SCALECITY: JOINT MULTI-SCALE AND MULTI-TASK URBAN REGION REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Urban region representation learning must accommodate tasks defined over different granularities and regional boundaries. Changes in these spatial configurations alter feature aggregation and target statistics, reflecting the scale and zoning effects of the Modifiable Areal Unit Problem (MAUP). We present ScaleCity, a multimodal urban region representation learning framework that jointly trains one model per city on mixed task–configuration samples. Joint training replaces separately fitted predictors with one model that supports multiple tasks and configurations. Graph-enhanced contrastive pretraining aligns street-view imagery, points-of-interest text, and remote-sensing features on atomic H3 cells. A prompt-conditioned hypernetwork adapts multimodal fusion and prediction to each task–configuration pair, while regional aggregation retains explicit extent information. Across three Chinese megacities, eight tasks, and nine configurations spanning grids, road-network partitions, and administrative regions, ScaleCity outperforms nine baselines in all task- and configuration-averaged R^2 comparisons. Further experiments demonstrate zero-shot prediction at unseen configurations within the same city, while few-shot experiments examine adaptation to unseen tasks. These findings support joint multi-scale, multi-task learning of urban region representations. Code is available at `https://github.com/far-rele/ScaleCity`.

1 INTRODUCTION

Urban region representation learning supports socioeconomic estimation and urban activity analysis (Huang et al., 2023; Balsebre et al., 2024). Multimodal approaches combine complementary sources such as points of interest (POI), street-view imagery (SVI), and satellite observations (Zou et al., 2025; Hao et al., 2025; Cao et al., 2026). Applications also require predictions over different spatial units, including grids, road-network partitions, and administrative regions (Li et al., 2023; Chen et al., 2024; Sun et al., 2025). Changes in granularity and boundaries alter feature aggregation and target statistics, reflecting the scale and zoning effects of the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984). We use *spatial configuration* to denote the combination of spatial granularity and partition family.

Recent methods advance task adaptation through urban prompts (Zhou et al., 2023; Jin et al., 2025) and task-aware multimodal modeling (Liu et al., 2026; Li et al., 2026a). Spatially flexible methods address how representations can be constructed for different regional boundaries. FlexiReg, for example, reuses fine-grained cell embeddings across partitions, while its reported downstream experiments fit task-specific predictors separately on each region set (Sun et al., 2025). ScaleCity instead jointly trains one predictive model with supervision from multiple tasks and spatial configurations. This setting requires a shared model to accommodate both changes in the prediction target and changes in the regions over which features and labels are aggregated. For example, regions of different sizes can have similar average characteristics but different population totals, making regional extent relevant alongside pooled content.

We introduce **ScaleCity**, a multimodal urban region representation learning framework that jointly trains one city-specific model on mixed task–configuration samples. Pretrained atomic-cell representations provide a reusable basis for regions with different granularities and boundaries. A Scale–Task Conditioned Hypernetwork (ST-HyperNet) uses task and configuration descriptions to adapt

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.

multimodal fusion and prediction within the shared downstream model. Regional aggregation retains explicit extent information alongside pooled content, allowing the model to account for targets that depend on regional size.

We evaluate ScaleCity across three Chinese megacities, eight prediction tasks, nine spatial configurations, and nine baselines. In the primary benchmark, ScaleCity outperforms all evaluated baselines in task- and configuration-averaged R^2. Further experiments assess housing-price (HP) prediction at unseen configurations without parameter updates and few-shot adaptation to NO_2 and PM_{10} using labeled data from a 10% spatial support split for parameter updates. These studies examine joint prediction, unseen-configuration prediction, and few-shot task adaptation within the same framework.

Our contributions are threefold: (i) a unified framework that jointly trains one city-specific model on mixed task–configuration samples, supporting multiple tasks, granularities, and partition families; (ii) a prompt-conditioned hypernetwork for task- and configuration-adaptive multimodal fusion and prediction, paired with extent-aware aggregation to account for regional size differences; and (iii) an evaluation demonstrating improvements over nine baselines in task- and configuration-averaged R^2 on the primary benchmark, with ablations and additional studies assessing component contributions, unseen-configuration prediction, and few-shot task adaptation.

2 Related Work

2.1 Urban Representation and Task Adaptation

Urban representation learning captures the semantic and spatial structure of cities (Tempelmeier et al., 2021; Li et al., 2022; Huang et al., 2023; Wu et al., 2022). Multimodal approaches integrate complementary observations through alignment and fusion (Wang et al., 2020; Zhang et al., 2023; Li et al., 2024; Sun et al., 2024; Li et al., 2026b); UrbanMMCL combines cross-modal alignment with adaptive multi-view graph contrastive learning (Cao et al., 2026), while image–text learning and urban pretraining further support reusable representations (Yan et al., 2024; Hao et al., 2025; Xiao et al., 2024a; Balsebre et al., 2024; Zhang et al., 2026). Task adaptation tailors these representations to different prediction objectives. HREP and GURPP introduce prompt-based adaptation (Zhou et al., 2023; Jin et al., 2025), while UrbanMoE and UrbanExpert make multimodal modeling task-dependent (Liu et al., 2026; Li et al., 2026a). Alongside task requirements, urban region representations must accommodate variation in the granularity and boundaries of the target regions.

2.2 Spatial Flexibility and Multi-Scale Learning

Changing regional granularity or boundaries changes the support over which observations and targets are aggregated, as reflected in MAUP (Openshaw, 1984). RegionDCL constructs region representations across partitions from building patterns (Li et al., 2023). FlexiReg reuses fine-grained representations across partitions, while its reported evaluation fits task-specific downstream models separately on each region set (Sun et al., 2025). ScaleCity jointly learns across tasks and spatial configurations, allowing one city-specific model to support multiple granularities and partition families.

2.3 Conditional Parameter Generation

Prompt learning adapts pretrained models through learned input or intermediate representations (Lester et al., 2021; Li & Liang, 2021; Jia et al., 2022; Zhou et al., 2022; Khattak et al., 2023), whereas hypernetworks provide a complementary mechanism by generating model parameters from a condition (Ha et al., 2017). HyperFormer and HyperPrompt generate task-specific adapters and attention prompts, respectively (Karimi Mahabadi et al., 2021; He et al., 2022); Polyhistor generates low-rank adapter factors for multi-task dense vision (Liu et al., 2022). ScaleCity brings conditional parameter generation to jointly trained urban region representations, adapting their fusion and readout to both task requirements and spatial configurations.

3 METHODOLOGY

ScaleCity learns urban region representations with task- and configuration-dependent fusion and prediction. Figure 1 shows four stages organized into two training phases: atomic-cell pretraining (Stage 1), followed by joint downstream optimization (Stages 2–4). In the second phase, one shared downstream model learns from mixed task–configuration samples while the pretrained cell embeddings remain frozen.

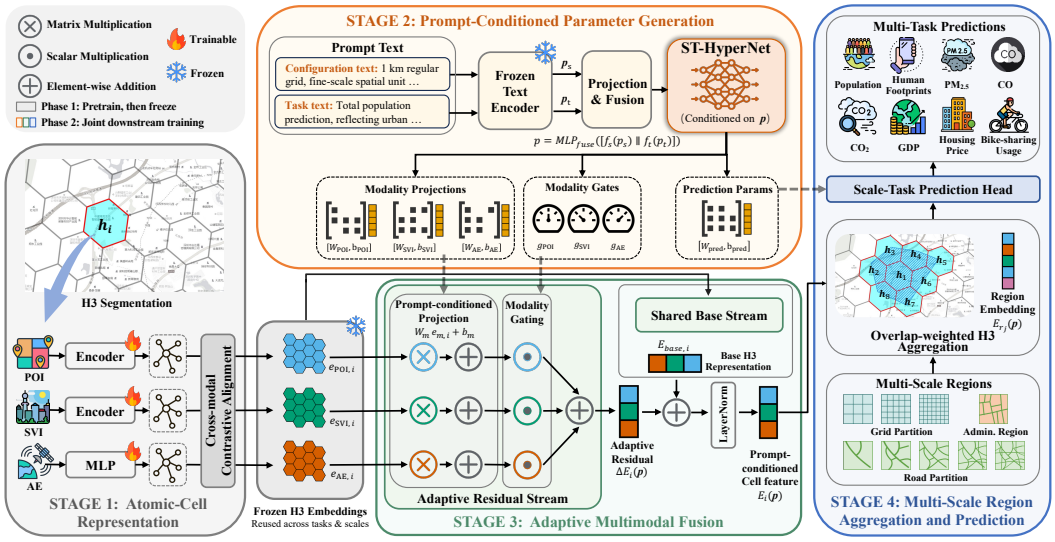

Figure 1: Overview of ScaleCity. Task–configuration prompts generate parameters for multimodal fusion and prediction. The fused cell features are aggregated over target regions and passed to the scale–task prediction head. Stages 2–4 jointly learn from mixed task–configuration samples using frozen Stage 1 embeddings.

3.1 ATOMIC-CELL REPRESENTATION

We represent a city by H3 cells $\mathcal{H} = \{h_i\}_{i=1}^{N}$. For each cell, CLIP (Radford et al., 2021) encodes street-view imagery (SVI), and MacBERT (Cui et al., 2020) encodes points-of-interest (POI) text, with valid observation embeddings averaged within each cell. A trainable projection maps the precomputed AlphaEarth (AE) features (Brown et al., 2025) into the graph feature space. Low-rank adapters (Hu et al., 2022) tune the image and text encoders during pretraining.

Each modality has a graph attention network (Veličković et al., 2018) whose edges connect one-hop and two-hop H3 neighbors, incorporating spatial context into the cell features. Let $z_{m,i}^{(L)}$ denote the feature of cell h_i after L graph layers for modality $m \in \{\mathrm{SVI}, \mathrm{POI}, \mathrm{AE}\}$. A modality-specific two-layer MLP P_m maps this feature into a shared d-dimensional space:

$$e_{m,i} = \mathrm{Norm}\Big(P_m(z_{m,i}^{(L)})\Big), \tag{1}$$

where Norm denotes ℓ_2 normalization. Pretraining aligns the three modality pairs through

$$\mathcal{L}_{\mathrm{pre}} = \sum_{m<n} \lambda_{m,n}\mathcal{L}_{m,n}, \tag{2}$$

where the sum runs over the three unordered modality pairs, $\mathcal{L}_{m,n}$ is a symmetric cross-modal contrastive loss (Radford et al., 2021), and $\lambda_{m,n}$ is its weight. The same cell's modality views form positives; other cells in the synchronized batch supply negatives. After pretraining, the atomic-cell encoder is frozen and its embeddings are reused in Stages 2–4. Encoder details and the full objective are in Appendix B.

3.2 Prompt-Conditioned Parameter Generation

We condition downstream computation on both the spatial configuration and the prediction task. A shared frozen text encoder maps the spatial-configuration and task descriptions separately to p_s and p_t. Separate projection heads f_s and f_t transform these embeddings, and a fusion MLP produces the joint condition

$$p = \mathrm{MLP}_{\mathrm{fuse}}([f_s(p_s) \,\|\, f_t(p_t)]) \,, \tag{3}$$

where $\|$ denotes concatenation. All regions in the same task–configuration pair share p.

The Scale–Task Conditioned Hypernetwork (ST-HyperNet) maps p to modality projections, fusion gates, and prediction-head parameters. Following the use of factorized parameter generation in multi-task adaptation (Liu et al., 2022), it generates two factors per modality whose product defines the adaptive projection:

$$\mathbf{W}_m(p) = \mathbf{A}_m(p)\mathbf{B}_m(p), \quad \mathbf{A}_m(p) \in \mathbb{R}^{d_{\mathrm{out}} \times r}, \quad \mathbf{B}_m(p) \in \mathbb{R}^{r \times d}, \tag{4}$$

where d_{out} is the fusion dimension and $r \ll \min(d, d_{\mathrm{out}})$ bounds the projection rank. This reduces the generated projection coefficients from $d_{\mathrm{out}}d$ to $r(d_{\mathrm{out}}+d)$. ST-HyperNet also generates modality biases $\mathbf{b}_m(p)$, softmax gates $\mathbf{g}(p)$ with $\sum_m g_m(p) = 1$, and prediction-head parameters $\mathbf{W}_{\mathrm{pred}}(p)$ and $\mathbf{b}_{\mathrm{pred}}(p)$. The joint condition thus controls both modality fusion and the final readout.

3.3 Adaptive Multimodal Fusion

We combine a shared base stream with a prompt-conditioned residual stream. The base stream maps the atomic embeddings $e_{m,i}$ through modality-specific projections $\phi_{\mathrm{base},m}$, concatenates their outputs, and applies an affine fusion map $(\mathbf{W}_{\mathrm{fuse}}, \mathbf{b}_{\mathrm{fuse}})$ to obtain $E_{\mathrm{base},i} \in \mathbb{R}^{d_{\mathrm{out}}}$ (Eq. 18). These parameters are trained downstream and shared across tasks and configurations.

The residual stream uses the generated projections to transform each modality and the gates to weight its contribution. The two streams are then added and normalized:

$$\Delta E_i(p) = \sum_m g_m(p) \left(\mathbf{W}_m(p) e_{m,i} + \mathbf{b}_m(p)\right), \tag{5}$$

$$E_i(p) = \mathrm{LayerNorm}(E_{\mathrm{base},i} + \Delta E_i(p)) \,. \tag{6}$$

The shared stream supplies a common representation, while the residual stream adjusts modality transformations and weights for each task–configuration pair. Their combination yields the conditional cell feature $E_i(p)$ used for region aggregation.

