POI-supervised **Space2Vec** (Mai et al., 2020) and **CaLLiPer** (Wang et al., 2025) coordinate encoders, and the pretrained **SatCLIP** (Klemmer et al., 2025) coordinate encoder. The POI and map group contains **Place2Vec** (Yan et al., 2017), **CityFM** (Balsebre et al., 2024), and **Urban2Vec** (Wang et al., 2020). The multimodal **MuseCL** (Yong & Zhou, 2024) is evaluated as a CityRep-compatible variant that uses POI semantics in place of its mobility branch. The raster group contains the released **TESSERA** (Feng et al., 2026) and **AlphaEarth** (Brown et al., 2025) features, plus **AETHER** (Liu et al., 2025a), which aligns AlphaEarth features with POI semantics.

Reproduced baselines use Foursquare POIs (Foursquare Labs, Inc., 2026), Mapillary imagery (Meta Platforms, Inc. & Mapillary, 2026), and OpenStreetMap entities (OpenStreetMap contributors, 2026) where applicable. Appendix C.2 specifies each model's inputs, checkpoint or reproduction status, embedding support, and alignment interface.

3.6 Evaluation

CityRep evaluates each representation after spatial alignment to the downstream task units. For a model m, city c, and task t, the aligned feature matrix $\mathbf{X}_{m,c,t}$ is paired with the task labels $\mathbf{y}_{c,t}$ and used to train a lightweight task head. To limit variation from downstream model engineering, CityRep uses the same predictor family and training protocol for all representation models. After alignment, each representation is evaluated by a task-specific prediction head:

$$\hat{\mathbf{y}}_{c,t} = g_{\theta,t}(\mathbf{X}_{m,c,t}), \tag{3}$$

where the output layer and loss are selected according to the task type.

Evaluation metrics. CityRep uses nine task-appropriate metrics across the three prediction types. For regression tasks, including road density, population, GDP, NTL, $PM_{2.5}$, and land surface temperature, we report R^2, mean absolute error (MAE), and root mean squared error (RMSE), with R^2 used as the primary metric. For land-use classification, we report macro F1, macro recall, and

Table 2: **Main CityRep results under spatial block splitting.** Entries report mean primary metrics across cities and seeds ± the standard deviation across the five seed-specific city averages. The seeds also determine spatial partitions. Type denotes the main data sources: *L* = location, *P* = POI, *R* = remote sensing, *S* = street-view imagery, and *O* = other urban/map data. **Bold**, underlined, and wavy-underlined values rank first, second, and third. *AGE reports KL divergence (lower is better) over the four cities with reliable age–sex coverage.

		Morphology		Demographics		Economy		Environment	
Model	Type	LUC (F1 ↑)	RDE (R^2 ↑)	POP (R^2 ↑)	AGE* (KL ↓)	GDP (R^2 ↑)	NTL (R^2 ↑)	$PM_{2.5}$ (R^2 ↑)	LST (R^2 ↑)
PE	*L*	$0.149_{\pm 0.019}$	$0.144_{\pm 0.022}$	$0.176_{\pm 0.043}$	$0.040_{\pm 0.002}$	$0.132_{\pm 0.044}$	$0.237_{\pm 0.112}$	$\underline{0.492}_{\pm 0.074}$	$0.378_{\pm 0.049}$
Place2Vec	*P*	$0.168_{\pm 0.005}$	$0.217_{\pm 0.010}$	$0.288_{\pm 0.018}$	$0.043_{\pm 0.002}$	$0.191_{\pm 0.051}$	$0.303_{\pm 0.064}$	$0.063_{\pm 0.054}$	$0.138_{\pm 0.038}$
Space2Vec	*LP*	$0.166_{\pm 0.014}$	$0.118_{\pm 0.016}$	$0.218_{\pm 0.031}$	$0.041_{\pm 0.001}$	$0.205_{\pm 0.042}$	$0.305_{\pm 0.033}$	$0.236_{\pm 0.093}$	$0.320_{\pm 0.026}$
CaLLiPer	*LP*	$0.189_{\pm 0.011}$	$0.161_{\pm 0.015}$	$0.298_{\pm 0.039}$	$\underline{0.038}_{\pm 0.003}$	$0.263_{\pm 0.042}$	$0.364_{\pm 0.057}$	$0.328_{\pm 0.084}$	$0.440_{\pm 0.050}$
CityFM	*PO*	$0.165_{\pm 0.015}$	$0.187_{\pm 0.015}$	$0.228_{\pm 0.045}$	$0.041_{\pm 0.003}$	$0.199_{\pm 0.038}$	$0.369_{\pm 0.091}$	$0.348_{\pm 0.086}$	$\underline{0.496}_{\pm 0.085}$
Urban2Vec	*PS*	$0.143_{\pm 0.007}$	$0.227_{\pm 0.012}$	$0.278_{\pm 0.034}$	$0.044_{\pm 0.002}$	$0.316_{\pm 0.048}$	$0.319_{\pm 0.060}$	$0.068_{\pm 0.047}$	$0.137_{\pm 0.075}$
MuseCL	*PRS*	$0.153_{\pm 0.007}$	$0.244_{\pm 0.014}$	$0.306_{\pm 0.031}$	$0.046_{\pm 0.002}$	$0.338_{\pm 0.057}$	$0.336_{\pm 0.048}$	$0.039_{\pm 0.038}$	$0.159_{\pm 0.077}$
SatCLIP	*LR*	$0.107_{\pm 0.005}$	$0.072_{\pm 0.021}$	$0.067_{\pm 0.049}$	$0.043_{\pm 0.003}$	$-0.003_{\pm 0.022}$	$0.076_{\pm 0.042}$	$0.032_{\pm 0.086}$	$0.184_{\pm 0.061}$
TESSERA	*R*	$\underline{0.322}_{\pm 0.006}$	$\mathbf{0.631}_{\pm 0.007}$	$\underline{0.675}_{\pm 0.022}$	$0.040_{\pm 0.001}$	$\underline{0.581}_{\pm 0.053}$	$\underline{0.595}_{\pm 0.035}$	$\underline{0.396}_{\pm 0.155}$	$0.465_{\pm 0.050}$
AlphaEarth	*R*	$\mathbf{0.346}_{\pm 0.004}$	$\underline{0.601}_{\pm 0.008}$	$\mathbf{0.695}_{\pm 0.023}$	$\underline{0.039}_{\pm 0.001}$	$\underline{0.612}_{\pm 0.047}$	$\mathbf{0.650}_{\pm 0.036}$	$0.390_{\pm 0.122}$	$\mathbf{0.509}_{\pm 0.061}$
AETHER	*PR*	$\underline{0.343}_{\pm 0.008}$	$\underline{0.566}_{\pm 0.007}$	$\underline{0.694}_{\pm 0.018}$	$\mathbf{0.037}_{\pm 0.001}$	$\mathbf{0.750}_{\pm 0.066}$	$\underline{0.537}_{\pm 0.057}$	$\mathbf{0.568}_{\pm 0.064}$	$\underline{0.500}_{\pm 0.052}$

macro precision, with macro F1 used as the primary metric because it gives equal weight to each class under imbalanced labels. For age-distribution prediction, we report KL divergence, Chebyshev distance, and L1 distance between the predicted and target distributions, with KL divergence used as the primary metric. Higher values are better for R^2, F1, recall, and precision, while lower values are better for MAE, RMSE, KL divergence, Chebyshev distance, and L1 distance. Regression targets are standardized using statistics from the training partition, and MAE and RMSE are reported in standardized target units. Formal metric definitions are provided in Appendix C.4.

Training and aggregation. All downstream predictors use the same MLP architecture and training protocol across models and tasks. The main probe has hidden dimension 1024 and is trained with Adam at learning rate 10^{-3}, weight decay 0, batch size 512, and a maximum of 100 epochs, with validation-loss early stopping (patience 10). Input embeddings are normalized to unit ℓ_2 norm row by row, without feature-wise z-standardization. For each model–task–city setting, we use the five fixed seeds defined in Section 3.3. The same seed determines the spatial block assignment, downstream-head initialization, and shuffled training minibatches, and does not affect the frozen representation. We average the primary test metric over seeds to obtain a city-level score. Table 2 reports, for each model and task, the mean primary metric across cities and seeds ± the standard deviation across the five seed-specific city averages. Because each seed also determines a spatial partition, this variability captures sensitivity to both partitioning and downstream-head training. To diagnose model–city interaction, we rank the 11 models within each city and compute Kendall's coefficient of concordance W across the city-specific rankings for each task. Larger W indicates that cities produce more similar model orderings. Rank-based diagnostic views are provided in Appendix D. Additional details on training and aggregation are provided in Appendix C.5; robustness to probe capacity and seed variation is reported in Appendix E.

4 Experiments

We evaluate eleven representation models on eight tasks across eight cities. The experiments examine four questions: which representations are the best under a unified protocol, whether performance is stable across cities and tasks, how much random splits overestimate generalization compared with spatial splits, and how well downstream supervision transfers to a city whose labels are withheld.

4.1 Main Performance

Table 2 shows that large-scale pretrained geospatial representations achieve the strongest overall transfer. AETHER, AlphaEarth, and TESSERA obtain the strongest results across most tasks.

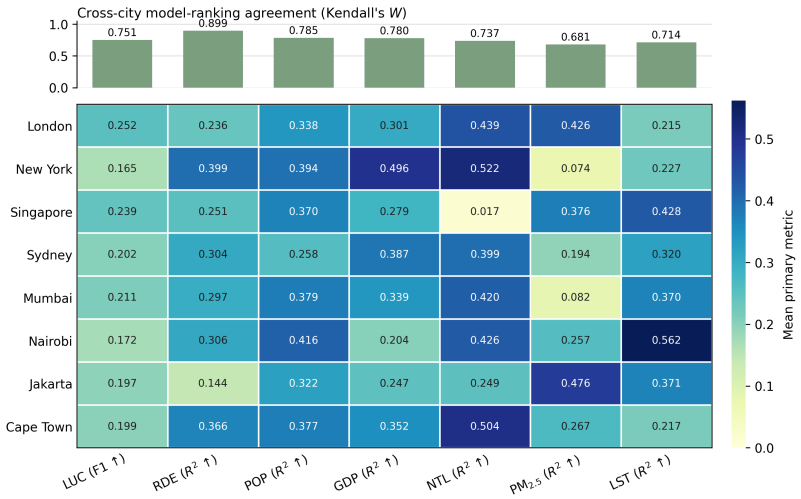


Figure 2: **Performance and model-ranking agreement across cities and tasks.** Heatmap cells report the city–task primary metric averaged over 11 models. Upper bars report Kendall's W across city-specific model rankings, where larger values indicate stronger agreement. LUC uses macro F1 and the six regression tasks use R^2. AGE is omitted because its main evaluation includes four cities.

This suggests that broad spatial coverage and large-scale pretraining are highly valuable when a representation is expected to support heterogeneous urban prediction tasks. These patterns are robust to downstream probe capacity, as shown by the linear-probe evaluation in Appendix E.2.

However, performance is not determined by pretraining scale alone. Several specialized or simple representations remain competitive on specific tasks. CityFM achieves strong performance on LST, indicating that map entities can provide useful signals for built intensity and thermal conditions. CaLLiPer is also competitive on AGE. Even the PE baseline performs strongly on the spatially smoother environmental task $PM_{2.5}$, suggesting that location-only signals can be informative when the target exhibits broad spatial gradients. These results show that simpler, or more targeted representations can still be effective when their encoded signals align well with the downstream phenomenon.

Another observation is that input modality alone is insufficient to explain performance across tasks. More modalities do not necessarily translate into stronger performance. At the same time, models with similar modalities exhibit distinct task-specific strengths. Among raster or raster-enhanced embeddings, AlphaEarth performs best on LUC, POP, NTL, and LST, TESSERA leads RDE, and AETHER is strongest on GDP, $PM_{2.5}$, and AGE. POI- and entity-based models also differ substantially, with CaLLiPer outperforming Place2Vec and Space2Vec overall and CityFM remaining competitive on LST. These results indicate that benchmark performance is shaped not only by input modality, but also by model architecture, pretraining objective, spatial support, and alignment strategy.

4.2 Performance across Cities and Tasks

Figure 2 shows two forms of cross-city variation. The heatmap captures differences in average task performance across cities, while the upper bars measure whether the relative ordering of representation models is consistent. City effects are strongly task dependent. For example, New York has the highest average performance on GDP and NTL but the lowest on $PM_{2.5}$, while Nairobi has the highest LST performance but the lowest GDP performance. There is therefore no single notion of an easy or difficult city across tasks.

Model comparisons also vary across cities. Kendall's W ranges from 0.681 to 0.899 across the seven tasks, indicating substantial but incomplete agreement among city-specific model rankings. RDE has the highest agreement, while $PM_{2.5}$ shows the largest variation in model ordering. These differences are visible in individual comparisons. On $PM_{2.5}$, CaLLiPer outperforms TESSERA in Singapore ($R^2 = 0.608$ versus 0.441), whereas TESSERA leads in Nairobi (0.425 versus 0.277). On GDP, AlphaEarth outperforms TESSERA in Jakarta ($R^2 = 0.597$ versus 0.403), whereas TESSERA leads

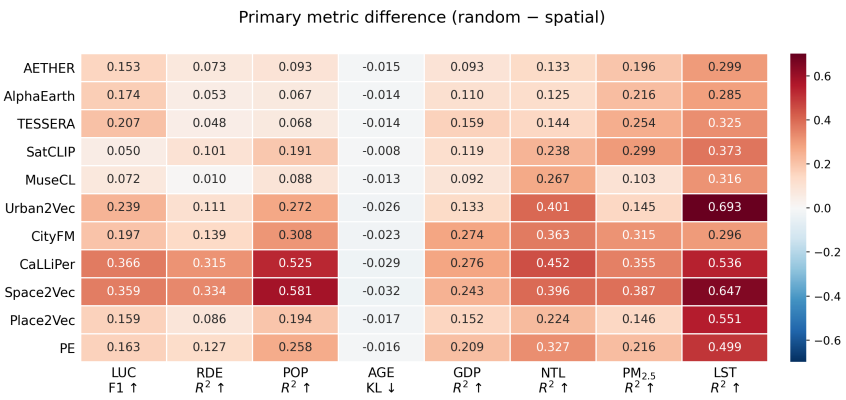


Figure 3: **Effect of random splitting on the primary metric.** Each cell reports random-split minus spatial-split performance. AGE* uses KL divergence on the four-city main evaluation subset, for which lower values are better.

in Cape Town (0.850 versus 0.703). The same set of representations can therefore lead to different comparative conclusions depending on the city used for evaluation.

This variation matters when representations are evaluated in only one or a few cities. A single-city benchmark captures one particular urban setting and can favor models that perform well in that setting. Evaluating multiple cities provides a clearer view of which performance differences persist across urban contexts and which are city dependent. Full city–task results with error bars are provided in Appendix D.1. Broader cross-city patterns are examined in Appendices F and G.

4.3 The Effect of Spatial Splits

Figure 3 compares random splits with the spatial block split used in the main benchmark. Random splits consistently produce higher apparent performance for most tasks and models. This confirms that random splitting is overly optimistic in urban prediction because nearby training and test samples can share strong spatial autocorrelation.

The inflation is largest for densely sampled tasks such as population and LST, where random splits are more likely to place test samples close to training samples. It is smaller for coarser tasks such as GDP. Coordinate-based models such as CaLLiPer and Space2Vec show large gains under random splitting on dense targets, suggesting that smooth coordinate encoders can exploit local interpolation when spatial separation is not enforced. These patterns demonstrate that the effect of split protocol depends on both task resolution and model inductive bias.

This result supports the need for spatially structured evaluation. Random splits measure interpolation within observed urban areas, but they do not reliably measure transfer to unseen areas. CityRep therefore uses spatial block splits as the main protocol and treats random splits as a diagnostic. Full random-split results across tasks, cities, and evaluation metrics are provided in Appendix D.1.

4.4 Cross-City Supervision Transfer Diagnostic

The main spatial-block protocol evaluates generalization to held-out regions within the same city. We additionally examine whether downstream relationships learned in one set of cities remain useful in another city without using its labels for training. This diagnostic concerns unseen downstream supervision rather than unseen geography, since the representation model or its pretraining data may have encountered the target city.

We evaluate all eight tasks and 11 representations. Seven tasks use all eight benchmark cities, while AGE uses the same four-city subset as the main benchmark. For each model, task, target city, and seed, we take the MLP-1024 heads trained independently on the other eligible cities and apply them

Table 3: **Cross-city supervision transfer diagnostic.** Entries report the mean over five seeds $\pm$ standard deviation. Metrics are macro F1 for LUC, KL divergence for AGE ($\downarrow$), and Pearson r for the regression tasks ($\uparrow$). Best and second-best unrounded means are **bold** and underlined.

