mathematics

MDPI

Article

RAMEN: Region-Adaptive Mixture of Ego-Networks for Multimodal Geospatial Fusion in Urban Region Representation

Genan Dai [1,2,†], Zitao Guo [3,†], Hu Huang [4,*], Jinzhou Cao [2,*], Liwen Jing [5] and Bowen Zhang [1,2]

[1] Guangdong Key Laboratory of Intelligent Computing for Public Service Provision, Peking University Shenzhen Graduate School, Shenzhen 518055, China
[2] Spatio-Temporal Intelligence Center and School of Artificial Intelligence, Shenzhen Technology University, Shenzhen 518118, China
[3] College of Applied Science, Shenzhen University, Shenzhen 518060, China
[4] School of Urban Planning and Design, Peking University Shenzhen Graduate School, Shenzhen 518055, China
[5] Pengcheng Laboratory, Shenzhen 518000, China
* Correspondence: huanghu@ustc.edu.cn (H.H.); caojinzhou@sztu.edu.cn (J.C.)
† These authors contributed equally to this work.

Abstract

Learning transferable region embeddings is fundamental to urban computing, supporting applications from economic forecasting to public safety. Existing multi-view fusion methods predominantly rely on coarse-grained fusion, applying uniform weights across an entire city or task, thereby neglecting spatial heterogeneity—the fact that the dominant view driving a region's function varies significantly across regions. To address this, we propose RAMEN, a two-stage framework for Region-adaptive Mixture of Ego-Networks learning. In the pre-training stage, a unified spatially aware Transformer distills universal urban semantics. In the fine-grained adaptation stage, a Mixture of Ego-Networks (MoEN) module employs a region-adaptive gating mechanism to dynamically allocate exclusive view weights for each region. A Region Ego-aware Spatial Transformer (REST) then aggregates these fused local subgraphs by explicitly injecting degree and physical distance priors, overcoming the over-smoothing limitations of traditional GNNs. Extensive experiments on real-world datasets for check-in, crime, service call and population prediction show that RAMEN consistently outperforms state-of-the-art baselines, achieving up to 35.3% MAE improvement. Visualizations of gating weights further suggest that RAMEN's fusion aligns well with real-world urban physical characteristics.

Keywords: urban region representation learning; multi-view fusion; ego-networks

MSC: 68T07

Academic Editor: Andrea Scozzari

Received: 17 June 2026
Revised: 20 July 2026
Accepted: 21 July 2026
Published: 24 July 2024

1. Introduction

Characterizing urban regions from heterogeneous geospatial data is a core problem in geospatial artificial intelligence (GeoAI), underpinning applications from public safety to urban planning [1–3]. No single source captures the full profile of a region. Remote sensing imagery records physical morphology and land cover, points of interest (POIs) reflect functional composition, and human mobility encodes connectivity and accessibility. Therefore, the learning of a transferable region representation depends on the fusion of these complementary modalities into a unified embedding that supports diverse downstream tasks, including crime, check-in, service demand, and population prediction [4,5].

Existing studies have made significant progress in fusing multi-source heterogeneous data to learn urban region embeddings. For instance, some approaches capture inter-view correlations through attention-based multi-view joint learning [6–8], while others enhance cross-view information consistency via contrastive learning [9–11]. However, the majority of these methods adopt "globally shared weights", meaning a uniform set of fusion weights is strictly applied across all regions within a city. While recent advancements have noticed this limitation and begun to manually or learnably adjust view weights based on specific downstream tasks (e.g., a check-in prediction task might inherently pay more attention to attractive POIs) [12,13], a critical flaw remains. Whether operating at the city level or the task level, these approaches fundamentally belong to coarse-grained fusion. They inevitably ignore the most crucial attribute of urban spaces: spatial heterogeneity.

Modeling the spatial heterogeneity of urban regions across different tasks—specifically, recognizing that the "dominant view" dictating a region's function varies drastically from one location to another—is of paramount importance. To illustrate this intuitively, consider the check-in prediction task (as conceptually illustrated in Figure 1). Region A, a suburban scenic area, draws a high volume of check-ins from its natural land cover, vegetation, water, and open space, which remote sensing imagery captures but sparse POI records and low transit flows do not; its check-in behavior is therefore driven by the physical environment rather than digital footprints. Conversely, Region B, a downtown transit hub, handles massive daily crowds where check-ins are largely a byproduct of immense transit volume, making its behavior highly mobility-driven. Forcing uniform view fusion weights upon these two distinct regions during prediction will inevitably neutralize their unique characteristics, leading to sub-optimal representations. Therefore, a fine-grained, region-adaptive fusion mechanism is urgently required.