3.4 Multi-Scale Region Aggregation and Prediction

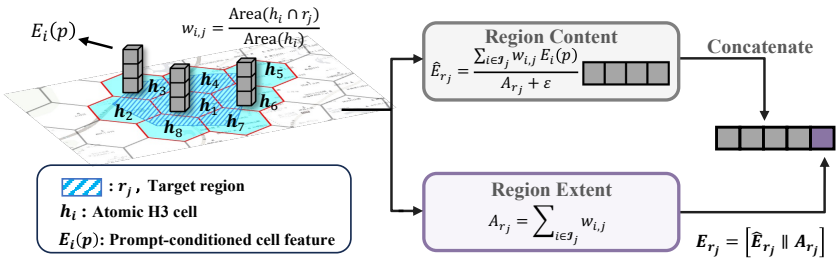

Figure 2: Content–extent region aggregation. Fractional cell overlaps yield both the normalized content feature $\hat{E}_{r_j}$ and the dimensionless effective atomic-cell count A_{r_j}. Concatenating them forms E_{r_j}; $\mathcal{I}_j$ indexes cells intersecting the target region r_j.

For a region–task query (r_j, k), let $p = p_{j,k}$ denote the condition for task k and the region's configuration. Let $\mathcal{I}_j$ index the atomic cells intersecting r_j. Define

$$w_{i,j} = \frac{\mathrm{Area}(h_i \cap r_j)}{\mathrm{Area}(h_i)}, \qquad A_{r_j} = \sum_{i \in \mathcal{I}_j} w_{i,j}. \tag{7}$$

Here $w_{i,j}$ is the fractional cell overlap, and A_{r_j} is the effective atomic-cell count, a dimensionless measure of region extent. We append it to the overlap-weighted mean to form the region representation (Fig. 2):

$$\hat{E}_{r_j} = \frac{\sum_{i \in \mathcal{I}_j} w_{i,j} E_i(p)}{A_{r_j} + \varepsilon}, \qquad E_{r_j} = [\hat{E}_{r_j} \,\|\, A_{r_j}], \tag{8}$$

where $\varepsilon > 0$ ensures numerical stability and $E_{r_j} \in \mathbb{R}^{d_{\text{out}}+1}$. The dependence of $\hat{E}_{r_j}$ and E_{r_j} on p is left implicit. Normalized pooling summarizes regional content but does not explicitly retain region extent. Appending A_{r_j} supplies size information, which is particularly relevant to additive targets such as total population. The predictor can therefore model extent-dependent target variation associated with the scale effect of MAUP.

The scale–task prediction head first maps E_{r_j} to a hidden feature H_{r_j} using a shared affine map $(\mathbf{W}_{\text{shared}}, \mathbf{b}_{\text{shared}})$, ReLU, and training-time dropout. Its prompt-generated readout produces the scalar prediction

$$\hat{y}_j^{(k)} = \mathbf{W}_{\text{pred}}(p) H_{r_j} + \mathbf{b}_{\text{pred}}(p). \tag{9}$$

We jointly train Stages 2–4 over observed task–configuration pairs using the mean squared error on a batch of region–task queries:

$$\mathcal{L}_{\text{down}} = \frac{1}{|\mathcal{B}|} \sum_{(r_j,k) \in \mathcal{B}} \left(\hat{y}_j^{(k)} - y_j^{(k)} \right)^2. \tag{10}$$

The loss is computed in z-score-standardized target space, using statistics fitted on the training fold.

Architecture and training details are provided in Appendix C.

4 Experiments

We evaluate (Q1) ScaleCity's performance across cities, tasks, and spatial configurations, (Q2) component contributions, (Q3) zero-shot prediction at unseen configurations, and (Q4) task- and configuration-dependent modality gates. Additional studies on few-shot task adaptation, city-specific evaluation in New York City and Singapore, and hyperparameter sensitivity are summarized in Section 4.6.

4.1 Experimental Setup

Benchmark. We evaluate on Beijing, Chengdu, and Shenzhen using eight targets: population (Pop.), Tencent human footprints (HFP), $PM_{2.5}$, CO, CO_2, GDP, housing price (HP), and bike-sharing (Bike). The nine spatial configurations comprise 1, 3, and 5 km grids, road-network partitions L1–L5, and L11 administrative street units, giving 72 task–configuration pairs per city. Partition construction and region counts are provided in Appendix A; data sources and target definitions are in Appendix C.1.

Baselines. We compare three unimodal methods (ViT, GeoVectors, and SpaBERT) and six multimodal methods (Urban2Vec, HAFusion, HREP, ComSRE, CityFM, and UrbanMFM). All methods use the same target regions and labels, with method-specific input sources and implementations detailed in Appendix C.4. ScaleCity jointly models all 72 task–configuration combinations within each city, whereas baselines fit separate predictors for each city–task–configuration combination.

Training and evaluation. ScaleCity pretrains its atomic-cell encoder once per city on city-wide unlabeled inputs, including evaluation regions, and reuses the frozen embeddings across downstream folds. Joint training samples region–task examples with replacement using configuration-dependent weights, favoring coarser levels within the grid and road-network families. All methods are scored using five-fold cross-validation. ScaleCity uses a common L11-anchor-to-fold mapping across configurations. Each region inherits the fold of the L11 anchor unit containing its centroid, with the largest-overlap unit used as a fallback. We report fold-averaged R^2; each task score averages available configurations, and Overall R^2 is the equal-weight mean of the eight task scores. Appendix C provides preprocessing, optimization, model-selection, and metric-aggregation details.

4.2 Performance Across Tasks and Spatial Configurations (Q1)

Overall comparison. Using one jointly trained model per city, ScaleCity leads all 24 task–city comparisons after averaging over configurations (Table 1). Relative to each city's strongest baseline, ScaleCity achieves absolute gains of 0.293–0.438 in Overall R^2, reaching 0.783, 0.671, and 0.644 in Beijing, Chengdu, and Shenzhen, respectively. Joint prediction thus retains a consistent advantage over separately fitted task–configuration predictors.

Table 1: Performance across three cities, eight tasks, and nine configurations (R^2 ↑). Task scores average available configurations after five-fold averaging; Overall is the mean of the eight task scores. The configuration range reports the min–max across configuration scores, each averaged over available tasks. Best and second-best task/Overall scores within each city are **bold** and underlined; shading identifies ScaleCity. BJ/CD/SZ: Beijing/Chengdu/Shenzhen; Uni./Multi.: unimodal/multimodal; EMNLP-F: Findings of EMNLP.

City	Group	Method	Population & activity		Atmospheric concentrations			Economy		Mobility	Overall	Configuration
			Pop.	HFP	$PM_{2.5}$	CO	CO_2	GDP	HP	Bike	R^2 ↑	R^2 (min–max)
BJ	Uni.	ViT (ICLR 2021)	0.422	0.443	0.413	0.167	0.335	0.250	0.489	0.243	0.345	0.172–0.520
		GeoVectors (CIKM 2021)	0.407	0.417	0.194	0.176	0.228	0.179	0.370	0.199	0.271	0.072–0.537
		SpaBERT (EMNLP-F 2022)	0.450	0.457	0.203	0.148	0.228	0.204	0.437	0.211	0.292	0.160–0.422
	Multi.	HAFusion (ICDE 2024)	0.476	0.540	0.093	0.096	0.201	0.266	0.367	0.130	0.271	0.131–0.518
		Urban2Vec (AAAI 2020)	0.358	0.353	0.210	0.094	0.251	0.129	0.433	0.146	0.247	0.040–0.481
		HREP (AAAI 2023)	0.355	0.357	0.055	0.088	0.084	0.159	-0.245	-0.072	0.098	-0.152–0.336
		CityFM (CIKM 2024)	0.360	0.369	0.128	0.087	0.167	0.228	0.264	0.157	0.220	0.088–0.346
		ComSRE (KDD 2026)	0.449	0.505	0.271	0.324	0.163	0.329	0.302	0.262	0.326	0.203–0.391
		UrbanMFM (TKDE 2026)	0.051	0.081	0.681	0.410	0.559	-0.065	0.405	-0.235	0.236	-0.406–0.506
		ScaleCity	**0.827**	**0.818**	**0.902**	**0.799**	**0.867**	**0.692**	**0.770**	**0.591**	**0.783**	0.653–0.891
CD	Uni.	ViT (ICLR 2021)	0.360	0.387	0.333	0.087	0.164	0.325	0.171	0.340	0.271	0.132–0.438
		GeoVectors (CIKM 2021)	0.385	0.387	0.485	0.169	0.045	0.339	0.180	0.247	0.280	0.103–0.472
		SpaBERT (EMNLP-F 2022)	0.511	0.495	0.492	0.165	0.116	0.433	0.202	0.305	0.340	0.181–0.442
	Multi.	HAFusion (ICDE 2024)	0.451	0.419	0.537	0.231	0.152	0.348	0.155	0.367	0.332	0.088–0.562
		Urban2Vec (AAAI 2020)	0.366	0.323	0.449	0.249	0.023	0.196	0.229	0.188	0.253	0.174–0.509
		HREP (AAAI 2023)	0.241	0.273	0.354	0.112	-0.003	0.178	-0.190	-0.156	0.101	-0.187–0.268
		CityFM (CIKM 2024)	0.182	0.244	0.273	0.031	0.046	0.212	0.156	0.243	0.174	0.083–0.254
		ComSRE (KDD 2026)	0.449	0.507	0.380	0.327	0.268	0.362	0.248	0.479	0.378	0.194–0.482
		UrbanMFM (TKDE 2026)	-0.049	-0.142	0.423	0.152	0.489	0.026	0.161	-0.005	0.132	-0.526–0.374
		ScaleCity	**0.622**	**0.677**	**0.864**	**0.725**	**0.774**	**0.629**	**0.494**	**0.580**	**0.671**	0.513–0.832
SZ	Uni.	ViT (ICLR 2021)	0.466	0.436	0.137	0.086	0.168	0.383	0.364	0.323	0.295	0.159–0.460
		GeoVectors (CIKM 2021)	0.326	0.309	-0.047	-0.034	0.053	0.229	0.297	0.302	0.179	0.053–0.302
		SpaBERT (EMNLP-F 2022)	0.418	0.367	-0.016	0.004	0.051	0.304	0.141	0.183	0.181	0.073–0.273
	Multi.	HAFusion (ICDE 2024)	0.453	0.340	0.267	0.018	0.125	0.235	0.226	0.253	0.240	0.102–0.418
		Urban2Vec (AAAI 2020)	0.454	0.450	0.068	0.069	0.222	0.241	0.277	0.171	0.244	0.143–0.349
		HREP (AAAI 2023)	0.160	0.202	0.380	-0.044	-0.171	0.158	-0.038	-0.021	0.078	-0.118–0.261
		CityFM (CIKM 2024)	0.281	0.251	0.015	0.048	0.081	0.213	0.140	0.167	0.150	0.072–0.241
		ComSRE (KDD 2026)	0.434	0.495	0.272	0.157	0.230	0.427	0.275	0.390	0.335	0.133–0.462
		UrbanMFM (TKDE 2026)	-0.077	0.021	0.351	-0.028	0.025	-0.132	0.419	-0.154	0.053	-0.431–0.245
		ScaleCity	**0.729**	**0.772**	**0.730**	**0.479**	**0.741**	**0.661**	**0.556**	**0.485**	**0.644**	0.498–0.755

Task and city variation. The strongest and weakest targets differ across cities. ScaleCity's highest-scoring task is $PM_{2.5}$ in Beijing and Chengdu, but HFP in Shenzhen. Its lowest-scoring task also varies: Bike in Beijing, HP in Chengdu, and CO in Shenzhen.

Performance across configurations. ScaleCity also leads all 27 city–configuration comparisons against the strongest baseline at each configuration (Fig. 3). Its task-averaged gains therefore extend across grid, road-network, and administrative partitions. L5 remains the lowest-scoring configuration in all three cities, identifying fine-grained prediction as a remaining challenge. Complete per-configuration scores and additional error metrics are in Appendix D.

4.3 Ablation Study (Q2)

Conditional fusion and prediction. Using prompt-generated modality projections and gates rather than trainable shared projections and uniform weights raises the three-city mean Overall R^2 by 0.099

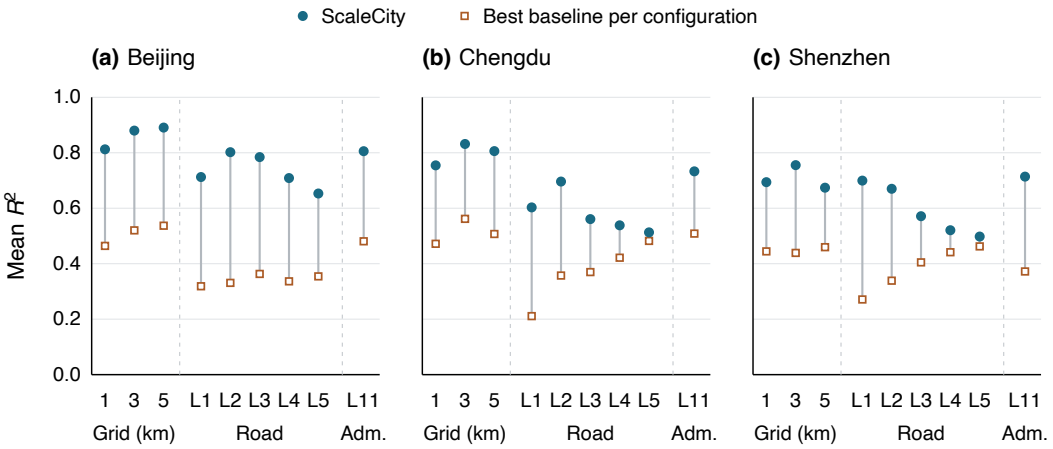

Figure 3: Performance across nine spatial configurations in (a) Beijing, (b) Chengdu, and (c) Shenzhen. Points show R^2 averaged over five folds, then available tasks. Grey lines pair ScaleCity with the best of nine baselines for each city–configuration pair. Grid labels give side lengths in kilometers.

(Table 2). Using the generated prediction head rather than one trainable head shared across tasks and configurations raises it by 0.101. Removing either fusion stream also reduces performance, with a larger effect for the adaptive residual stream. These comparisons support conditional computation in both feature fusion and prediction.

