UrbanExpert: Task-Conditioned Multi-Modal Fusion via Semantic Expert Routing for Urban Socioeconomic Prediction

Zechen Li
School of Software,
Shandong University
Jinan, China
lizc@mail.sdu.edu.cn

Hongwei Jia
Graduate School of Comprehensive
Human Sciences,
University of Tsukuba
Tsukuba, Japan
jia.hongwei.tkb_ct@u.tsukuba.ac.jp

Weiming Huang
School of Geography,
University of Leeds
Leeds, UK
W.Huang@leeds.ac.uk

Kai Zhao
AI Research,
Workday AI
Pleasanton, USA
kaizhaofrank@gmail.com

Meng Chen*
School of Software,
Shandong University
Jinan, China
mchen@sdu.edu.cn

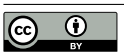
Abstract

Predicting urban indicators from multi-modal sensing data requires fusing heterogeneous modalities, yet different prediction tasks rely on different feature combinations while semantically related tasks share common cues. Existing approaches either train task-specific models that cannot share knowledge across tasks, or learn a single task-agnostic representation that ignores inter-task semantic correlations. Both paradigms also assume complete modality observations, which rarely holds in practice. We propose UrbanExpert, a multi-task framework that explicitly conditions multi-modal fusion on task semantics while accommodating incomplete modality observations. UrbanExpert first encodes each available modality into a shared token sequence via pretrained encoders, and recovers missing modality representations through a cross-modal reconstruction module to handle incomplete observations. A task-conditioned Mixture-of-Experts module then routes each task to a relevant subset of experts based on the semantic affinity between the task description and learnable expert prototypes, producing task-conditioned region embeddings. Task-specific prediction heads map the embeddings to target indicators, and the model is jointly optimized with prediction, reconstruction, and routing regularization objectives. Experiments on two real-world urban datasets show that UrbanExpert consistently outperforms both single-task and multi-task baselines, and that the learned routing structure generalizes effectively to unseen tasks and cities.

CCS Concepts

• **Information systems** → **Spatial-temporal systems**; **Data mining**.

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
KDD 2026, Jeju Island, Republic of Korea.
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818168

Keywords

Urban Region Embedding; Urban Socioeconomic Prediction; Missing Modality; Multi-Modal Fusion; Mixture-of-Experts

ACM Reference Format:
Zechen Li, Hongwei Jia, Weiming Huang, Kai Zhao, and Meng Chen. 2026. 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 (KDD 2026), August 9–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818168

1 Introduction

Multi-modal urban sensing data, including satellite imagery, street-view photographs, points-of-interest (POIs), and map tiles, provides rich signals for predicting diverse urban indicators such as socioeconomic conditions, environmental quality, and transportation dynamics [4, 10, 13, 35]. A common pipeline for such predictions first encodes each modality with a dedicated encoder, fuses the multi-modal features into a region representation, and maps the result to the target indicator via a prediction head. Within this pipeline, the fusion stage is critical: different prediction tasks naturally rely on varying combinations of modalities and feature patterns, while semantically related tasks such as income and education often share common cues. This raises a central question: *how should we design a fusion mechanism that is flexible enough to serve diverse yet overlapping prediction tasks?*

Existing methods have explored two paradigms to address this challenge. Task-specific modeling trains a dedicated model per prediction target, optimizing the fusion strategy directly for each task [12, 41] (Figure 1 (a)). Self-supervised pretraining learns a task-agnostic region encoder through pretext objectives such as contrastive alignment [14, 15] or spatial adjacency constraints [13, 34], yielding general-purpose embeddings shared across all downstream tasks (Figure 1 (b)). Although effective in many settings, both paradigms face two significant limitations.

Limitation 1: Task-semantic correlations are not exploited. Urban prediction tasks are semantically interdependent: income, education, commercial activity, and the built environment co-evolve within interconnected urban systems, so semantically related tasks

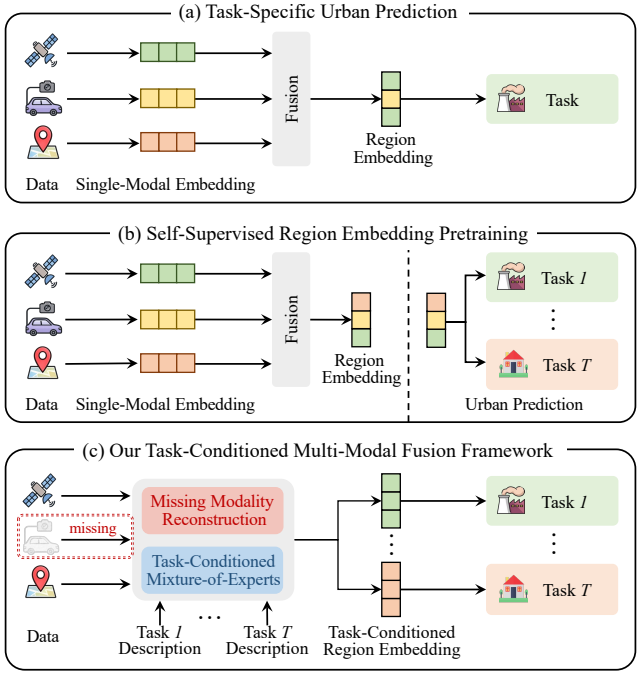


Figure 1: Comparison of various multi-modal urban prediction paradigms.

impose similar demands on the fused representation. Task-specific models learn each target in isolation and cannot share discovered patterns across related tasks. Self-supervised approaches project all tasks onto a single shared representation without distinguishing genuine semantic correlations from incidental ones. Neither paradigm can exploit the proximity between a new target and an existing task: the former requires full retraining, while the latter offers no task-aware adaptation.

Limitation 2: Multi-modal observations are often incomplete. Data acquisition constraints introduce systematic gaps: remote areas frequently lack street-view imagery, functionally homogeneous districts yield sparse POI records, and crowdsourced map data remains incomplete in many cities. Most existing methods assume complete multi-modal observations [14, 27, 42], and their fixed fusion pipelines degrade considerably when inputs are partially available. A practical framework must therefore accommodate arbitrary subsets of observed modalities without retraining.

Motivated by these observations, we propose **UrbanExpert**, a multi-task framework that addresses both limitations within a unified architecture, as shown in Figure 1 (c). UrbanExpert first employs *multi-modal encoding*, which transforms each available modality into a fixed-length token sequence in a shared latent space via pretrained encoders. To address Limitation 1, we introduce a *task-conditioned Mixture-of-Experts* module that encodes each task's natural-language description into a semantic query and routes it to a relevant subset of experts based on semantic affinity with learnable expert prototypes; semantically related tasks thereby activate overlapping experts and share predictive factors, while distant tasks are served separately. To address Limitation 2, a *missing modality reconstruction* module is placed before the MoE layer to recover absent modality representations from observed ones through cross-modal decoders, ensuring that the expert module always operates on a complete multi-modal sequence regardless of data availability. Finally, *task-specific prediction heads* map the resulting task-conditioned region embeddings to scalar predictions, and the model is jointly optimized via a composite objective combining prediction, reconstruction, and routing regularization losses.

Our contributions can be summarized as follows:

- We propose UrbanExpert, a multi-task urban prediction framework that explicitly conditions multi-modal fusion on task semantics and handles incomplete modality observations within a unified architecture. Unlike task-specific models or task-agnostic region embeddings, UrbanExpert adaptively selects and composes modality features according to each task's semantics, allowing the learned routing structure to generalize to unseen tasks.
- We introduce a semantics-driven MoE mechanism that unifies task-level expert selection and instance-level feature extraction through shared expert prototypes. This design demonstrates that task semantics can serve as an effective and learnable prior for guiding multi-modal fusion, offering a general principle for task-conditioned representation learning that can be extended to other multi-modal, multi-task prediction settings.
- We conduct extensive experiments on two real-world urban datasets spanning multiple prediction tasks. UrbanExpert consistently outperforms both single-task and multi-task baselines. Empirical analysis reveals that experts self-organize into interpretable modality specialists and that the learned routing structure generalizes effectively to held-out tasks and unseen cities. Our code is publicly available at https://github.com/AIMUrban/UrbanExpert.