	Morphology		Demographics		Economy		Environment	
Model	LUC (F1 $\uparrow$)	RDE (r $\uparrow$)	POP (r $\uparrow$)	AGE* (KL $\downarrow$)	GDP (r $\uparrow$)	NTL (r $\uparrow$)	$PM_{2.5}$ (r $\uparrow$)	LST (r $\uparrow$)
PE	$0.040_{\pm 0.002}$	$0.051_{\pm 0.008}$	$-0.030_{\pm 0.043}$	$0.083_{\pm 0.004}$	$-0.038_{\pm 0.016}$	$0.076_{\pm 0.059}$	$0.145_{\pm 0.063}$	$0.106_{\pm 0.032}$
Place2Vec	$0.028_{\pm 0.002}$	$-0.284_{\pm 0.022}$	$-0.144_{\pm 0.023}$	$0.121_{\pm 0.020}$	$-0.240_{\pm 0.044}$	$-0.151_{\pm 0.051}$	$-0.032_{\pm 0.046}$	$-0.153_{\pm 0.022}$
Space2Vec	$0.045_{\pm 0.001}$	$0.034_{\pm 0.021}$	$0.048_{\pm 0.020}$	$0.066_{\pm 0.001}$	$0.045_{\pm 0.049}$	$0.015_{\pm 0.037}$	$0.107_{\pm 0.056}$	$-0.041_{\pm 0.028}$
CaLLiPer	$0.060_{\pm 0.002}$	$0.170_{\pm 0.008}$	$0.218_{\pm 0.030}$	$\underline{0.062}_{\pm 0.001}$	$0.175_{\pm 0.029}$	$0.191_{\pm 0.017}$	$0.018_{\pm 0.027}$	$0.116_{\pm 0.014}$
CityFM	$0.031_{\pm 0.001}$	$-0.057_{\pm 0.073}$	$-0.026_{\pm 0.018}$	$0.071_{\pm 0.002}$	$0.104_{\pm 0.021}$	$0.105_{\pm 0.073}$	$0.188_{\pm 0.036}$	$-0.019_{\pm 0.048}$
Urban2Vec	$0.052_{\pm 0.004}$	$-0.046_{\pm 0.034}$	$0.002_{\pm 0.037}$	$0.073_{\pm 0.002}$	$0.275_{\pm 0.034}$	$0.296_{\pm 0.020}$	$0.099_{\pm 0.019}$	$-0.043_{\pm 0.029}$
MuseCL	$0.053_{\pm 0.002}$	$0.084_{\pm 0.026}$	$0.126_{\pm 0.029}$	$0.094_{\pm 0.028}$	$0.361_{\pm 0.049}$	$0.207_{\pm 0.031}$	$0.078_{\pm 0.044}$	$-0.060_{\pm 0.012}$
SatCLIP	$0.021_{\pm 0.007}$	$-0.026_{\pm 0.022}$	$-0.057_{\pm 0.014}$	$0.400_{\pm 0.108}$	$-0.082_{\pm 0.058}$	$-0.088_{\pm 0.072}$	$0.018_{\pm 0.152}$	$0.049_{\pm 0.065}$
TESSERA	$\mathbf{0.158}_{\pm 0.006}$	$\mathbf{0.637}_{\pm 0.008}$	$\mathbf{0.587}_{\pm 0.016}$	$0.091_{\pm 0.007}$	$\mathbf{0.646}_{\pm 0.014}$	$\underline{0.706}_{\pm 0.012}$	$\underline{0.194}_{\pm 0.043}$	$\underline{0.386}_{\pm 0.013}$
AlphaEarth	$\underline{0.154}_{\pm 0.005}$	$\underline{0.586}_{\pm 0.019}$	$\underline{0.502}_{\pm 0.030}$	$0.125_{\pm 0.050}$	$\underline{0.646}_{\pm 0.010}$	$\mathbf{0.735}_{\pm 0.010}$	$0.192_{\pm 0.017}$	$\mathbf{0.560}_{\pm 0.007}$
AETHER	$0.071_{\pm 0.002}$	$0.234_{\pm 0.019}$	$0.417_{\pm 0.014}$	$\mathbf{0.059}_{\pm 0.001}$	$0.493_{\pm 0.018}$	$0.380_{\pm 0.012}$	$\mathbf{0.230}_{\pm 0.007}$	$0.244_{\pm 0.026}$

to the target city's frozen representations. Predictions from these source-city heads are then averaged. No target-city labels are used to train the heads or construct the ensemble.

For the six regression tasks, labels are standardized using statistics from each source city's training partition. We report Pearson correlation (r) between the averaged predictions and target-city labels, which measures whether relative spatial variation transfers across cities rather than whether absolute target values can be recovered. For LUC, we average the source heads' softmax probabilities and report macro F1. For AGE, we average their predicted distributions and report KL divergence. Table 3 reports the mean and standard deviation across the same five seeds used in the main protocol, with each seed-specific score first averaged over target cities.

The results show a clear difference between broadly pretrained raster representations and most earlier city-specific representations. TESSERA achieves the highest mean on LUC, RDE, POP, and GDP, while AlphaEarth leads NTL and LST. AETHER performs best on AGE and $PM_{2.5}$. TESSERA and AlphaEarth also occupy the top two positions on most of the remaining tasks. In contrast, several earlier representations show weak or even negative cross-city correlations on some regression tasks. These results suggest that large-scale geospatial pretraining produces representations whose downstream structure transfers more consistently across cities.

This diagnostic is more demanding for representations that, by design, learn separate embedding spaces for each city. A predictor trained on one city may not transfer directly when another city's embeddings are expressed in a different feature basis. The cross-city results should therefore be viewed as a complement to the main within-city evaluation rather than a replacement for it. They measure whether downstream structure can be informative across cities without target supervision, under the representation spaces produced by each method.

5 Conclusion

We presented CityRep, a unified benchmark for evaluating urban representations across heterogeneous spatial supports, varying data modalities, diverse cities, and downstream tasks. Our results highlight four key findings: multi-city and multi-task evaluation is necessary because model performance varies substantially across urban contexts and target domains; spatially structured splits are essential because random splits can overestimate spatial generalization performance; broadly pretrained raster representations preserve the strongest transferable relative spatial signal in the cross-city source-head diagnostic; and future urban representation learning could move toward larger-scale pretrained models that better integrate complementary modalities such as remote sensing, POIs, street-view, map entities, and locations. By making these comparisons systematic and reproducible, CityRep provides a foundation for measuring progress toward general-purpose urban representations and for developing urban foundation models.

Reproducibility Statement

The benchmark tasks, spatial alignment rules, spatial block splits, baseline interfaces, and evaluation metrics are described in the main text and appendix. The code is publicly available. The appendix also documents source data, label harmonization, training protocols, and city–task results. These details, together with the benchmark artifacts, support reproduction of the reported comparisons.

Ethics Statement

CityRep uses geospatial datasets and city-specific land-use sources subject to their respective access and redistribution terms. Appendix A discusses source-data limitations, uneven coverage, and the risks of using benchmark scores for policy decisions. The benchmark is intended for research evaluation; deployment in planning or resource allocation requires separate validation and stakeholder review.

AI Use Statement

An LLM-assisted procedure generated the initial mapping of city-specific land-use labels to the benchmark taxonomy. The authors manually audited the resulting mappings and corrected ambiguous or inconsistent categories, as described in the appendix. Generative AI also assisted with experiment implementation and execution, as well as manuscript review and editing.

REFERENCES

American Planning Association. Land based classification standards. https://www.planning.org/lbcs/, 2000. Accessed 2026-05-05.

James R. Anderson, Ernest E. Hardy, John T. Roach, and Richard E. Witmer. A land use and land cover classification system for use with remote sensor data. Professional Paper 964, U.S. Geological Survey, 1976.

Pasquale Balsebre, Weiming Huang, Gao Cong, and Yi Li. City foundation models for learning general purpose representations from openstreetmap. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pp. 87–97, 2024. doi: 10.1145/3627673.3679662.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16726–16736, 2023. doi: 10.1109/ICCV51070.2023.01538.

M. Bondarenko, R. Priyatikanto, N. Tejedor-Garavito, W. Zhang, T. McKeen, A. Cunningham, T. Woods, J. Hilton, D. Cihan, B. Nosatiuk, T. Brinkhoff, A. Tatem, and A. Sorichetta. Constrained estimates of 2015–2030 total number of people per grid square broken down by gender and age groupings at a resolution of 3 arc (approximately 100m at the equator), R2024B version v1. WorldPop, School of Geography and Environmental Science, University of Southampton, 2025a. URL https://doi.org/10.5258/SOTON/WP00805. Global Demographic Data Project, funded by the Bill and Melinda Gates Foundation, grant INV-045237; Creative Commons Attribution 4.0; accessed 2026-05-06.

M. Bondarenko, R. Priyatikanto, N. Tejedor-Garavito, W. Zhang, T. McKeen, A. Cunningham, T. Woods, J. Hilton, D. Cihan, B. Nosatiuk, T. Brinkhoff, A. Tatem, and A. Sorichetta. Constrained estimates of 2015–2030 total number of people per grid square at a resolution of 3 arc (approximately 100m at the equator), R2024B version v1. WorldPop, School of Geography and Environmental Science, University of Southampton, 2025b. URL https://doi.org/10.5258/SOTON/WP00803. Global Demographic Data Project, funded by the Bill and Melinda Gates Foundation, grant INV-045237; Creative Commons Attribution 4.0; accessed 2026-05-06.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025. doi: 10.48550/arXiv.2507.22291.

City of Cape Town. Zoning. Cape Town Open Data Portal feature service, 2019. URL https://odp-cctegis.opendata.arcgis.com/datasets/zoning-1. Zoning / split-zoning feature service; published 2019-02-10; dataset-specific or custom licence; accessed 2026-05-06.

Earth Observation Group, Payne Institute for Public Policy, Colorado School of Mines. VIIRS Nighttime Lights Annual V2.2. https://eogdata.mines.edu/products/vnl/, 2024. Accessed: 2026-05-06.

Christopher D. Elvidge, Mikhail Zhizhin, Tilottama Ghosh, Feng-Chi Hsu, and Jay Taneja. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing*, 13(5):922, 2021. doi: 10.3390/rs13050922.

Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: geographical latent representation for predicting future visitors. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 102–108. AAAI Press, 2017. doi: 10.1609/aaai.v31i1.10500.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, Toby Jackson, James Ball, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34818–34831, 2026.

Foursquare Labs, Inc. FSQ OS Places. Foursquare Open Source Places data product, 2026. URL https://opensource.foursquare.com/. Open POI dataset; Apache License 2.0; preserve Foursquare attribution and NOTICE file; accessed 2026-05-06.

Weiming Huang, Lizhen Cui, Meng Chen, Daokun Zhang, and Yao Yao. Estimating urban functional distributions with semantics preserved poi embedding. *International Journal of Geographical Information Science*, 36(10): 1905–1930, 2022. doi: 10.1080/13658816.2022.2040510.

Weiming Huang, Daokun Zhang, Gengchen Mai, Xu Guo, and Lizhen Cui. Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 134–145, 2023. doi: 10.1016/j.isprsjprs.2022.11.021.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: global, general-purpose location embeddings with satellite imagery. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i4.32457. URL https://doi.org/10.1609/aaai.v39i4.32457.

Matti Kummu, Maria Kosonen, and Sina Masoumzadeh Sayyar. Downscaled gridded global dataset for gross domestic product (GDP) per capita PPP over 1990–2022. *Scientific Data*, 12:178, 2025. doi: 10.1038/s41597-025-04487-x. URL https://doi.org/10.1038/s41597-025-04487-x.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David 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 Xiao Xiang Zhu. Geo-bench: toward foundation models for earth monitoring. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Junyuan Liu, Quan Qin, Guangsheng Dong, Xinglei Wang, Jiazhuang Feng, Zichao Zeng, and Tao Cheng. Beyond alphaearth: toward human-centered spatial representation via poi-guided contrastive learning. *arXiv preprint arXiv:2510.09894*, 2025a. doi: 10.48550/arXiv.2510.09894.

Junyuan Liu, Xinglei Wang, and Tao Cheng. Enriching location representation with detailed semantic information. In *13th International Conference on Geographic Information Science (GIScience 2025)*, volume 346 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 3:1–3:15, 2025b. doi: 10.4230/LIPIcs.GIScience.2025.3.

Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJljdh4KDH.

Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023. doi: 10.1016/j.isprsjprs.2023.05.006.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, Dávid Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. Pangaea: Assessing geospatial foundation models capabilities through a global and inclusive benchmark. *IEEE Geoscience and Remote Sensing Magazine*, 14:245–285, 2026. doi: 10.1109/MGRS.2025.3628194. URL https://doi.org/10.1109/MGRS.2025.3628194.

Meta Platforms, Inc. and Mapillary. Mapillary Open Street-Level Imagery. Mapillary platform / API, 2026. URL https://ai.meta.com/ai-for-good/datasets/open-street-level-imagery/. Street-level imagery and map-data platform; accessed via Mapillary API on 2026-05-06; images are shared under CC BY-SA and individual image/user attribution should be preserved where required.

Hanna Meyer, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss. Importance of spatial predictor variable selection in machine learning applications: Moving from data reproduction to spatial prediction. *Ecological Modelling*, 411:108815, 2019. doi: 10.1016/j.ecolmodel.2019.108815.

Julia Moska, Oleksii Furman, Kacper Kozaczko, Szymon Leszkiewicz, Jakub Polczyk, Piotr Gramacki, and Piotr Szymański. OBSR: Open benchmark for spatial representations. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '25, 2025. doi: 10.1145/3748636.3762783. URL https://doi.org/10.1145/3748636.3762783.

Municipal Corporation of Greater Mumbai. Development Plan for Greater Mumbai 2014–2034: Existing Land Use Maps and Report. Municipal Corporation of Greater Mumbai, 2012. URL https://www.loginmumbai.org/sources.html. Existing Land Use 2012 source; cited via UDRI/Login Mumbai data-source register; accessed 2026-05-06.

New York City Department of City Planning. Primary Land Use Tax Lot Output (PLUTO), version 25v4. NYC Open Data / BYTES of the BIG APPLE, 2026. URL https://catalog.data.gov/dataset/primary-land-use-tax-lot-output-pluto. CSV tax-lot records; current version 25v4; last modified 2026-02-20; accessed 2026-05-06.

NSW Department of Planning, Housing and Infrastructure. Environmental Planning Instrument – Land Zoning. NSW Planning Portal Open Data, 2022. URL https://www.planningportal.nsw.gov.au/opendata/dataset/environment-planning-instrument-local-environmental-plan-land-zoning. Vector land-zoning dataset; update frequency weekly; Creative Commons Attribution; accessed 2026-05-06.

OpenStreetMap contributors. OpenStreetMap road-network extract for the CityRep benchmark cities. OpenStreetMap database, 2026. URL https://www.openstreetmap.org/copyright. Drivable highway ways extracted for London, New York, Singapore, Sydney, Mumbai, Nairobi, Jakarta, and Cape Town; licensed under the Open Data Commons Open Database License (ODbL) 1.0; accessed/extracted 2026-05-06.

Pemerintah Provinsi DKI Jakarta. Peraturan Gubernur Nomor 31 Tahun 2022 tentang Rencana Detail Tata Ruang Wilayah Perencanaan Provinsi Daerah Khusus Ibukota Jakarta. JDIH Provinsi DKI Jakarta, 2022. URL https://jdih.jakarta.go.id/dokumen/detail/4933/peraturan-gubernur-nomor-31-tahun-2022-tentang-rencana-detail-tata-ruang-wilayah-perencanaan-provinsi-daerah-khusus-ibukota-jakarta. Peraturan Gubernur No. 31; ditetapkan 2022-06-27; BD Provinsi DKI Jakarta Tahun 2022 Nomor 63003; accessed 2026-05-06.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten F. Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1):4540, 2020. doi: 10.1038/s41467-020-18321-y. URL https://doi.org/10.1038/s41467-020-18321-y.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. Torchgeo: Deep learning with geospatial data. *ACM Transactions on Spatial Algorithms and Systems*, 11(4), 2025. doi: 10.1145/3707459. URL https://doi.org/10.1145/3707459. Article 15.

Uber Technologies, Inc. H3 indexes points and shapes into a hexagonal grid, 2026. URL https://h3geo.org/. https://h3geo.org/. Data available under the Apache 2.0 License. Accessed 06/05/2026.

Urban Redevelopment Authority. Master Plan 2019 Land Use layer (2025). data.gov.sg, 2023. URL https://data.gov.sg/datasets/d_90d86daa5bfaa371668b84fa5f01424f/view. Indicative land-use polygon layer; data from Dec. 2019; extracted from the Master Plan 2019 approved on 2019-11-18; last updated 2025-12-05; accessed 2026-05-06.

Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232, 2019. doi: 10.1111/2041-210X.13107.

A. van Donkelaar, M. S. Hammer, L. Bindle, M. Brauer, J. R. Brook, M. J. Garay, N. C. Hsu, O. V. Kalashnikova, R. A. Kahn, C. Lee, R. C. Levy, A. Lyapustin, A. M. Sawyer, and R. V. Martin. Global Annual PM2.5 Grids from MODIS, MISR, SeaWiFS and VIIRS Aerosol Optical Depth (AOD), 1998–2022, V5.GL.04, Version 5.04. NASA Socioeconomic Data and Applications Center (SEDAC), 2024. URL https://doi.org/10.7927/AS2R-9P42. Data set; accessed 2026-05-06.

Zhengming Wan, Simon Hook, and Glynn Hulley. MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061. NASA Land Processes Distributed Active Archive Center, 2021. URL https://doi.org/10.5067/MODIS/MOD11A2.061. Data set; accessed 2026-05-06.

Xinglei Wang, Tao Cheng, Stephen Law, Zichao Zeng, Lu Yin, and Junyuan Liu. Multi-modal contrastive learning of urban space representations from poi data. *Computers, Environment and Urban Systems*, 120: 102299, 2025. doi: 10.1016/J.COMPENVURBSYS.2025.102299.

Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1013–1020, 2020. doi: 10.1609/aaai.v34i01.5450.

Ya Wen, Jixuan Cai, Qiyao Ma, Linyan Li, Xinhuan Chen, Chris Webster, and Yulun Zhou. MoRA: Mobility as the backbone for geospatial representation learning at scale. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IlBr5JJsCj.

World Bank Data Catalog. Nairobi, Kenya Land Use 2010. World Bank Data Catalog, 2010. URL https://datacatalog.worldbank.org/search/dataset/0041743/nairobi-kenya-land-use-2010. Land-use shapefile; source: Columbia University's Center for Sustainable Urban Development; shapefile last updated 2020-06-25; metadata last updated 2023-01-19; Creative Commons Attribution 4.0; accessed 2026-05-06.

Nemin Wu, Qian Cao, Zhangyu Wang, Zeping Liu, Yanlin Qi, Jielu Zhang, Joshua Ni, Xiaobai Yao, Hongxu Ma, Lan Mu, Stefano Ermon, Tanuja Ganu, Akshay Nambi, Ni Lao, and Gengchen Mai. Torchspatial: A location encoding framework and benchmark for spatial representation learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 81437–81460, 2024. URL https://arxiv.org/abs/2406.15658.

Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450354905. doi: 10.1145/3139958.3140054. URL https://doi.org/10.1145/3139958.3140054.

Xixian Yong and Xiao Zhou. Musecl: predicting urban socioeconomic indicators via multi-semantic contrastive learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/834. URL https://doi.org/10.24963/ijcai.2024/834.

Appendix Contents

A Discussion **16**

A.1 Limitations 16

A.2 Broader Impacts 16

B Details of Downstream Task Data **16**

B.1 Raw Data 16

B.2 Land Use Label Mapping 21

C Additional Experimental Details **25**

C.1 Spatial Split 25

C.2 Details of Baselines Reproduction 26

C.3 Aggregation for Sparse Entity Representations 29

C.4 Evaluation Metrics 29

C.5 Training Protocol 31

D Comprehensive Results **32**

D.1 Results Across Tasks and Cities 32

E Robustness to Benchmark Design Choices **37**

E.1 Raster Pooling 37

E.2 Downstream Probe Capacity 37

E.3 Results with a 20-by-20 Spatial Split 38

E.4 Seed Stability 38

F Exploratory Extension to 26 Cities **39**

G Analysis of Performance Heterogeneity Across Tasks and Cities **41**

A Discussion

A.1 Limitations

CityRep is designed to make urban representation evaluation broader and more comparable, but several limitations and scope boundaries remain. First, the benchmark is constrained by data quality in both downstream task construction and baseline reproduction. The downstream labels are derived from public geospatial products and city-specific land-use sources, whose uncertainty, update cycles, spatial resolutions, and semantic definitions vary across regions. CityRep should therefore be interpreted as an evaluation of representations under the released data products, rather than as a definitive measurement of the underlying urban phenomena. At the same time, reproducing existing representation models depends on the availability and quality of their required pretraining inputs. Some original models rely on proprietary, city-specific, or unavailable modalities, so public reproductions may not capture every advantage of the original models under their original data environments. Second, spatial splitting remains a methodological choice rather than a settled standard. CityRep uses block-based spatial splits to reduce leakage and evaluate generalization to held-out urban areas, but other partitioning strategies are possible, such as administrative regions, spatial clustering, distance buffers, or task-specific regionalization. The most appropriate split may depend on task resolution, spatial autocorrelation, city morphology, and the intended generalization target. Further research is needed to understand how different spatial validation designs affect conclusions in urban representation learning. Third, CityRep does not fully investigate how multimodal data scarcity shapes global urban embeddings. Urban representations may depend on remote sensing, POIs, street-view imagery, road networks, mobility data, and administrative records, but these modalities are unevenly available across cities and regions. Such an imbalance may affect what representations learn and may introduce geographic or socioeconomic biases into models intended for broad deployment. The main benchmark evaluates representations at the intra-city level, while the complementary cross-city source-head diagnostic provides a deliberately stringent test of downstream-supervision transfer. Standard pooled-source training, calibrated absolute-value prediction, and target-city adaptation remain to be investigated further. A systematic study of how missing or uneven multimodal data influences global urban representation learning remains an important direction for future work.

A.2 Broader Impacts

CityRep is intended as an evaluation tool for understanding what urban representations encode and how reliably they transfer across tasks and cities. A standardized benchmark can reduce duplicated preprocessing effort, make model claims more comparable, and encourage evaluation protocols that better reflect spatial generalization. It may also help researchers identify where existing urban representations fail, especially in cities with limited data availability or urban forms underrepresented in current foundation models.

At the same time, urban prediction systems can influence planning, resource allocation, environmental assessment, and infrastructure decisions. Benchmark scores should therefore not be treated as evidence that a model is ready for direct policy deployment. Downstream use requires task-specific validation, uncertainty analysis, stakeholder review, and careful consideration of social context. The benchmark uses aggregated spatial labels, public geospatial layers, and derived task units; nevertheless, urban data can still encode socioeconomic inequalities and uneven data coverage. We therefore release CityRep as a research benchmark rather than a decision-making system, and we encourage users to report failures, biases, and city-specific limitations alongside aggregate scores.

B Details of Downstream Task Data

B.1 Raw Data

The released task data are processed into a common city–task registry, but the source layers retain different spatial supports, semantics, and access terms. Land use is a point-level classification task derived from city-specific zoning or land-use sources and harmonized into a shared taxonomy (New York City Department of City Planning, 2026; Urban Redevelopment Authority, 2023; NSW Department of Planning, Housing and Infrastructure, 2022; City of Cape Town, 2019; Pemerintah Provinsi DKI Jakarta, 2022; Municipal Corporation of Greater Mumbai, 2012; World Bank Data Catalog,

Table 4: Raw data sources used to construct CityRep downstream tasks. Task abbreviations follow the main text: LUC = land-use classification, RDE = road-density estimation, POP = population prediction, AGE = age-distribution prediction, GDP = gross-domestic-product prediction, NTL = nighttime-lights prediction, $PM_{2.5}$ = $PM_{2.5}$ prediction, and LST = land-surface-temperature prediction. Native support describes the source product before CityRep processing; released support describes the spatial support used by the benchmark task files. Years in the raw-source column denote the actual source or CityRep task version used by the released benchmark files.

Task	Raw source	Native support	Released support	License / terms
Morphology				
LUC	City-specific official or public zoning / land-use layers (New York City Department of City Planning, 2026; Urban Redevelopment Authority, 2023; NSW Department of Planning, Housing and Infrastructure, 2022; City of Cape Town, 2019; Pemerintah Provinsi DKI Jakarta, 2022; Municipal Corporation of Greater Mumbai, 2012; World Bank Data Catalog, 2010)	Vector polygons or parcel-level records	Harmonized sampled point labels, subject to source terms	Varies by city; CityRep releases labels only where redistribution is permitted by source terms.
RDE	OpenStreetMap drivable highway ways extracted for the CityRep 2026 road-density task version (OpenStreetMap contributors, 2026)	Vector road geometries	Road length density on the task grid	ODbL 1.0; OpenStreetMap attribution required.
Demographics				
POP	WorldPop constrained population counts, R2024B, 2024 (Bondarenko et al., 2025b)	3 arc-second raster, approximately 100m at the equator	Same task raster support	CC BY 4.0.
AGE	WorldPop age–sex structured population counts, R2024B, 2024 (Bondarenko et al., 2025a)	3 arc-second rasters by age and sex group	Ten-bin age distribution on the task raster support	CC BY 4.0.
Economy				
GDP	Gridded GDP total, 2024 band (Kummu et al., 2025)	30 arc-second gridded GDP product	City-clipped task grid	CC BY 4.0 for the public Kummu GDP product.
NTL	VIIRS Nighttime Lights Annual V2.2 average masked radiance composite, 2024 (Earth Observation Group, Payne Institute for Public Policy, Colorado School of Mines, 2024; Elvidge et al., 2021)	Annual cloud-free radiance composite at approximately 15 arc-second resolution	City-clipped nighttime-lights task grid	Public domain / open EOG product; cite product page and source paper.
Environment				
$PM_{2.5}$	SEDAC/CIESIN Global Annual $PM_{2.5}$ Grids, V5.GL.04, 2022 (van Donkelaar et al., 2024)	0.01-degree WGS84 raster	City-clipped task grid	CIESIN open data policy; cite DOI and follow SEDAC/NASA use terms.
LST	MODIS/Terra MOD11A2.061 8-day daytime land surface temperature annual mean, 2024 (Wan et al., 2021)	8-day, 1km MODIS sinusoidal grid	Annual daytime mean on the CityRep task support	NASA/LP DAAC data are openly shared without restriction; cite DOI.

2010). Because land-use redistribution terms vary by city, CityRep publicly releases harmonized land-use labels only where permitted by the corresponding source terms. Road density is computed from OpenStreetMap road geometries by aggregating drivable road length to the task grid (OpenStreetMap contributors, 2026). Population and age distribution are derived from WorldPop products, with population represented as a scalar count and age represented as a distribution over ten bins (0–4, 5–14, 15–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, and 85+) (Bondarenko et al., 2025b;a). GDP uses a gridded economic output product (Kummu et al., 2025), nighttime lights use the VIIRS Nighttime Lights Annual V2.2 average masked radiance composite (Earth Observation Group, Payne Institute for Public Policy, Colorado School of Mines, 2024; Elvidge et al., 2021), $PM_{2.5}$ uses global annual concentration grids (van Donkelaar et al., 2024), and LST uses MODIS/Terra daytime land surface temperature aggregated to an annual mean (Wan et al., 2021).

Figures 4–11 visualize the processed downstream labels for all eight cities. The visualizations are included to show both the diversity of task supports and the differences in city extent. Some tasks are dense and fine grained, such as road density, population, and LST. Others are coarser, such as GDP, nighttime lights, and $PM_{2.5}$. CityRep preserves these task-specific supports instead of forcing all labels onto a single global grid, because the alignment module is responsible for mapping each representation to the target task unit.

Table 4 summarizes the raw data sources used to construct the eight downstream tasks. Because the benchmark preserves task-specific spatial supports, the "resolution/support" column reports the native or effective support of the released task labels rather than forcing all sources into a single grid.

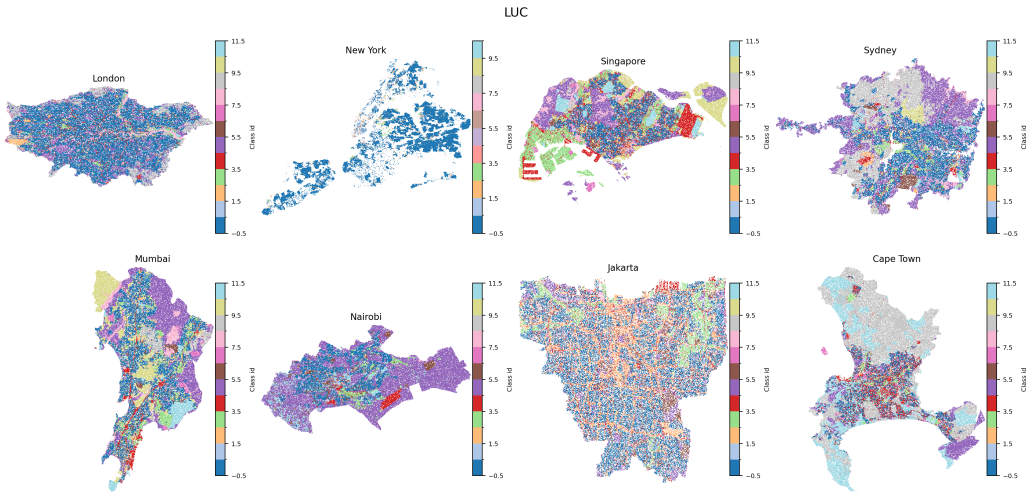


Figure 4: Land-use (LUC) labels across the eight benchmark cities.

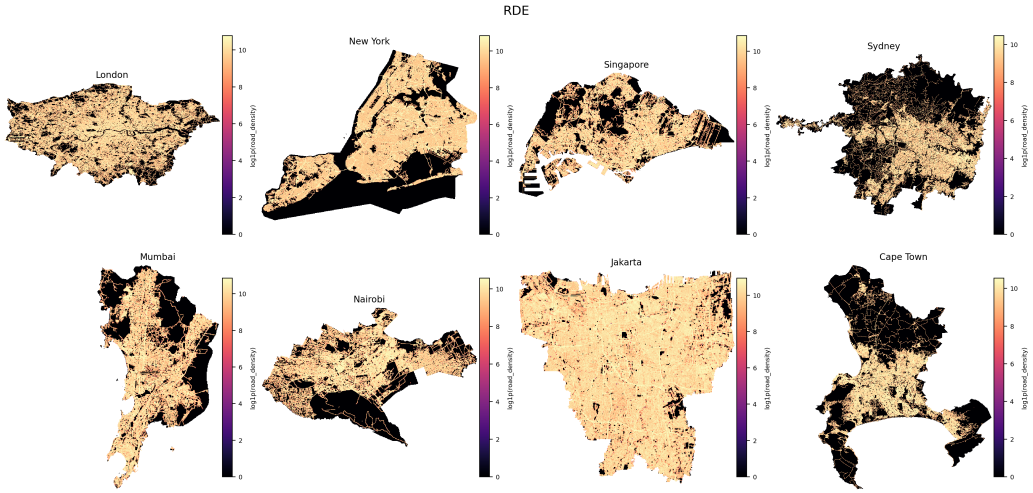


Figure 5: Road-density (RDE) labels across the eight benchmark cities.

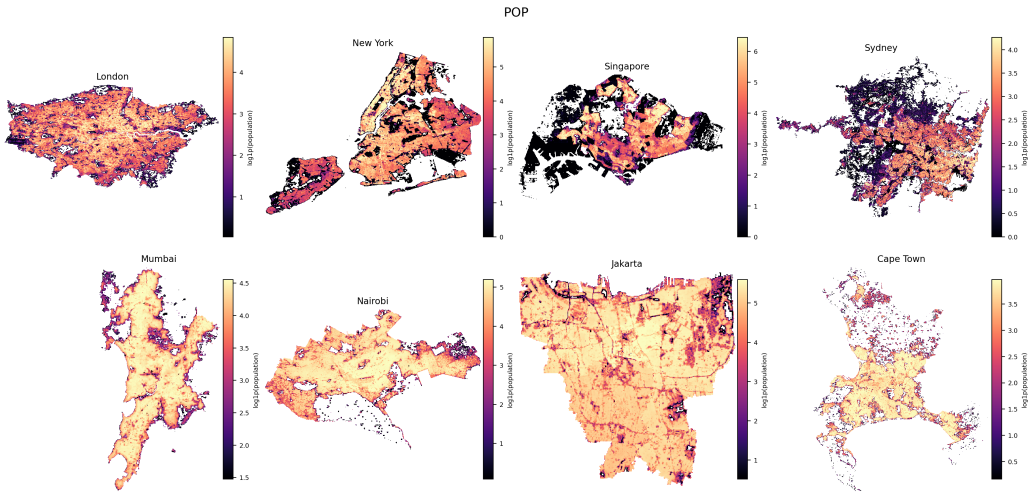


Figure 6: Population (POP) labels across the eight benchmark cities.

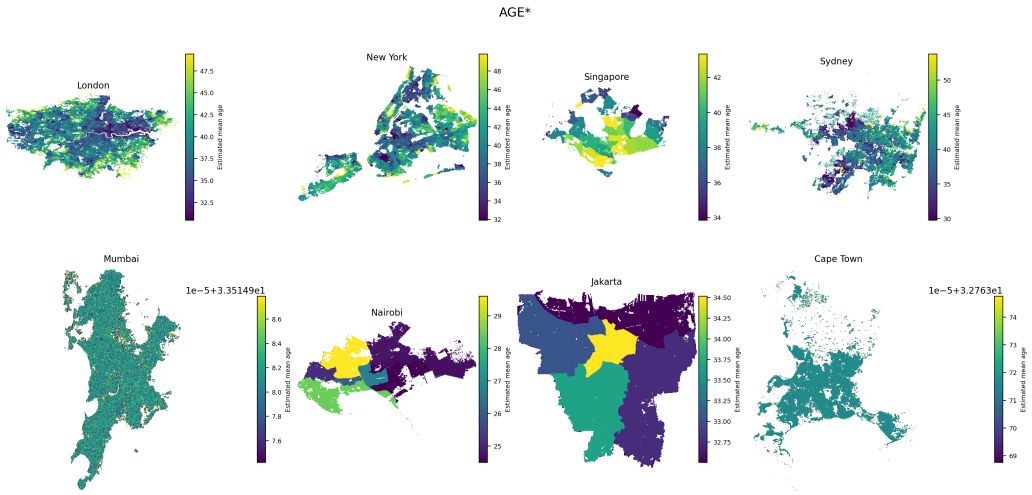


Figure 7: Age-distribution (AGE) labels across the eight benchmark cities.

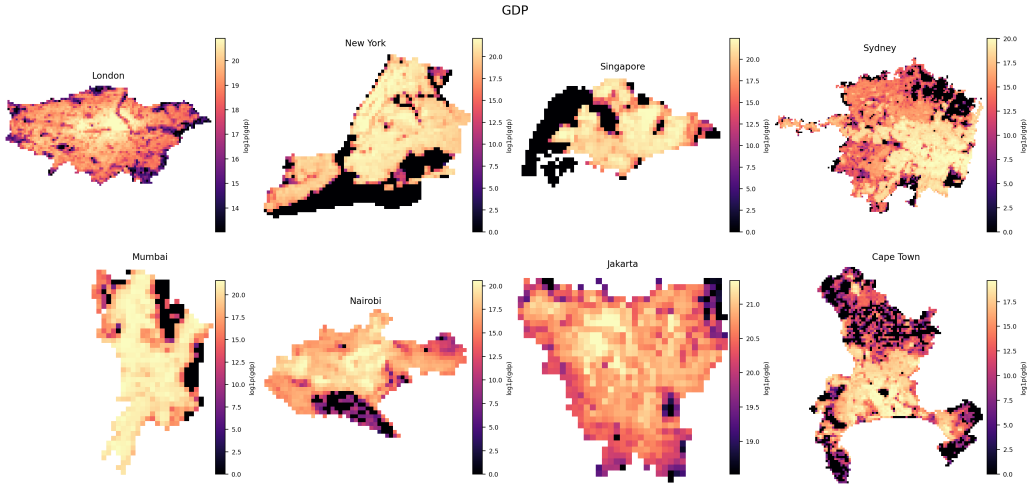


Figure 8: Gross domestic product (GDP) labels across the eight benchmark cities.