Region extent. The full model's largest mean advantage among the reported ablations is over the variant without A_{r_j} (0.180). Extensive targets, such as total population, are affected more on average than intensive targets, such as pollutant concentrations and average housing prices (Appendix E.3). Regional totals depend on both local characteristics and region size, whereas intensive targets describe levels or averages. The region-extent covariate complements normalized pooling with size information, helping capture the size-related variation in aggregated targets associated with MAUP.

Table 2: Component ablations, measured by Overall R^2 ↑ over eight tasks and nine configurations. Best and second-best values in each city are **bold** and underlined; shading marks the full model. Variant definitions are in Appendix E.1.

Variant	Beijing	Chengdu	Shenzhen
Full model	**0.783**	**0.671**	**0.644**
Inputs and spatial encoding			
w/o POI	0.738	0.647	0.578
w/o SVI	0.776	0.655	0.614
w/o AE	0.672	0.578	0.526
w/o GAT	0.766	0.629	0.608
Downstream fusion and prediction			
w/o Dynamic Routing ($\mathbf{W}_m, g_m$)	0.710	0.561	0.532
w/o Dyn. Pred. Head	0.632	0.585	0.578
w/o Adaptive Residual Stream	0.719	0.602	0.574
w/o Shared Base Stream	0.778	0.654	0.613
w/o Separate Prompt Encoding	0.779	0.664	0.632
w/o Region-Extent Covariate	0.634	0.502	0.422

Inputs and spatial context. The full model outperforms all three modality-removal variants, with its largest mean advantage over the no-AE variant (0.107). Retaining GAT-based neighborhood propagation improves mean Overall R^2 by 0.032 relative to the no-GAT variant, supporting the contribution of spatial context alongside multimodal observations.

Prompt design. Separately encoding scale and task descriptions improves mean Overall R^2 by 0.008 over concatenated-text encoding, with both descriptions retained. Beyond encoding, descrip-

tive prompts and symbolic identifiers yield similar joint-prediction scores, whereas constant prompts yield much lower scores. Descriptive prompts achieve higher mean performance than symbolic identifiers in few-shot task adaptation (Appendix G).

4.4 Zero-Shot Prediction at Unseen Configurations (Q3)

Unlike the joint-prediction benchmark, this experiment excludes entire target configurations from supervised training. We test *zero-shot prediction at unseen configurations* on HP without inference-time parameter updates. Table 3 covers omitted granularities within represented partition families and transfer across families, using city-specific models trained on the listed source sets.

Table 3: Five-fold mean $R^2 \uparrow$ for zero-shot HP prediction within each city. Blocks list the complete source sets; target configurations are excluded from training and evaluated without parameter updates. L1–L5 are road-network configurations; L11 is administrative.

Target configuration	Beijing	Chengdu	Shenzhen
I. Unseen granularities within represented families			
Source: 1/5 km grids; L1, L3, L5; L11			
3 km grid	0.703	0.750	0.790
L2	0.892	0.789	0.834
L4	0.761	0.641	0.746
II. Grids + administrative → road-network			
Source: 1/3/5 km grids; L11			
L1	0.569	0.674	0.700
L2	0.839	0.661	0.629
L3	0.818	0.529	0.626
L4	0.740	0.474	0.540
L5	0.677	0.412	0.529
III. Road-network → grids + administrative			
Source: L1–L5			
L11	0.937	0.729	0.847
1 km grid	0.804	0.652	0.653
3 km grid	0.543	0.643	0.726
5 km grid	0.434	0.558	0.637

Unseen granularities. On the omitted 3 km grid, L2, and L4 configurations, ScaleCity achieves R^2 of 0.641–0.892 across the three cities, demonstrating prediction at unseen granularities within represented partition families.

Unseen partition families. Training on grids and L11 (Setting II) yields R^2 of 0.412–0.839 on road-network targets. In the reverse direction (Setting III), scores range from 0.434 to 0.937, with L11 highest in each city. For the common L2 and L4 targets, Setting I achieves higher R^2 than Setting II in all three cities.

Together, these HP results show that a city-specific model supports unseen granularities and partition families without target-configuration supervision or parameter updates. Appendix F.1 provides the evaluation protocol.

4.5 Task- and Configuration-Dependent Modality Gating (Q4)

Figure 4 examines how ST-HyperNet allocates residual-stream modality weights $\mathbf{g}(p)$ across training, tasks, and configurations (Eq. 5). Recording and aggregation details are in Appendix F.2.

Training dynamics. The mean POI gate increases during training, while the AE gate decreases (Fig. 4a). Late-training weights average 0.399 for POI, 0.309 for SVI, and 0.292 for AE. The learned residual allocation thus places greater average weight on POI than on either visual modality. The larger AE ablation effect concerns removal from the full pipeline, whereas these gates describe weighting within the adaptive residual stream.

Task deviations. Relative to each city's modality-wise mean, population and human-activity targets (Pop. and HFP) shift toward AE, atmospheric concentrations ($PM_{2.5}$, CO, and CO_2) toward SVI,

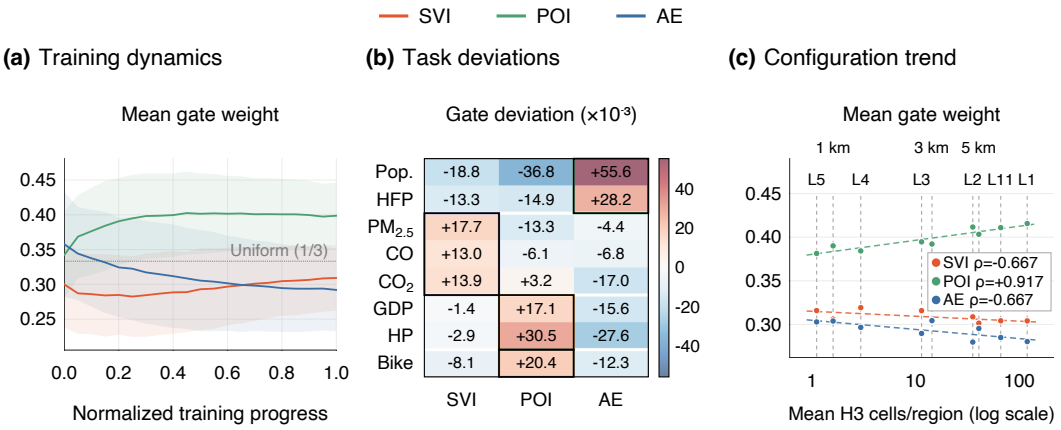

Figure 4: Modality gates across training, tasks, and configurations. (a) Mean trajectories with shading for ± 1 standard deviation over tasks, configurations, folds, and cities. (b) Task-wise deviations from each city's modality mean ($\times 10^{-3}$); outlined cells mark the largest positive deviation per task. (c) Mean gates over tasks, folds, and cities versus aggregation size; colored dashed lines show log-linear fits, and ρ denotes Spearman correlation across nine configurations. Panels (b,c) average the final three recorded epochs.

and economic and mobility targets (GDP, HP, and Bike) toward POI (Fig. 4b). The overall preference for POI therefore coexists with task-specific adjustments toward other modalities.

Configuration trend. Across nine configurations, larger aggregation size is associated with more POI weight (Spearman $\rho = 0.917$) and less SVI and AE weight ($\rho = -0.667$ each; Fig. 4c), indicating a shift toward POI in residual fusion at coarser granularities.

4.6 Additional Results

Few-shot task adaptation. Using labeled data from a 10% spatial support split for parameter updates, ScaleCity adapts to unseen NO_2 and PM_{10} targets with mean $R^2 = 0.747$ across cities and configurations. The existing downstream architecture accommodates both targets with frozen atomic representations and no additional task-specific modules (Appendix H).

City-specific evaluation. Training separately in New York City and Singapore yields Overall R^2 of 0.629 and 0.500, respectively, over four tasks and seven configurations. The best-performing partition differs between the cities, while the finest road-network partition, L5, has the lowest score in both (Appendix I).

Hyperparameter sensitivity. The tested prompt encoders yield mean Overall R^2 values within 0.005 of each other. H3 resolution 8, embedding dimension $d = 512$, and equal alignment-loss weights outperform their respective tested alternatives in all three cities, supporting the default configuration (Appendix J).

5 Conclusion

ScaleCity unifies multi-scale, multi-task urban region representation learning through joint training on mixed task–configuration samples. Prompt-conditioned fusion and prediction allow one city-specific model to serve multiple granularities and partition families, with explicit region extent retained during aggregation. Across three cities, eight tasks, and nine configurations, it outperforms the evaluated baselines in task- and configuration-averaged R^2. Ablations support the contributions of conditional computation and region extent; HP experiments demonstrate prediction at unseen configurations without parameter updates, and few-shot experiments extend the model to NO_2 and PM_{10}. Future work will focus on improving fine-grained prediction and characterizing the effects of regional boundary changes.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting and grammar enhancement, literature search and reference checking, and consistency checking of reported results. All AI-edited text was manually reviewed and verified by the authors. The authors take responsibility for the scientific claims, experimental results, and final content of this work.

REPRODUCIBILITY STATEMENT

Appendix B details the model formulation, and Appendix C documents data construction, model and training settings, evaluation protocols, and baseline implementations. Default prompts and symbolic identifiers appear in Appendix G.3; experiment-specific protocols accompany the supplementary results. Code is available on GitHub (`https://github.com/far-rele/ScaleCity`).

REFERENCES

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. URL https://doi.org/10.1145/3627673.3679662.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025. URL https://arxiv.org/abs/2507.22291.

Jinzhou Cao, Jiashi Chen, Xiangxu Wang, Weiming Huang, Dongsheng Chen, Tianhong Zhao, Wei Tu, and Qingquan Li. UrbanMMCL: Urban region representations via multi-modal and multi-graph self-supervised contrastive learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232:75–93, 2026. doi: 10.1016/j.isprsjprs.2025.11.012. URL https://doi.org/10.1016/j.isprsjprs.2025.11.012.

Liyue Chen, Jiangyi Fang, Tengfei Liu, Shaosheng Cao, and Leye Wang. A unified model for spatio-temporal prediction queries with arbitrary modifiable areal units. In *Proceedings of the 40th IEEE International Conference on Data Engineering*, pp. 1352–1365, 2024. doi: 10.1109/ICDE60146.2024.00111. URL https://doi.org/10.1109/ICDE60146.2024.00111.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 657–668, 2020. doi: 10.18653/v1/2020.findings-emnlp.58. URL https://aclanthology.org/2020.findings-emnlp.58/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

David Ha, Andrew Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkpACe1lx.

Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. UrbanVLP: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28061–28069, 2025. doi: 10.1609/aaai.v39i27.35024. URL https://ojs.aaai.org/index.php/AAAI/article/view/35024.

Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, Yaguang Li, Zhao Chen, Donald Metzler, Heng-Tze Cheng, and Ed H. Chi. HyperPrompt: Prompt-based task-conditioning of transformers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8678–8690, 2022. URL https://proceedings.mlr.press/v162/he22f.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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. URL https://doi.org/10.1016/j.isprsjprs.2022.11.021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision – ECCV 2022*, volume 13693 of *Lecture Notes in Computer Science*, pp. 709–727. Springer, 2022. doi: 10.1007/978-3-031-19827-4_41. URL https://link.springer.com/chapter/10.1007/978-3-031-19827-4_41.

Jiahui Jin, Yifan Song, Dong Kan, Haojia Zhu, Xiangguo Sun, Zhicheng Li, Xigang Sun, and Jinghui Zhang. Urban region pre-training and prompting: A graph-based approach. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 1071–1082, 2025. doi: 10.1145/3711896.3737176. URL https://doi.org/10.1145/3711896.3737176.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 565–576, 2021. doi: 10.18653/v1/2021.acl-long.47. URL https://aclanthology.org/2021.acl-long.47/.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023. doi: 10.1109/CVPR52729.2023.01832. URL https://openaccess.thecvf.com/content/CVPR2023/html/Khattak_MaPLe_Multi-Modal_Prompt_Learning_CVPR_2023_paper.html.

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://www.nature.com/articles/s41597-025-04487-x.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Yi Li, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. Urban region representation learning with OpenStreetMap building footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1363–1373, 2023. doi: 10.1145/3580305.3599538. URL https://doi.org/10.1145/3580305.3599538.

Zechen Li, Weiming Huang, Kai Zhao, Min Yang, Yongshun Gong, and Meng Chen. Urban region embedding via multi-view contrastive prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8724–8732, 2024. doi: 10.1609/aaai.v38i8.28718. URL https://ojs.aaai.org/index.php/AAAI/article/view/28718.

Zechen Li, Hongwei Jia, Weiming Huang, Kai Zhao, and Meng Chen. UrbanExpert: Task-conditioned multi-modal fusion via semantic expert routing for urban socioeconomic prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2026a. doi: 10.1145/3770855.3818168. URL https://doi.org/10.1145/3770855.3818168.

Zechen Li, Hongwei Jia, Kai Zhao, Weiming Huang, and Meng Chen. Multi-view urban region embedding via commonality-specificity disentanglement. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 748–758, 2026b. doi: 10.1145/3770854.3780234. URL https://doi.org/10.1145/3770854.3780234.

Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. SpaBERT: A pretrained language model from geographic data for geo-entity representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2757–2769, 2022. doi: 10.18653/v1/2022.findings-emnlp.200. URL https://aclanthology.org/2022.findings-emnlp.200/.

Pingping Liu, Jiamiao Liu, Zijian Zhang, Hao Miao, Qi Jiang, Qingliang Li, Qiuzhan Zhou, and Irwin King. UrbanMoE: A sparse multi-modal mixture-of-experts framework for multi-task urban region profiling. In *Proceedings of the ACM Web Conference 2026*, 2026. doi: 10.1145/3774904.3792701. URL https://doi.org/10.1145/3774904.3792701.

Yen-Cheng Liu, Chih-Yao Ma, Junjiao Tian, Zijian He, and Zsolt Kira. Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36889–36901, 2022. doi: 10.52202/068431-2673. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/efb02f96766a3b599c76852abf4d42dd-Abstract-Conference.html.