2 Problem Formulation

Definition 1 (Multi-Modal Region Data). *Let $\mathcal{R} = \{r_1, \ldots, r_n\}$ denote a set of n urban regions. Each region r_i is associated with M sensing modalities (e.g., satellite imagery, street-view photographs, POI records, and map tiles). We write $\mathbf{X}_i^{(m)}$ for the raw observation of modality m. Because data availability varies across regions, we let $\mathcal{M}_i \subseteq \{1, \ldots, M\}$ denote the set of observed modalities for region r_i, so that its available observations are $\mathcal{X}_i = \{\mathbf{X}_i^{(m)}\}_{m \in \mathcal{M}_i}$.*

Definition 2 (Urban Prediction Tasks). *A prediction task τ_j requires estimating a real-valued urban indicator (e.g., median household income or vegetation index) for every region. Each task is paired with a natural-language description t_j that specifies the prediction objective and its semantic context. We consider a task set $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$, and denote the ground-truth label of region r_i under task τ_j as $y_i^{(j)} \in \mathbb{R}$.*

Problem 1 (Multi-Task Urban Prediction under Incomplete Observations). *Given a set of regions $\mathcal{R}$ with potentially incomplete multi-modal observations $\{\mathcal{X}_i\}_{i=1}^n$ and a set of prediction tasks $\mathcal{T}$ with associated descriptions $\{t_j\}_{j=1}^T$, the goal is to learn: (i) a fusion function $f(\cdot \mid t_j) : \mathcal{X}_i \mapsto \mathbf{h}_i^{(j)} \in \mathbb{R}^d$ that produces a task-conditioned region representation from available modalities, guided*

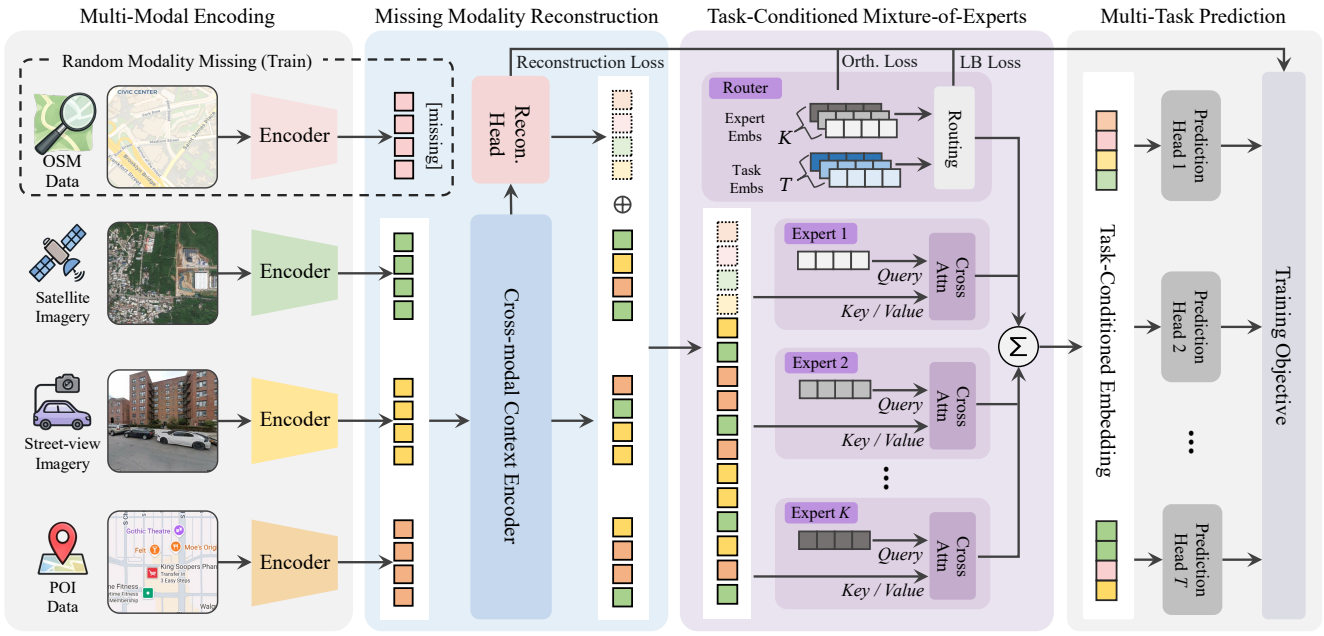


Figure 2: Overview of UrbanExpert. Multi-Modal Encoding: pretrained encoders extract fixed-length token representations from each available modality; Missing Modality Reconstruction: a reconstruction module recovers missing modality representations from observed ones; Task-Conditioned MoE: task descriptions are then matched against learnable expert prototypes to route each task to a relevant expert subset, whose outputs are aggregated into task-conditioned region embeddings; Multi-Task Prediction: task-specific heads produce scalar predictions.

by task semantics; and (ii) task-specific prediction heads $\{g_j\}_{j=1}^{T}$ *such that* $\hat{y}_i^{(j)} = g_j(\mathbf{h}_i^{(j)})$ *accurately predicts the target indicator, while exploiting semantic correlations among tasks to improve both multi-task performance and generalization to unseen tasks.*

3 Method

As illustrated in Figure 2, UrbanExpert addresses Problem 1 through four components. (1) **Multi-Modal Encoding**: pretrained modality-specific encoders transform each available modality into a fixed-length token sequence in a shared latent space; (2) **Missing Modality Reconstruction**: during training, modalities are randomly masked and cross-modal decoders reconstruct the missing representations, enabling robust inference under incomplete observations; (3) **Task-Conditioned Mixture-of-Experts**: task descriptions are encoded into semantic queries that route each task to a relevant subset of experts via learnable expert prototypes, producing task-conditioned region embeddings; (4) **Multi-Task Prediction**: task-specific heads map the embeddings to scalar predictions, and the model is optimized via a composite loss combining prediction, reconstruction, and routing regularization.

3.1 Multi-Modal Encoding

3.1.1 Pretrained Feature Extraction. For each region r_i, we consider four sensing modalities $m \in \{\text{sat}, \text{sv}, \text{poi}, \text{osm}\}$, corresponding to satellite imagery, street-view imagery, POI data, and OpenStreetMap (OSM) tiles. For each observed modality $m \in \mathcal{M}_i$, a pretrained encoder extracts a variable-length token sequence:

$$\mathbf{Z}_i^{(m)} = \text{Encoder}^{(m)}\left(\mathbf{X}_i^{(m)}\right) \in \mathbb{R}^{N_m \times d_m}, \tag{1}$$

where N_m is the number of tokens and d_m the native embedding dimension of modality m. Specifically, satellite imagery is encoded via AlphaEarth Foundations [1] at 100 m resolution, with each grid cell as a token; street-view images are encoded by CLIP ViT-L/14, using the `[CLS]` token per image; POI attributes are summarized into templated sentences and encoded by a pretrained text encoder [39]; and OSM tiles are encoded by DINOv2 [23], yielding one visual token per tile. Further details are provided in Appendix A.2.

Since d_m may differ across modalities, a modality-specific linear projection maps all tokens into a shared d_h-dimensional latent space:

$$\hat{\mathbf{Z}}_i^{(m)} = \text{Proj}^{(m)}\left(\mathbf{Z}_i^{(m)}\right) \in \mathbb{R}^{N_m \times d_h}. \tag{2}$$

3.1.2 Latent Token Aggregation. Because N_m varies across regions, we compress each modality's token sequence into a fixed-length representation via Perceiver-style cross-attention [9]. A set of P learnable latent queries $\mathbf{H}_0^{(m)} \in \mathbb{R}^{P \times d_h}$ attend to the projected tokens:

$$\mathbf{H}_i^{(m)} = \text{CrossAttn}\left(\mathbf{H}_0^{(m)}, \hat{\mathbf{Z}}_i^{(m)}\right) \in \mathbb{R}^{P \times d_h}, \tag{3}$$

yielding a fixed-size representation per observed modality. The P latent tokens from all observed modalities are then concatenated and passed to the reconstruction module.

3.2 Missing Modality Reconstruction

Real-world data acquisition constraints (e.g., limited street-view coverage in remote areas) cause region-dependent modality availability, i.e., $\mathcal{M}_i \subsetneq \{1, \ldots, M\}$ for some regions. To handle arbitrary missing patterns without retraining, we adopt a reconstruction-based strategy that infers missing modality representations from observed ones.