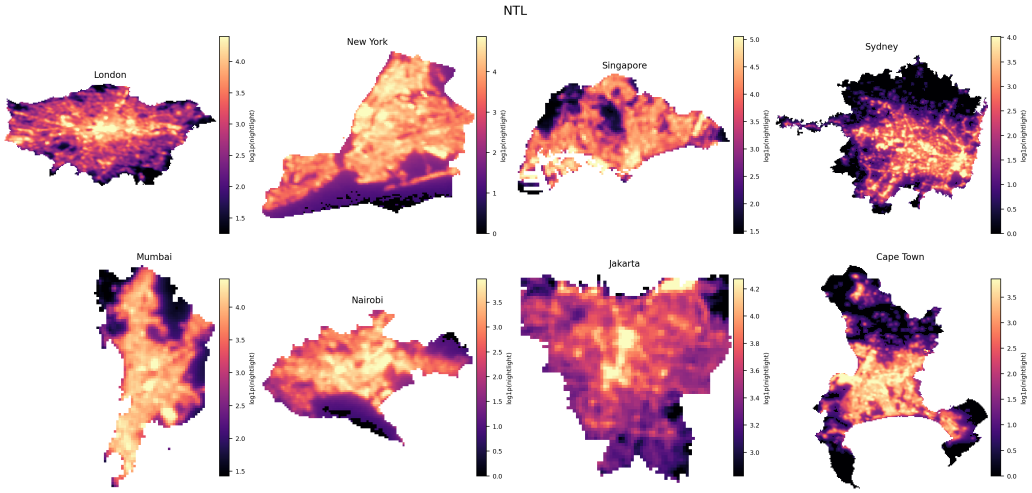


Figure 9: Nighttime lights (NTL) labels across the eight benchmark cities.

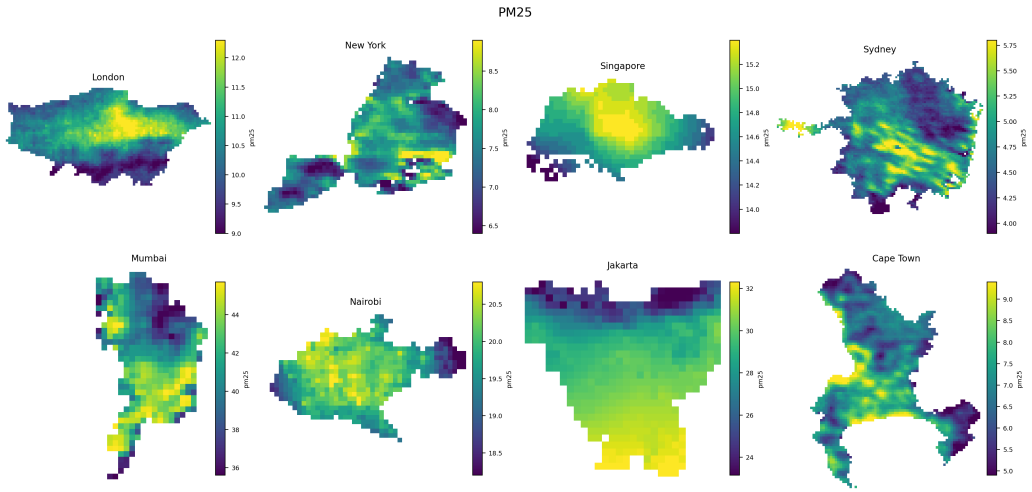


Figure 10: $PM_{2.5}$ labels across the eight benchmark cities.

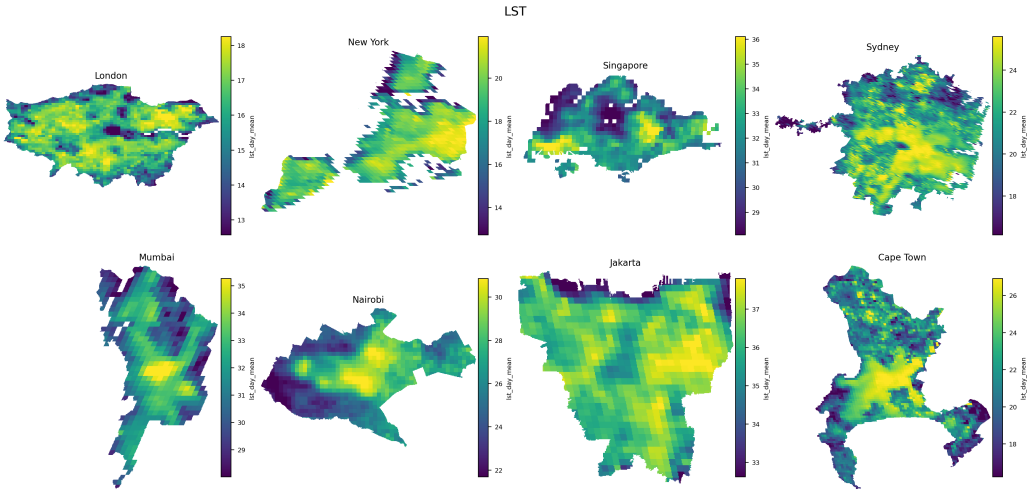


Figure 11: Land-surface-temperature labels across the eight benchmark cities.

B.2 Land Use Label Mapping

We harmonize city-specific land-use labels into twelve benchmark classes: Residential, Mixed Use, Commercial, Industrial, Transportation, Green / Recreation, Institutional / Civic, Utilities, Water, Agriculture / Rural, Vacant / Reserve, and Other. These classes are informed by established land-use classification schemes, including the Land-Based Classification Standards (LBCS) developed by the American Planning Association and partner agencies (American Planning Association, 2000), and the USGS Anderson land-use/land-cover classification system (Anderson et al., 1976). Both schemes distinguish broad urban functions such as residential, commercial, industrial, transportation, utilities, institutional/public uses, recreation, agriculture, and water. We adapt these high-level categories to a compact benchmark taxonomy so that labels from heterogeneous municipal planning systems can be compared across cities.

The mapping is city-specific because local planning systems use different category names and different levels of detail. For example, a source category may explicitly identify utilities in one city but group the same function under public facilities in another. We therefore maintain a mapping table for each city and export both a readable mapping table and audit files with source categories, standardized labels, and counts.

We used an LLM-assisted procedure to produce the initial harmonization from source labels to the twelve benchmark classes. The resulting mappings were then manually audited by the authors. Ambiguous or inconsistent categories were corrected before final task reconstruction. Categories that were duplicated or semantically equivalent in the source mapping were also deduplicated. The purpose of this review is not to impose a universal planning ontology, but to ensure that the benchmark classes are consistent enough for cross-city evaluation.

Table 5 reports the readable version of the city-specific label mapping used in the benchmark. For compactness, the Sydney mapping is omitted from this appendix table; the complete mapping, including Sydney, is provided in the released code.

Table 5: Mapping of city-specific land-use source labels, excluding Sydney.

City	Source label	Mapped class	City	Source label	Mapped class
London	High density residential with retail and commercial sites	Residential	London	Low density residential with amenities (suburbs and small villages / hamlets)	Residential
London	Medium density residential with high streets and amenities	Residential	London	Urban centres - mainly commercial/retail with residential pockets	Mixed Use
London	Business parks	Commercial	London	Large complex buildings various use (travel/recreation/ retail)	Commercial
London	Retail parks	Commercial	London	Industrial areas	Industrial
London	Principle Transport	Transportation	London	Coastal dunes	Green / Recreation
London	Coniferous and undifferentiated woodland	Green / Recreation	London	Deciduous woodland	Green / Recreation
London	Open or heath and moor land	Green / Recreation	London	Recreational land	Green / Recreation
London	Wetlands	Green / Recreation	London	Coastal water	Water
London	Inland Water	Water	London	Agriculture - mainly crops	Agriculture / Rural
London	Agriculture - mixed use	Agriculture / Rural	London	Farms	Agriculture / Rural
London	Glasshouses	Agriculture / Rural	London	Orchards	Agriculture / Rural
London	Mining and spoil areas	Other	New York	Multi-Family Elevator Buildings	Residential
New York	Multi-Family Walk-Up Buildings	Residential	New York	One & Two Family Buildings	Residential

Continued on next page

Table 5: City-specific land-use source labels mapped to the common taxonomy, excluding Sydney (continued).

City	Source label	Mapped class	City	Source label	Mapped class
New York	Mixed Residential & Commercial Buildings	Mixed Use	New York	Commercial & Office Buildings	Commercial
New York	Industrial & Manufacturing	Industrial	New York	Parking Facilities	Transportation
New York	Transportation & Utility	Transportation	New York	Open Space & Outdoor Recreation	Green / Recreation
New York	Public Facilities & Institutions	Institutional / Civic	New York	Vacant Land	Vacant / Reserve
Singapore	RESIDENTIAL	Residential	Singapore	COMMERCIAL & RESIDENTIAL	Mixed Use
Singapore	COMMERCIAL / INSTITUTION	Mixed Use	Singapore	RESIDENTIAL / INSTITUTION	Mixed Use
Singapore	RESIDENTIAL WITH COMMERCIAL AT 1ST STOREY	Mixed Use	Singapore	WHITE	Mixed Use
Singapore	COMMERCIAL	Commercial	Singapore	HOTEL	Commercial
Singapore	BUSINESS 1	Industrial	Singapore	BUSINESS 1 - WHITE	Industrial
Singapore	BUSINESS 2	Industrial	Singapore	BUSINESS 2 - WHITE	Industrial
Singapore	BUSINESS PARK	Industrial	Singapore	BUSINESS PARK - WHITE	Industrial
Singapore	LIGHT RAPID TRANSIT	Transportation	Singapore	MASS RAPID TRANSIT	Transportation
Singapore	PORT / AIRPORT	Transportation	Singapore	ROAD	Transportation
Singapore	TRANSPORT FACILITIES	Transportation	Singapore	BEACH AREA	Green / Recreation
Singapore	CEMETERY	Green / Recreation	Singapore	OPEN SPACE	Green / Recreation
Singapore	PARK	Green / Recreation	Singapore	SPORTS & RECREATION	Green / Recreation
Singapore	CIVIC & COMMUNITY INSTITUTION	Institutional / Civic	Singapore	EDUCATIONAL INSTITUTION	Institutional / Civic
Singapore	HEALTH & MEDICAL CARE	Institutional / Civic	Singapore	PLACE OF WORSHIP	Institutional / Civic
Singapore	UTILITY	Utilities	Singapore	WATERBODY	Water
Singapore	AGRICULTURE	Agriculture / Rural	Singapore	RESERVE SITE	Vacant / Reserve
Singapore	SPECIAL USE	Other	Mumbai	Residential	Residential
Mumbai	Slum / Cluster	Residential	Mumbai	Urban Villages	Residential
Mumbai	Commercial Activities	Commercial	Mumbai	Informal Market	Commercial
Mumbai	Municipal Market	Commercial	Mumbai	Other Offices	Commercial
Mumbai	Industrial Use	Industrial	Mumbai	Transport	Transportation
Mumbai	Cemetery	Green / Recreation	Mumbai	Natural Areas	Green / Recreation
Mumbai	Open Spaces	Green / Recreation	Mumbai	Swimming Pool	Green / Recreation
Mumbai	Educational Amenities	Institutional / Civic	Mumbai	Government Office	Institutional / Civic
Mumbai	Law and Order	Institutional / Civic	Mumbai	Medical Amenities	Institutional / Civic
Mumbai	Municipal Chowkies	Institutional / Civic	Mumbai	Municipal Office	Institutional / Civic
Mumbai	Social Amenities	Institutional / Civic	Mumbai	Communication	Utilities
Mumbai	Public Utility and Facity	Utilities	Mumbai	Town Duty / Octroi Office	Utilities
Mumbai	Water	Water	Mumbai	Primary Activity	Agriculture / Rural
Mumbai	Area Under SPA	Vacant / Reserve	Mumbai	Under Construction	Vacant / Reserve

Continued on next page

Table 5: City-specific land-use source labels mapped to the common taxonomy, excluding Sydney (continued).

City	Source label	Mapped class	City	Source label	Mapped class
Mumbai	Vacant	Vacant / Reserve	Mumbai	Unclassified	Other
Nairobi	res_slum	Residential	Nairobi	residential	Residential
Nairobi	mixed CI	Mixed Use	Nairobi	mixed RC	Mixed Use
Nairobi	commercial	Commercial	Nairobi	industrial	Industrial
Nairobi	transportation	Transportation	Nairobi	open space	Green / Recreation
Nairobi	open_space	Green / Recreation	Nairobi	recreational	Green / Recreation
Nairobi	institutional	Institutional / Civic	Nairobi	water	Water
Nairobi	no_structures	Vacant / Reserve	Nairobi	unknown	Other
Jakarta	Zona Perumahan	Residential	Jakarta	Zona Pariwisata	Commercial
Jakarta	Zona Perdagangan dan Jasa	Commercial	Jakarta	Zona Perkantoran	Commercial
Jakarta	Zona Kawasan Peruntukan Industri	Industrial	Jakarta	Zona Badan Jalan	Transportation
Jakarta	Zona Transportasi	Transportation	Jakarta	Zona Ekosistem Mangrove	Green / Recreation
Jakarta	Zona Hutan Lindung	Green / Recreation	Jakarta	Zona Hutan produksi	Green / Recreation
Jakarta	Zona Konservasi	Green / Recreation	Jakarta	Zona Perlindungan Setempat	Green / Recreation
Jakarta	Zona Ruang Terbuka Hijau	Green / Recreation	Jakarta	Zona Pelayanan Umum	Institutional / Civic
Jakarta	Zona Pertahanan dan Keamanan	Institutional / Civic	Jakarta	Zona Pembangkit Tenaga Listrik	Utilities
Jakarta	Zona Badan Air	Water	Jakarta	Zona Perikanan	Agriculture / Rural
Jakarta	Zona Pertanian	Agriculture / Rural	Jakarta	0	Other
Cape Town	General Residential 1 : Group Housing	Residential	Cape Town	General Residential 2	Residential
Cape Town	General Residential 3	Residential	Cape Town	General Residential 4	Residential
Cape Town	General Residential 5	Residential	Cape Town	General Residential 6	Residential
Cape Town	Residential 1 : Conventional Housing	Residential	Cape Town	Residential 2 : Incremental Housing	Residential
Cape Town	Mixed Use 1	Mixed Use	Cape Town	Mixed Use 2	Mixed Use
Cape Town	Mixed Use 3	Mixed Use	Cape Town	General Business 1	Commercial
Cape Town	General Business 2	Commercial	Cape Town	General Business 3	Commercial
Cape Town	General Business 4	Commercial	Cape Town	General Business 5	Commercial
Cape Town	General Business 6	Commercial	Cape Town	General Business 7	Commercial
Cape Town	Local Business 1 : Intermediate Business	Commercial	Cape Town	Local Business 2 : Local Business	Commercial
Cape Town	General Industrial 1	Industrial	Cape Town	General Industrial 2	Industrial
Cape Town	Risk Industry	Industrial	Cape Town	Transport 1 : Transport Use	Transportation
Cape Town	Transport 2 : Public Road and Public Parking	Transportation	Cape Town	Open Space 1 : Environmental Conservation	Green / Recreation

Continued on next page

Table 5: City-specific land-use source labels mapped to the common taxonomy, excluding Sydney (continued).

City	Source label	Mapped class	City	Source label	Mapped class
Cape Town	Open Space 2 : Public Open Space	Green / Recreation	Cape Town	Open Space 3: Special Open Space	Green / Recreation
Cape Town	Community 1 : Local	Institutional / Civic	Cape Town	Community 2 : Regional	Institutional / Civic
Cape Town	Utility	Utilities	Cape Town	Agricultural	Agriculture / Rural
Cape Town	Rural	Agriculture / Rural	Cape Town	Limited Use Zone	Other
Cape Town	None	Other			

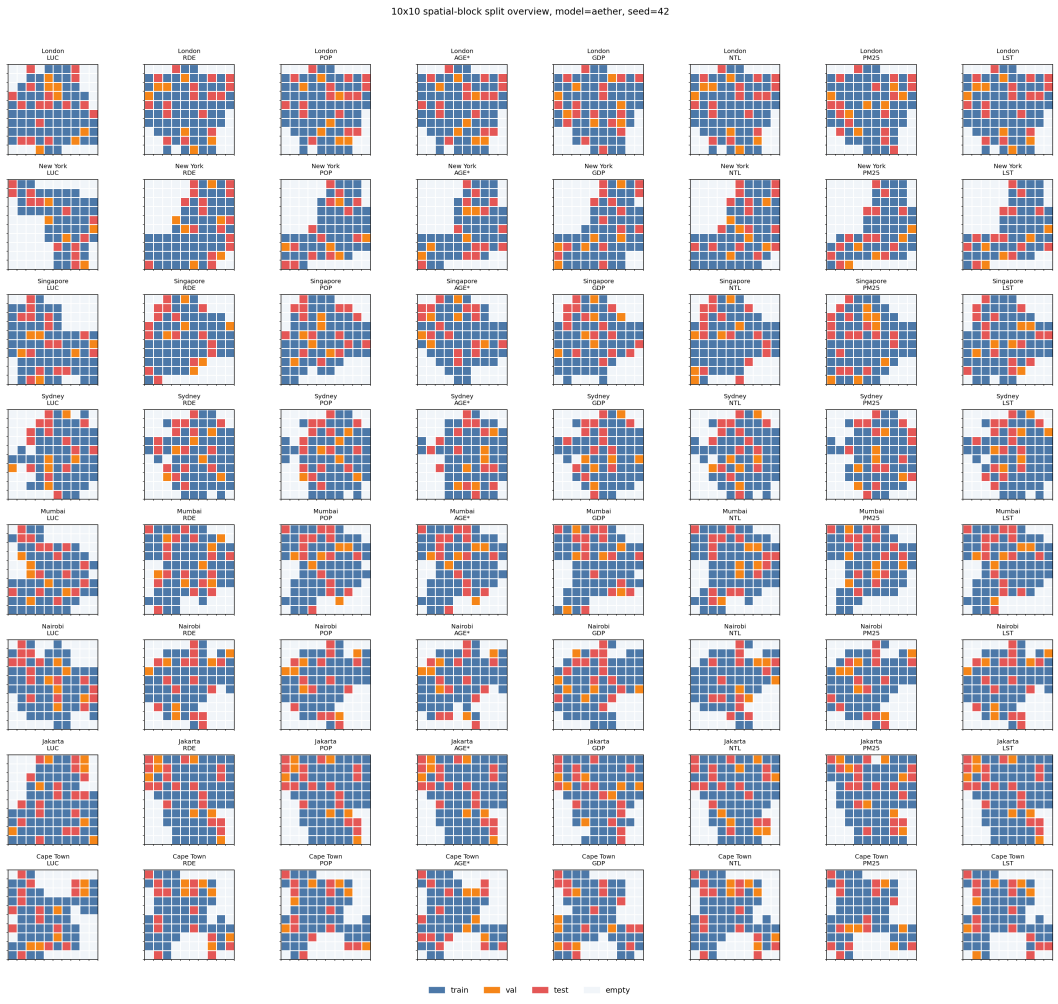


Figure 12: Example 10-by-10 spatial block split assignments for AETHER under seed 42. Blocks, rather than individual samples, are assigned to train, validation, and test splits.