Stan Openshaw. *The Modifiable Areal Unit Problem*. Number 38 in Concepts and Techniques in Modern Geography. Geo Books, Norwich, 1984. ISBN 0860941345.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Fengze Sun, Jianzhong Qi, Yanchuan Chang, Xiaoliang Fan, Shanika Karunasekera, and Egemen Tanin. Urban region representation learning with attentive fusion. In *Proceedings of the 40th IEEE International Conference on Data Engineering*, pp. 4409–4421, 2024. doi: 10.1109/ICDE60146.2024.00336. URL https://doi.org/10.1109/ICDE60146.2024.00336.

Fengze Sun, Yanchuan Chang, Egemen Tanin, Shanika Karunasekera, and Jianzhong Qi. FlexiReg: Flexible urban region representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 2702–2713, 2025. doi: 10.1145/3711896.3736965. URL https://doi.org/10.1145/3711896.3736965.

Jintong Tang, Liyan Xu, Hongbin Yu, Hezhishi Jiang, Dejie He, Tianshu Li, Wanchen Xiao, Xinying Zheng, Keyi Liu, Yiqin Li, Shijie Li, Qian Huang, Jun Zhang, Yinsheng Zhou, Lun Wu, and Yu Liu. A dataset of multi-level street-block divisions of 985 cities worldwide. *Scientific Data*, 12:456, 2025. doi: 10.1038/s41597-025-04704-7. URL https://www.nature.com/articles/s41597-025-04704-7.

Nicolas Tempelmeier, Simon Gottschalk, and Elena Demidova. GeoVectors: A linked open corpus of OpenStreetMap embeddings on world scale. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 4604–4612, 2021. doi: 10.1145/3459637.3482004. URL https://doi.org/10.1145/3459637.3482004.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Jiawei Wang. Global 1 km resolution atmospheric carbon dioxide concentration dataset (2003–2023). National Tibetan Plateau / Third Pole Environment Data Center, 2025. URL https://doi.org/10.11888/Atmos.tpdc.302399. Dataset.

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. URL https://ojs.aaai.org/index.php/AAAI/article/view/5450.

Jing Wei, Zhanqing Li, Maureen Cribb, Wei Huang, Wenhao Xue, Lin Sun, Jianping Guo, Yiran Peng, Jing Li, Alexei Lyapustin, Lei Liu, Hao Wu, and Yimeng Song. Improved 1 km resolution $PM_{2.5}$ estimates across china using enhanced space–time extremely randomized trees. *Atmospheric Chemistry and Physics*, 20(6):3273–3289, 2020. doi: 10.5194/acp-20-3273-2020. URL https://acp.copernicus.org/articles/20/3273/2020/.

Jing Wei, Zhanqing Li, Alexei Lyapustin, Jun Wang, Oleg Dubovik, Joel Schwartz, Lin Sun, Chi Li, Song Liu, and Tong Zhu. First close insight into global daily gapless 1 km $PM_{2.5}$ pollution, variability, and health impact. *Nature Communications*, 14:8349, 2023a. doi: 10.1038/s41467-023-43862-3. URL https://doi.org/10.1038/s41467-023-43862-3.

Jing Wei, Zhanqing Li, Jun Wang, Can Li, Pawan Gupta, and Maureen Cribb. Ground-level gaseous pollutants (NO_2, SO_2, and CO) in china: Daily seamless mapping and spatiotemporal variations. *Atmospheric Chemistry and Physics*, 23(2):1511–1532, 2023b. doi: 10.5194/acp-23-1511-2023. URL https://acp.copernicus.org/articles/23/1511/2023/.

Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. Multi-graph fusion networks for urban region embedding. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2312–2318, 2022. doi: 10.24963/ijcai.2022/321. URL https://www.ijcai.org/proceedings/2022/321.

Congxi Xiao, Jingbo Zhou, Yixiong Xiao, Jizhou Huang, and Hui Xiong. ReFound: Crafting a foundation model for urban region understanding upon language and visual foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3527–3538, 2024a. doi: 10.1145/3637528.3671992. URL https://doi.org/10.1145/3637528.3671992.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024b. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Xinliang Xu. China GDP spatial distribution kilometer-grid dataset. Resource and Environment Science Data Registration and Publishing System, 2017. URL https://www.resdc.cn/DOI/DOI.aspx?DOIID=33. Dataset (in Chinese).

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. UrbanCLIP: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*, pp. 4006–4017, 2024. doi: 10.1145/3589334.3645378. URL https://doi.org/10.1145/3589334.3645378.

Liang Zhang, Cheng Long, and Gao Cong. Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9031–9036, 2023. doi: 10.1109/TKDE.2022.3220874. URL https://doi.org/10.1109/TKDE.2022.3220874.

Zhaoqi Zhang, Miao Xie, Pasquale Balsebre, Weiming Huang, Siqiang Luo, and Gao Cong. UrbanMFM: Spatial graph-based multiscale foundation models for learning generalized urban representation. *IEEE Transactions on Knowledge and Data Engineering*, 38(3):2064–2078, 2026. doi: 10.1109/TKDE.2026.3656202. URL https://doi.org/10.1109/TKDE.2026.3656202.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022. doi: 10.1007/s11263-022-01653-1. URL https://link.springer.com/article/10.1007/s11263-022-01653-1.

Silin Zhou, Dan He, Lisi Chen, Shuo Shang, and Peng Han. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4981–4989, 2023. doi: 10.1609/aaai.v37i4.25625. URL https://ojs.aaai.org/index.php/AAAI/article/view/25625.

Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, and Yuxuan Liang. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion*, 113:102606, 2025. doi: 10.1016/j.inffus.2024.102606. URL https://doi.org/10.1016/j.inffus.2024.102606.

A Spatial Partitions and Region Statistics

A.1 Spatial Units and Partition Construction

Resolution-8 H3 cells provide the common atomic support. The benchmark uses nine target configurations from three partition families: regular grids with side lengths of 1, 3, and 5 km; road-network regions L1–L5, ordered from coarse to fine; and L11 administrative street units. Regular grids are constructed within each city's study area. Road-network partitions are taken from MSDCW (Tang et al., 2025). L1 uses motorway and trunk roads; L2–L5 successively add primary, secondary, tertiary, and residential roads. The source construction applies road buffering, overlay, and polygon refinement. L11 uses official administrative street boundaries in the Chinese cities. In the international experiments, L11 corresponds to police precincts in New York City[1] and Neighbourhood Police Centre boundaries in Singapore[2]. Figure 5 illustrates the nine configurations.

A.2 Region and Input Statistics

Table 4 lists total region counts by city and configuration, while Table 5 reports H3 cell, POI, and SVI counts.

Table 4: Region counts by city and spatial configuration.

Family	Configuration	Beijing	Chengdu	Shenzhen	NYC	Singapore
Regular grid	1 km	16,481	14,485	1,964	779	806
	3 km	1,831	1,612	220	85	93
	5 km	654	584	75	31	38
Road network	L1	207	114	102	88	53
	L2	610	442	301	268	238
	L3	1,923	1,358	945	1,188	456
	L4	7,309	5,668	2,835	2,197	778
	L5	23,402	12,787	7,064	29,978	4,208
Administrative	L11	331	316	74	78	140

Table 5: H3 cell, POI, and SVI counts by city.

City	H3 cells (res. 8)	POI records	SVI images
Beijing	31,072	1,295,588	43,406
Chengdu	20,222	1,189,679	29,024
Shenzhen	2,984	1,060,760	30,442
New York City	1,054	500,211	1,820,780
Singapore	943	335,095	594,872

[1] https://www.nyc.gov/content/planning/pages/resources/datasets/police-precincts

[2] https://data.gov.sg/datasets/d_89b44df21fccc4f51390eaff16aa1fe8/view

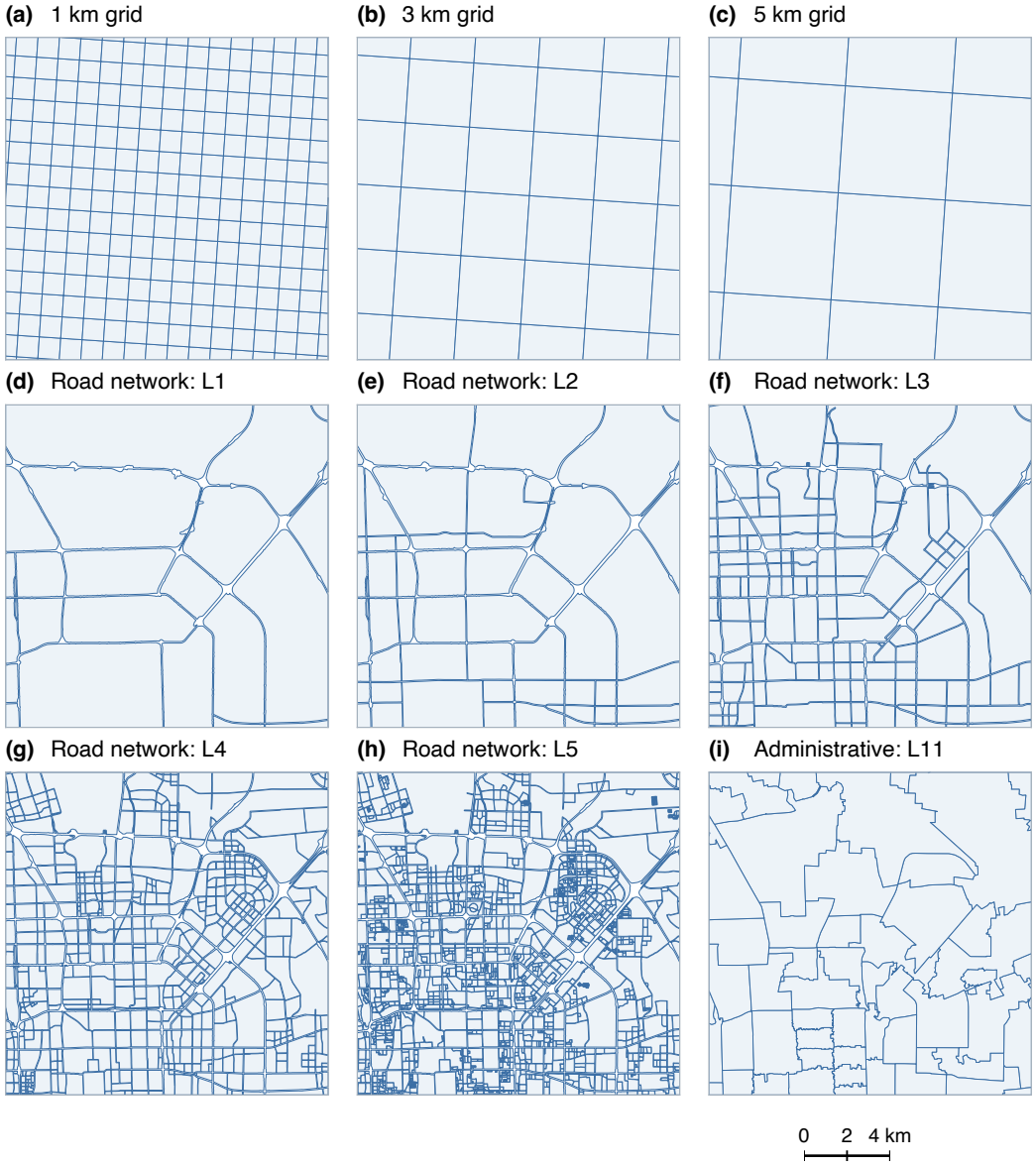

Figure 5: Nine spatial configurations over the same 15×15 km window in Beijing: regular grids (a–c), road-network partitions from coarse L1 to fine L5 (d–h), and L11 administrative units (i).

B Supplementary Method Details

B.1 Input Encoding and Spatial Context

For $m \in \{\text{SVI}, \text{POI}\}$, encoder f_m maps each observation $x_{i,q}^m$ in cell h_i to a feature vector. With validity mask $M_{i,q}^m \in \{0, 1\}$, masked mean pooling gives

$$z_{m,i}^{(0)} = \frac{\sum_q M_{i,q}^m f_m(x_{i,q}^m)}{\sum_q M_{i,q}^m + \varepsilon}. \tag{11}$$

Masks exclude invalid or padded observations; $\varepsilon > 0$ keeps pooling defined when all observations are invalid, yielding a zero vector. For remote sensing, a trainable projection maps the precomputed

AE vector x_i^{AE} into the graph feature space:

$$z_{\mathrm{AE},i}^{(0)} = \mathrm{GELU}\big(\mathrm{LayerNorm}\big(\mathbf{W}_{\mathrm{lift}} x_i^{\mathrm{AE}} + \mathbf{b}_{\mathrm{lift}}\big)\big). \tag{12}$$

Training-time dropout follows this projection.

Each modality uses L multi-head GAT blocks over neighboring H3 cells. Let $\mathcal{N}(i)$ denote the nonself neighbors; self-loops are included in attention. At layer ℓ, head a uses projection $\mathbf{U}_m^{(\ell,a)}$ and attention vector $\mathbf{a}_m^{(\ell,a)}$ to compute

$$\begin{aligned} c_{m,ij}^{(\ell,a)} &= \mathrm{LeakyReLU}\Big(\mathbf{a}_m^{(\ell,a)\top}\Big[\mathbf{U}_m^{(\ell,a)} z_{m,i}^{(\ell)} \,\|\, \mathbf{U}_m^{(\ell,a)} z_{m,j}^{(\ell)}\Big]\Big), \\ \alpha_{m,ij}^{(\ell,a)} &= \frac{\exp(c_{m,ij}^{(\ell,a)})}{\sum_{v\in\mathcal{N}(i)\cup\{i\}} \exp(c_{m,iv}^{(\ell,a)})}. \end{aligned} \tag{13}$$

Training-time dropout on the softmax coefficients gives $\tilde{\alpha}_{m,ij}^{(\ell,a)}$. Each head outputs

$$u_{m,i}^{(\ell,a)} = \mathrm{ELU}\left(\sum_{j\in\mathcal{N}(i)\cup\{i\}} \tilde{\alpha}_{m,ij}^{(\ell,a)} \mathbf{U}_m^{(\ell,a)} z_{m,j}^{(\ell)}\right). \tag{14}$$

The concatenated outputs pass through a width-preserving linear projection, residual addition, and LayerNorm to produce $z_{m,i}^{(\ell+1)}$, with dropout on the residual path during training.