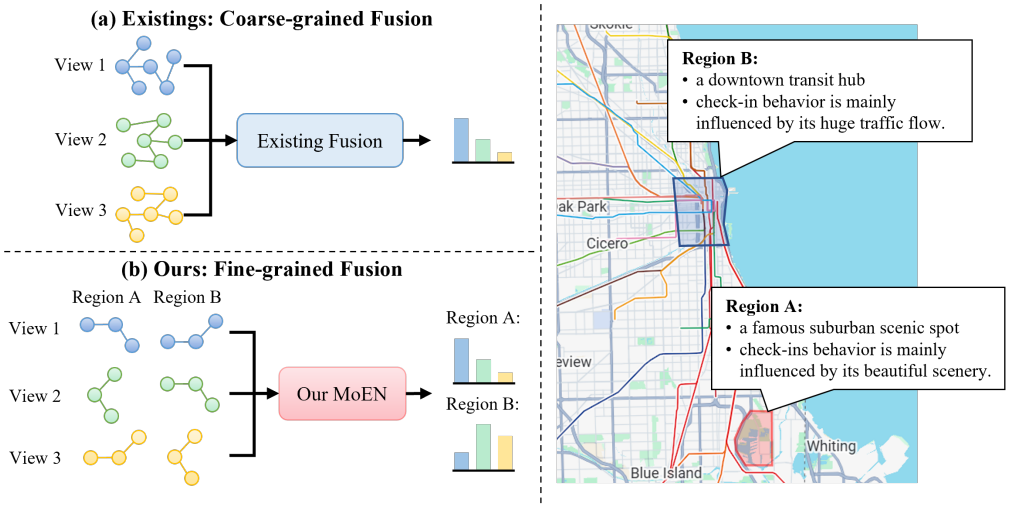


Figure 1. Example of region-level spatial heterogeneity.

To achieve such fine-grained, region-level fusion, the ego-network—comprising a target region and its immediate spatial context—serves as the most natural microscopic granularity. Ideally, we need to dynamically learn a fused region ego-network for each area and employ Graph Neural Networks (GNNs) to model this fine-grained spatial heterogeneity. However, traditional message-passing GNNs encounter severe bottlenecks in this scenario. First, they aggregate neighbors based on simple topological structures, failing to capture the complex spatial relationships (e.g., physical distance or functional similarity) between the central node and its specific neighbors. Second, they lack the ability to explicitly model the connectivity degree of different regional ego-networks.

To address the aforementioned challenges, we propose a novel framework: Region-Adaptive Mixture of Ego-Networks for Multimodal Geospatial Fusion in Urban Region Representation (RAMEN). First, we design the Mixture of Ego-Networks (MoEN) fusion module. Through a region-adaptive gating mechanism, this module dynamically generates exclusive view weights for each region, enabling the precise and fine-grained fusion of multi-view ego-networks. Second, to overcome the limitations of traditional GNNs on ego-networks, we introduce a Region Ego-aware Spatial Transformer (REST) module. By injecting connectivity degree and distance priors (e.g., physical distance or cosine similarity) directly into the attention mechanism, this module achieves highly precise aggregation of the fused ego-networks. Furthermore, to maintain architectural uniformity and adequately model physical spatial relations, this Transformer module is also repurposed as the front-end single-view feature encoder.

The main contributions of this work are summarized as follows:

- We reveal the critical limitations of "coarse-grained fusion" in existing multi-view urban region representation methods, explicitly identifying the problem of region-level spatial heterogeneity.
- We propose an innovative Mixture of Ego-Networks module, achieving genuine fine-grained, region-level multi-view fusion through a region-adaptive gating mechanism.
- We tailor a unified Region Ego-aware Spatial Transformer module, effectively overcoming the structural defects of traditional GNNs when processing the spatial topology of ego-networks.
- Extensive experiments on multiple real-world city datasets demonstrate that our method not only achieves state-of-the-art (SOTA) performance but also exhibits strong regional interpretability.

The remainder of this paper is organized as follows: Section 2 comprehensively reviews the pertinent literature on multi-view urban region representation and the application of ego-networks in graph learning. Section 3 formalizes the research problem and elaborates on the proposed RAMEN framework, specifically detailing the mathematical formulations of the Region-Adaptive Gating mechanism and the REST module. Section 4 presents a rigorous empirical evaluation against state-of-the-art baselines across diverse downstream tasks, accompanied by in-depth ablation, a case study and parameter-sensitivity analyses. Finally, Section 5 concludes this work and discusses potential avenues for future research.

2. Related Work

2.1. Multi-View Urban Region Representation Learning

Early endeavors in urban region representation primarily focused on single-view learning, relying on isolated data sources such as human mobility trajectories [4,14], satellite or street-view imagery [2,15,16], and Points of Interest (POIs) [17]. Such uni-modal approaches inevitably yield partial urban profiles and generally perform well only on a narrow set of specific tasks.

To enrich region representations, recent studies have shifted towards multi-view learning, fusing diverse data modalities to comprehensively profile urban spaces [18,19]. For instance, methods like MV-PN [20] and CGAL [21] model intra-region structural information and inter-region spatial autocorrelations from POI and mobility views, but they often rely on naive concatenation or basic MLP assemblies for multi-view fusion. To discern the varying importance of different views, models such as MVURE [5], Region2Vec [8], and HA-Fusion [7] employ attention mechanisms to capture cross-view correlations. Concurrently, inspired by contrastive learning paradigms, approaches like ReMVC [9], ReCP [10], and UrbanMMCL [1] align multi-view semantic spaces to ensure cross-modal consistency. More recently, the focus has expanded to task-specific adaptation, with works like HREP [22],

FlexiReg [23], and GURPP [12] exploring prompt tuning to align general embeddings with downstream objectives.