3.2.1 Cross-Modal Context Encoder. For region r_i, let $\mathcal{M}_i^{\text{obs}}$ and $\mathcal{M}_i^{\text{mask}} = \{1, \ldots, M\} \setminus \mathcal{M}_i^{\text{obs}}$ denote the observed and masked modality sets, respectively. We concatenate the latent tokens from all observed modalities into a single sequence $\mathbf{H}_i$ and apply a self-attention block to capture cross-modal interactions:

$$\mathbf{H}_i' = \text{SelfAttn}(\mathbf{H}_i). \tag{4}$$

3.2.2 Reconstruction Decoder. Since urban modalities provide complementary signals (e.g., POI distributions and street-view appearances jointly characterize neighborhood function and form), it is often feasible to infer missing modalities from the observed context [33]. We maintain a dedicated cross-attention decoder for each modality. For each masked modality $m \in \mathcal{M}_i^{\text{mask}}$, the decoder uses P learnable query tokens $\mathbf{R}^{(m)} \in \mathbb{R}^{P \times d_h}$ to attend to the observed context $\mathbf{H}_i'$:

$$\hat{\mathbf{H}}_i^{(m)} = \text{Decoder}^{(m)}\left(\mathbf{R}^{(m)},\ \mathbf{H}_i'\right) \in \mathbb{R}^{P \times d_h}, \tag{5}$$

where each $\text{Decoder}^{(m)}$ consists of a multi-head cross-attention layer followed by a feed-forward network. During training, each modality is independently masked with a fixed probability, exposing the model to diverse missing configurations. At inference time, the same mechanism handles naturally missing modalities.

3.2.3 Reconstruction Loss. The reconstruction module is trained by minimizing the mean squared error between reconstructed and original (pre-masking) tokens:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{M}_i^{\text{mask}}|} \sum_{m \in \mathcal{M}_i^{\text{mask}}} \left\| \hat{\mathbf{H}}_i^{(m)} - \mathbf{H}_i^{(m)} \right\|_F^2. \tag{6}$$

After reconstruction, observed and reconstructed tokens are assembled into a complete multi-modal sequence $\tilde{\mathbf{H}}_i$ by placing each modality's tokens in their canonical positions.

3.3 Task-Conditioned Mixture-of-Experts

Different prediction tasks rely on different cross-modal feature patterns, yet semantically related tasks often share partially overlapping patterns. To produce task-conditioned representations while enabling principled cross-task sharing, we employ a Mixture-of-Experts (MoE) module in which task descriptions drive both expert selection and expert behavior.

3.3.1 Task-Semantic Query. Each task τ_j is associated with a natural-language description t_j that specifies (i) the target indicator, (ii) the urban characteristics it reflects, and (iii) observable cues that may be relevant for prediction. We generate these descriptions with an LLM under a shared schema from the task name and target definition (see Appendix E). The resulting descriptions provide a task-level semantic prior for modeling relationships among prediction tasks and guiding semantics-aware expert routing.

We encode t_j with a pretrained text encoder [39] to obtain $\mathbf{t}_j \in \mathbb{R}^{d_t}$, and project it into a compact routing query:

$$\mathbf{q}_j = \text{MLP}(\mathbf{t}_j) \in \mathbb{R}^{d_a}, \tag{7}$$

where $d_a \ll d_t$ constrains routing to a task-relevant subspace.

3.3.2 Semantic Routing. Each expert $k \in \{1, \ldots, N\}$ maintains a learnable prototype $\mathbf{e}_k \in \mathbb{R}^{d_a}$ that encodes its functional specialization. Expert prototypes are initialized to be approximately orthogonal to promote differentiated specialization from the outset. The affinity between task τ_j and expert k is computed via a learned projection over their normalized element-wise interaction:

$$a_{j,k} = \text{MLP}\left(\frac{\mathbf{q}_j \odot \mathbf{e}_k}{\|\mathbf{q}_j \odot \mathbf{e}_k\|_2}\right), \tag{8}$$

where $\odot$ denotes element-wise multiplication.

Only the top-K experts with the highest scores are activated per task, encouraging semantically related tasks to be routed to overlapping expert subsets and inducing implicit parameter sharing. The routing weights are obtained by applying softmax over the selected experts:

$$w_{j,k} = \begin{cases} \text{Softmax}_{k' \in \text{Top-K}(j)}(a_{j,k}), & k \in \text{Top-K}(j), \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

3.3.3 Cross-Attention Experts. Each expert aggregates task-relevant signals from the multi-modal token sequence $\tilde{\mathbf{H}}_i$ via cross-attention pooling. Crucially, the cross-attention query of expert k is derived from its prototype through a coupling MLP, while keys and values are computed from the token sequence:

$$\tilde{\mathbf{e}}_k = \text{MLP}(\mathbf{e}_k) \in \mathbb{R}^{d_h}, \quad \mathbf{K}_{i,k} = \tilde{\mathbf{H}}_i \mathbf{W}_{K_k}, \quad \mathbf{V}_{i,k} = \tilde{\mathbf{H}}_i \mathbf{W}_{V_k}. \tag{10}$$

By sharing the prototype $\mathbf{e}_k$ between routing (Eq. 8) and attention-query construction, the model couples the criterion for selecting an expert with how that expert aggregates multi-modal tokens, imposing a shared semantic constraint on both routing and token aggregation.

The expert output is then:

$$E_k(\tilde{\mathbf{H}}_i) = \text{MLP}_k\left(\text{softmax}\left(\frac{\tilde{\mathbf{e}}_k \mathbf{K}_{i,k}^\top}{\sqrt{d_h}}\right)\mathbf{V}_{i,k}\right) \in \mathbb{R}^d. \tag{11}$$

This design yields a two-level adaptation: semantic routing performs *task-level* expert selection that is shared across all regions, while cross-attention performs *instance-level* token aggregation tailored to each region's multi-modal composition.

3.3.4 Task-Conditioned Region Embedding. For task τ_j and region r_i, the expert outputs are aggregated via weighted summation to obtain the task-conditioned region embedding:

$$\mathbf{h}_i^{(j)} = \sum_{k=1}^{N} w_{j,k} \cdot E_k(\tilde{\mathbf{H}}_i). \tag{12}$$

3.3.5 Expert Regularization. We introduce two regularization terms. First, to encourage expert diversity, we enforce approximate mutual orthogonality among the ℓ_2-normalized expert queries $\bar{\mathbf{e}}_k = \tilde{\mathbf{e}}_k / \|\tilde{\mathbf{e}}_k\|$:

$$\mathcal{L}_{\text{ortho}} = \sum_{k \neq k'} \left(\bar{\mathbf{e}}_k^\top \bar{\mathbf{e}}_{k'}\right)^2. \tag{13}$$

Second, to mitigate expert collapse, we employ a standard load-balancing objective [6]:

$$\mathcal{L}_{\text{lb}} = N \sum_{k=1}^{N} f_k \cdot p_k, \tag{14}$$

where f_k is the fraction of tasks routed to expert k and p_k is its average soft routing probability.

3.4 Multi-Task Prediction

Given the task-conditioned embedding $\mathbf{h}_i^{(j)}$, a task-specific prediction head g_j produces a scalar prediction:

$$\hat{y}_i^{(j)} = g_j(\mathbf{h}_i^{(j)}). \tag{15}$$

For each task τ_j, the per-task loss is the mean squared error over all regions with available labels:

$$\ell_j = \frac{1}{|\mathcal{R}_j|} \sum_{r_i \in \mathcal{R}_j} \left(\hat{y}_i^{(j)} - y_i^{(j)}\right)^2, \tag{16}$$

where $\mathcal{R}_j \subseteq \mathcal{R}$ denotes the set of regions with ground-truth labels for task τ_j. Since tasks may differ substantially in label scale and intrinsic difficulty, we adopt uncertainty-based weighting [11] to balance per-task contributions automatically:

$$\mathcal{L}_{\text{pred}} = \sum_{j=1}^{|\mathcal{T}|} \left(\frac{1}{2\sigma_j^2} \ell_j + \log \sigma_j \right), \tag{17}$$

where σ_j is a learnable scalar for each task, jointly optimized with all model parameters. Intuitively, σ_j captures the task's aleatoric uncertainty: a larger σ_j down-weights the contribution of a noisier or harder task, while the $\log \sigma_j$ term prevents the trivial solution of $\sigma_j \to \infty$.