An alignment MLP P_m with a hidden-layer ReLU maps the graph features to the shared embedding space. Final ℓ_2 normalization yields $e_{m,i}$.

B.2 Cross-Modal Contrastive Alignment

Let $\mathcal{B}_{\mathrm{pre}}$ denote the synchronized batch from one forward pass. For modality pair (m, n), the directional loss is

$$\mathcal{L}_{m\to n} = -\frac{1}{|\mathcal{B}_{\mathrm{pre}}|} \sum_{i\in\mathcal{B}_{\mathrm{pre}}} \log \frac{\exp(e_{m,i}^{\top} e_{n,i}/\tau)}{\sum_{j\in\mathcal{B}_{\mathrm{pre}}} \exp(e_{m,i}^{\top} e_{n,j}/\tau)}. \tag{15}$$

Here τ is the contrastive temperature. Averaging both directions gives

$$\mathcal{L}_{m,n} = \tfrac{1}{2}\left(\mathcal{L}_{m\to n} + \mathcal{L}_{n\to m}\right). \tag{16}$$

B.3 Prompt Conditioning and Parameter Generation

A frozen text encoder represents P_s and P_t separately using final-layer `[CLS]` features without ℓ_2 normalization. Each projection head f_s or f_t applies a linear map, GELU, and LayerNorm. Their concatenated outputs pass through a fusion MLP comprising a linear map, GELU, LayerNorm, and a final linear map to form p.

ST-HyperNet maps p through a shared condition encoder with GELU and LayerNorm after each linear layer. Separate generation heads use GELU in hidden layers to produce the adaptive parameters. The modality gates are

$$\mathbf{g}(p) = \mathrm{Softmax}(\mathrm{MLP}_{\mathrm{gate}}(p)) \in \mathbb{R}^3. \tag{17}$$

Here $\mathrm{MLP}_{\mathrm{gate}}$ includes the shared encoder and gate head; its outputs correspond to SVI, POI, and AE.

B.4 Shared Fusion and Regional Readout

The shared base stream is

$$E_{\mathrm{base},i} = \mathbf{W}_{\mathrm{fuse}}\left[\phi_{\mathrm{base,SVI}}(e_{\mathrm{SVI},i}) \,\|\, \phi_{\mathrm{base,POI}}(e_{\mathrm{POI},i}) \,\|\, \phi_{\mathrm{base,AE}}(e_{\mathrm{AE},i})\right] + \mathbf{b}_{\mathrm{fuse}}. \tag{18}$$

Regions with no contributing atomic cells are excluded. The shared map preceding the generated readout applies ReLU and training-time dropout:

$$H_{r_j} = \mathrm{Dropout}\big(\mathrm{ReLU}\big(\mathbf{W}_{\mathrm{shared}} E_{r_j} + \mathbf{b}_{\mathrm{shared}}\big)\big). \tag{19}$$

C EXPERIMENTAL SETUP

C.1 DATA SOURCES AND PREPROCESSING

Pretraining inputs. Stage 1 learns atomic-cell representations from unlabeled SVI, POI, and AE features. Beijing, Chengdu, and Shenzhen use Amap[3] POI and Baidu SVI; New York City and Singapore use Foursquare Open Source Places[4] and Google Street View. SVI dates to 2020 in the Chinese cities, 2020–2022 in New York City, and predominantly 2021 in Singapore; POI data are from 2020. All cities use the 2020 Satellite Embedding V1[5] layer from AlphaEarth Foundations (Brown et al., 2025). SVI is deduplicated by panorama identifier and sampled without replacement up to 32 images per H3 cell, with standard CLIP preprocessing. POI text combines names and hierarchical categories, capped at 64 records per cell and 64 tokens per record; foreign category labels are translated into Chinese with NLLB-200-3.3B for the POI encoder. AE features are aggregated over H3 cells in Google Earth Engine at a 10 m sampling scale.

Prediction targets. Source data have their own spatial support, such as raster cells or geolocated records; labels for each target configuration are constructed using the aggregation rules in Table 6. Totals on source cells are apportioned by fractional overlap, whereas areal densities are integrated over the intersecting area. In particular, the Chinese GDP raster records 10^4 CNY/km^2; multiplying by overlap areas in km^2 and summing yields regional GDP in 10^4 CNY. Concentrations and per-unit prices retain their units under the listed mean aggregations.

CO_2 denotes column-averaged atmospheric concentration. Beijing Bike data come from daily bicycle-location records; Chengdu and Shenzhen use local bike-sharing platform records. New York City and Singapore use WorldPop, global XCO_2, and the 2020 30-arc-second total-GDP (PPP) product (Kummu et al., 2025), with overlap-weighted sums for totals and means for concentrations. Their $PM_{2.5}$ labels use the 2020 annual 1 km GlobalHighPM$_{2.5}$ product (Wei et al., 2023a). Targets are z-score standardized using training-fold statistics, without logarithmic transformation, clipping, or outlier removal.

Table 6: Primary-benchmark label sources, regional aggregation rules, and resulting target units.

Target	Source	Period	Regional aggregation	Target unit
Population	WorldPop[a], 100 m	2020	Overlap-weighted sum	People
HFP	Tencent human footprints	2020	Sum of human-footprint values	Person-visits
$PM_{2.5}$	ChinaHighPM$_{2.5}$ (Wei et al., 2020)	2020	Area-weighted mean	μg/m^3
CO	ChinaHighCO (Wei et al., 2023b)	2020	Area-weighted mean	mg/m^3
CO_2	Global 1 km XCO_2 (Wang, 2025)	2020	Mean raster value	ppm
GDP	China 1 km GDP (Xu, 2017)	2020	Area integral of GDP density	10^4 CNY
HP	Lianjia[b] transaction records	2020-11	Mean price per square metre	CNY/m^2
Bike	City-specific platform records	2020-11	Sum of bike-sharing counts	Person-trips

[a] https://hub.worldpop.org/

[b] https://www.lianjia.com/

[3] https://lbs.amap.com/

[4] https://opensource.foursquare.com/os-places/

[5] https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_SATELLITE_EMBEDDING_V1_ANNUAL

C.2 Data Access and Evaluation

Stage 1 pretrains once per city on city-wide unlabeled inputs, including held-out-region features, without partitioning the graph by downstream folds. Its frozen atomic embeddings are reused across ScaleCity's five L11-anchored folds. Each target region inherits the fold of the L11 anchor unit containing its centroid, falling back to the unit with the largest overlap. The L11-anchor-to-fold mapping is shared across configurations. The same held-out fold is used for model selection and final scoring.

C.3 Model Configuration and Training

Tables 7–9 list model and training settings. SVI uses CLIP's post-LayerNorm `[CLS]` features before visual projection. Stage 1 trains LoRA adapters, graph layers, and modality projections, keeping backbone weights frozen. LoRA targets attention and feed-forward projections in CLIP and MacBERT. All downstream modules are jointly trained with the atomic embeddings and prompt encoder fixed.

Table 7: Default model configuration.

Setting	Value
Stage 1: atomic-cell encoder	
Observation encoders	CLIP ViT-B/32 (SVI); MacBERT (POI); both 768-dimensional
AE lift	$64 \to 256$
Graph neighborhood	One-hop and two-hop H3 neighbors
GAT depth / heads	$L = 2$ blocks per modality; 4 heads
GAT head width / LeakyReLU slope	192 (SVI, POI), 64 (AE) / 0.2
Alignment MLP widths	$768 \to 768 \to 512$ (SVI, POI); $256 \to 256 \to 512$ (AE), with $d = 512$
Encoder dropout	0.1 at the AE lift, GAT attention, and residual path
LoRA rank / α / dropout	16 / 32 / 0.1
Contrastive temperature / pair weights	Fixed $\tau = 0.07$; equal modality-pair weights
Stages 2–4: conditional prediction	
Prompt encoder	BGE-base-en-v1.5 (Xiao et al., 2024b); 768 dimensions, at most 128 tokens
Prompt projection / fusion MLP	$768 \to 384$ per description / $768 \to 768 \to 768$
Shared condition encoder	$768 \to 512 \to 256$
Fused cell / factorization widths	$d_{\text{out}} = 256$ / $r = 8$
Shared base stream	Three affine $512 \to 256$ modality projections; concatenation followed by affine $768 \to 256$ fusion, without activation or dropout
Shared region-to-latent map	$257 \to 64$
Downstream dropout	0.3 at the shared readout hidden layer only

Table 8: ST-HyperNet head widths. Factor outputs cover all three modalities.

Generated quantity	Head widths
Gate logits	$256 \to 64 \to 3$
$\mathbf{A}_m$	$256 \to 256 \to 256 \to (3r \cdot 256)$
$\mathbf{B}_m$	$256 \to 256 \to 256 \to (3 \cdot 512r)$
Modality biases $\mathbf{b}_m$	$256 \to 128 \to (3 \cdot 256)$
Readout weights $\mathbf{W}_{\text{pred}}$	$256 \to 128 \to 64 \to 64$
Readout bias $\mathbf{b}_{\text{pred}}$	$256 \to 32 \to 1$

Downstream sampling draws region–task examples with replacement, with one dataset-length set of draws per epoch. Weights are $5, 4, 3, 2, 1$ for L1–L5, 5 for L11, and $2, 3, 5$ for the 1, 3, and 5 km grids, shared across tasks without task-frequency balancing. Stage 1 uses seed 42 plus GPU rank; downstream splits and distributed sampling use 2026, with other downstream randomness unseeded.

Table 9: Training settings. The contrastive batch is formed per forward pass; the effective optimization batch includes gradient accumulation.

Setting	Value
Stage 1: atomic-cell pretraining	
Optimizer; learning rate / weight decay	AdamW; 10^{-4} / 10^{-4} for all parameter groups
Epochs / learning-rate schedule	40 / constant
Per-GPU / global contrastive batch	4 / 16 center cells on 4 GPUs
Accumulation / effective batch	4 forward passes / 64 center cells per full update
Checkpoint selection	Lowest epoch-average training loss
Stages 2–4: joint downstream training	
Optimizer; learning rate / weight decay	Adam; 10^{-4} / 10^{-3}
Batch size	256 region–task examples
Learning-rate scheduler	ReduceLROnPlateau; factor 0.5, patience 5
Maximum epochs / stopping patience	100 / 10
Scheduler and checkpoint metric	Validation Overall R^2 (maximize)

For the primary benchmark, results are from one five-fold run. On four A6000 GPUs, downstream training takes 29.97 s per epoch and averages 38 epochs per fold.

C.4 Baseline Implementations

Baseline inputs are listed in Table 10. HAFusion and HREP learn representations per city–configuration pair, and ComSRE per city–task–configuration combination; the other baselines reuse encoders across configurations. Representations are cached across folds, with a separate predictor fitted for each city–task–configuration–fold combination.

Unimodal baselines.

- **ViT** (Dosovitskiy et al., 2021) uses frozen CLIP ViT-B/32 pre-projection features (Radford et al., 2021), followed by regional mean pooling and an MLP.
- **GeoVectors** (Tempelmeier et al., 2021) concatenates frozen fastText features with city-trained Node2Vec-style embeddings, followed by regional mean pooling and an MLP.
- **SpaBERT** (Li et al., 2022) continues BERT-base-uncased masked-language pretraining for five epochs on pooled POI descriptions from the three primary-benchmark cities. Mean-pooled valid-token features feed an MLP.

Multimodal baselines.

- **Urban2Vec** (Wang et al., 2020) fine-tunes ImageNet-pretrained InceptionV3 with two-stage neighborhood-based margin-ranking training. Regionally mean-pooled features feed an RBF SVR ($C = 1$, `gamma=auto`).
- **HAFusion** (Sun et al., 2024) and **ComSRE** (Li et al., 2026b) use learned region embeddings with Ridge regression.
- **HREP** (Zhou et al., 2023) combines a frozen Ridge-initialized readout ($\alpha = 1$) with trainable region-specific RegionPrompt parameters.
- **CityFM** (Balsebre et al., 2024) uses frozen city-pretrained embeddings, regional mean pooling, and an MLP, with positional encoding set to zero.
- **UrbanMFM** (Zhang et al., 2026) replaces satellite imagery with AE. Regional mean pooling and modality averaging replace the untrained GatedFusion module; a gated MLP uses these features and positional encoding.

Training and selection. Hyperparameters use published, default, or fixed values without systematic search; all methods use one five-fold evaluation run. ViT selects predictors by validation loss; GeoVectors, SpaBERT, and CityFM use validation R^2; UrbanMFM uses its final predictor after 20

epochs. HAFusion and ComSRE use Ridge with $\alpha = 1$ and clip negative predictions to zero. Before cross-fold caching, HAFusion selects representations using ten-fold population Ridge scores, and HREP using five-fold Ridge scores across all task labels.

Table 10: Input sources for each method. OSM: OpenStreetMap.

Method	Input sources
ViT	SVI
GeoVectors	OSM entity tags and node locations
SpaBERT	POI descriptions
Urban2Vec	SVI and POI
HAFusion	POI, land use, and inter-region mobility flows
HREP	POI, inflow/outflow mobility, and geographic neighbors
ComSRE	POI and source/destination mobility
CityFM	OSM tags, building footprints, and coordinates
UrbanMFM	AE, POI, and SVI
ScaleCity	SVI, POI, AE, and spatial-configuration/task prompts

C.5 Metric Aggregation

For the primary benchmark, task scores average five folds and then available configurations; Overall R^2 equally averages the eight task scores. The configuration range reports the minimum and maximum R^2 across configuration scores, each averaged over available tasks. MAE and RMSE are computed after inverse transformation to the original target units and use the same fold and configuration averaging. All R^2 scores and differences are rounded to three decimal places after calculation.