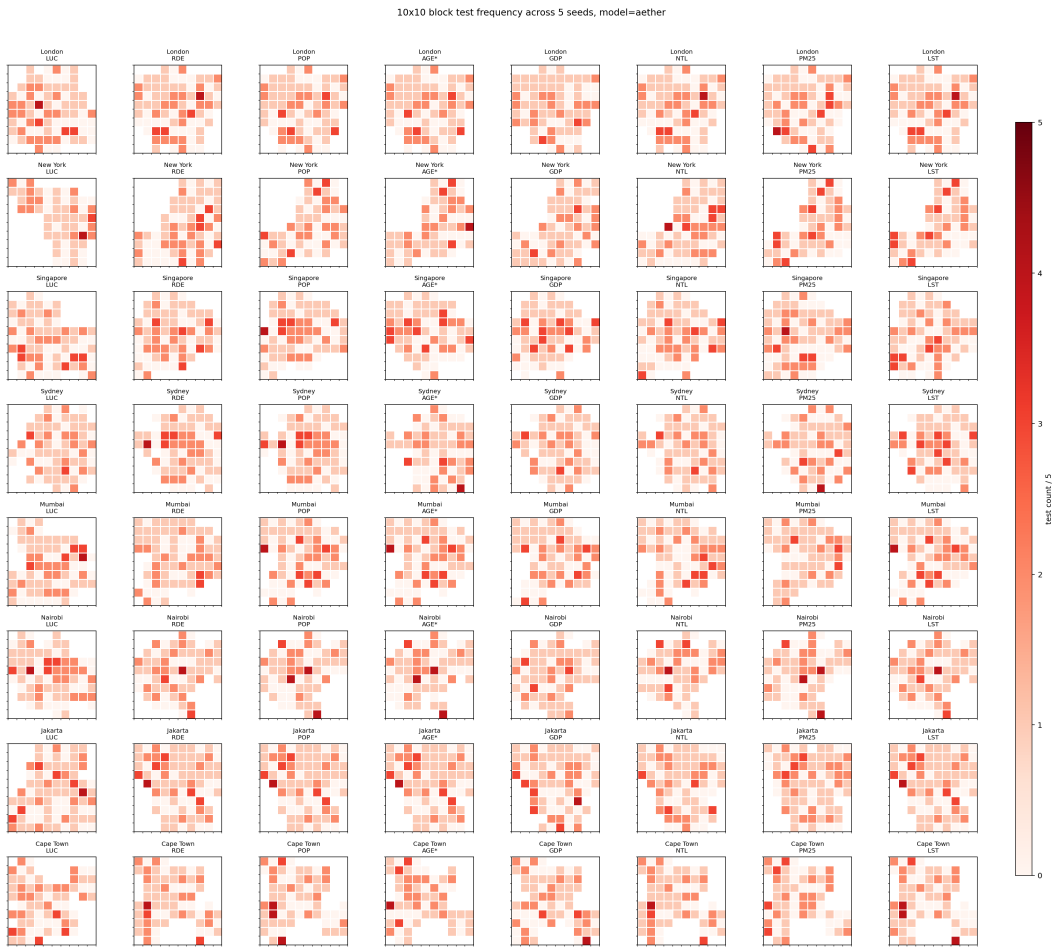


Figure 13: Test-block frequency across the five seed-specific spatial partitions. The figure shows how often each spatial block is assigned to the test partition.

C Additional Experimental Details

C.1 Spatial Split

The main benchmark uses spatial block splitting because urban labels are strongly spatially autocorrelated. If nearby samples from the same neighborhood are placed into both training and test sets, a downstream predictor can achieve high performance through local interpolation rather than through representation transfer. This is especially problematic for dense raster tasks such as population, road density, and land surface temperature, where adjacent cells often share similar values.

For each city–task dataset, we construct a 10×10 block grid over the valid task extent. Blocks are assigned, rather than individual task units, to train, validation, and test partitions. The test partition contains 20% of the blocks, and 10% of the remaining training blocks are used for validation. We repeat the procedure with five fixed random seeds, $\{42, 24, 7, 0, 100\}$, and all models evaluated on the same city–task pair use the same split assignments. Figure 12 illustrates the resulting block layout for the main protocol, and Figure 13 shows how often each block is used as test across seeds.

We also run a random split with the same ratios and seeds as a diagnostic comparison. Figures 14 and 15 show that random splitting often increases apparent performance and can alter rankings. This motivates our choice to use spatial splits as the main benchmark protocol and to treat random splits as a measure of spatial leakage sensitivity.

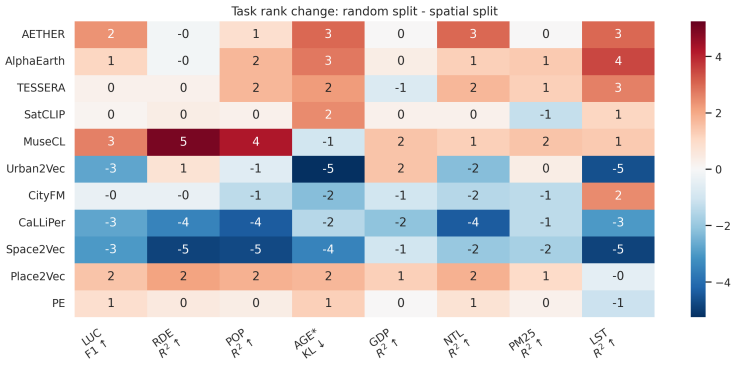


Figure 14: Task-level rank changes between random and spatial splits by model and task. Negative values indicate better rank under random splitting.

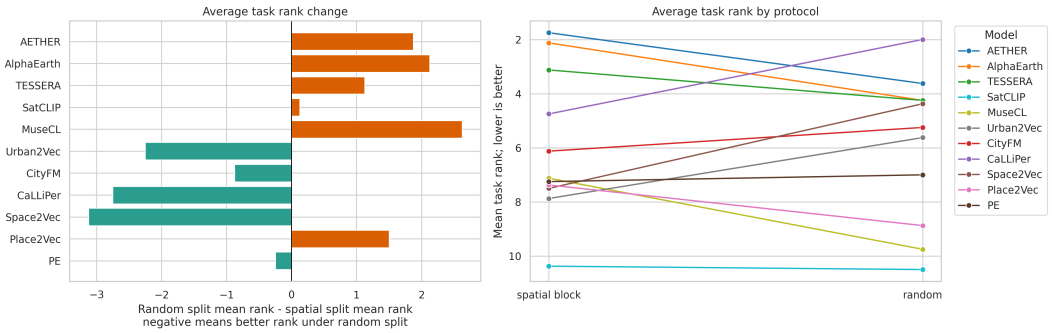


Figure 15: Overall mean-rank change between random and spatial splits across tasks. Negative values indicate better rank under random splitting.

C.2 Details of Baselines Reproduction

Table 6: Reproduction and interface details for the eleven CityRep baselines. The table distinguishes released feature products, official checkpoints, and CityRep-compatible reproductions. "Dim." is the feature dimension used by the downstream evaluator. "Support" is the spatial interface registered in CityRep before alignment to each task. All models are evaluated with the same downstream predictor, spatial split protocol, and city–task files.

Model	Reproduction status	CityRep input data	Dim. / support	Checkpoint, feature, or code source
Remote-sensing raster representations				
AlphaEarth	Released feature product	Google AlphaEarth / Satellite Embedding annual raster product, 2024.	64-d raster. Native raster sampled at land-use points; task-grid rasters use mean aggregation or coordinate sampling.	Official feature product from Google Earth Engine / Source Cooperative AlphaEarth Foundations COGs; CityRep wrapper code is used for cropping and task-grid export.
Implementation note. AlphaEarth is evaluated as an off-the-shelf released feature product. Results reflect the public remote-sensing embedding together with CityRep's raster alignment rule.				
TESSERA	Official package / feature export	TESSERA / GeoTessera 2024 remote-sensing embeddings from Sentinel-1/Sentinel-2 temporal-spectral inputs.	128-d raster. Reference-grid raster for point sampling; task-grid rasters for raster tasks.	Official GeoTessera code/package: `https://github.com/ucam-eo/geotessera`; official pretrained embedding export.
Implementation note. TESSERA is evaluated as frozen remote-sensing features exported with the official package/API. CityRep standardizes the task-grid aggregation, point sampling, and downstream evaluator.				

Continued on next page

Model	Reproduction status	CityRep input data	Dim. / support	Checkpoint, feature, or code source
AETHER	Pretrained AETHER reproduction	AlphaEarth 64-d raster embeddings plus Foursquare POI semantic text; POI text encoded with `Qwen/Qwen3-Embedding-8B` and projected into the shared space.	128-d raster. Native AETHER rasters and task-specific mean rasters registered for all 8 cities and 8 tasks.	AETHER implementation details: (Liu et al., 2025a); provenance of the pretrained AETHER checkpoint evaluated in CityRep is described with the benchmark artifacts.
Implementation note. AETHER is reproduced for the benchmark cities following the original training setting: AlphaEarth pixels and Foursquare POI descriptions are used with AE–AE and AE–POI contrastive objectives under a fixed 100-epoch schedule. No CityRep downstream labels, target rasters, splits, validation or test results, or benchmark metrics are used for representation training or checkpoint selection. The resulting frozen embeddings use the same alignment and downstream evaluation protocol as the other baselines.				
Map, POI, street-view, and region representations				
CityFM	Mixed official and reproduced	OSM-derived road, polygon, tag, and surface features prepared in CityFM format.	2560-d H3 region table, resolution 8. Task samples are mapped to H3 cells.	Official CityFM code: `https://github.com/PasqualeTurin/CityFM`; official weights used where available, with CityRep-compatible training/export for other cities.
Implementation note. We use official CityFM weights where available and produce CityFM-compatible embeddings for the remaining benchmark cities with the same registered input interface. The released export uses components that can be prepared consistently across all CityRep cities.				
Place2Vec	CityRep-compatible reproduction	Foursquare Open Places, second-level POI categories.	128-d H3 region table, resolution 8. POI category embeddings are pooled into H3 cells.	Reproduce code according to the paper; CityRep wrapper prepares Foursquare POIs and H3 exports.
Implementation note. CityRep uses Foursquare Open Places as the common POI source across cities. POI counts provide the available activity signal for constructing a reproducible Place2Vec-style region representation.				
Urban2Vec	CityRep-compatible reproduction	Mapillary street-view imagery plus Foursquare POI categories aggregated to H3 regions.	128-d H3 region table, resolution 8. Variant registered as street-view plus POI.	Urban2Vec open-source code: `https://github.com/wangzhecheng/urban2vec_`; CityRep wrapper adapts inputs to Mapillary, Foursquare, and H3 regions.
Implementation note. Urban2Vec is reproduced with the public street-view and POI inputs available in CityRep. The registered embeddings use the union of available modality support so that regions with one available modality can still be evaluated.				
MuseCL	CityRep-compatible variant	Mapillary street-view branch, AlphaEarth-based remote branch, and Foursquare POI text branch.	128-d H3 region table, resolution 8. Region embeddings are looked up by H3 cell.	Official MuseCL code: `https://github.com/XixianYong/MuseCL`; CityRep wrapper provides cross-city asset preparation and H3-region export.
Implementation note. MuseCL is implemented as a CityRep-compatible multimodal region encoder using modalities that can be registered consistently across the benchmark. The mobility-related activity signal is represented with available POI-type information.				
Coordinate and POI-supervised location encoders				
PE	Deterministic location baseline	Longitude–latitude coordinates at each CityRep task unit.	192-d entity/coordinate table. Fixed Sphere2Vec-`sphereC` position encoding with 64 frequencies; task samples are matched by `sample_id` after evaluating at point coordinates or raster-cell centers.	CityRep implementation following the Sphere2Vec-`sphereC` positional basis: `https://github.com/gengchenmai/sphere2vec`; no learned checkpoint.
Implementation note. PE is a non-learned location-only baseline. It is included to measure how much downstream performance can be explained by deterministic coordinate encoding alone under the same spatial split and downstream evaluator.				

Continued on next page

Model	Reproduction status	CityRep input data	Dim. / support	Checkpoint, feature, or code source
CaLLiPer	CityRep-compatible reproduction	Foursquare POI coordinate–text pairs; POI descriptions include names and categories.	128-d coordinate encoder. Land-use uses point-table inference; raster tasks query the encoder at task pixel centers.	Official CaLLiPer code: `https://github.com/xlwang233/CaLLiPer`; CityRep wrapper retrains/exports coordinate embeddings on benchmark POI inputs.
Implementation note. CaLLiPer is trained with the common Foursquare input corpus used by CityRep so that the same city coverage and task alignment protocol can be applied to all benchmark cities.				
Space2Vec	CityRep-compatible reproduction	Foursquare POI coordinates and second-level category labels.	128-d coordinate encoder. Land-use uses point-table inference; raster tasks query the encoder at task pixel centers.	Official Space2Vec code: `https://github.com/gengchenmai/space2vec`; CityRep wrapper prepares POI classification data and task-grid exports.
Implementation note. Space2Vec is trained under a uniform Foursquare-based POI classification setup for all cities. This provides a consistent coordinate-encoder baseline with the same task sampling and raster-center inference interface used by other coordinate models.				
SatCLIP	Official checkpoint	Longitude–latitude coordinates only at CityRep task locations.	256-d coordinate encoder. Land-use and raster tasks are queried directly at sample coordinates or raster-cell centers.	Official SatCLIP code: `https://github.com/microsoft/satclip`; official checkpoint `microsoft/SatCLIP-ViT16-L40`.
Implementation note. SatCLIP is evaluated as an off-the-shelf coordinate checkpoint. It uses only location coordinates at CityRep task units and does not use CityRep POI, map, street-view, or task-label data.				

Table 7: Average feature coverage of H3-first aggregation and direct task-unit aggregation for sparse entity-based representations. Coverage is the fraction of downstream task units with a valid aligned representation. Numbers in parentheses on H3-first rows report Δ = H3-first minus Task-unit direct. Green indicates higher coverage, and red indicates lower coverage. AGE uses the same four-city subset as the main benchmark; all other columns use all eight cities.

Model	Variant	LUC	RDE	POP	AGE	GDP	NTL	$PM_{2.5}$	LST
CityFM	H3-first	0.817 (+0.488)	0.775 (+0.479)	0.912 (+0.554)	0.987 (+0.602)	0.781 (-0.003)	0.776 (+0.134)	0.797 (-0.039)	0.814 (+0.505)
	Task-unit direct	0.329	0.296	0.358	0.385	0.784	0.642	0.836	0.309
Place2Vec	H3-first	0.840 (+0.482)	0.806 (+0.474)	0.927 (+0.513)	0.992 (+0.442)	0.810 (-0.003)	0.806 (+0.118)	0.827 (-0.029)	0.836 (+0.474)
	Task-unit direct	0.358	0.332	0.414	0.551	0.813	0.688	0.856	0.362

Table 8: Effect of H3-first aggregation compared with direct task-unit aggregation for sparse entity-based representations. Numbers in parentheses on H3-first rows report Δ = H3-first minus Task-unit direct. Green indicates improvement according to the metric direction, and red indicates degradation. AGE uses the same four-city subset as the main benchmark; all other columns use all eight cities.

Model	Variant	LUC F1↑	RDE R^2 ↑	POP R^2 ↑	AGE KL↓	GDP R^2 ↑	NTL R^2 ↑	$PM_{2.5}$ R^2 ↑	LST R^2 ↑
CityFM	H3-first	0.165 (+0.041)	0.187 (+0.057)	0.228 (+0.093)	0.041 (-0.004)	0.199 (+0.001)	0.369 (+0.011)	0.348 (-0.038)	0.496 (+0.385)
	Task-unit direct	0.125	0.130	0.135	0.045	0.198	0.358	0.386	0.111
Place2Vec	H3-first	0.168 (+0.034)	0.217 (+0.037)	0.288 (+0.135)	0.043 (-0.006)	0.191 (+0.006)	0.303 (+0.041)	0.063 (-0.022)	0.138 (+0.159)
	Task-unit direct	0.134	0.180	0.154	0.048	0.185	0.262	0.085	-0.022

C.3 Aggregation for Sparse Entity Representations

CityRep uses an H3-first aggregation strategy for sparse entity-based representations such as POI and map-entity embeddings. Instead of directly aggregating sparse entities to each downstream task unit, we first aggregate entities to H3 resolution-8 cells and then align the resulting cell embeddings to task units. This intermediate support reduces missing or unstable features when task units are small or when entities are unevenly distributed.

Table 8 compares this strategy with direct task-unit aggregation for CityFM and Place2Vec. Task averages follow the main benchmark aggregation: AGE uses London, New York, Singapore, and Sydney, while the other tasks use all eight cities. The H3-first strategy improves most task metrics for both models, supporting our choice to use an intermediate spatial support before downstream alignment.