Despite their tremendous success, a critical limitation persists: most of these methods rely on globally shared weights or merely distinguish view importance at the task level. Consequently, they overlook the fine-grained spatial heterogeneity inherent in urban spaces—specifically, the phenomenon whereby different regions exhibit vastly different view dependencies, even within the exact same task. The MoEN proposed in this paper is precisely designed to bridge this gap.

2.2. Ego-Networks and Mixture of Experts for Graph Learning

In graph learning, capturing micro-level spatial heterogeneity is crucial, and ego-centric networks are naturally suited for this characteristic. Recently, the superiority of ego-centric modeling has been validated in broader graph-learning domains. For instance, in spatiotemporal forecasting, ENRLF [24] constructs ego-centric networks for sample-level targets, effectively filtering out irrelevant long-distance noise while preserving critical micro-dependencies. Similarly, CS-GNN [25] utilizes ego-networks to distill local structural diversity, successfully mitigating structural distribution shifts during graph knowledge transfer. These advances corroborate that, to effectively model spatial heterogeneity at the urban region level, the ego-network—a subgraph comprising a central region and its direct neighbors—serves as the most natural and effective granularity for delineating micro-level urban features.

Concurrently, the Mixture-of-Experts (MoE) architecture has demonstrated immense potential as a dynamic routing mechanism for the processing of heterogeneous and multimodal data. Recent works, such as UniGraph2 [26], I^2MoE [27], and HMoE [28], have successfully leveraged MoE to dynamically align multimodal features and quantify modal interactions. In the specific context of graph learning, MoEGCL [29] takes a pioneering step by integrating MoE with ego-networks to achieve sample-level fusion for multi-view node clustering.

However, existing MoE-based graph methods exhibit critical limitations when applied to complex urban predictive tasks. Models like UniGraph2 and I^2MoE primarily route at the flat node-feature level, largely ignoring the local topological structures inherent in spatial data. Furthermore, while MoEGCL incorporates ego-graphs, it is strictly designed for unsupervised node clustering without task-specific guidance, and it fails to overcome the inherent bottlenecks of traditional GNNs when processing ego-networks. In contrast, our proposed MoEN module explicitly leverages task-specific supervision signals to generate unique, region-specific fusion weights for each ego-network. To further complement this, we tailor a REST module to execute fine-grained message passing specifically on ego-networks, thereby maximizing the preservation of fine-grained spatial heterogeneity for specific downstream tasks.

3. Methodology

3.1. Preliminaries

Definition 1 (Urban Region)**.** *A city is partitioned into a set of N non-overlapping regions ($\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$) based on administrative boundaries or census tracts. Each region ($r_i \in \mathcal{R}$) is associated with multi-view data capturing functional, visual, and social characteristics.*

Definition 2 (Geographic Topology Graph)**.** *We construct a spatial adjacency matrix ($\mathbf{A} \in \mathbb{R}^{N \times N}$) based on geographical distance. Specifically, for any pair (r_i, r_j), we compute the Manhattan distance (d_{ij}) between the centroids. The adjacency weight is defined via a Gaussian*

kernel: $A_{ij} = \exp(-d_{ij}^2/2\sigma^2)$ *if* $d_{ij} \leq \tau$ *and* $A_{ij} = 0$ *otherwise, where* σ *is the standard deviation and* τ *is a distance threshold to maintain sparsity.*

Definition 3 (Multi-View Features). *We denote the input feature matrices for POIs, Satellite Imagery (SI), and mobility as* $\mathbf{X}^{poi} \in \mathbb{R}^{N\times P}$, $\mathbf{X}^{si} \in \mathbb{R}^{N\times S}$, *and* $\mathbf{X}^{mob} \in \mathbb{R}^{N\times N}$, *respectively. Here, P represents the total number of distinct POI categories, and S denotes the dimensionality of the visual feature vectors extracted from satellite imagery. Specifically,* $\mathbf{X}^{poi}$ *encapsulates regional functional attributes,* $\mathbf{X}^{si}$ *captures visual urban morphology and land-use patterns, and* $\mathbf{X}^{mob}$ *reflects the integrated connectivity and accessibility between regions.*

Definition 4 (Region Ego-Network). *The ego-network of a target region* (r_i) *under view v is defined as a local subgraph* $(\mathcal{G}_i^v = (\mathcal{V}_i, \mathcal{E}_i^v)$, *where* $\mathcal{V}_i$ *comprises* r_i *and its Top-k nearest neighbors based on view-specific similarities). We denote the local adjacency matrix of this ego-network as* $\mathbf{S}_i^{(v)} \in \mathbb{R}^{|\mathcal{V}_i|\times|\mathcal{V}_i|}$.