D Additional Performance Results (Q1)

This appendix provides the complete numerical breakdowns for Q1.

Per-configuration performance. Table 11 lists the per-configuration R^2 of every model.

Table 11: Per-configuration R^2 across nine configurations and three cities. Best and second-best values within each city are **bold** and underlined; shading identifies ScaleCity.

City	Method	1 km	3 km	5 km	L1	L2	L3	L4	L5	L11
Beijing	ViT	0.428	0.520	0.491	0.319	0.331	0.227	0.208	0.172	0.409
	GeoVectors	0.375	0.470	0.537	0.218	0.212	0.122	0.074	0.072	0.361
	SpaBERT	0.422	0.406	0.374	0.170	0.219	0.252	0.229	0.160	0.399
	HAFusion	0.345	0.518	0.344	0.188	0.216	0.184	0.178	0.131	0.338
	Urban2Vec	0.274	0.329	0.321	0.209	0.204	0.191	0.171	0.040	0.481
	HREP	-0.152	0.309	0.305	0.077	0.062	0.009	0.078	-0.143	0.336
	CityFM	0.284	0.346	0.316	0.163	0.225	0.170	0.129	0.088	0.260
	ComSRE	0.391	0.376	0.390	0.203	0.312	0.363	0.336	0.354	0.203
	UrbanMFM	0.464	0.506	0.236	-0.406	0.218	0.263	0.282	0.282	0.277
	ScaleCity (Ours)	**0.812**	**0.880**	**0.891**	**0.713**	**0.802**	**0.784**	**0.709**	**0.653**	**0.806**
Chengdu	ViT	0.358	0.438	0.401	0.195	0.209	0.212	0.186	0.132	0.306
	GeoVectors	0.325	0.461	0.472	0.211	0.243	0.184	0.103	0.103	0.414
	SpaBERT	0.391	0.442	0.380	0.181	0.338	0.370	0.304	0.235	0.417
	HAFusion	0.297	0.562	0.507	0.088	0.293	0.356	0.344	0.255	0.289
	Urban2Vec	0.209	0.222	0.181	0.192	0.305	0.288	0.197	0.174	0.509
	HREP	-0.187	0.217	0.187	0.069	0.139	0.056	0.156	0.005	0.268
	CityFM	0.246	0.253	0.254	0.130	0.141	0.136	0.097	0.083	0.221
	ComSRE	0.472	0.440	0.385	0.194	0.357	0.364	0.422	0.482	0.282
	UrbanMFM	0.374	0.344	0.259	-0.526	0.044	0.209	0.210	0.221	0.052
	ScaleCity (Ours)	**0.755**	**0.832**	**0.806**	**0.603**	**0.696**	**0.561**	**0.538**	**0.513**	**0.733**
Shenzhen	ViT	0.321	0.439	0.460	0.271	0.254	0.202	0.181	0.159	0.372
	GeoVectors	0.237	0.302	0.232	0.220	0.119	0.053	0.065	0.104	0.282
	SpaBERT	0.273	0.218	0.206	0.073	0.145	0.183	0.201	0.194	0.142
	HAFusion	0.287	0.418	0.102	0.213	0.259	0.257	0.310	0.169	0.141
	Urban2Vec	0.213	0.337	0.256	0.143	0.230	0.238	0.215	0.215	0.349
	HREP	0.243	0.192	0.195	-0.118	-0.048	-0.079	0.013	0.044	0.261
	CityFM	0.169	0.237	0.226	0.241	0.113	0.092	0.078	0.072	0.117
	ComSRE	0.444	0.334	0.257	0.198	0.339	0.405	0.441	0.462	0.133
	UrbanMFM	0.245	0.195	-0.073	-0.431	0.123	0.160	0.148	0.148	-0.035
	ScaleCity (Ours)	**0.694**	**0.755**	**0.674**	**0.700**	**0.670**	**0.571**	**0.521**	**0.498**	**0.714**

Mean absolute error. Table 12 reports per-task MAE. ScaleCity ranks first in all 24 task–city comparisons.

Table 12: Per-task MAE ↓ of all models in the three cities. Each entry averages the available spatial configurations for that task; lower is better. Best and second-best values in each city–task column are **bold** and underlined; shading identifies ScaleCity.

City	Method	Pop.	HFP	$PM_{2.5}$	CO	CO_2	GDP	HP	Bike
Beijing	ViT	32,346	2,132.6	1.427	0.0264	0.1643	697,230	18,088	1,169,666
	GeoVectors	23,059	1,444.5	2.297	0.0305	0.2292	456,460	19,766	1,001,660
	SpaBERT	24,957	1,672.2	2.268	0.0305	0.2254	485,816	18,108	1,016,735
	HAFusion	20,481	1,272.8	2.528	0.0335	0.2344	403,729	18,492	974,878
	Urban2Vec	22,169	1,503.5	2.332	0.0322	0.2258	454,602	17,415	976,237
	HREP	22,638	1,491.6	2.216	0.0284	0.2154	494,473	29,844	1,317,618
	CityFM	22,805	1,513.2	2.579	0.0337	0.2455	453,605	21,693	1,098,812
	ComSRE	18,004	1,892.7	1.258	0.0258	0.1407	604,955	20,353	916,754
	UrbanMFM	42,905	3,167.4	0.7031	0.0192	0.0993	960,032	18,473	1,392,938
	ScaleCity (Ours)	**12,920**	**978.8**	**0.4952**	**0.0115**	**0.0689**	**318,843**	**10,352**	**585,105**
Chengdu	ViT	41,804	3,131.0	1.293	0.0165	0.1598	374,734	3,950.0	1,691.8
	GeoVectors	29,160	2,036.2	1.594	0.0211	0.2919	237,131	4,093.3	1,147.2
	SpaBERT	28,753	2,174.0	1.553	0.0210	0.2645	250,257	4,103.2	1,254.5
	HAFusion	29,395	1,994.8	1.568	0.0207	0.2949	233,481	4,068.2	1,018.2
	Urban2Vec	26,389	1,899.0	1.676	0.0215	0.3241	242,893	3,671.8	1,034.9
	HREP	32,311	2,483.1	1.793	0.0217	0.3514	265,337	5,288.8	1,630.7
	CityFM	34,520	2,404.8	2.079	0.0241	0.3359	270,386	4,231.3	1,192.2
	ComSRE	24,194	2,522.2	1.334	0.0134	0.1599	294,458	3,907.7	1,310.0
	UrbanMFM	56,380	4,403.5	1.163	0.0121	0.1125	436,160	3,935.9	2,208.9
	ScaleCity (Ours)	**21,722**	**1,677.9**	**0.5429**	**0.0077**	**0.0748**	**191,635**	**2,759.9**	**888.1**
Shenzhen	ViT	56,668	5,285.0	0.7336	0.0181	0.2128	913,109	12,622	482.0
	GeoVectors	61,544	5,687.9	0.8641	0.0221	0.2257	903,692	13,262	474.4
	SpaBERT	61,375	5,780.2	0.8390	0.0211	0.2213	898,293	14,915	591.3
	HAFusion	45,501	5,218.3	0.6501	0.0220	0.2119	849,376	13,282	554.7
	Urban2Vec	51,627	4,687.6	0.8182	0.0186	0.2098	926,996	12,924	529.6
	HREP	57,710	5,034.1	0.6177	0.0264	0.2559	956,786	16,698	659.6
	CityFM	54,792	5,173.4	0.8736	0.0234	0.2331	862,000	15,200	542.3
	ComSRE	43,843	4,662.7	0.6922	0.0173	0.1761	723,990	12,606	415.5
	UrbanMFM	76,699	6,357.5	0.5384	0.0206	0.2045	1,264,195	11,779	767.3
	ScaleCity (Ours)	**25,763**	**2,316.3**	**0.3173**	**0.0122**	**0.1049**	**499,528**	**9,090.5**	**323.4**

Root mean squared error. Table 13 reports per-task RMSE. ScaleCity ranks first in 23 of the 24 task–city comparisons.

Table 13: Per-task RMSE ↓ of all models in the three cities. Each entry averages the available spatial configurations for that task; lower is better. Best and second-best values in each city–task column are **bold** and underlined; shading identifies ScaleCity.

City	Method	Pop.	HFP	$PM_{2.5}$	CO	CO_2	GDP	HP	Bike
Beijing	ViT	50,174	3,584.8	1.917	0.0334	0.2203	1,404,742	24,379	2,122,078
	GeoVectors	41,817	2,835.2	3.011	0.0376	0.2903	1,095,554	26,528	1,852,014
	SpaBERT	44,510	3,066.9	2.973	0.0378	0.2908	1,159,041	24,576	2,048,690
	HAFusion	38,485	2,648.2	3.272	0.0405	0.3001	981,750	25,384	2,043,126
	Urban2Vec	46,424	3,269.3	3.033	0.0393	0.2898	1,222,424	24,981	2,113,841
	HREP	39,677	2,742.2	2.876	0.0355	0.2707	1,022,820	37,812	2,349,686
	CityFM	41,381	2,926.7	3.300	0.0405	0.3153	1,036,063	28,809	2,021,209
	ComSRE	25,876	2,888.8	1.700	0.0312	0.1810	1,181,394	26,383	1,725,216
	UrbanMFM	63,743	4,690.5	0.9812	0.0243	0.1304	1,635,424	24,995	2,253,250
	ScaleCity (Ours)	**20,646**	**1,665.1**	**0.7691**	**0.0163**	**0.0976**	**659,057**	**14,996**	**1,122,981**
Chengdu	ViT	60,901	5,342.0	1.750	0.0213	0.2196	543,122	5,707.3	3,201.7
	GeoVectors	46,757	3,871.4	2.072	0.0272	0.3928	404,412	5,912.2	2,661.1
	SpaBERT	47,796	4,069.7	2.014	0.0268	0.3563	423,049	5,839.6	2,785.3
	HAFusion	48,705	4,135.3	2.070	0.0265	0.3979	396,762	5,831.8	2,463.5
	Urban2Vec	51,560	4,385.0	2.183	0.0269	0.4247	468,081	5,610.8	2,834.5
	HREP	50,699	4,191.8	2.279	0.0277	0.4455	409,269	7,022.8	3,279.0
	CityFM	53,484	4,277.3	2.652	0.0308	0.4394	432,496	5,972.2	2,661.9
	ComSRE	**31,909**	4,145.4	1.704	0.0173	0.2064	424,415	5,361.2	2,248.8
	UrbanMFM	78,734	6,423.5	1.494	0.0185	0.1444	623,876	5,404.3	3,428.7
	ScaleCity (Ours)	33,452	**2,749.9**	**0.7114**	**0.0106**	**0.1057**	**280,985**	**3,949.0**	**1,580.8**
Shenzhen	ViT	76,104	7,217.5	0.9848	0.0273	0.2603	1,402,045	17,343	759.1
	GeoVectors	83,581	7,726.4	1.166	0.0327	0.2787	1,463,013	17,683	712.9
	SpaBERT	87,169	8,319.5	1.110	0.0309	0.2764	1,437,548	19,389	923.5
	HAFusion	69,282	7,821.7	0.9873	0.0353	0.2711	1,387,557	17,950	903.8
	Urban2Vec	77,527	7,227.3	1.128	0.0288	0.2571	1,593,150	17,348	937.6
	HREP	84,562	7,563.0	0.9271	0.0401	0.3235	1,427,627	21,315	1,021.5
	CityFM	80,219	7,503.0	1.156	0.0357	0.2871	1,403,300	19,536	888.4
	ComSRE	54,600	6,152.4	0.8962	0.0252	0.2213	1,136,737	16,505	612.5
	UrbanMFM	104,833	8,670.5	0.7515	0.0300	0.2597	1,767,254	15,411	1,077.1
	ScaleCity (Ours)	**36,328**	**3,291.9**	**0.4834**	**0.0186**	**0.1376**	**791,001**	**12,688**	**509.8**

D.1 Aggregation Size and Predictive Performance

Across the nine jointly trained configurations, larger estimated aggregation sizes generally correspond to higher R^2 (Fig. 6; Table 11). Among road-network partitions, mean R^2 rises from 0.555 at L5 to 0.723 at L2. Performance peaks at intermediate granularities: the 3 km grid among grids and L2 among road-network partitions. Fine-grained prediction is therefore particularly challenging, while the best-performing granularity depends on the partition family.

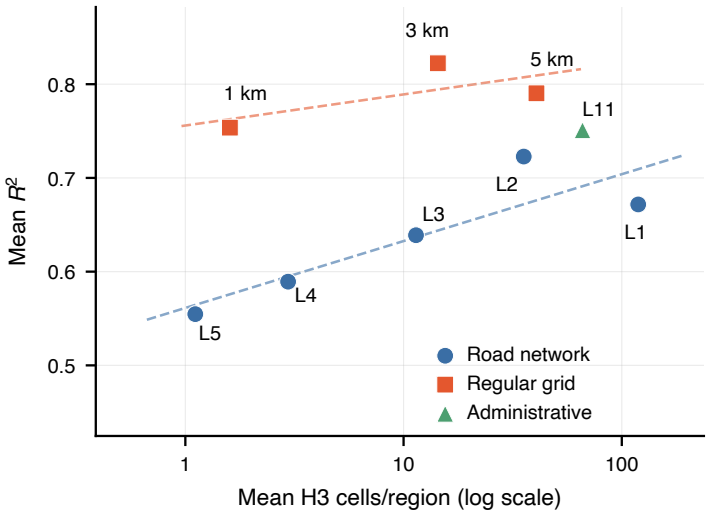

Figure 6: Performance versus aggregation size across nine jointly trained configurations. R^2 is averaged over five folds, eight tasks, and then three cities. The logarithmic horizontal axis shows the city-averaged ratio of total H3 resolution-8 cells to configuration-specific regions. Dashed lines show separate log-linear fits for grid and road-network configurations.

E Additional Ablation Analyses (Q2)

E.1 Ablation Variants

The variants in Table 2 use the primary benchmark's optimization settings. Input and graph ablations repeat Stage 1 pretraining; downstream ablations reuse the full model's atomic-cell embeddings.