The complete objective combines prediction, reconstruction, and regularization:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{r}} \mathcal{L}_{\text{recon}} + \lambda_{\text{o}} \mathcal{L}_{\text{ortho}} + \lambda_{\text{lb}} \mathcal{L}_{\text{lb}}, \tag{18}$$

where λ_{r}, λ_{o}, and λ_{lb} are hyperparameters controlling the relative contribution of each auxiliary loss.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets and Implementation Details. We evaluate UrbanExpert on two real-world urban datasets from New York City (NYC) and London. Each dataset provides four sensing modalities: satellite embeddings [1], street-view imagery from Google Street View, points of interest (POI) from OpenStreetMap, and cartographic map tiles rendered from OpenStreetMap. Each dataset contains around 15 region-level prediction tasks spanning socioeconomic, health, environmental, and transportation indicators. Additional dataset statistics and implementation details are provided in Appendix A.1 and B, respectively.

4.1.2 Evaluation Protocol. We train UrbanExpert jointly on five supervised tasks covering distinct urban domains: education (bachelor's degree rate), health (obesity rate for NYC and life expectancy for London), environment (NDVI), population (population density), and transport (public transport ratio). Performance is evaluated with 5-fold cross-validation using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2), and we report the mean score across folds. Generalization to additional indicators is assessed on the remaining held-out tasks (Section 4.5).

To evaluate robustness under realistic deployment conditions, we report results under two test conditions:

- **Complete-modality setting**: all four modalities are available for every test region, providing a clean benchmark that isolates architectural contributions from the masking protocol.
- **Incomplete-modality setting**: for each region and modality, the input is independently dropped with probability p=0.1 at test time, simulating realistic data availability with variable modality combinations.

4.1.3 Baselines. We compare UrbanExpert with five representative region embedding methods: **HREP** [43], **ReCP** [14], **HAFusion** [27], **FlexiReg** [26] and **GURPP** [10]. All baselines use the same pretrained features for the four modalities, extracted by the encoders described in Section 3.1.1. We exclude methods like [13, 34, 38], which are designed for fixed modality configurations and are not directly compatible with our four-modality and missing-modality setting. Detailed baseline descriptions are provided in Appendix C. Additionally, we include two UrbanExpert variants as ablations: **UrbanExpert (ST)**, a single-task variant trained independently per task, and **UrbanExpert (w/o Recon)**, which removes the cross-modal reconstruction module and fuses only observed modalities.

4.2 Performance Comparison

Table 1 presents per-task performance under the incomplete modality setting on both the NYC and London datasets. Results under the complete-modality setting are provided in Appendix D.1. We highlight three key observations from the results:

- Existing region embedding methods struggle under missing modalities. Across both cities and all tasks, these methods yield modest R^2 scores and, in several cases, negative R^2 under modality dropout. Graph-based methods that explicitly model inter-region dependencies (HREP, HAFusion and GURPP) are particularly vulnerable: once features in partially observed regions are corrupted, message passing propagates noise and leads to severe performance degradation. By comparison, ReCP and FlexiReg are somewhat more stable, with ReCP achieving the strongest baseline results on NYC, likely because its contrastive objective encourages representations that are more robust to random modality dropout. Nevertheless, their performance still plateaus around $R^2 \leq 0.5$ on most tasks, indicating that they remain limited by the absence of an explicit mechanism for handling missing inputs.
- Multi-task learning improves over single-task training. Compared with its single-task counterpart UrbanExpert (ST), the full multi-task model achieves better results on the majority of tasks across both cities. This improvement suggests that shared routing and backbone parameters allow each task to leverage complementary supervision from related tasks, yielding stronger task-specific representations.
- UrbanExpert consistently outperforms all baselines, with statistically significant improvements confirmed by paired t-tests

Table 1: Performance comparison under the incomplete modality setting (p=0.1) with five downstream tasks. Best results are in bold and second-best are underlined.

NYC	Education			Health			Environment			Population			Transport		
Method	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑
HREP	13.36	18.51	0.26	3.93	5.22	0.37	0.033	0.057	0.78	7982	13341	−0.26	12.71	18.31	−0.33
ReCP	12.35	16.33	0.43	3.68	4.77	0.47	0.034	0.054	0.81	5735	8357	0.52	10.97	14.34	0.20
HAFusion	11.77	15.40	0.49	3.72	4.78	0.47	0.051	0.075	0.63	5264	7852	0.58	10.01	12.93	0.36
FlexiReg	14.67	18.91	0.24	4.50	5.55	0.29	0.078	0.107	0.24	7388	10288	0.27	11.90	15.02	0.13
GURPP	12.17	16.35	0.43	3.93	5.27	0.34	0.060	0.083	0.54	7384	11193	0.08	10.07	13.20	0.32
UrbanExpert (ST)	10.31	13.53	0.61	3.08	4.09	0.61	0.030	0.050	0.82	5224	7545	0.61	9.44	12.22	0.41
UrbanExpert (w/o Recon)	8.81	11.86	0.69	2.79	3.75	0.67	0.033	0.042	0.88	4719	6948	**0.77**	9.10	11.38	**0.49**
UrbanExpert	**7.74**	**9.67**	**0.70**	**2.34**	**2.93**	**0.69**	**0.023**	**0.036**	**0.91**	**3870**	**5875**	0.76	**8.86**	**11.34**	**0.49**

London	Education			Health			Environment			Population			Transport		
Method	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑	MAE↓	RMSE↓	R^2↑
HREP	9.74	11.62	0.13	2.18	2.74	0.05	0.063	0.081	0.43	2549	3401	0.49	6.91	8.38	0.09
ReCP	8.46	10.53	0.28	2.16	2.75	0.04	0.057	0.090	0.27	2500	4551	0.07	6.26	7.91	0.18
HAFusion	9.68	11.72	0.12	2.21	2.76	0.04	0.075	0.096	0.19	2521	3497	0.44	6.89	8.42	0.08
FlexiReg	9.55	11.52	0.15	2.26	2.83	−0.02	0.082	0.103	0.06	2870	3961	0.32	7.23	9.03	−0.07
GURPP	7.96	9.86	0.37	2.31	2.95	−0.11	0.080	0.106	−0.01	2724	4035	0.27	6.74	8.57	0.02
UrbanExpert (ST)	5.10	6.74	0.71	1.80	2.25	0.39	0.038	0.048	0.83	2425	3031	0.80	3.57	4.62	**0.71**
UrbanExpert (w/o Recon)	5.72	7.35	0.65	1.92	2.45	0.35	0.031	0.042	0.81	**1385**	**1948**	**0.81**	4.12	5.30	0.64
UrbanExpert	**4.11**	**5.14**	**0.72**	**1.70**	**2.17**	**0.40**	**0.025**	**0.034**	**0.84**	1420	2035	**0.81**	**2.84**	**3.55**	**0.71**

(p<0.01). The full model achieves the best or second-best R^2 on all ten task–city combinations. Moreover, even without the reconstruction module (w/o Recon), UrbanExpert remains competitive with baselines, confirming that cross-task knowledge sharing in the MoE framework contributes materially beyond the reconstruction mechanism alone.

4.3 Ablation Study

To isolate the contribution of each component, we evaluate several model variants under the incomplete modality setting, reporting R^2 across tasks in Figure 3. The figure shows results on three tasks (education, environment, and population).

Single Expert collapses the MoE into one shared cross-attention expert shared across all tasks. The consistent performance decline across tasks indicates that a single fusion pathway is insufficient, as different tasks inherently depend on distinct combinations of multi-modal features.

Uniform Routing retains all N experts but replaces learned routing with uniform weights over the selected experts ($1/K$), removing task-specific gating. Its performance is consistently worse than UrbanExpert and close to the Single Expert variant, suggesting that naive expert averaging is insufficient to fully exploit the MoE capacity. These results further indicate that semantics-adaptive routing is crucial for learning specialized experts and thereby improving task prediction performance.

w/o $\mathcal{L}_{\text{lb}}$ removes the load-balancing loss. The performance degradation suggests that, without this regularizer, the router concentrates its assignments on a small subset of experts, leading to systematic underutilization of the remaining experts and a reduced effective model capacity.