Table 7 further compares the average feature coverage of H3-first aggregation and direct task-unit aggregation across downstream tasks. Coverage is defined as the fraction of task units that receive a valid non-missing representation after alignment. H3-first aggregation substantially improves coverage for both CityFM and Place2Vec, especially on LST and POP, explaining why the intermediate H3 support leads to more stable downstream evaluation.

C.4 Evaluation Metrics

CityRep evaluates downstream prediction performance using task-appropriate metrics for regression, classification, and distribution prediction. This appendix provides the formal definitions of all reported metrics.

Regression metrics. For regression tasks, including road density, population, GDP, NTL, $PM_{2.5}$, and land-surface temperature, we report coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE).

The coefficient of determination is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4}$$

where y_i is the ground-truth target, $\hat{y}_i$ is the prediction, and $\bar{y}$ is the mean target value.

Mean absolute error is defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{5}$$

Root mean squared error is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{6}$$

Among these metrics, R^2 is used as the primary metric because it measures explained variance and is less sensitive to target scale differences across tasks and cities. MAE and RMSE are computed on targets standardized using training-partition statistics and are therefore expressed in standardized units.

Classification metrics. For land-use classification, we report macro F1, macro precision, and macro recall. Let C denote the number of classes. For class c, precision and recall are defined as

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \tag{7}$$

where TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c.

The class-wise F1 score is

$$F1_c = \frac{2P_cR_c}{P_c + R_c}. \tag{8}$$

Macro precision, macro recall, and macro F1 are computed by averaging across classes:

$$\mathrm{MacroPrecision} = \frac{1}{C}\sum_{c=1}^{C} P_c, \tag{9}$$

$$\mathrm{MacroRecall} = \frac{1}{C}\sum_{c=1}^{C} R_c, \tag{10}$$

$$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C} F1_c. \tag{11}$$

Macro F1 is used as the primary metric because the land-use labels are imbalanced across cities and categories, and macro averaging gives equal weight to each class.

Distribution-prediction metrics. For age-distribution prediction, the target for task unit i is a probability distribution $\mathbf{p}_i = (p_{i1}, \ldots, p_{iK})$ over K age groups, and $\mathbf{q}_i = (q_{i1}, \ldots, q_{iK})$ is the predicted distribution. Before evaluation, probabilities are clipped below at ϵ and renormalized to sum to one.

We report the mean KL divergence:

$$D_{\mathrm{KL}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\log\frac{p_{ik}}{q_{ik}}, \tag{12}$$

where ϵ is a small numerical constant used for stability in the clipping step.

We also report the mean Chebyshev distance:

$$D_{\mathrm{Chebyshev}} = \frac{1}{n}\sum_{i=1}^{n}\max_k |p_{ik} - q_{ik}|, \tag{13}$$

and the mean L1 distance:

$$D_{L1} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} |p_{ik} - q_{ik}|. \tag{14}$$

KL divergence is used as the primary metric because it measures distributional mismatch while accounting for relative probability mass differences across age groups.

Metric direction. Higher values indicate better performance for R^2, macro precision, macro recall, and macro F1. Lower values indicate better performance for MAE, RMSE, KL divergence, Chebyshev distance, and L1 distance.

C.5 Training Protocol

All downstream predictors use an MLP with hidden dimension 1024 and the same optimization settings across models and tasks: Adam, learning rate 10^{-3}, weight decay 0, batch size 512, a maximum of 100 epochs, and validation-loss early stopping with patience 10. Each aligned input embedding is normalized row-wise to unit ℓ_2 norm; input features are not standardized feature-wise to z scores. Regression targets are standardized using statistics computed from the training partition. These choices keep the downstream model sufficiently expressive to use the embeddings, while limiting task-specific tuning.

All experiments were conducted using PyTorch on a workstation equipped with two NVIDIA RTX PRO 6000 Blackwell GPUs, each with 96GB memory, and an AMD Ryzen Threadripper PRO 9975WX CPU with 32 cores and 64 threads. Spatial preprocessing and alignment were performed on the CPU using standard geospatial raster and vector processing libraries, while downstream predictors were trained on GPUs. Most downstream evaluation runs used a single GPU and completed within several hours per city–task pair, depending on task resolution and embedding size.

Aggregation across seeds and cities. Let $s_{m,t,c,k}$ denote the primary test score for model m, task t, city c, and seed k. In our benchmark instantiation, we run five fixed random seeds, $\{42, 24, 7, 0, 100\}$; each seed also determines the spatial block assignments. We first average the primary metric over seeds to obtain a city-level score:

$$s_{m,t,c} = \frac{1}{5} \sum_{k=1}^{5} s_{m,t,c,k}. \tag{15}$$

For each seed, we also average the primary metric across the cities included for task t:

$$\bar{s}_{m,t,k} = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} s_{m,t,c,k}. \tag{16}$$

The reported average is equivalently the mean of the city-level scores or the mean of these seed-specific city averages:

$$\mathrm{Avg}_{m,t} = \frac{1}{5} \sum_{k=1}^{5} \bar{s}_{m,t,k} = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} s_{m,t,c}. \tag{17}$$

The variability reported with each main-table mean is the standard deviation of the five seed-specific city averages:

$$\text{S-Std.}_{m,t} = \sqrt{\frac{1}{5} \sum_{k=1}^{5} \left(\bar{s}_{m,t,k} - \mathrm{Avg}_{m,t} \right)^2}. \tag{18}$$

Here $\mathcal{C}_t$ denotes the set of cities included for task t. The average score measures overall task utility, while S-Std. measures the sensitivity of aggregate performance to the random seeds, including the spatial partitions they determine. Cross-city model-ordering consistency is analyzed separately using the city-level ranks defined below. Since different tasks use different primary metrics and metric scales, CityRep uses raw task metrics for the main result columns and uses rank-based summaries only as diagnostic views for comparing models across tasks.

Overall rank. The overall rank used in the appendix diagnostics is computed from city-level rankings rather than from normalized metric values. For each task–city pair (t, c), we rank all models according to the seed-averaged city-level score $s_{m,t,c}$, with rank 1 assigned to the best model. Higher values are better for R^2 and macro F1, while lower values are better for KL divergence in the age-distribution task. Ties receive the best shared rank. This gives a city-level rank $r_{m,t,c}$.

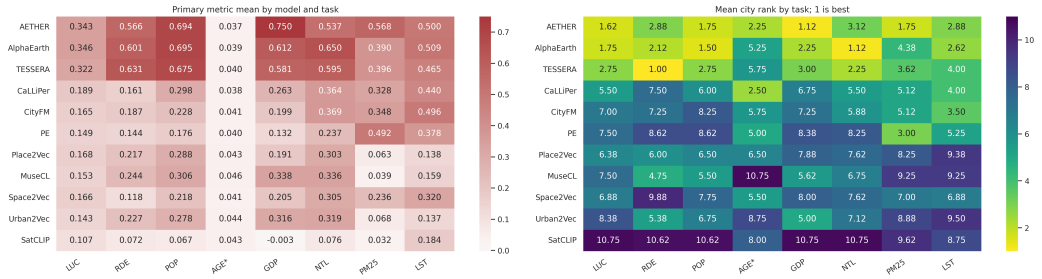


Figure 16: Spatial-split primary metric values and rank diagnostics for all evaluated models.

For the cross-city agreement diagnostic in Figure 2, we treat cities as raters of the $M = 11$ models and compute Kendall's coefficient of concordance for each task:

$$W_t = \frac{\chi^2_{F,t}}{|\mathcal{C}_t|(M-1)}, \tag{19}$$

where $\chi^2_{F,t}$ is the Friedman statistic computed from the city-specific model rankings $r_{m,t,c}$. Kendall's W ranges from 0 to 1, with larger values indicating stronger agreement in model ordering across cities and $W = 1$ indicating identical rankings. Figure 2 reports the seven tasks evaluated in all eight cities; AGE is omitted because its main aggregation uses only four cities.

We then average city-level ranks within each task:

$$R_{m,t} = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} r_{m,t,c}, \tag{20}$$

where $\mathcal{C}_t$ is the set of evaluated cities for task t. For all tasks except age distribution, $\mathcal{C}_t$ contains the eight benchmark cities. For age distribution, $\mathcal{C}_t$ contains only London, New York, Singapore, and Sydney because the remaining four cities are excluded from the main AGE aggregation due to source-data quality concerns.

Finally, the overall rank is the equal-task average:

$$R_m = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} R_{m,t}. \tag{21}$$

Lower values indicate better overall performance. This aggregation avoids comparing raw metric scales across tasks and prevents tasks with more evaluated cities or denser samples from dominating the overall ranking.

D Comprehensive Results

Figure 16 provides spatial-split diagnostics corresponding to the main benchmark results, and Figure 17 provides the same view under random splitting. These figures are diagnostic rather than replacements for the main table: raw primary metrics remain task-specific, while rank views are used to compare model ordering across heterogeneous metrics. We omit separate mean-city-rank figures here because that heatmap is already included in the right panel of each primary-metric diagnostic figure.

D.1 Results Across Tasks and Cities

This section provides the full task-level metric plots for both spatial and random splits. Figures 18 and 19 show the land-use results; Figures 20 and 21 show the road-density results; Figures 22 and 23 show the population results; Figures 24 and 25 show the age-distribution results; Figures 26 and 27 show the GDP results; Figures 28 and 29 show the NTL results; Figures 30 and 31 show the $PM_{2.5}$ results; and Figures 32 and 33 show the land-surface-temperature results. Unlike the main table, which reports one primary metric per task, these figures include secondary metrics as well: macro

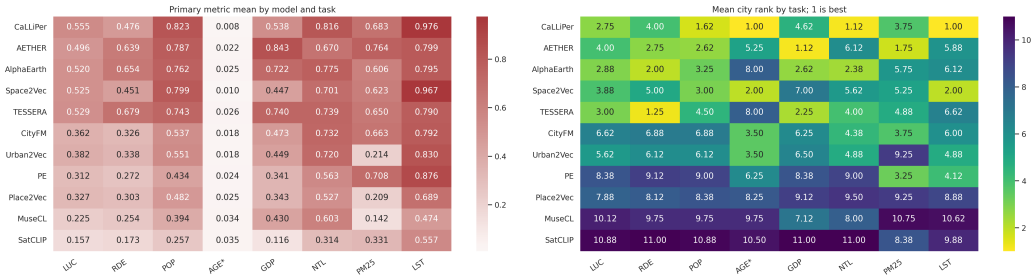


Figure 17: Random-split primary metric values and rank diagnostics for all evaluated models.

precision and macro recall for land use, MAE and RMSE for regression tasks, and L1 distance and Chebyshev distance for age-distribution prediction. Bars report the mean across five seeds, and error bars report the standard deviation across those seeds. The AGE plots retain all eight cities as descriptive diagnostics, whereas the main AGE aggregation and all aggregate result tables use only London, New York, Singapore, and Sydney. These views are useful for checking whether a model's ranking is driven by one metric alone or whether it is stable across multiple error summaries.

The paired layout also shows how the split protocol changes both apparent performance and uncertainty. Spatial splits generally yield lower apparent performance on spatially smooth targets, while random splits can make local interpolation easier. In addition, spatial splits often produce larger error bars than random splits, especially for coarse raster tasks such as GDP, NTL, and $PM_{2.5}$. This does not necessarily indicate only optimization instability. Under the spatial protocol, different seeds hold out different geographic blocks, so the standard deviation combines model stochasticity with variation in the difficulty and target variance of the held-out regions. This effect is particularly visible for R^2, whose denominator depends on the variance of the test targets; if a held-out block has a narrow target range, a moderate prediction error can lead to a very low or negative R^2. MAE and RMSE are therefore reported alongside R^2 to help distinguish unstable explained-variance estimates from changes in absolute prediction error. The difference is not uniform across tasks or models, which is why we report full task-level plots rather than only a single aggregate rank. Spatial and random figures are shown separately to keep labels and city-level bars readable.

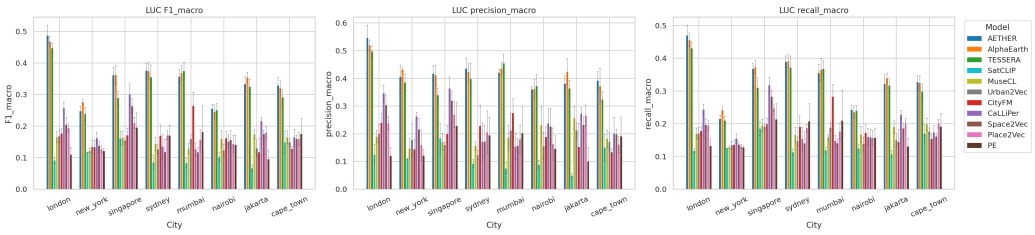


Figure 18: Land-use classification: full metric bars under spatial block splits.

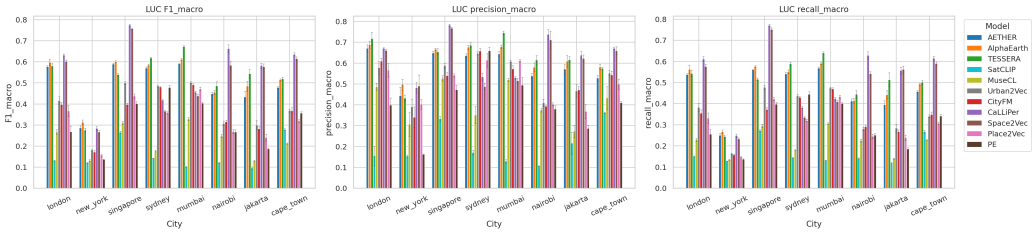


Figure 19: Land-use classification: full metric bars under random splits.

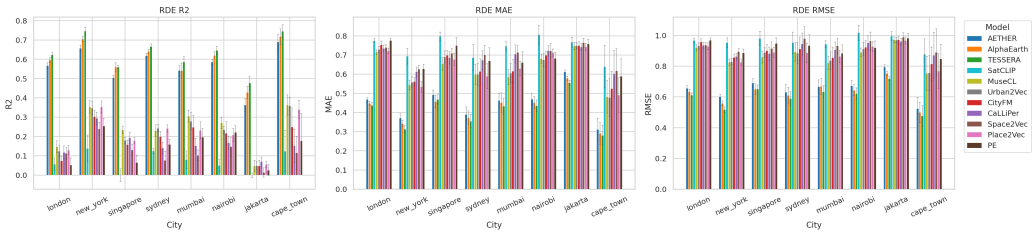


Figure 20: Road-density regression: full metric bars under spatial block splits.

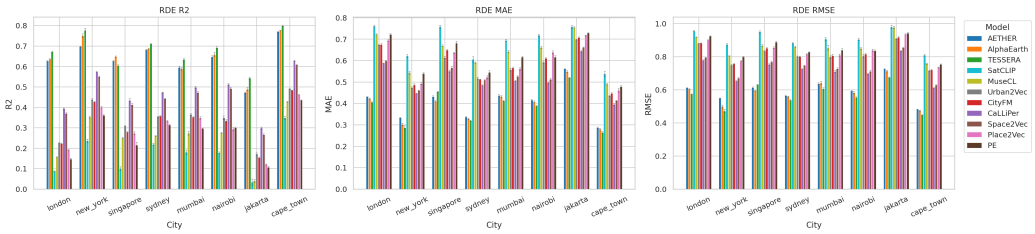


Figure 21: Road-density regression: full metric bars under random splits.

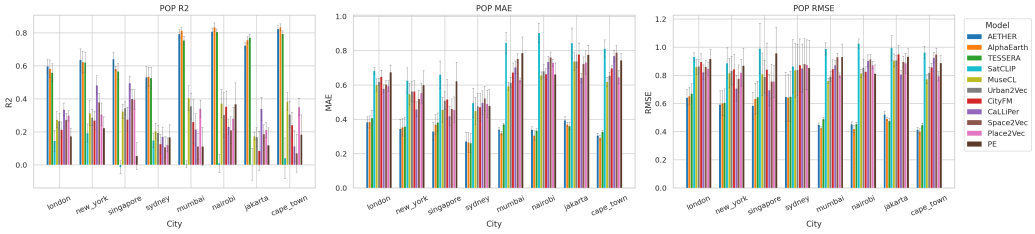


Figure 22: Population regression: full metric bars under spatial block splits.

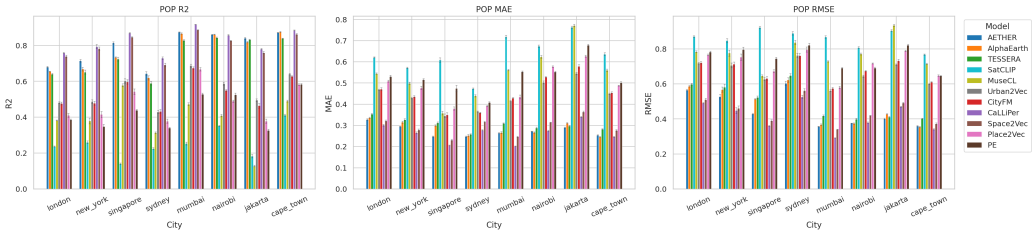


Figure 23: Population regression: full metric bars under random splits.

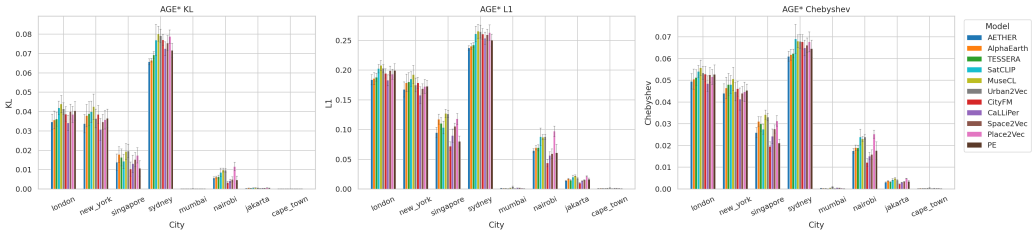


Figure 24: Age-distribution prediction across all eight cities: full metric bars under spatial block splits. The main AGE aggregation and aggregate tables use only London, New York, Singapore, and Sydney.