- **w/o POI**, **w/o SVI**, and **w/o AE** remove the corresponding modality. Stage 1 is retrained on the remaining two modalities, followed by downstream training.
- **w/o GAT** removes all neighborhood message passing. Atomic representations are relearned without GAT, followed by downstream training.
- **w/o Dynamic Routing ($\mathbf{W}_m$, g_m)** replaces prompt-generated modality projections with three trainable, modality-specific projections shared across all tasks and configurations, and fixes the gates to $g_m = 1/3$.
- **w/o Dyn. Pred. Head** replaces the generated final readout with one trainable static head shared across all tasks and configurations, retaining the shared map that produces H_{r_j} (Eq. 9).
- **w/o Adaptive Residual Stream** and **w/o Shared Base Stream** remove $\Delta E_i(p)$ and $E_{\text{base},i}$, respectively, retaining the other stream in Eq. 6.
- **w/o Separate Prompt Encoding** encodes the concatenated scale and task descriptions as one sequence instead of encoding them separately (Eq. 3). Both descriptions are retained.
- **w/o Region-Extent Covariate** removes A_{r_j} from the predictor input while preserving overlap-weighted mean aggregation and its normalization (Eq. 8).

E.2 Additional Component Comparisons

Separate prompt encoding improves the three-city mean Overall R^2 by 0.008 over concatenated-text encoding, showing a modest benefit from representing the two descriptions separately. The dynamic-routing ablation instead compares prompt-generated projections and gates jointly against shared projections with uniform weighting, assessing conditional fusion as a whole.

The relative effects vary across cities. In Beijing, the full model's largest advantage is over the variant without the dynamic prediction head (0.152), slightly exceeding its advantage over extent removal (0.149); in Chengdu and Shenzhen, its largest advantage is over extent removal. In Chengdu, the full model has a larger advantage over the no-GAT variant than over the no-POI variant (0.041 versus 0.024), with the reverse ordering in Beijing and Shenzhen. The three-city averages therefore summarize contributions whose relative importance depends on the city.

E.3 Effect of Region Extent on Extensive and Intensive Targets

Extensive targets measure regional totals, whereas intensive targets describe levels such as concentrations and mean prices. Retaining A_{r_j} rather than omitting it raises group-mean R^2 on irregular partitions by 0.403–0.503 for extensive targets, compared with 0.006–0.046 for intensive targets (Table 14). Extensive-target gains are smaller on grids (0.032–0.068). Explicit extent information is thus especially valuable for predicting totals over regions whose sizes vary within a configuration.

Table 14: Effect of removing the region-extent covariate: $\Delta R^2 = R^2_{\text{full}} - R^2_{\text{ablated}}$; negative values favor the ablated model. Entries average six irregular configurations (L1–L5 and L11; Irreg.) or three grids (1/3/5 km; Grid). Bold rows give equal-weight means across the four tasks of each type.

		Beijing		Chengdu		Shenzhen	
Target type	**Task**	**Irreg.**	**Grid**	**Irreg.**	**Grid**	**Irreg.**	**Grid**
Extensive	Pop.	+0.570	+0.043	+0.473	−0.017	+0.584	+0.033
	GDP	+0.355	+0.053	+0.412	+0.069	+0.570	+0.100
	HFP	+0.504	+0.050	+0.520	+0.014	+0.603	+0.048
	Bike	+0.182	+0.099	+0.300	+0.061	+0.256	+0.091
	Group mean	**+0.403**	**+0.061**	**+0.426**	**+0.032**	**+0.503**	**+0.068**
Intensive	$PM_{2.5}$	+0.009	0.000	+0.024	+0.042	+0.122	+0.300
	CO	+0.008	+0.024	+0.026	+0.064	+0.050	+0.201
	CO_2	+0.019	+0.016	+0.073	+0.022	+0.069	+0.080
	HP	−0.011	+0.032	+0.045	+0.044	−0.057	+0.092
	Group mean	**+0.006**	**+0.018**	**+0.042**	**+0.043**	**+0.046**	**+0.168**

F Unseen-Configuration Prediction and Gate Analysis Protocols

F.1 Evaluation on Unseen Configurations (Q3)

For each city and source set in Table 3, we train on HP labels from the source configurations using the primary benchmark's L11-anchored, shuffled five-fold procedure. Scheduling, early stopping, and checkpoint selection use only source-configuration validation R^2. Target normalization and inverse transformation use source-configuration statistics only. Each fold's model evaluates the target configurations over the same city geography without applying the source-fold assignments to target scoring or updating parameters. Table 3 reports the mean target R^2 across the five trained folds.

F.2 Gate Recording and Aggregation (Q4)

Figure 4 summarizes how residual-stream modality gates vary during training and across tasks and configurations. We record all 72 task–configuration gate vectors per epoch in each city and fold. Task and configuration summaries average the final three recorded epochs before training stops; predictive scores use the validation-selected checkpoint.

Training dynamics. Each trajectory is interpolated to 21 normalized progress points. At each point, curves and shaded bands report modality-wise means and ± 1 standard deviation over tasks, configurations, folds, and cities.

Task-wise deviations. We subtract each city's modality mean before averaging:

$$\Delta g_{t,m} = \mathbb{E}_{c,s,v}\left[g_m^{(c,t,s,v)} - \bar{g}_{c,m}\right], \qquad \bar{g}_{c,m} = \mathbb{E}_{t,s,v}\left[g_m^{(c,t,s,v)}\right]. \tag{20}$$

Here c, t, s, and v index cities, tasks, configurations, and folds; $g_m^{(c,t,s,v)}$ is the final-three-epoch mean. The heatmap displays deviations in units of 10^{-3}, highlighting task preferences relative to each city's overall allocation.

Configuration-level associations. Gates average over tasks, folds, and cities. Estimated aggregation size is each city's total H3-cell count divided by its configuration-specific region count, averaged across cities. Spearman correlations use these nine configuration points with two-sided approximate p-values: $\rho = 0.917$ ($p < 0.001$) for POI and $\rho = -0.667$ ($p \approx 0.050$) for both SVI and AE. Dashed lines show log-linear fits.

G Prompt Design Sensitivity

We examine whether descriptive prompts improve joint prediction and few-shot task adaptation beyond distinct condition identifiers.

G.1 Prompt Variants and Evaluation Protocol

We compare six designs (Table 15). A uses detailed descriptions; B paraphrases A with the same information; C gives concise definitions; and D uses short semantic labels. E replaces descriptions with distinct symbolic identifiers, while F uses one fixed scale string and one fixed task string for every input. A versus B tests wording, A–D vary descriptive detail, and E versus F tests whether condition identities remain distinguishable.

All schemes use the same frozen BGE-base-en-v1.5 encoder, cached atomic embeddings, data splits, and training settings. Each scheme retrains Stages 2–4 and is evaluated with the primary benchmark's five-fold protocol. For few-shot adaptation to NO_2 and PM_{10}, each scheme starts from its own jointly trained eight-task model and retains its prompt design, following the ten support-fold rotations in Appendix H. These comparisons use one run per scheme, without multi-seed repetitions.

Table 15: Representative inputs for A–F. In F, every spatial configuration uses `H3 grid` and every task uses `Population`.

Scheme	Scale: 1 km	Scale: L11	Task: $PM_{2.5}$
A	`Scale: A 1 km regular grid, a fine spatial unit for analyzing neighborhood-level spatial heterogeneity.`	`Scale: Subdistrict-level administrative areas, relatively large grassroots administrative units.`	`Task: Predict the spatial distribution of fine particulate matter (PM2.5) concentrations, reflecting air pollution associated with traffic emissions, industrial activity, and weather.`
B	`Spatial granularity: A square grid with 1 km sides, resolving spatial variation at the neighborhood level.`	`Spatial granularity: Relatively large local administrative areas at the subdistrict level.`	`Prediction target: The spatial distribution of fine particulate matter (PM2.5) concentrations, an air-pollution indicator associated with vehicle exhaust, industrial emissions, and meteorological conditions.`
C	`A 1 km regular grid cell.`	`Subdistrict-level administrative boundaries.`	`Predict fine particulate matter (PM2.5) concentration.`
D	`1 km grid`	`Subdistrict L11`	`PM2.5`
E	`A`	`I`	`2`
F	`H3 grid`	`H3 grid`	`Population`

G.2 Joint Prediction and Few-Shot Adaptation

Distinct condition inputs support joint prediction. Across A–E, per-city Overall R^2 varies by at most 0.015 (Fig. 7(a)). Detailed descriptions (A) and symbolic identifiers (E) achieve three-city means of 0.699 and 0.695, respectively, whereas constant prompts (F) yield approximately -0.030 to 0.100 across cities. Distinct condition identities thus support accurate prediction on the jointly trained task–configuration pairs, with only small differences among the descriptive and symbolic designs.

Descriptive prompts show a larger average advantage in few-shot adaptation. A outperforms E for both NO_2 and PM_{10} after averaging the three cities (Fig. 7(b)). Across cities and targets, A achieves the highest mean R^2 among the six designs (0.747), exceeding E (0.652) by 0.095. The strongest design can still vary by target: D slightly exceeds A in the three-city mean for PM_{10}. These comparisons show a larger average benefit from descriptive prompts in few-shot task adaptation than in joint prediction on the base tasks.

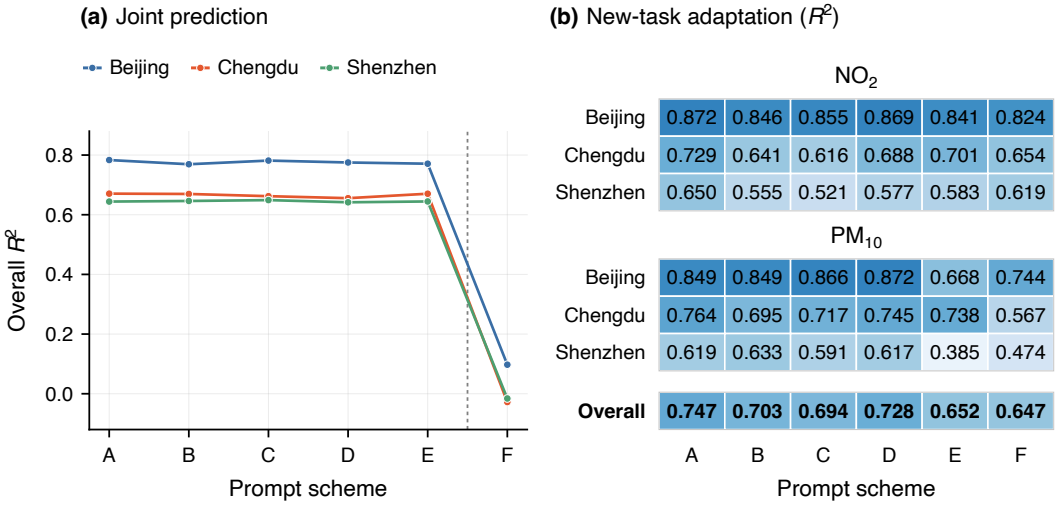

Figure 7: Prompt-design sensitivity. (a) Per-city Overall R^2 over eight base tasks and nine configurations after five-fold averaging. (b) Few-shot adaptation to NO_2 and PM_{10} from each scheme's own jointly trained model. Scores average ten support-fold rotations and nine configurations. The Overall row averages both targets and all three cities. Darker cells indicate higher R^2 on a common color scale.

G.3 Default Prompts and Symbolic Identifiers

Tables 16 and 17 list the Scheme-A descriptions and corresponding Scheme-E identifiers. In table order, Scheme E assigns letters A–I to the nine spatial configurations and numbers 1–8 to the eight base tasks, with 9 and 10 assigned to NO_2 and PM_{10}, respectively. Each identifier remains fixed during training and adaptation.

Table 16: Scheme-A descriptions and Scheme-E identifiers for the nine spatial configurations. Area ranges are approximate descriptors in the prompts.

Key	Code	Scheme-A description
1km	A	Scale: A 1 km regular grid, a fine spatial unit for analyzing neighborhood-level spatial heterogeneity.
3km	B	Scale: A 3 km regular grid, a medium-sized spatial unit for analyzing community-level spatial patterns.
5km	C	Scale: A 5 km regular grid, a larger spatial unit for analyzing broad patterns of urban functional zones.
L1	D	Scale: Large urban functional zones partitioned by an A-level road network, about 10--50 square kilometers, enclosed by expressways and arterial roads.
L2	E	Scale: Medium-sized urban areas partitioned by a B-level road network, about 3--10 square kilometers, enclosed by arterial and secondary roads.
L3	F	Scale: Urban clusters partitioned by a C-level road network, about 1--3 square kilometers, enclosed by secondary and branch roads.
L4	G	Scale: Urban blocks partitioned by a D-level road network, about 0.3--1 square kilometers, enclosed by branch and local roads.
L5	H	Scale: Small parcels partitioned by an E-level road network, about 0.05--0.3 square kilometers, the smallest road-network units enclosed by local roads and minor paths.
L11	I	Scale: Subdistrict-level administrative areas, relatively large grassroots administrative units.

Table 17: Scheme-A descriptions and Scheme-E identifiers for the eight base tasks and two adaptation targets, NO2 and PM10. HFP denotes Tencent human footprints.

Task	Code	Scheme-A description
population	1	Task: Predict regional resident population counts, reflecting the spatial distribution of urban residents and associated with residential land and public services.
PM2.5	2	Task: Predict the spatial distribution of fine particulate matter (PM2.5) concentrations, reflecting air pollution associated with traffic emissions, industrial activity, and weather.
CO	3	Task: Predict the spatial concentration of carbon monoxide (CO), mainly arising from incomplete combustion and associated with vehicle density and road traffic.
CO2	4	Task: Predict regional CO2 levels associated with energy use, transport, and building density.
GDP	5	Task: Predict regional gross domestic product (GDP), reflecting the spatial concentration of economic activity and associated with commercial land, employment density, and infrastructure.
HFP	6	Task: Predict regional totals of human-footprint values derived from Tencent mobile-device records, reflecting aggregate human activity.
HP	7	Task: Predict regional mean housing prices per square meter, reflecting spatial variation in land value and residential attractiveness, associated with transport access, public services, and environmental quality.
bike	8	Task: Predict regional bike-sharing ride volume, reflecting short-distance travel demand and cycling activity, associated with metro stations, commercial areas, and residential areas.
NO2	9	Task: Predict the spatial concentration of nitrogen dioxide (NO2), mainly arising from vehicle exhaust and combustion and indicating urban traffic pollution.
PM10	10	Task: Predict the spatial concentration of inhalable particles (PM10), associated with dust, construction, and industrial emissions.