Definition 5 (Problem Formulation). *Given* $\mathcal{R}$ *and multi-view features, we aim to learn a unified mapping function* $(\mathcal{F} : \mathcal{R} \to \mathbb{R}^d)$, *assigning each region* (r_i) *a fine-grained embedding* $(\mathbf{h}_i \in \mathbb{R}^d)$ *that effectively captures spatial heterogeneity for diverse downstream tasks.*

3.2. Framework Overview

As illustrated in Figure 2, the proposed framework is designed as a hierarchical architecture that progressively evolves urban representations from coarse-grained universal profiles to fine-grained task-specific descriptors. Specifically, the framework first learns coarse-grained universal embeddings from multi-view heterogeneous data, which capture fundamental cross-modal correlations and task-agnostic urban semantics across the entire city. This foundational layer serves as a coarse-grained knowledge base that encapsulates the global consensus of urban functions.

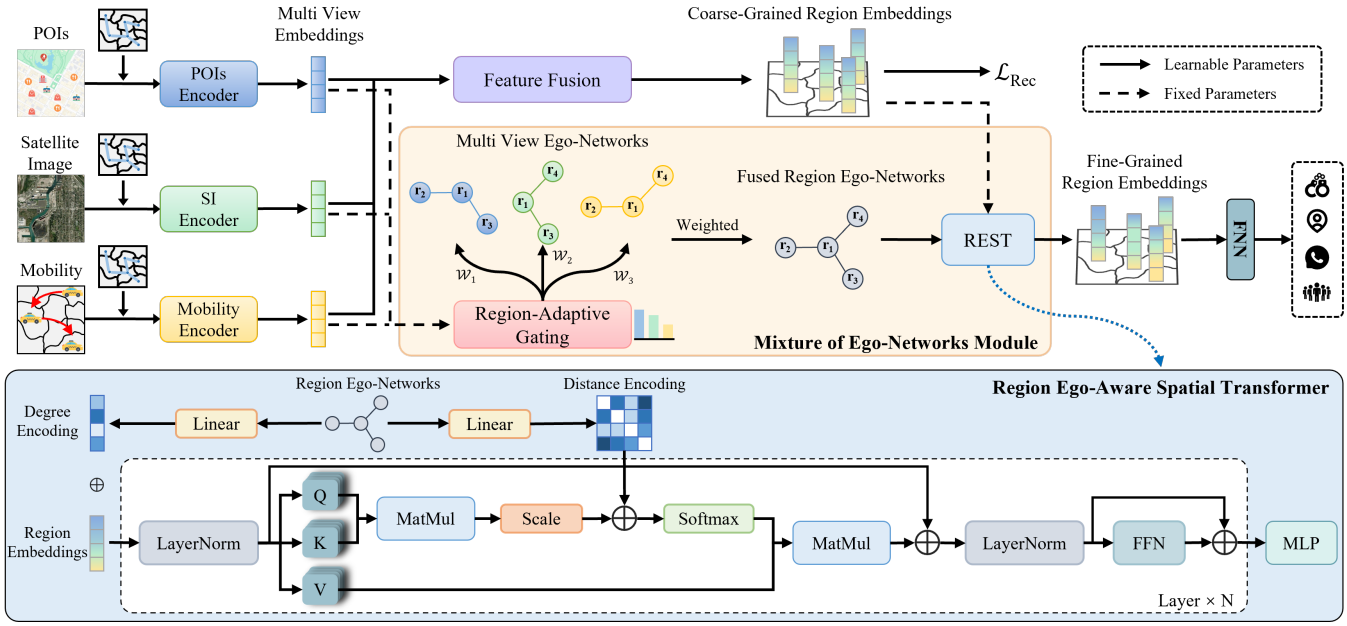


Figure 2. The model framework.

Building upon these base representations, the framework transitions into a micro-level refinement phase to untangle spatial heterogeneity. This is achieved through the Mixture of Ego-Networks (MoEN) module, which adaptively re-weights multi-view local topologies via a region-adaptive gating mechanism, thereby shifting the focus from global

patterns to region-level view dependencies. Subsequently, these dynamically fused ego-networks are processed by the Region Ego-Aware Spatial Transformer (REST) to capture fine-grained structural dependencies and micro-spatial priors. By seamlessly progressing from coarse-grained semantic learning to fine-grained, region-adaptive modeling, the framework produces highly expressive embeddings tailored to diverse urban tasks.

3.3. Coarse-Grained Region Representation Learning

To capture the spatial autocorrelation within each view, heterogeneous inputs are first projected into a unified space using MLPs. To maintain architectural uniformity, we leverage the REST module (detailed in Section 3.5) as the backbone view encoder, which extracts fundamental spatial dependencies by injecting distance-based priors.