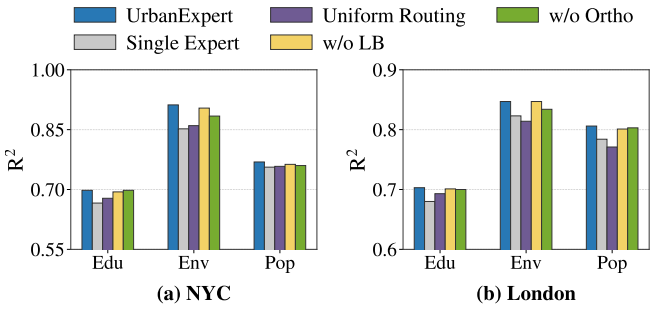


Figure 3: Ablation study: mean R^2 of model variants under the incomplete modality setting on NYC and London.

w/o $\mathcal{L}_{\text{ortho}}$ disables orthogonality regularization on expert queries. The performance drop, especially on the environment task, highlights that orthogonality encourages experts to attend to complementary feature subspaces, thereby improving expert diversity and overall capacity.

4.4 Expert Specialization Analysis

To examine whether task-conditioned routing induces interpretable expert specialization, we visualize expert behaviors on the NYC dataset. The expert–modality attention distributions in Figure 4(a) are computed by aggregating the cross-attention weights from Eq. (11) over all regions and grouping tokens by modality. The task–expert routing weights in Figure 4(b) and Figure 4(c) correspond to $w_{j,k}$ from Eq. (9). These results lead to the following observations.

Table 2: Generalization to held-out tasks (NYC, complete modality setting). Performance is measured by R^2; best in bold, second-best underlined.

Domain	Task	HREP	ReCP	HAFusion	FlexiReg	GURPP	Uniform Routing	UrbanExpert (ST)	UrbanExpert
Economy	Median Income	0.41	0.48	0.47	0.40	0.50	0.58	0.59	**0.68**
	Median Rent	0.33	0.43	0.40	0.33	0.43	0.55	0.53	**0.64**
	Poverty Rate	0.27	0.33	0.34	0.27	0.33	0.30	0.37	**0.52**
	Unemployment	0.10	0.05	0.07	0.09	0.12	0.10	0.11	**0.18**
Safety	Crime Rate	−0.15	−0.65	−0.25	−0.03	−0.01	0.30	0.25	**0.45**
Environment	Building Height	0.68	0.80	0.77	0.76	0.78	0.76	**0.85**	0.80
Transport	Carpool	0.14	0.01	0.10	0.17	0.17	0.18	0.14	**0.20**
	Drive	0.71	0.74	0.74	0.71	0.78	0.70	0.77	**0.80**
	Walk/Bike	0.38	0.38	0.40	0.39	0.44	0.43	0.47	**0.55**
	Work from Home	0.40	0.38	0.37	0.37	0.46	0.40	0.46	**0.56**

- **Experts self-organize into modality specialists.** Figure 4(a) shows that experts develop distinct modality-attention profiles without explicit supervision. For instance, E1 allocates 54.1% of its attention to satellite tokens, serving as a satellite specialist, while E2 exhibits the most balanced distribution across all four modalities. This emergent specialization arises from the coupling bridge: distinct expert prototypes induce distinct attention behaviors.
- **Tasks route to modality-appropriate experts.** Figure 4(b) confirms that tasks preferentially activate experts whose modality preferences align with their semantics. Environment task (Env) routes predominantly to E1 (44.3%) and E4 (27.6%), both satellite-leaning experts, consistent with the domain knowledge that vegetation indices are primarily derived from spectral imagery. Education (Edu) and transport (Trans) share similar routing patterns through E5, E6, and E7, the experts with higher street-view and POI attention, which may indicate that these experts capture shared factors such as neighborhood development level or amenity accessibility.
- **Inter-task routing structure generalizes across cities.** Comparing Figure 4(b) and Figure 4(c) reveals that semantically related tasks exhibit similar expert-activation patterns across NYC and London, suggesting that the learned inter-task structure is not city-specific. This cross-city consistency provides a foundation for the transfer results reported in Section 4.5.

We further examine the sensitivity of the routing mechanism to task descriptions in Appendix E.2. The results show that performance is insensitive to moderate wording changes, but drops when task descriptions are assigned to the wrong tasks.

4.5 Generalizability

4.5.1 Generalization to Unseen Tasks. To evaluate whether the learned routing mechanism transfers beyond the five training tasks, we consider 10 held-out NYC tasks that are unseen during multi-task training. For each held-out task, we feed its natural-language description into the frozen model to obtain task-conditioned embeddings and train only a Ridge regression head for prediction. To decouple generalization from missing-modality effects, we report results under the complete modality setting.

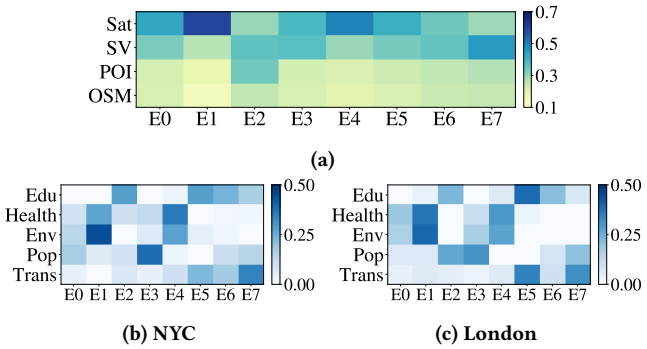


Figure 4: Expert specialization analysis. (a) Average expert–modality attention distribution across the four sensing modalities. (b) Task–expert routing weights on NYC. (c) Task–expert routing weights on London.

As shown in Table 2, UrbanExpert achieves substantially higher R^2 than region embedding baselines (HREP, ReCP, HAFusion, FlexiReg and GURPP) on most held-out tasks, indicating that the task-conditioned embeddings produced by UrbanExpert encode task-relevant semantic information that a single pretrained region embedding cannot capture, and that the model effectively exploits task relatedness to enhance downstream predictions. UrbanExpert further surpasses the single-task baseline UrbanExpert (ST) on 9 of 10 held-out tasks, demonstrating the strong transferability of representations learned via multi-task training. The gains are especially pronounced on difficult targets: for crime rate, UrbanExpert achieves $R^2 = 0.45$, whereas all region embedding baselines produce negative R^2 values. This result highlights that semantically related training tasks can provide complementary signals that substantially improve predictions on challenging held-out targets. Comparing UrbanExpert with the Uniform Routing variant further confirms that learned semantic routing, rather than the MoE architecture alone, is responsible for the transfer gains. Consistent findings on the London dataset are reported in Appendix D.2.

Table 3: Cross-city transfer between NYC and London measured by Spearman ρ.

(a) NYC→London								
	Education		Environment		Population		Transport	
Method	0-shot	Ridge	0-shot	Ridge	0-shot	Ridge	0-shot	Ridge
ReCP	0.01	0.74	0.16	0.73	-0.09	0.85	0.04	0.70
FlexiReg	0.12	0.76	0.27	0.74	0.35	0.86	0.26	0.68
UrbanExpert	**0.54**	**0.82**	**0.87**	**0.95**	**0.83**	**0.93**	**0.30**	**0.76**
(b) London→NYC								
	Education		Environment		Population		Transport	
Method	0-shot	Ridge	0-shot	Ridge	0-shot	Ridge	0-shot	Ridge
ReCP	0.04	0.55	0.29	0.87	0.01	0.74	0.07	0.58
FlexiReg	0.13	0.47	0.07	0.46	0.16	0.54	0.06	0.41
UrbanExpert	**0.54**	**0.75**	**0.80**	**0.95**	**0.52**	**0.89**	**0.16**	**0.67**

4.5.2 Generalization to Unseen Cities. To assess whether the learned task–modality dependencies transfer across urban contexts, we conduct cross-city transfer experiments in both directions between NYC and London. For each direction, the model is trained on the source city and evaluated on the target city across four shared tasks under two protocols: *zero-shot* (reusing the source-city prediction heads) and *Ridge* (freezing the backbone and fitting a Ridge regressor on target-city labels). Performance is measured by Spearman ρ, which quantifies the rank correlation between predicted and ground-truth values, with higher values indicating better agreement in the induced rankings.