H Few-Shot Adaptation to Unseen Tasks

Few-shot adaptation reuses the jointly trained model for NO_2 and PM_{10}, which are excluded from the eight-task benchmark, without adding task-specific modules.

Protocol. Within each city, we rotate ten L11 folds, using one fold (10% of L11 anchor units) as the support set for parameter updates and the other nine for query evaluation. Regions at all nine configurations inherit their parent L11 anchor unit's assignment; coarse regions can cross support–query boundaries. Each rotation starts from the same city-specific jointly trained checkpoint. All downstream parameters are updated, while the atomic embeddings and BGE encoder remain frozen. Target-normalization statistics are fitted on the support set only.

Results. Adaptation achieves an overall mean R^2 of 0.747 across the two unseen targets, three cities, and nine configurations (Fig. 8(a)). The existing downstream architecture thus accommodates both additional targets. For both targets, Beijing has higher mean scores and smaller variation across configurations than Chengdu and Shenzhen (Fig. 8(b)).

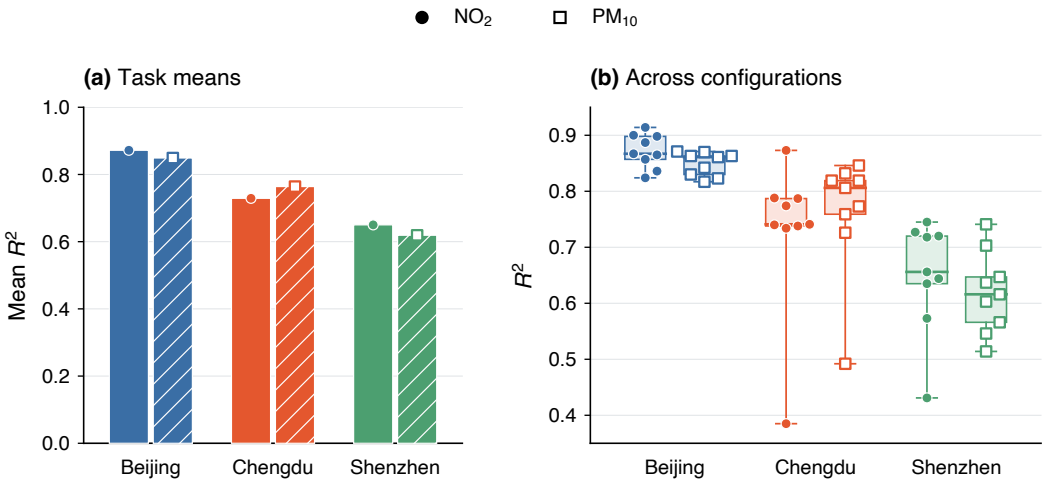

Figure 8: Few-shot adaptation to unseen NO_2 and PM_{10} targets. Scores average ten support-fold rotations, each using 10% of L11 units for support under the protocol above. (a) Mean R^2 over nine configurations. (b) Configuration scores per city–target pair ($n = 9$); boxes show medians and interquartile ranges, with min–max whiskers. Colors identify cities; markers distinguish targets.

I City-Specific Evaluation in New York City and Singapore

We evaluate joint prediction of Population, $PM_{2.5}$, CO_2, and GDP in New York City and Singapore. ScaleCity is trained from scratch separately in each city using the primary benchmark's architecture and training hyperparameters.

Protocol. L11 anchor units are police precincts in New York City and Neighbourhood Police Centre areas in Singapore. Five-fold evaluation includes anchors overlapping available H3 features: all 78 New York City units and 42 of 140 Singapore units. L1 and the 5 km grid are excluded; the seven evaluated configurations are L2–L5, L11, and the 1 and 3 km grids. Overall R^2 averages five folds, four tasks, and seven configurations.

Results. Overall R^2 is 0.629 in New York City and 0.500 in Singapore (Fig. 9), with $PM_{2.5}$ the highest-scoring task in both cities. The best-performing configuration is the 3 km grid in New York City (0.743) and L11 in Singapore (0.653); L5 has the lowest score in both cities (0.385 and 0.320). The city-specific models retain the fine-grained prediction challenge observed in the primary benchmark, while their best-performing partitions differ.

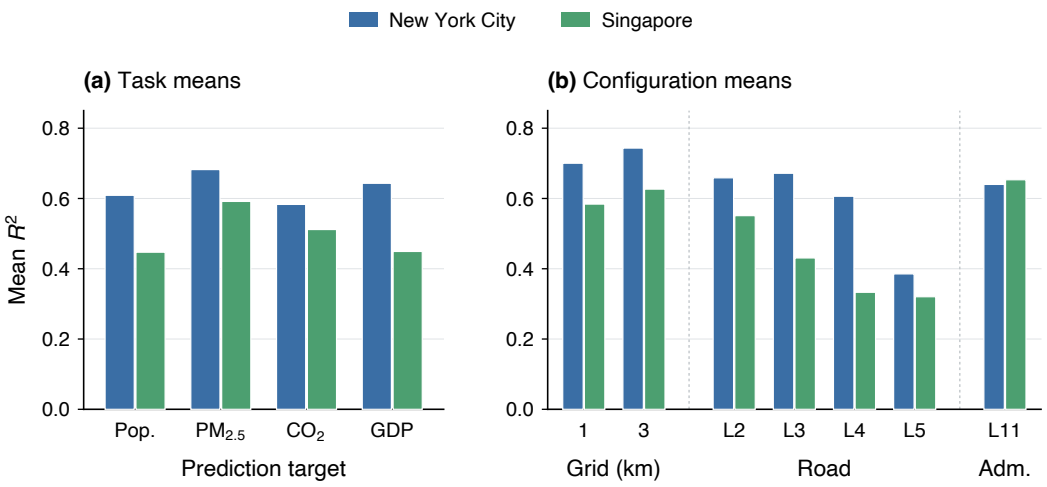

Figure 9: City-specific evaluation in New York City and Singapore, with training from scratch in each city. (a) Mean R^2 over seven configurations for each task. (b) Mean R^2 over four tasks for each configuration. Scores average five folds.

J Hyperparameter Sensitivity

Table 18 compares alternatives for atomic resolution, aligned embedding dimension d, prompt encoder, and contrastive-loss weights. Stage 1 alternatives require new atomic embeddings and downstream training; prompt-encoder variants reuse the cached embeddings and retrain Stages 2–4. Scores follow the primary benchmark's five-fold evaluation and Overall R^2 aggregation.

Table 18: Hyperparameter sensitivity, measured by Overall R^2 ↑ over eight tasks and nine configurations; Mean averages the three cities. Best and second-best scores, ranked before rounding, are **bold** and underlined. Shading identifies the default configuration.

Parameter axis	Configuration	Beijing	Chengdu	Shenzhen	Mean
H3 Resolution	Res = 7 (coarse)	0.682	0.609	0.586	0.626
	Res = 9 (fine)	0.727	0.635	0.627	0.663
	Res = 8 (default)	**0.783**	**0.671**	**0.644**	**0.699**
Embedding	256	0.756	0.656	0.609	0.674
	512 (default)	**0.783**	**0.671**	**0.644**	**0.699**
Prompt Text Encoder	MacBERT	0.781	0.667	0.636	0.695
	RoBERTa-wwm	0.779	0.667	0.642	0.696
	M3E	0.780	0.667	0.642	0.696
	BGE-en (default)	**0.783**	**0.671**	**0.644**	**0.699**
Loss Weights ($\lambda_1 : \lambda_2 : \lambda_3$)	2 : 1 : 1 (POI–SVI↑)	0.768	0.657	0.638	0.687
	1 : 2 : 1 (POI–AE↑)	0.765	0.661	0.637	0.688
	1 : 1 : 2 (SVI–AE↑)	0.767	0.668	0.611	0.682
	1 : 1 : 1 (default)	**0.783**	**0.671**	**0.644**	**0.699**

Atomic resolution. H3 resolution 8 achieves the highest Overall R^2 in all three cities. Its three-city mean of 0.699 exceeds 0.626 at resolution 7 and 0.663 at resolution 9. Thus, the intermediate atomic resolution performs best among the tested choices.

Embedding dimension. Increasing d from 256 to 512 improves accuracy in every city, raising mean Overall R^2 from 0.674 to 0.699 and supporting the default $d = 512$.

Prompt encoder. The four tested encoders yield mean Overall R^2 values from 0.695 to 0.699. BGE-base-en-v1.5 (BGE-en in the table) achieves the highest mean score; the small range indicates limited sensitivity to encoder choice in this comparison.

Alignment weights. Equal weights for the three modality-pair losses yield the highest score in every city. Equal weighting improves mean Overall R^2 by 0.012–0.017 over doubling one pair's weight, with the largest advantage over the SVI–AE-weighted variant. We retain equal alignment weights during pretraining and learn task- and configuration-dependent modality allocation downstream.

K NOTATION SUMMARY

Table 19 summarizes the core notation. Symbols used only in local implementation formulas are defined where they appear.

Table 19: Core notation, grouped by stage. Dimensions follow the column-vector convention; numerical widths refer to the default configuration.

Symbol	Description
Stage 1: atomic-cell pretraining	
$\mathcal{H} = \{h_1, \ldots, h_N\}$	Atomic H3 cells in a city; N is the cell count.
m, n	Modality indices in $\{\mathrm{SVI}, \mathrm{POI}, \mathrm{AE}\}$.
$z_{m,i}^{(\ell)},\ P_m$	Modality-m cell feature at graph layer ℓ and its alignment MLP.
$e_{m,i} \in \mathbb{R}^d$	ℓ_2-normalized aligned embedding of cell h_i for modality m.
$\mathcal{B}_{\mathrm{pre}},\ \tau$	Synchronized batch of cell indices for one contrastive forward pass, and contrastive temperature.
$\lambda_{m,n}$	Modality-pair loss weight; $(\lambda_1, \lambda_2, \lambda_3)$ denote SVI–POI, POI–AE, and SVI–AE weights.
$\mathcal{L}_{m,n},\ \mathcal{L}_{\mathrm{pre}}$	Symmetric contrastive loss for a modality pair, and total pretraining loss.
Stage 2: prompt-conditioned parameter generation	
$P_s,\ P_t$	Spatial-configuration and task prompt texts.
$p_s,\ p_t \in \mathbb{R}^{768}$	Frozen BGE text-encoder features, using the final-layer `[CLS]` token without ℓ_2 normalization.
$f_s,\ f_t$	Separate projection heads for configuration and task prompt features.
$p = p_{j,k} \in \mathbb{R}^{768}$	Fused condition for query (r_j, k); shared by all regions in the same task–configuration pair.
$\mathbf{W}_m(p),\ \mathbf{b}_m(p)$	Generated modality projection in $\mathbb{R}^{d_{\mathrm{out}} \times d}$ and bias in $\mathbb{R}^{d_{\mathrm{out}}}$.
$\mathbf{A}_m(p),\ \mathbf{B}_m(p),\ r$	Factors of shapes $d_{\mathrm{out}} \times r$ and $r \times d$, with $\mathbf{W}_m = \mathbf{A}_m\mathbf{B}_m$ and rank at most r.
$\mathbf{g}(p)$	Nonnegative modality gates satisfying $\sum_m g_m(p) = 1$.
Stage 3: adaptive multimodal fusion	
$\phi_{\mathrm{base},m},\ \mathbf{W}_{\mathrm{fuse}},\ \mathbf{b}_{\mathrm{fuse}}$	Trainable base-stream affine modality projections and fusion parameters, shared across conditions.
$E_{\mathrm{base},i},\ \Delta E_i(p)$	Shared base and prompt-conditioned residual features, both in $\mathbb{R}^{d_{\mathrm{out}}}$.
$E_i(p) \in \mathbb{R}^{d_{\mathrm{out}}}$	Fused cell feature after adding the two streams and applying LayerNorm.
Stage 4: region aggregation and prediction	
$(r_j, k),\ \mathcal{I}_j$	Region–task query and the indices of atomic cells intersecting region r_j.
$w_{i,j}$	Fractional cell overlap, $\mathrm{Area}(h_i \cap r_j)/\,\mathrm{Area}(h_i)$.
$A_{r_j} = \sum_{i \in \mathcal{I}_j} w_{i,j}$	Effective atomic-cell count; the dimensionless region-extent covariate appended to the pooled feature.
$\hat{E}_{r_j},\ E_{r_j}$	Overlap-weighted mean in $\mathbb{R}^{d_{\mathrm{out}}}$ and augmented region feature $[\hat{E}_{r_j} \,\|\, A_{r_j}]$ in $\mathbb{R}^{d_{\mathrm{out}}+1}$.
$\mathbf{W}_{\mathrm{shared}},\ \mathbf{b}_{\mathrm{shared}}$	Shared affine map from the augmented region feature to a 64-dimensional hidden feature.
$H_{r_j} \in \mathbb{R}^{64}$	Shared affine-map output after ReLU and training-time dropout.
$\mathbf{W}_{\mathrm{pred}}(p),\ \mathbf{b}_{\mathrm{pred}}(p)$	Generated scalar-readout parameters: a 1×64 row vector and a scalar bias.
$y_j^{(k)},\ \hat{y}_j^{(k)}$	Target and prediction for region r_j and task k; standardized in the primary-benchmark training objective.
$\mathcal{B},\ \mathcal{L}_{\mathrm{down}}$	Batch of region–task queries and the downstream mean squared error over that batch.