After obtaining view-specific embeddings ($\mathbf{h}_i^v$), we adopt an attention-based fusion mechanism following [7] to derive the coarse-grained embedding ($\mathbf{h}_i^{coarse}$). This representation is pre-trained using a multi-task reconstruction loss ($\mathcal{L}_{Rec}$). Specifically, we apply view-specific decoders to $\mathbf{h}_i^{coarse}$ to generate reconstructed outputs ($\hat{\mathbf{x}}_i^v$).

- **Similarity Reconstruction**: We reconstruct the pairwise similarity matrix by computing the inner product of POI and SI outputs:

$$\mathcal{L}_u = \frac{1}{N^2} \sum_{i,j} \|\text{sim}(\hat{\mathbf{x}}_i^u, \hat{\mathbf{x}}_j^u) - \text{sim}(\mathbf{x}_i^u, \mathbf{x}_j^u)\|^2 \tag{1}$$

  where $u \in \{poi, si\}$ and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity between original features.
- **Mobility Reconstruction**: We follow [7] to reconstruct region-to-region transition probabilities using KL divergence:

$$\mathcal{L}_{mob} = \sum_{i,j} P_{ij} \log\left(\frac{P_{ij}}{\hat{P}_{ij}}\right) \tag{2}$$

  where P_{ij} and $\hat{P}_{ij}$ represent the empirical and predicted transition probabilities from region r_i to r_j, respectively.

The total reconstruction loss is $\mathcal{L}_{Rec} = \lambda_1 \mathcal{L}_{poi} + \lambda_2 \mathcal{L}_{si} + \lambda_3 \mathcal{L}_{mob}$. Once pre-trained, these parameters are strictly frozen during the fine-grained learning stage.

3.4. Mixture of Ego-Networks Module (MoEN)

We first construct view-specific ego-networks for each region (r_i) using K-Nearest Neighbors based on feature similarity, yielding the local adjacency matrices ($\mathbf{S}_i^{(v)}$). To address spatial heterogeneity, where different regions exhibit varying view dependencies, we introduce region-adaptive gating. This module takes the concatenated multi-view embeddings as input and computes region-specific view weights ($\mathbf{w}_i = [w_i^{poi}, w_i^{si}, w_i^{mob}]$) as follows:

$$\mathbf{w}_i = \text{Softmax}(\text{MLP}([\mathbf{h}_i^{poi} \| \mathbf{h}_i^{si} \| \mathbf{h}_i^{mob}])) \tag{3}$$

The fused regional ego-network topology ($\mathbf{S}_i^{fused}$) for region r_i is derived by the weighted summation of its multi-view adjacency matrices:

$$\mathbf{S}_i^{fused} = \sum_{v \in \{poi, si, mob\}} w_i^v \mathbf{S}_i^{(v)} \tag{4}$$

Unlike global weighting schemes, this gating mechanism dynamically evaluates the task-specific importance of each view solely based on the micro-profile of region r_i, enabling genuine fine-grained spatial heterogeneity modeling.

However, during the training process, the region-adaptive gating mechanism may suffer from the view dominance problem. The model might greedily assign an overwhelmingly high weight (e.g., a near one-hot vector) to a single dominant view that converges faster, thereby suppressing the informative signals from other complementary views. To prevent such weight collapse and encourage a balanced yet adaptive fusion, we introduce an entropy-based regularization mechanism, which is jointly optimized during the downstream task adaptation (detailed in Section 3.6).

3.5. Region Ego-Aware Spatial Transformer (REST)

To overcome the inability of traditional GNNs to capture fine-grained spatial heterogeneity and distance-weighted dependencies, we propose the REST module. Note that this REST module is instantiated independently from the frozen view encoders in the pre-training stage, aiming specifically to aggregate the fused ego-networks.

3.5.1. Spatial Encodings

For each region (r_i), we first distill two types of spatial priors from the dynamically fused ego-network ($\mathcal{G}_i^{fused}$).

Degree Encoding: To characterize regional hub-ness within the local context, we compute the node degree ($deg_i = \sum_{j \in \mathcal{V}_i} S_{ij}^{fused}$). To ensure numerical stability and scale-invariance across different urban graphs, we perform max-normalization:

$$\overline{deg}_i = \frac{deg_i}{\max_k(deg_k) + \epsilon} \tag{5}$$

where ϵ is a small constant. This normalized degree is then mapped via a linear projection layer to a d-dimensional vector ($\mathbf{e}_i^{deg} = \text{Linear}(\overline{deg}_i)$), which is added to the coarse-grained region embedding ($\mathbf{h}_i^{coarse}$) to form the initial hidden state: $\mathbf{z}_i^{(0)} = \mathbf{h}_i^{coarse} + \mathbf{e}_i^{deg}$.

Distance Encoding: To capture explicit pairwise spatial correlations, we map the topological edge weights in $\mathbf{S}_i^{fused}$ to a distance-aware attention bias:

$$B_{ij} = \text{Linear}(S_{ij}^{fused}) \tag{6}$$

This bias provides an explicit structural prior for the subsequent information aggregation.