We observe from Table 3 that region embedding baselines transfer poorly in the zero-shot setting. ReCP is particularly affected by cross-city domain shift, likely because its contrastive objective yields city-specific feature distributions, while FlexiReg is slightly more robust due to street-view alignment but remains limited by task-agnostic fusion. UrbanExpert, in contrast, achieves substantially higher zero-shot ρ, suggesting that the learned task–modality dependencies exhibit meaningful cross-city consistency and can be effectively captured by semantic routing. Transferability also varies across tasks: environment and population transfer particularly well, indicating that physically grounded targets rely on more city-invariant feature patterns, whereas transport is harder to transfer, potentially due to differences between the two transit systems. Under the Ridge protocol, all methods improve, and UrbanExpert maintains its advantage across all four tasks.

4.6 Robustness Under Missing Modalities

To systematically evaluate robustness, we vary the per-modality drop probability $p \in \{0.0, 0.1, 0.2, 0.3\}$ at test time and report mean R^2 on NYC, where p=0.3 corresponds to a severe setting in which about 76% of regions lack at least one modality.

Figure 5(a) compares UrbanExpert with the strongest baselines (ReCP and HREP) under increasing missingness. As missingness increases, both baselines deteriorate sharply. HREP's R^2 drops below zero and ReCP falls close to zero at p=0.3, indicating a near collapse of predictive performance. In contrast, UrbanExpert degrades much more gracefully: at p=0.3 it still achieves a mean R^2 of 0.62, substantially outperforming both baselines and confirming that explicitly

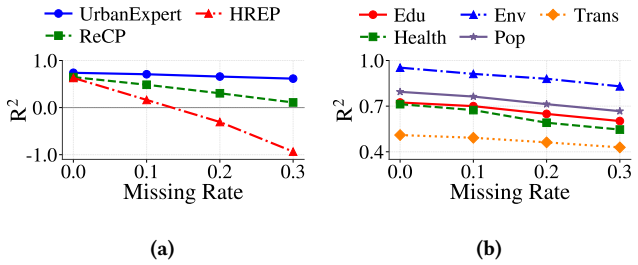


Figure 5: Robustness to missing modalities on NYC. (a) Mean R^2 versus drop probability for UrbanExpert and baselines. (b) Per-task R^2 versus drop probability.

modeling cross-modal reconstruction provides a clear advantage when inputs are incomplete.

Figure 5(b) further reports per-task R^2 across different drop probabilities. Environment and population density are the most robust, retaining approximately 85% of their complete-modality performance at p=0.3, suggesting that their predictive signal is largely carried by modalities that can be accurately reconstructed, making these targets less sensitive to modality dropout. Education and health show moderate sensitivity, while transport is most affected, indicating that the signal relevant for transport is more easily disrupted by missing inputs and harder to recover from partially observed modalities.

5 Related Work

Urban Region Representation Learning. Learning region embeddings has attracted growing attention as a way to support diverse downstream urban prediction tasks. Some efforts primarily leverage human mobility records to model urban dynamics [28, 29, 40]. Driven by privacy considerations and data accessibility, recent work has shifted toward publicly available modalities such as POIs, satellite imagery, and street-view photographs [13, 34, 37]. To integrate heterogeneous data sources, graph-based methods construct region similarity graphs and employ GNNs to propagate information across correlated regions [3, 42], while contrastive learning approaches align representations of the same region across modalities to enhance cross-view consistency [14, 15]. More recently, large-scale pre-trained models have been explored to inject commonsense or geographic knowledge into region representations [38]. However, these methods produce a single shared embedding for all downstream tasks without accounting for the fact that different prediction targets depend on different feature combinations. Additionally, prompt-based approaches such as HREP [43] and FlexiReg [26] introduce task-aware adaptation on top of a fixed encoder, but do not reconfigure the underlying fusion pathways. In contrast, UrbanExpert explicitly conditions the fusion strategy on task semantics, allowing each task to activate a distinct subset of experts and compose modality-specific signals accordingly.

Multi-Modal Learning with Incomplete Observations. Multimodal fusion has been widely studied in vision-language and urban computing settings [36]. Most methods assume that all modalities are available at both training and inference time, yet in practice

not every modality is observed for every region, making complete multi-modal coverage rarely guaranteed across large-scale urban areas. Existing strategies for handling missing modalities include modality dropout during training [18] and generative reconstruction of absent inputs [33]. The proposed UrbanExpert addresses this challenge through a dedicated cross-modal reconstruction module that leverages inter-modality correlations to recover missing representations during both training and inference.

Multi-Task Learning and Mixture-of-Experts. Multi-task learning aims to improve generalization by sharing representations across related tasks. A central challenge is balancing shared and task-specific capacity: hard parameter sharing risks negative transfer among dissimilar tasks, while fully separate models forgo cross-task knowledge reuse. Mixture-of-Experts (MoE) architectures offer a principled middle ground by routing different inputs or tasks to specialized sub-networks [6, 25]. Recent MoE variants in multi-task settings typically route based on input features [16] or task ID [8], but do not exploit the semantic relationships among tasks. UrbanExpert routes based on task-semantic similarity to learnable expert prototypes, so that semantically related tasks naturally share experts while distant tasks are served separately, enabling both effective knowledge sharing and task-specific specialization.

6 Conclusion

We presented UrbanExpert, a multi-task framework for multi-modal urban prediction. The core idea is twofold: a cross-modal reconstruction mechanism models inter-modality semantic correlations and recovers missing modality representations, while a task-conditioned Mixture-of-Experts architecture enables adaptive fusion of heterogeneous features guided by task semantics. The model is jointly optimized through a multi-task objective combining prediction, reconstruction, and routing regularization losses to improve overall region representation quality. Experiments on two real-world urban datasets demonstrated consistent improvements over existing baselines. Beyond quantitative gains, our analysis revealed that experts spontaneously specialize toward distinct modality groups, and that the learned routing structure transfers effectively to held-out tasks and unseen cities, suggesting that task semantics provides a generalizable inductive bias for multi-modal fusion. Future work could explore automated task description generation or leverage large language models to produce richer semantic queries, potentially broadening the applicability of semantic routing to a wider range of urban and non-urban prediction problems.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62572274.

References

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

[2] Centers for Disease Control and Prevention. 2023. PLACES: Local data for better health. https://www.cdc.gov/places/.

[3] Meng Chen, Hongwei Jia, Zechen Li, Weiming Huang, Kai Zhao, Yongshun Gong, Haoran Xu, and Hongjun Dai. 2025. Region embedding with adaptive correlation discovery for predicting urban socioeconomic indicators. *IEEE Transactions on Knowledge and Data Engineering* (2025).

[4] Meng Chen, Zechen Li, Hongwei Jia, Xin Shao, Jun Zhao, Qiang Gao, Min Yang, and Yilong Yin. 2025. MGRL4RE: A multi-graph representation learning approach for urban region embedding. *ACM Transactions on Intelligent Systems and Technology* 16, 2 (2025), 1–23.

[5] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment* 120 (2012), 25–36.

[6] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[7] Mordechai Haklay and Patrick Weber. 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing* 7, 4 (2008), 12–18.

[8] Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. 2021. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems* 34 (2021), 29335–29347.

[9] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*. PMLR, 4651–4664.

[10] Jiahui Jin, Yifan Song, Dong Kan, Haojia Zhu, Xiangguo Sun, Zhicheng Li, Xigang Sun, and Jinghui Zhang. 2025. Urban region pre-training and prompting: a graph-based approach. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 1071–1082.

[11] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7482–7491.

[12] Stephen Law, Brooks Paige, and Chris Russell. 2019. Take a look around: using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology* 10, 5 (2019), 1–19.

[13] Yi Li, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. 2023. Urban region representation learning with OpenStreetMap building footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1363–1373.

[14] Zechen Li, Weiming Huang, Kai Zhao, Min Yang, Yongshun Gong, and Meng Chen. 2024. Urban region embedding via multi-view contrastive prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8724–8732.

[15] Zechen Li, Hongwei Jia, Kai Zhao, Weiming Huang, and Meng Chen. 2026. 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*. 748–758.

[16] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1930–1939.

[17] Ministry of Housing, Communities and Local Government. 2019. English Indices of deprivation 2019. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.

[18] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. 2015. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2015), 1692–1706.

[19] New York City Police Department. 2024. NYPD Complaint Data (Historic). https://opendata.cityofnewyork.us/.