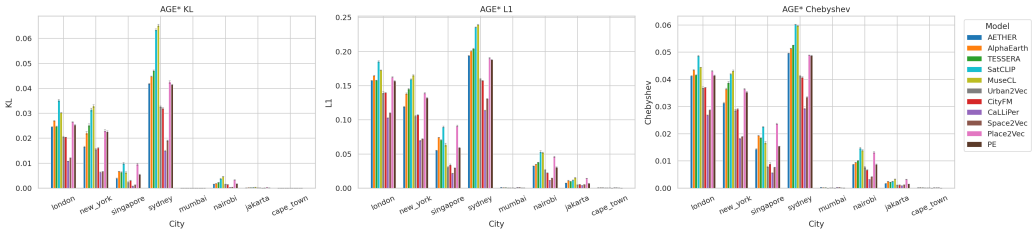


Figure 25: Age-distribution prediction across all eight cities: full metric bars under random splits. The main AGE aggregation and aggregate tables use only London, New York, Singapore, and Sydney.

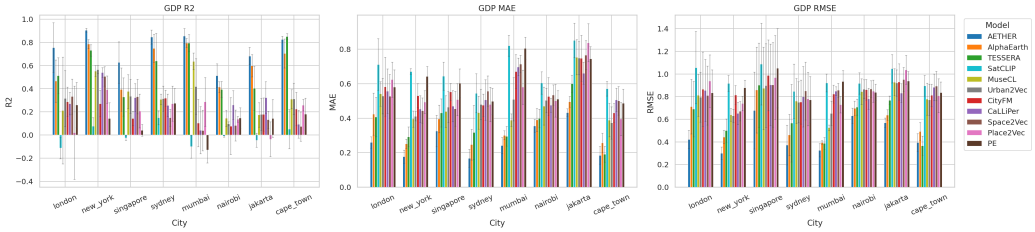


Figure 26: GDP regression: full metric bars under spatial block splits.

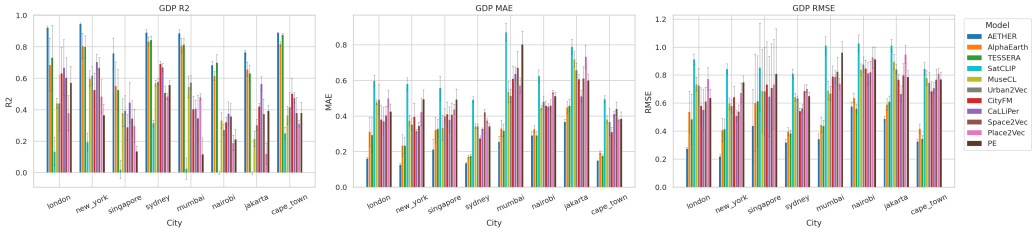


Figure 27: GDP regression: full metric bars under random splits.

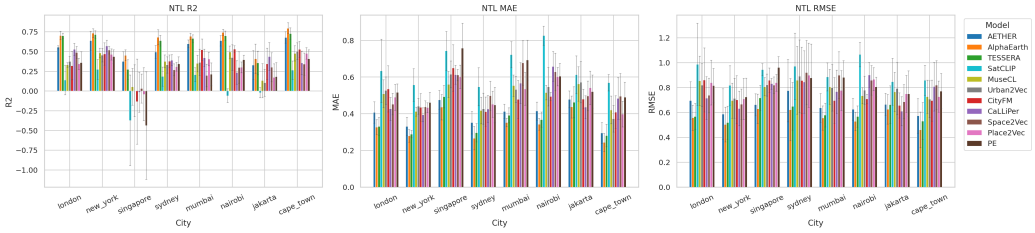


Figure 28: NTL regression: full metric bars under spatial block splits.

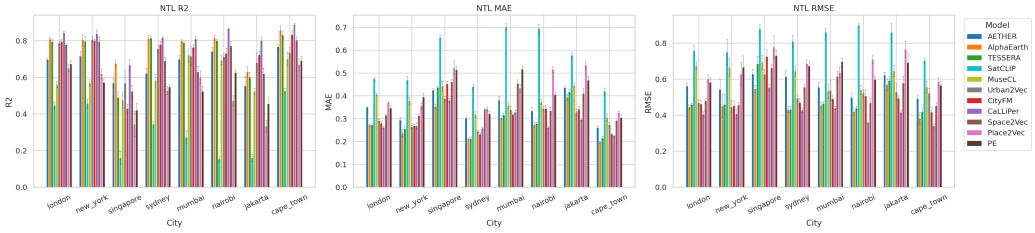


Figure 29: NTL regression: full metric bars under random splits.

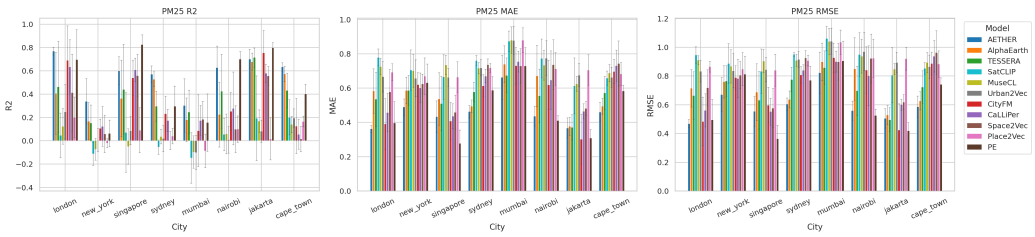


Figure 30: $PM_{2.5}$ regression: full metric bars under spatial block splits.

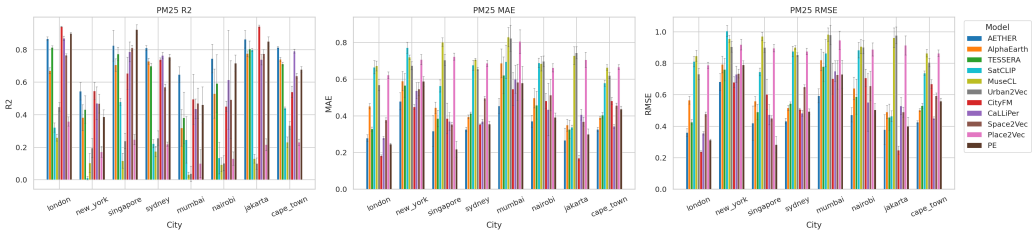


Figure 31: $PM_{2.5}$ regression: full metric bars under random splits.

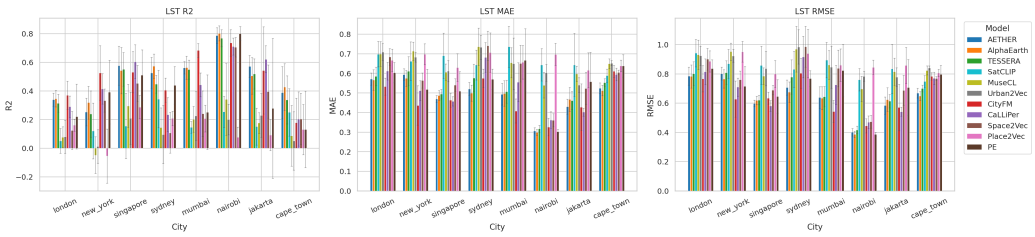


Figure 32: Land-surface-temperature regression: full metric bars under spatial block splits.

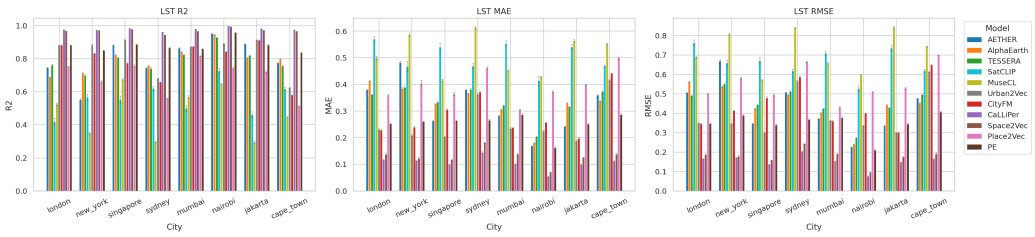


Figure 33: Land-surface-temperature regression: full metric bars under random splits.

Table 9: Raster-pooling robustness for the three affected models. Entries are mean task ranks after recomputing the 11-model leaderboard; lower is better.

Model	Mean pooling	Max pooling	Overall position
AETHER	1.750	1.875	$1 \rightarrow 1$
AlphaEarth	2.125	2.125	$2 \rightarrow 2$
TESSERA	3.125	4.000	$3 \rightarrow 3$

Table 10: Mean task rank under alternative downstream probes. Ranks use the complete 11-model comparison; lower is better.

Model	MLP-1024	MLP-512	MLP-256	Linear
AETHER	1.750	1.875	1.750	1.625
AlphaEarth	2.125	1.875	2.125	2.500
TESSERA	3.125	3.250	2.875	2.625
CaLLiPer	4.750	4.875	4.750	4.625
CityFM	6.125	5.750	6.250	7.875
MuseCL	7.125	7.625	7.500	6.000
PE	7.250	6.875	7.000	6.750
Place2Vec	7.375	6.875	6.750	7.500
Space2Vec	7.500	8.250	8.000	7.500
Urban2Vec	7.875	8.250	7.875	7.375
SatCLIP	10.375	10.250	10.375	11.000

E Robustness to Benchmark Design Choices

We test whether the main comparisons depend on raster pooling, downstream probe capacity, spatial block granularity, or the seed-specific spatial partition. These checks use the same benchmark tasks and model roster unless stated otherwise.

E.1 Raster Pooling

Mean pooling is the default when multiple raster embeddings fall within an areal task unit. To test this choice, we reran all affected rows for AETHER, AlphaEarth, and TESSERA with max pooling, holding task units, splits, downstream heads, and five seeds fixed, and recomputed rankings against all 11 models. Table 9 shows that the order of these three leading models is preserved, although their mean task ranks change modestly.

E.2 Downstream Probe Capacity

Probe-capacity comparison. In addition to the MLP-1024 main probe, we evaluated MLP-512, MLP-256, and a linear probe across all 11 models, eight cities, eight tasks, and five seeds under the same spatial block protocol. Table 10 reports mean task ranks (lower is better). Spearman correlations between the complete 11-model rank vectors obtained with MLP-1024 and the alternative probes are $\rho = 0.966$ for MLP-512, $\rho = 0.955$ for MLP-256, and $\rho = 0.834$ for the linear probe. AETHER, AlphaEarth, and TESSERA remain the leading three models under every probe, while the top-five model set is unchanged across the MLP widths. These results indicate that probe capacity affects the relative rankings to some extent, particularly when moving from an MLP to a linear probe, but the overall model ordering remains highly consistent across probe choices.

Linear-probe evaluation. To examine whether the main benchmark conclusions depend on the capacity of the downstream MLP evaluator, we additionally evaluate all representations with a linear probe under the same spatial block splits, city–task datasets, metrics, and aggregation protocol. Table 11 reports the resulting average performance and seed sensitivity. Overall, the task-level pattern is broadly consistent with the main benchmark: AETHER, AlphaEarth, and TESSERA remain the strongest models across tasks. The three raster or raster-enhanced representations remain the strongest overall, suggesting that their advantage is not solely due to the nonlinear downstream head. At the

Table 11: **Linear-probe CityRep results under spatial block splitting.** Entries report mean primary metrics across cities and seeds ± the standard deviation across the five seed-specific city averages. The seeds also determine spatial partitions. Type denotes the main data sources: *L* = location, *P* = POI, *R* = remote sensing, *S* = street-view imagery, and *O* = other urban/map data. **Bold**, underlined, and wavy-underlined values rank first, second, and third. *AGE reports KL divergence (lower is better) over the four cities with reliable age–sex coverage.

		Morphology		Demographics		Economy		Environment	
Model	Type	LUC (F1 ↑)	RDE (R^2 ↑)	POP (R^2 ↑)	AGE* (KL ↓)	GDP (R^2 ↑)	NTL (R^2 ↑)	$PM_{2.5}$ (R^2 ↑)	LST (R^2 ↑)
PE	*L*	$0.125_{\pm 0.018}$	$0.111_{\pm 0.045}$	$0.130_{\pm 0.074}$	$0.042_{\pm 0.004}$	$0.183_{\pm 0.132}$	$0.220_{\pm 0.170}$	$0.536_{\pm 0.103}$	$0.152_{\pm 0.242}$
Place2Vec	*P*	$0.137_{\pm 0.013}$	$0.159_{\pm 0.032}$	$0.217_{\pm 0.056}$	$0.046_{\pm 0.004}$	$0.200_{\pm 0.121}$	$0.266_{\pm 0.098}$	$0.084_{\pm 0.114}$	$0.077_{\pm 0.095}$
Space2Vec	*LP*	$0.128_{\pm 0.013}$	$0.107_{\pm 0.036}$	$0.180_{\pm 0.050}$	$0.048_{\pm 0.007}$	$0.202_{\pm 0.152}$	$0.282_{\pm 0.122}$	$0.271_{\pm 0.099}$	$0.135_{\pm 0.151}$
CaLLiPer	*LP*	$0.150_{\pm 0.013}$	$0.137_{\pm 0.034}$	$0.250_{\pm 0.056}$	$0.045_{\pm 0.005}$	$0.243_{\pm 0.149}$	$0.315_{\pm 0.094}$	$0.335_{\pm 0.107}$	$0.245_{\pm 0.132}$
CityFM	*PO*	$0.119_{\pm 0.018}$	$0.139_{\pm 0.037}$	$0.141_{\pm 0.073}$	$0.053_{\pm 0.009}$	$0.229_{\pm 0.178}$	$0.284_{\pm 0.140}$	$0.348_{\pm 0.126}$	$0.134_{\pm 0.197}$
Urban2Vec	*PS*	$0.119_{\pm 0.012}$	$0.192_{\pm 0.030}$	$0.232_{\pm 0.046}$	$0.045_{\pm 0.004}$	$0.268_{\pm 0.131}$	$0.226_{\pm 0.139}$	$0.061_{\pm 0.101}$	$0.126_{\pm 0.099}$
MuseCL	*PRS*	$0.131_{\pm 0.014}$	$0.206_{\pm 0.030}$	$0.267_{\pm 0.048}$	$0.046_{\pm 0.004}$	$0.321_{\pm 0.135}$	$0.347_{\pm 0.076}$	$0.054_{\pm 0.089}$	$0.118_{\pm 0.110}$
SatCLIP	*LR*	$0.092_{\pm 0.012}$	$0.054_{\pm 0.037}$	$0.022_{\pm 0.044}$	$0.048_{\pm 0.005}$	$-0.015_{\pm 0.097}$	$0.046_{\pm 0.149}$	$0.033_{\pm 0.104}$	$0.073_{\pm 0.132}$
TESSERA	*R*	$0.258_{\pm 0.019}$	$\mathbf{0.538}_{\pm 0.027}$	$0.592_{\pm 0.039}$	$0.043_{\pm 0.005}$	$0.596_{\pm 0.094}$	$0.595_{\pm 0.072}$	$0.445_{\pm 0.108}$	$0.373_{\pm 0.102}$
AlphaEarth	*R*	$0.254_{\pm 0.018}$	$0.514_{\pm 0.029}$	$0.615_{\pm 0.037}$	$0.042_{\pm 0.005}$	$0.598_{\pm 0.083}$	$\mathbf{0.632}_{\pm 0.074}$	$0.436_{\pm 0.095}$	$\mathbf{0.442}_{\pm 0.104}$
AETHER	*PR*	$\mathbf{0.259}_{\pm 0.017}$	$0.494_{\pm 0.024}$	$\mathbf{0.637}_{\pm 0.041}$	$\mathbf{0.040}_{\pm 0.004}$	$\mathbf{0.741}_{\pm 0.059}$	$0.488_{\pm 0.089}$	$\mathbf{0.562}_{\pm 0.099}$	$0.433_{\pm 0.101}$
Mean over models	–	$0.161_{\pm 0.015}$	$0.241_{\pm 0.033}$	$0.298_{\pm 0.051}$	$0.045_{\pm 0.005}$	$0.324_{\pm 0.121}$	$0.337_{\pm 0.111}$	$0.288_{\pm 0.104}$	$0.210_{\pm 0.133}$

task level, AETHER performs best on LUC, POP, AGE, GDP, and $PM_{2.5}$, AlphaEarth leads on NTL and LST, and TESSERA achieves the best result on RDE. As expected, the linear probe generally yields lower absolute scores than the MLP evaluator, especially on tasks that require more nonlinear decision boundaries or heterogeneous spatial signals. These results indicate that CityRep rankings are reasonably stable under a lower-capacity downstream evaluator, while also showing that some tasks benefit from nonlinear probing.

E.3 Results with a 20-by-20 Spatial Split

To examine whether the benchmark conclusions are sensitive to the spatial block granularity, we additionally evaluate all baselines using a finer 20×20 spatial split while keeping the same downstream models, task files, evaluation metrics, and five fixed random seeds. In Figure 34, the 20×20 split produces higher raw scores than the main 10×10 split for most regression and classification tasks, especially for LST, NTL, and $PM_{2.5}$. This is expected because smaller spatial blocks reduce the distance between training and test regions and therefore make the spatial generalization setting less strict.

The results are in Table 12. Importantly, the relative ordering of model families remains broadly consistent. Remote-sensing raster representations remain the strongest group overall, while coordinate and region models improve more noticeably on several tasks under the finer split. These results support the main benchmark choice of the 10×10 split as the stricter protocol, and show that the main findings are not an artifact of a single spatial partitioning resolution.