3.5.2. Ego-Aware Attention Mechanism

Inside each REST layer, message passing is executed strictly within the bounded receptive field of the ego-network. To enhance training stability, we adopt a pre-normalization architecture as illustrated in Figure 2.

Let $\mathbf{z}_i^{(l)} \in \mathbb{R}^d$ denote the representation of region r_i at the l-th layer. The input is first normalized and projected by weight matrices ($\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d_k}$) to obtain the query, key, and value matrices:

$$\mathbf{q}_i = \text{LN}(\mathbf{z}_i^{(l)})\mathbf{W}_Q, \quad \mathbf{k}_j = \text{LN}(\mathbf{z}_j^{(l)})\mathbf{W}_K, \quad \mathbf{v}_j = \text{LN}(\mathbf{z}_j^{(l)})\mathbf{W}_V \tag{7}$$

To explicitly model fine-grained spatial heterogeneity, we incorporate the distance-aware attention bias (B_{ij}) into the similarity calculation. The pre-softmax attention score (e_{ij}) and the final attention weight (α_{ij}) between region r_i and its neighbor ($r_j \in \mathcal{V}_i$) are formulated as follows:

$$e_{ij} = \frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_{key}}} + \beta B_{ij}, \quad \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{u \in \mathcal{V}_i} \exp(e_{iu})} \tag{8}$$

where d_{key} is the dimension of the key vectors and β is a hyperparameter controlling the importance of the spatial inductive bias. Crucially, the softmax normalization is strictly constrained to the neighborhood ($\mathcal{V}_i$) defined by the fused ego-network.

The final output of the ego-aware attention mechanism for region r_i is computed as the weighted sum of the value vectors:

$$\text{Attn}(\mathbf{z}_i^{(l)}) = \sum_{j \in \mathcal{V}_i} \alpha_{ij} \mathbf{v}_j \tag{9}$$

Following the attention stage, this output is added to the un-normalized input via a residual connection to form an intermediate state ($\mathbf{z'}_i^{(l)}$), which is then processed by a Feed-Forward Network (FFN):

$$\mathbf{z'}_i^{(l)} = \mathbf{z}_i^{(l)} + \text{Attn}(\mathbf{z}_i^{(l)}) \tag{10}$$

$$\mathbf{z}_i^{(l+1)} = \mathbf{z'}_i^{(l)} + \text{FFN}(\text{LN}(\mathbf{z'}_i^{(l)})) \tag{11}$$

where $\text{LN}(\cdot)$ denotes layer normalization. After L layers of such spatially aware aggregation, the final output is mapped via an MLP head to derive the task-specific fine-grained representation: $\mathbf{h}_i^{fine} = \text{MLP}(\mathbf{z}_i^{(L)})$.

3.6. Task Adaptation and Optimization

The fine-grained embeddings ($\mathbf{h}_i^{fine}$) are fed into a task-specific FNN for final prediction. For regression tasks (e.g., check-in or crime prediction), we minimize the mean squared error (MSE) as the supervision loss:

$$\mathcal{L}_{task} = \frac{1}{|\mathcal{R}_{train}|} \sum_{i \in \mathcal{R}_{train}} (y_i - \hat{y}_i)^2, \tag{12}$$

where y_i and $\hat{y}_i$ denote the ground truth and the predicted value for region r_i, respectively, and $\mathcal{R}_{train}$ represents the set of training regions.

Furthermore, to mitigate the view dominance issue discussed in Section 3.4, we explicitly regularize the gating weights ($\mathbf{w}_i$) by minimizing their negative entropy. This prevents the gating distribution from degenerating into a one-hot state. The gating entropy regularization loss ($\mathcal{L}_{reg}$) is formulated as follows:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i=1}^{N} \sum_{v \in \{poi, si, mob\}} w_i^v \log(w_i^v + \epsilon), \tag{13}$$

where ϵ is a small constant to ensure numerical stability. Minimizing $\mathcal{L}_{reg}$ is mathematically equivalent to maximizing the information entropy of the gating weights, thereby forcing the model to continuously explore and extract valuable semantics from all available views rather than overly relying on a single modality.

The overall objective function for the fine-grained learning stage is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \gamma \mathcal{L}_{reg}, \tag{14}$$

where γ is a hyperparameter balancing downstream prediction and gating regularization.

3.7. Time-Complexity Analysis

We conducted a time-complexity analysis for the key modules. In the fine-grained adaptation stage, the coarse-grained encoder is frozen, and the main computational cost comes from MoEN and REST. For each region, MoEN computes region-adaptive view

weights via an MLP and fuses multi-view adjacency matrices restricted to its top-K ego-network. REST then performs ego-aware attention strictly within each region's bounded receptive field rather than over the full graph. The overall complexity is expressed as follows: $O(Nd^2 + N \cdot K \cdot d)$, where $O(Nd^2)$ comes from linear projections and FFN layers and O(NKd) comes from ego-network-restricted attention. Provided that K and d are treated as fixed constants independent of N, the overall complexity scales linearly with the number of regions (N). This linear scalability in N provides a favorable advantage over baselines that rely on full graph-level attention (which can scale as ($O(N^2d)$) in the worst case).