[20] Office for Health Improvement and Disparities. 2024. Public Health Profiles (Fingertips). https://fingertips.phe.org.uk/. Indicator 93283.

[21] Office for National Statistics. 2020. Income Estimates for Small Areas, England and Wales. https://www.ons.gov.uk/.

[22] Office for National Statistics. 2022. Census 2021, England and Wales. https://www.nomisweb.co.uk/.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal* (2024).

[24] Martino Pesaresi and Panagiotis Politis. 2023. GHS-BUILT-H R2023A-GHS building height, derived from AW3D30, SRTM30, and Sentinel2 composite (2018). *European Commission, Joint Research Centre (JRC)* (2023).

[25] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

[26] Fengze Sun, Yanchuan Chang, Egemen Tanin, Shanika Karunasekera, and Jianzhong Qi. 2025. FlexiReg: Flexible urban region representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[27] Fengze Sun, Jianzhong Qi, Yanchuan Chang, Xiaoliang Fan, Shanika Karunasekera, and Egemen Tanin. 2024. Urban region representation learning with attentive fusion. In *Proceedings of the IEEE 40th International Conference on Data Engineering*. 4409–4421.
[28] Tianao Sun, Meng Chen, Bowen Zhang, Genan Dai, Weiming Huang, and Kai Zhao. 2025. SILO: Semantic integration for location prediction with large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 2756–2767.
[29] Tianao Sun, Ke Fu, Weiming Huang, Kai Zhao, Yongshun Gong, and Meng Chen. 2024. Going where, by whom, and at what time: Next location prediction considering user preference and temporal regularity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2784–2793.
[30] Andrew J Tatem. 2017. WorldPop, open data for spatial demography. *Scientific data* 4, 1 (2017), 170004.
[31] UK Home Office. 2024. Police Recorded Crime (Street-Level Data). https://data.police.uk/.
[32] U.S. Census Bureau. 2023. American Community Survey (ACS) 5-Year Estimates. https://www.census.gov/programs-surveys/acs.
[33] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15878–15887.
[34] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1013–1020.
[35] Ronghui Xu, Jihao Chen, Jindong Tian, Chenjuan Guo, and Bin Yang. 2026. MoST: A Foundation Model for Multi-modality Spatio-temporal Traffic Prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 1716–1727.
[36] Ronghui Xu, Hanyin Cheng, Chenjuan Guo, Hongfan Gao, Jilin Hu, Sean Bin Yang, and Bin Yang. 2025. Mm-path: Multi-modal, multi-granularity path representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 1703–1714.
[37] Ronghui Xu, Weiming Huang, Jun Zhao, Meng Chen, and Liqiang Nie. 2023. A spatial and adversarial representation learning approach for land use classification with POIs. *ACM Transactions on Intelligent Systems and Technology* 14, 6 (2023), 1–25.
[38] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*. 4006–4017.
[39] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
[40] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. 2018. Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.
[41] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* 11, 1 (2020), 2583.
[42] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2021. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 4431–4437.
[43] Silin Zhou, Dan He, Lisi Chen, Shuo Shang, and Peng Han. 2023. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4981–4989.

A Datasets and Preprocessing

A.1 Datasets

The NYC dataset contains 2,324 census tracts, the standard small-area statistical units delineated by the U.S. Census Bureau. The London dataset contains 1,000 Middle Layer Super Output Areas (MSOAs), the corresponding units defined by the U.K. Office for National Statistics. Both serve as the prediction regions in our experiments. Table 4 lists all prediction tasks across both datasets, including domain categorization and data sources.

Table 4: Prediction tasks and data sources across the NYC and London datasets.

Domain	Task	NYC Source	London Source
Socioeconomic	Education	ACS 2023 [32]	Census 2021 [22]
	Income	ACS 2023 [32]	ONS Income Estimates [21]
	Rent	ACS 2023 [32]	–
	Unemployment	ACS 2023 [32]	Census 2021 [22]
	Poverty	ACS 2023 [32]	IMD 2019 [17]
Health	Obesity	PLACES [2]	–
	Life expectancy	–	OHID Fingertips [20]
Environment	NDVI	Sentinel-2 [5]	Sentinel-2 [5]
	Population density	WorldPop [30]	WorldPop [30]
	Building height	[24]	[24]
Safety	Crime rate	NYC OpenData [19]	UK Police [31]
Transport	Public transport	ACS 2023 [32]	Census 2021 [22]
	Drive	ACS 2023 [32]	Census 2021 [22]
	Walk/bike	ACS 2023 [32]	Census 2021 [22]
	Work from home	ACS 2023 [32]	Census 2021 [22]
	Carpool	ACS 2023 [32]	Census 2021 [22]

A.2 Multi-Modal Encoding Details

This section provides detailed preprocessing pipelines for each modality introduced in Section 3.1.

Satellite Embeddings. We use pretrained satellite embeddings from AlphaEarth Foundations [1], a global geospatial foundation model that fuses multi-source Earth-observation data. We aggregate the native 10 m embeddings to a 100 m grid, with each cell treated as one token, yielding a variable-length sequence $\mathbf{Z}_i^{(\text{sat})} \in \mathbb{R}^{N_{\text{sat}} \times 64}$, where N_{sat} is the number of cells in the region.

Street-View Imagery. We collect up to 10 street-view images per region via the Google Street View Static API[1]. Each image is encoded by CLIP ViT-L/14, and the `[CLS]` token serves as the per-image representation, yielding $\mathbf{Z}_i^{(\text{sv})} \in \mathbb{R}^{N_{\text{sv}} \times 768}$.

POI Text. POI records from OpenStreetMap [7] are summarized into structured textual templates describing category distributions and functional characteristics per region. Each template sentence is encoded by Qwen3-8B [39], producing $\mathbf{Z}_i^{(\text{poi})} \in \mathbb{R}^{N_{\text{poi}} \times 1024}$.

OpenStreetMap Tiles. Cartographic basemap tiles are rendered from OpenStreetMap, covering building footprints, road networks, and land cover. Each tile is encoded by DINOv2 ViT-B/14 [23], yielding $\mathbf{Z}_i^{(\text{osm})} \in \mathbb{R}^{N_{\text{osm}} \times 768}$.

B Model Configuration

For missing modality handling, we apply per-modality independent masking with probability $p = 0.1$ during training. The task-specific prediction heads are implemented as linear heads with L_2 regularization (Ridge-style). Hyperparameters are selected via a coarse-to-fine grid search on the validation set. In the Multi-Modal Encoding module, each modality is projected to a hidden dimension of $d_h = 384$ and compressed into $P = 6$ latent tokens using a Perceiver-style cross-attention layer. In the Missing Modality Reconstruction module, we use a single-layer Transformer decoder with 4 attention heads to reconstruct masked modality tokens. The Task-Conditioned Mixture-of-Experts module comprises $N = 8$

[1] https://developers.google.com/maps/documentation/streetview/overview

Table 5: Performance comparison under the complete modality setting. Best results are in bold and second-best are underlined.

NYC	Education			Health			Environment			Population			Transport		
Method	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
HREP	10.80	14.02	0.582	3.42	4.43	0.544	0.022	0.030	0.941	5009	7101	0.654	9.33	12.16	0.430
ReCP	10.55	13.89	0.590	3.28	4.22	0.586	0.024	0.033	0.928	4446	6342	0.724	9.56	12.43	0.405
HAFusion	10.96	14.36	0.560	3.47	4.43	0.545	0.042	0.062	0.751	4704	6894	0.673	9.46	12.23	0.423
FlexiReg	13.79	17.86	0.320	4.30	5.29	0.344	0.072	0.100	0.347	6752	9322	0.405	11.35	14.28	0.214
GURPP	10.13	13.20	0.629	2.99	3.89	0.648	0.045	0.060	0.766	4652	6529	0.708	8.84	11.52	0.488
UrbanExpert (w/o Recon)	**8.22**	**11.07**	**0.737**	**2.43**	**3.31**	**0.746**	**0.018**	**0.025**	**0.959**	**3394**	**5123**	**0.820**	**8.52**	**10.97**	**0.524**