E.4 Seed Stability

The cross-city model-ranking agreement in Figure 2 measures consistency across cities, whereas the main table reports seed sensitivity. To assess sensitivity to the spatial partition, we reconstruct the complete 11-model leaderboard independently for each of the five fixed random seeds, which also determine the spatial partitions, using the benchmark's hierarchical task and city aggregation. Kendall's coefficient of concordance is $W = 0.948$ across the five seed-specific rankings; mean pairwise Spearman correlation is $\rho = 0.935$ across the ten seed pairs. The top-five ordering—AETHER, AlphaEarth, TESSERA, CaLLiPer, and CityFM—is identical under all five seeds. Table 13 reports the underlying mean city ranks and leaderboard positions. Mid-table models can exchange

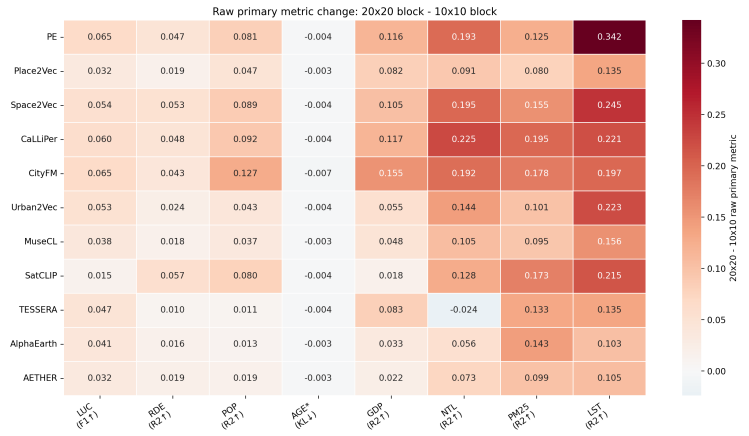


Figure 34: Raw primary metric change from the main 10×10 spatial split to the 20×20 spatial split. Values are computed as 20×20 block result minus 10×10 block result. Positive values indicate higher raw metric values; for AGE, the metric is KL divergence, so lower values are better.

Table 12: **CityRep results under the** 20×20 **spatial block split.** Entries report mean primary metrics across cities and seeds ± the standard deviation across the five seed-specific city averages. The seeds also determine spatial partitions. Type denotes the main data sources: *L* = location, *P* = POI, *R* = remote sensing, *S* = street-view imagery, and *O* = other urban/map data. **Bold**, underlined, and wavy-underlined values rank first, second, and third. *AGE reports KL divergence (lower is better) over the four cities with reliable age–sex coverage.

		Morphology		Demographics		Economy		Environment	
Model	Type	LUC (F1 ↑)	RDE (R^2 ↑)	POP (R^2 ↑)	AGE* (KL ↓)	GDP (R^2 ↑)	NTL (R^2 ↑)	$PM_{2.5}$ (R^2 ↑)	LST (R^2 ↑)
PE	*L*	$0.213_{\pm 0.031}$	$0.191_{\pm 0.034}$	$0.257_{\pm 0.071}$	$0.036_{\pm 0.004}$	$0.248_{\pm 0.087}$	$0.430_{\pm 0.096}$	$\underline{0.617}_{\pm 0.091}$	$\mathbf{0.720}_{\pm 0.054}$
Place2Vec	*P*	$0.200_{\pm 0.026}$	$0.236_{\pm 0.026}$	$0.336_{\pm 0.042}$	$0.040_{\pm 0.003}$	$0.273_{\pm 0.114}$	$0.394_{\pm 0.075}$	$0.143_{\pm 0.079}$	$0.273_{\pm 0.067}$
Space2Vec	*LP*	$0.220_{\pm 0.025}$	$0.171_{\pm 0.031}$	$0.306_{\pm 0.044}$	$0.037_{\pm 0.004}$	$0.309_{\pm 0.080}$	$0.500_{\pm 0.068}$	$0.391_{\pm 0.126}$	$0.565_{\pm 0.060}$
CaLLiPer	*LP*	$0.249_{\pm 0.027}$	$0.209_{\pm 0.031}$	$0.390_{\pm 0.047}$	$\mathbf{0.033}_{\pm 0.003}$	$0.379_{\pm 0.049}$	$\underline{0.589}_{\pm 0.053}$	$0.523_{\pm 0.125}$	$\underline{0.661}_{\pm 0.048}$
CityFM	*PO*	$0.230_{\pm 0.037}$	$0.230_{\pm 0.031}$	$0.355_{\pm 0.053}$	$\underline{0.035}_{\pm 0.004}$	$0.354_{\pm 0.086}$	$0.560_{\pm 0.072}$	$0.526_{\pm 0.093}$	$\underline{0.693}_{\pm 0.057}$
Urban2Vec	*PS*	$0.196_{\pm 0.026}$	$0.251_{\pm 0.024}$	$0.321_{\pm 0.056}$	$0.040_{\pm 0.004}$	$0.371_{\pm 0.079}$	$0.463_{\pm 0.069}$	$0.169_{\pm 0.067}$	$0.360_{\pm 0.073}$
MuseCL	*PRS*	$0.191_{\pm 0.021}$	$0.262_{\pm 0.025}$	$0.343_{\pm 0.040}$	$0.044_{\pm 0.005}$	$0.386_{\pm 0.095}$	$0.441_{\pm 0.059}$	$0.134_{\pm 0.067}$	$0.315_{\pm 0.089}$
SatCLIP	*LR*	$0.122_{\pm 0.014}$	$0.129_{\pm 0.027}$	$0.147_{\pm 0.063}$	$0.040_{\pm 0.005}$	$0.016_{\pm 0.072}$	$0.205_{\pm 0.082}$	$0.205_{\pm 0.114}$	$0.399_{\pm 0.113}$
TESSERA	*R*	$\underline{0.368}_{\pm 0.025}$	$\mathbf{0.640}_{\pm 0.016}$	$\underline{0.685}_{\pm 0.033}$	$0.037_{\pm 0.004}$	$\underline{0.664}_{\pm 0.085}$	$0.572_{\pm 0.053}$	$0.529_{\pm 0.123}$	$0.600_{\pm 0.047}$
AlphaEarth	*R*	$\mathbf{0.387}_{\pm 0.026}$	$\underline{0.617}_{\pm 0.016}$	$\underline{0.708}_{\pm 0.026}$	$0.036_{\pm 0.004}$	$\underline{0.645}_{\pm 0.079}$	$\mathbf{0.706}_{\pm 0.060}$	$\underline{0.533}_{\pm 0.112}$	$0.612_{\pm 0.049}$
AETHER	*PR*	$\underline{0.375}_{\pm 0.024}$	$\underline{0.585}_{\pm 0.018}$	$\mathbf{0.713}_{\pm 0.020}$	$\underline{0.034}_{\pm 0.004}$	$\mathbf{0.772}_{\pm 0.043}$	$\underline{0.610}_{\pm 0.057}$	$\mathbf{0.667}_{\pm 0.091}$	$0.605_{\pm 0.043}$
Mean over models	–	$0.250_{\pm 0.026}$	$0.320_{\pm 0.025}$	$0.415_{\pm 0.045}$	$0.037_{\pm 0.004}$	$0.402_{\pm 0.079}$	$0.497_{\pm 0.068}$	$0.403_{\pm 0.099}$	$0.527_{\pm 0.063}$

positions, so these results support the stability of the leading group without implying that every adjacent rank difference is meaningful.

F Exploratory Extension to 26 Cities

To examine whether CityRep can scale beyond the eight benchmark cities, we construct an exploratory 26-city extension using the same task registry, alignment pipeline, spatial split protocol, and evaluation code. The extension includes the eight main benchmark cities and 18 additional cities for which the global raster-based task layers can be prepared. In this exploratory setting, we evaluate two raster representation models, AlphaEarth and TESSERA, on the six tasks available for all 26 cities: road density, population, GDP, NTL, $PM_{2.5}$, and land surface temperature.

Table 13: Seed-specific overall mean city ranks and leaderboard positions. Lower rank is better.

Model	Seed 0	Seed 7	Seed 24	Seed 42	Seed 100	Position
AETHER	2.516	2.344	2.594	2.172	2.344	1/1/1/1/1
AlphaEarth	2.625	2.891	2.781	2.813	2.781	2/2/2/2/2
TESSERA	3.438	3.141	3.172	3.219	3.141	3/3/3/3/3
CaLLiPer	5.531	5.469	5.453	5.484	5.422	4/4/4/4/4
CityFM	5.875	6.570	6.188	6.531	5.898	5/5/5/5/5
PE	7.188	7.055	6.734	6.828	6.523	7.5/6/6/6/6
MuseCL	7.125	7.125	7.719	7.125	7.523	6/8/10/7/8
Place2Vec	7.188	7.234	7.250	7.609	7.617	7.5/9/8/10/9
Space2Vec	7.594	7.469	7.188	7.156	7.313	10/10/7/8/7
Urban2Vec	7.359	7.109	7.563	7.453	8.008	9/7/9/9/10
SatCLIP	9.563	9.594	9.359	9.609	9.430	11/11/11/11/11

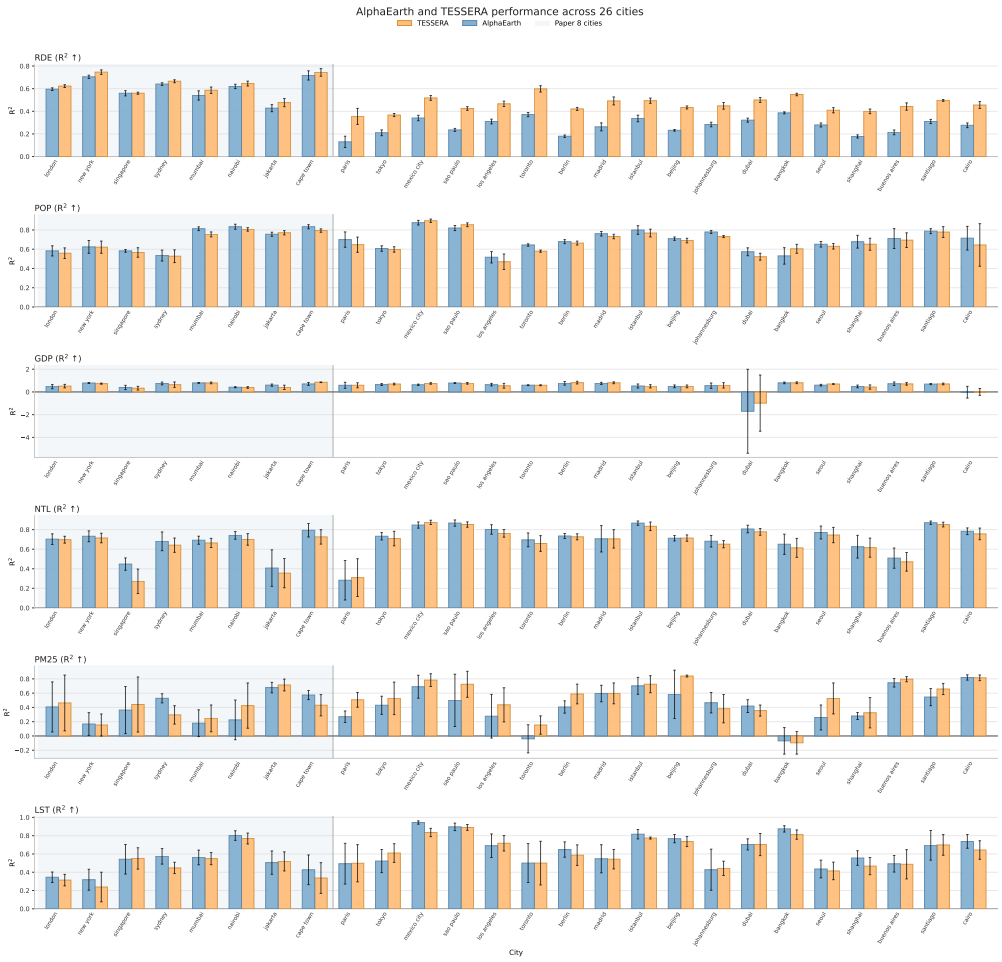


Figure 35: Exploratory 26-city extension using AlphaEarth and TESSERA. Bars report the mean primary metric across five fixed random seeds, which also determine the spatial partitions, for each city–task pair; error bars report the standard deviation across seeds. The shaded region marks the eight cities used in the main benchmark setting. The extension uses the six globally available raster-based tasks: RDE, POP, GDP, NTL, $PM_{2.5}$, and LST. Higher is better for all reported metrics.

Figure 35 compares AlphaEarth and TESSERA across all available city–task pairs. Overall, the two models exhibit broadly consistent cross-city trends: cities that are difficult for one model often also

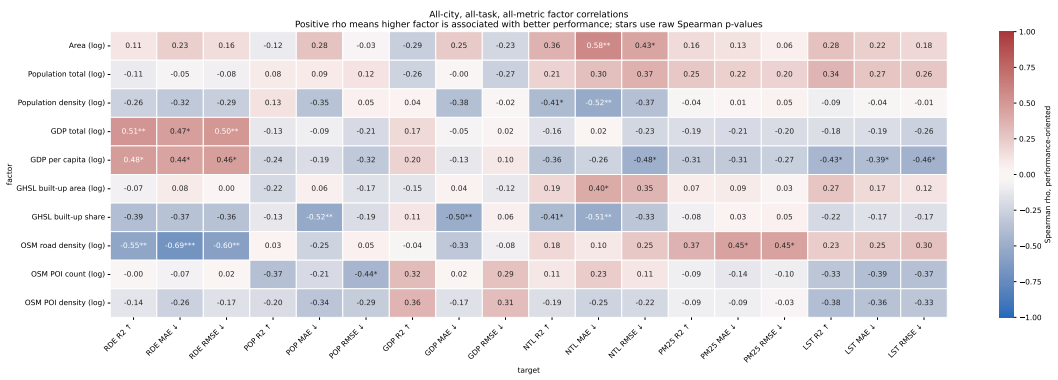


Figure 36: Exploratory correlation analysis between city-level factors and downstream performance across the 26-city remote-sensing extension. Each cell reports the Spearman correlation between a city factor and the two-model average performance of AlphaEarth and TESSERA for a task-metric pair. Metrics are oriented so that higher values always indicate better performance; MAE and RMSE are sign-flipped before averaging. Positive correlations therefore mean that larger factor values are associated with better performance. Stars denote raw, unadjusted Spearman p-values. The analysis is descriptive and intended to identify sources of performance heterogeneity rather than establish causal effects.

tend to be difficult for the other, suggesting that part of the performance variation reflects city- and task-specific data characteristics rather than purely model-specific behavior. At the same time, the two models show different strengths across individual tasks. TESSERA performs strongly on several morphology-related and environmental tasks, whereas AlphaEarth is competitive or stronger on some demographic and thermal tasks. This highlights the necessity of multi-task evaluation: results from any single task are insufficient for comprehensively understanding model behavior.

Although the multi-city evaluation reveals overall patterns of consistency, observations from an individual city can sometimes contradict the aggregate trend. For example, on the population task, TESSERA performs better in Mexico City and São Paulo, even though AlphaEarth outperforms TESSERA in most cities overall. We therefore treat this extension as evidence for the value of systematic multi-city evaluation. The 26-city extension is not used as the main leaderboard because not all baselines currently provide complete embeddings or reproducible checkpoints for the expanded city set, and tasks such as land use and age distribution require additional city-specific auditing. Nevertheless, the experiment demonstrates that CityRep's registry-based design can support broader evaluation once new city–task files and model embeddings are registered.

G Analysis of Performance Heterogeneity Across Tasks and Cities

To better characterize this heterogeneity, we conduct an exploratory city-factor analysis in the 26-city remote-sensing extension. For each city, we construct a common set of interpretable descriptors: total area, population, population density, total GDP, GDP per capita, GHSL built-up area, GHSL built-up share, OSM road density, OSM POI count, and OSM POI density. We then combine AlphaEarth and TESSERA results to obtain a two-model average performance signal for each city-task-metric pair. Metrics are first converted to a common performance orientation, so that larger values always indicate better prediction: R^2 is kept unchanged, while MAE and RMSE are sign-flipped. Because both models are evaluated on the same task and metric, we average the raw performance-oriented values directly rather than normalizing across cities. Finally, we compute Spearman correlations between each city factor and each task-metric performance signal.

Figure 36 reports these correlations. The heatmap shows that city difficulty across all tasks is not explained by a single factor such as city size, population, or economic development. Some associations are task-specific and occasionally counterintuitive. GDP-related factors are positively associated with road-density prediction performance, but they do not uniformly improve performance on socioeco-

nomic or environmental targets. Population density and built-up share are often negatively associated with performance on several metrics, suggesting that compact and highly built-up environments can be more difficult for current representations. OSM road density is strongly negatively associated with road-density prediction performance, indicating that dense street systems remain challenging even when the target itself is derived from road structure. POI-based factors show weaker and less consistent correlations, which may reflect differences in OSM completeness across cities as well as the fact that POI distributions only partially capture the spatial signals needed by the downstream tasks.

A useful interpretation is that development-related factors involve two competing effects. On the one hand, more developed cities often have more complex urban structure, stronger functional specialization, denser infrastructure, and more heterogeneous economic activity. These properties can make tasks related to urban function and socioeconomic characteristics harder, because the mapping from observable geospatial patterns to targets such as GDP, nighttime lights, or other functional indicators may be less direct. On the other hand, more developed cities also tend to have richer and higher-quality supporting data, including better coverage of roads, POIs, and other spatial layers, which can improve both model fitting and evaluation reliability.

The opposite tension may hold for cities in less developed contexts. Their urban form may in some cases be simpler or more spatially regular, which could make certain prediction problems easier in principle. However, this potential advantage can be offset by weaker data coverage, lower source-data quality, and greater measurement uncertainty. We therefore do not interpret the correlations as showing that either developed or developing regions are uniformly easier. Instead, the observed performance reflects a trade-off between urban complexity and data availability, together with task-specific differences in what each target requires from the representation.

This analysis supports the design choice of CityRep. A single aggregate leaderboard score can hide large differences across cities, tasks, and evaluation settings. Robust evaluation of urban foundation models should therefore report not only average performance, but also spatial-split results, random-split results, city-level variation, and coarse geographic stratifications.