4. Experiments

In this section, we conduct comprehensive experiments on real-world urban datasets to thoroughly validate the superiority of the proposed framework. Specifically, our evaluation is driven by the following five core research questions:

- **RQ1**: Does our framework achieve new state-of-the-art (SOTA) performance across diverse downstream urban computing tasks compared to existing representation paradigms?
- **RQ2**: To what extent do our core innovations—the Mixture of Ego-Networks (MoEN) module and the Region Ego-Aware Spatial Transformer (REST)—contribute to the final representation quality?
- **RQ3**: Is the RAMEN framework resilient to the absence of specific data modalities, and can the MoEN effectively compensate for information loss across various task scenarios?
- **RQ4**: Is the framework resilient to perturbations in critical hyperparameters, such as the topological depth of ego-networks (K-hop) and the dimensionality of feature embeddings?
- **RQ5**: Does the model genuinely untangle spatial heterogeneity? More importantly, are the fine-grained view weights dynamically allocated by the model consistent with observable urban physical characteristics?

4.1. Experimental Setup

4.1.1. Datasets

To comprehensively evaluate the proposed framework, we construct multi-view datasets from two major metropolises: Manhattan, New York City (NYC), and Chicago (CHI). The spatial granularity of the regions is rigorously defined by official administrative boundaries. Specifically, the study area in NYC is delineated by census tracts in Manhattan, whereas CHI is partitioned based on official community area boundaries obtained from the government open data portal.

To profile the multifaceted nature of urban spaces, we extract features from three distinct heterogeneous views: Points of Interest (POIs), high-resolution satellite imagery, and human mobility trajectories. Furthermore, to rigorously assess the quality and transferability of the learned region representations, we curate four diverse downstream prediction tasks using longitudinal urban records. These target variables encompass crime incidents, user check-in volumes, public service requests (e.g., 311 calls), and demographic population statistics. The detailed descriptive statistics of all datasets are summarized in Table 1.

Table 1. Dataset statistics.

Dataset	NYC	CHI	Source
Regions	180	77	Open Portal [30,31]
POIs	24,496	57,891	OpenStreetMap [32]
POI categories	26	26	OpenStreetMap [32]
Satellite images	180	77	Google Maps [33]
Taxi trips	10,953,879	3,381,807	Open Portal [30,31]
Land-use categories	11	12	OpenStreetMap [32]
Crime records	35,335	18,200	Open Portal [30,31]
Check-in counts	106,902	167,232	Foursquare [34]
Service call records	516,187	24,350	Open Portal [30,31]
Population counts	1,540,692	2,508,984	WorldPop [35]

4.1.2. Baselines

To comprehensively evaluate the effectiveness of our proposed framework, we compare it against eight state-of-the-art (SOTA) urban region representation methods. Based on their fusion paradigms and task-adaptation strategies, we categorize these baselines into two distinct groups:

(1) Coarse-Grained Fusion Methods: These methods focus on generating task-agnostic universal representations. They typically employ intra-graph message passing or cross-modal contrastive alignment. However, a common limitation is their reliance on globally shared fusion weights across the entire city, inherently ignoring regional spatial heterogeneity.

- **MGFN** [36] constructs multi-period mobility graphs to capture spatial–temporal mobility patterns, employing intra-graph message passing and an inter-graph cross-attention mechanism for feature aggregation.
- **MVURE** [5] is a multi-view joint learning framework that leverages Graph Attention Networks (GATs) to encode mobility and regional attributes (e.g., POIs), followed by a self-attention module for cross-view information sharing and global weighted fusion.
- **HAFusion** [7] proposes a hybrid attentive fusion (DAFusion) mechanism that first aggregates multi-view embeddings via view-aware attention and subsequently captures high-order inter-region correlations through region-aware self-attention.
- **ReCP** [10] is an unsupervised paradigm that enforces cross-modal consistency by maximizing mutual information through inter-view contrastive learning, coupled with intra-view feature reconstruction and conditional entropy minimization.
- **UrbanCLIP** [11] is an alignment-based method that leverages Large Language Models (LLMs) to generate textual descriptions for satellite imagery, aligning vision and language modalities through contrastive pre-training to enrich regional visual profiles.
- **ComSRE** [37] is a commonality-specificity disentanglement framework that learns multi-view urban region embeddings by combining region-specific attentive fusion, view-to-commonality contrastive alignment, and differential orthogonality constraints to preserve both shared cross-view semantics and view-specific signals.

(2) Task-Adaptive Coarse-Grained Methods: Recognizing the gap between universal representations and downstream objectives, these recent approaches introduce task-

specific adaptations (e.g., prompt tuning). Nevertheless, their prompts or fusion strategies remain coarse-grained—tuned solely at the task level—therefore failing to accommodate the fine-grained, region-specific variations within the same task.