London	Education			Health			Environment			Population			Transport		
Method	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
HREP	9.61	11.46	0.156	2.15	2.70	0.077	0.058	0.073	0.533	2050	2793	0.659	6.75	8.24	0.116
ReCP	7.75	9.60	0.407	2.05	2.60	0.144	0.038	0.049	0.792	**1264**	**1856**	**0.847**	5.80	7.24	0.314
HAFusion	9.54	11.58	0.136	2.17	2.71	0.068	0.071	0.091	0.271	2143	2896	0.591	6.68	8.19	0.125
FlexiReg	9.45	11.40	0.165	2.24	2.80	0.007	0.085	0.107	−0.002	2818	3938	0.323	7.03	8.73	0.002
GURPP	7.54	9.31	0.442	2.05	2.58	0.158	0.061	0.077	0.475	1609	2225	0.781	5.72	7.14	0.336
UrbanExpert (w/o Recon)	**4.65**	**6.02**	**0.766**	**1.64**	**2.11**	**0.434**	**0.026**	**0.034**	**0.901**	1442	1968	0.827	**3.36**	**4.34**	**0.749**

cross-attention experts with top-$K = 4$ sparse routing. Each expert produces an output of dimension 224 through a 2-layer output network. The total loss combines: (1) uncertainty-weighted multi-task regression loss, (2) expert query orthogonality loss (weight 0.1), (3) Switch-style load balancing loss (weight 1.0), and (4) reconstruction loss (weight 0.1). For optimization, we use the AdamW optimizer with differentiated learning rates: 10^{-4} for base network and router parameters, 10^{-3} for expert prototypes, and 0.025 for the uncertainty parameters σ_j.

C Baseline Descriptions

- **HREP** [43] constructs a heterogeneous region graph from mobility, POI, and geographic proximity, and applies relation-aware GCN with attention-based fusion. Prefix-tuning prompts enable downstream task adaptation.
- **ReCP** [14] learns consistent region representations from POI and mobility views via intra-view contrastive learning and inter-view consistency constraints.
- **HAFusion** [27] employs a dual-feature attentive fusion module to capture inter-region correlations across views, with hybrid attentive learning over intra- and inter-view interactions.
- **FlexiReg** [26] learns grid embeddings from public data (POI, land use, satellite, street view), aggregates them into region embeddings, and applies a prompt enhancer for task-specific adaptation.
- **GURPP** [10] constructs a heterogeneous urban region graph from POIs, roads, junctions, and typed spatial relations, learns subgraph-centric region embeddings via pre-training, and adapts them with graph-based prompts for downstream tasks.

D Additional Results

D.1 Performance Comparison under the Complete-modality Setting

Table 5 reports results under the complete-modality setting, where all four modalities are available for each region. We compare **UrbanExpert (w/o Recon)** with four region embedding baselines and three multi-task learning baselines. Since no modality is missing at test time, this comparison mainly evaluates whether task-conditioned fusion improves representation quality beyond a shared task-agnostic embedding.

On NYC, UrbanExpert (w/o Recon) achieves the highest average R^2 across the five tasks, with clear gains on socioeconomic and transport-related targets. ReCP and HREP remain competitive on the environment task, suggesting that task-agnostic embeddings can capture strong satellite-derived signals for NDVI prediction. However, their performance is less consistent on education, health, population, and transport, where different modality combinations may be required.

On London, UrbanExpert (w/o Recon) again obtains the best average R^2, with notable gains on transport and health. These results show that task-conditioned fusion is useful even when all modalities are observed, as different targets still depend on different combinations of the same input modalities.

D.2 Generalization to Held-Out Tasks on London

We further evaluate representation transferability on the London dataset using nine held-out tasks that are not included during multi-task training. As shown in Table 6, UrbanExpert achieves the best R^2 on all nine downstream tasks, with an average R^2 of 0.75, compared with 0.48 for the strongest baseline. The gains are especially clear on socioeconomic indicators, suggesting that task-conditioned embeddings learned from the supervised tasks transfer effectively to related but unseen targets. UrbanExpert also improves transport-related tasks, indicating that the learned routing mechanism can be reused beyond the original training targets.

Crime prediction remains the most difficult held-out task. Although UrbanExpert obtains the best result, its R^2 is much lower than on other indicators, suggesting that safety-related outcomes may depend on factors only partially captured by the available visual and POI modalities. Still, UrbanExpert yields positive transfer, whereas several baselines produce near-zero or negative R^2.

Table 6: Generalization to held-out tasks (London, complete modality setting). Performance is measured by R^2; best in bold, second-best underlined.

Domain	Task	ReCP	HREP	FlexiReg	HAFusion	GURPP	UrbanExpert
Economy	Median Income	0.29	<u>0.33</u>	0.08	-0.01	0.27	**0.84**
	Unemployment	<u>0.26</u>	<u>0.26</u>	0.15	0.11	0.21	**0.68**
	Poverty Rate	0.21	<u>0.24</u>	0.11	0.10	0.21	**0.72**
Safety	Crime Rate	-0.11	<u>0.16</u>	0.11	0.11	0.07	**0.29**
Environment	Bldg Height	0.81	<u>0.82</u>	0.71	0.64	0.69	**0.84**
Transport	Carpool	0.62	<u>0.66</u>	0.51	0.34	0.65	**0.83**
	Drive	0.79	<u>0.80</u>	0.68	0.50	0.79	**0.93**
	Walk/Bike	0.55	0.57	0.52	0.41	<u>0.58</u>	**0.69**
	Work from Home	0.42	<u>0.47</u>	0.20	0.06	0.43	**0.89**

Table 7: Natural-language task descriptions for semantic routing.

Task	Description
Education	Predict the proportion of residents holding bachelor's degrees or higher. Education level indicates human capital concentration and professional workforce presence. It correlates with proximity to educational institutions, cultural amenities, professional service establishments, and residential environment quality.
Health	Predict the adult obesity prevalence at the census tract level. Obesity prevalence reflects community health behaviors shaped by the built environment. It correlates with fast food concentration, healthy food access, park availability, street walkability, and active transportation infrastructure.
Environment	Predict the vegetation greenness index of a region. Vegetation index measures the presence and health of plant cover. It indicates environmental quality and green space availability. Higher vegetation areas typically feature parks, tree-lined streets, and lower building coverage.
Population	Predict the population density of a region. Population density reflects the intensity of human settlement. Higher density areas exhibit vertical development, concentrated building footprints, extensive road networks, and diverse commercial activities.
Transport	Predict the public transportation usage ratio within a region. Public transport ratio measures mass transit adoption relative to private vehicle use. It reflects transit infrastructure including subway stations and bus stops, as well as urban density and mixed land use patterns.

E Task Descriptions for Semantic Routing

E.1 Description Construction and Examples

Table 7 provides examples of the task descriptions used for task semantic encoding (Section 3.3). The descriptions are generated by an LLM from each task name and target definition, following a shared schema that specifies the prediction objective, indicator interpretation, and potentially correlated observable urban features. We then manually check the generated descriptions for factual consistency. The full set of descriptions is available in our repository.

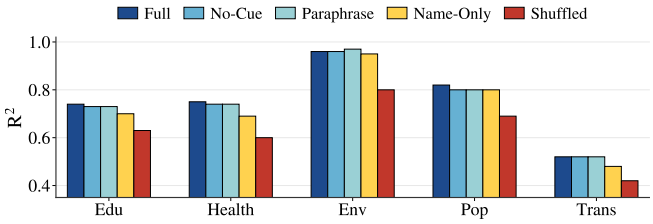


Figure 6: Sensitivity to task description wording (NYC, R^2).

E.2 Sensitivity to Task Descriptions

To assess how sensitive the routing mechanism is to the specific wording of descriptions, we evaluate five variants on the NYC dataset: *Full* (original LLM-generated descriptions), *No-Cue* (removing explicit modality hints from descriptions), *Paraphrase* (LLM-generated alternative phrasings), *Name-Only* (using only the task name as the description), and *Shuffled* (randomly mismatching descriptions across tasks).

As shown in Figure 6, Full, No-Cue, and Paraphrase yield similar performance, indicating that UrbanExpert is not overly sensitive to variation in phrasing and does not depend on explicit modality cues. Name-Only also remains substantially better than Shuffled, suggesting that task names already preserve useful semantic information for routing. In contrast, mismatched descriptions break the correspondence between tasks and descriptions and cause a clear performance drop. Together, these results indicate that the improvements stem from meaningful semantic routing, rather than from the MoE architecture or description-specific prompt engineering.