- **HREP** [22] constructs a heterogeneous urban graph and extracts generic embeddings via relation-aware GCNs. Crucially, it introduces a continuous prompt-learning strategy (prefix tuning) to adapt the frozen pre-trained embeddings to different downstream tasks.
- **FlexiReg** [23] learns fine-grained grid-level embeddings and employs an adaptive aggregation module for arbitrary regional partitioning. It utilizes prompt enhancement driven by street-view and textual alignments to flexibly adapt representations for specific downstream applications.

4.1.3. Implementation Details

RAMEN is trained following a two-stage paradigm. In the first stage, coarse-grained representation learning is conducted for 3000 epochs with a learning rate of 5×10^{-4} to capture generalizable urban knowledge. In the second stage, fine-grained task adaptation is performed for 1500 epochs to model task-specific spatial heterogeneity. During the fine-grained adaptation stage, all parameters learned in the coarse-grained stage are frozen, which preserves the stability of the pre-trained general urban knowledge while allowing the subsequent optimization to focus on the region-adaptive fusion and spatial aggregation modules.

For each downstream prediction task, we conduct region-level ten-fold cross-validation following [22,23]. Under each random seed, all regions are randomly shuffled and partitioned into ten mutually exclusive folds. In each fold, 90% of the regions constitute the training pool, while the remaining 10% are held out as the test set. The training pool is further divided into 80% for model training and 20% as an independent validation subset. Hyperparameters are selected exclusively according to performance on the validation subset, whereas the test fold remains unseen during both model optimization and hyperparameter selection. The test fold is used only for the final performance evaluation. We repeat the complete ten-fold cross-validation procedure using five different random seeds and report the average performance across the five independent runs. For statistical significance testing, each random-seed run is treated as one statistical observation: for each seed, the results from the ten test folds are first averaged into a single aggregated value per method; paired t-tests are then conducted using the five seed-level observations, with RAMEN and each comparison method paired under the same random seed.

All experiments are implemented using PyTorch 2.4.0 and conducted on an NVIDIA GeForce RTX 4090 GPU. We investigate the key hyperparameters of RAMEN over predefined search spaces: the spatial inductive bias coefficient ($\beta \in \{0.0, 0.1, \ldots, 1.0\}$), the ego-network size ($K \in \{5, 10, 15\}$), the number of REST layers ($L \in \{2, 3, 4, 5\}$), and the second-stage learning rate ($\eta \sim \text{LogUniform}(10^{-4}, 10^{-2})$). The final configurations are outlined as follows: for the MoEN module, K is set to 10 (i.e., top-10 neighbors are selected for each region and each data view); for REST, (β, L) is set to $(0.5, 3)$ for NYC and $(0.2, 2)$ for CHI; and the second-stage learning rate is 5×10^{-3} for NYC and 1×10^{-3} for CHI. Following the settings adopted in [7,22,23], the dimensionality of the final region representations is fixed at 144.

4.2. Overall Performance (RQ1)

Table 2 presents the comprehensive predictive performance of all evaluated methods across the four downstream tasks in New York and Chicago. Overall, our proposed RAMEN consistently outperforms all baselines across all metrics, with statistically significant improvements (p-value < 0.05) on the majority of task–metric combinations. Only a few

metrics on Service Call CHI and Population CHI do not reach significance, though RAMEN still achieves the best absolute values.

Notably, RAMEN yields larger relative improvements on tasks that exhibit stronger spatial heterogeneity. For example, in the check-in prediction task, RAMEN reduces RMSE by 18.6% in NYC and MAE by 35.3% in Chicago compared to the strongest baseline. These results provide empirical support for RQ1, indicating the effectiveness of our fine-grained representation paradigm.

Table 2. Performance comparison on four downstream tasks. Best: bold; second: underline. ↓ indicates lower is better; ↑ indicates higher is better. Values marked with * indicate statistically significant improvements over the second-best result (paired *t*-test, p-value < 0.05). The improvement is calculated against the best baseline.

Crime	**NYC**			**CHI**		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MVURE	67.4 ± 0.9	90.5 ± 1.3	0.625 ± 0.017	100.4 ± 6.6	129.2 ± 7.3	0.461 ± 0.062
MGFN	73.2 ± 2.6	91.4 ± 2.9	0.618 ± 0.014	107.4 ± 5.4	137.9 ± 5.2	0.386 ± 0.047
ReCP	81.4 ± 2.0	101.3 ± 1.9	0.483 ± 0.025	86.9 ± 5.5	120.1 ± 7.1	0.534 ± 0.057
HAFusion	56.1 ± 0.7	76.1 ± 2.0	0.734 ± 0.014	77.8 ± 1.2	107.1 ± 4.4	0.631 ± 0.033
UrbanCLIP	97.4 ± 2.6	126.1 ± 1.9	0.267 ± 0.012	101.6 ± 0.6	134.7 ± 1.7	0.416 ± 0.006
ComSRE	72.3 ± 1.1	99.6 ± 6.5	0.545 ± 0.059	75.2 ± 4.2	100.3 ± 6.2	0.675 ± 0.040
HREP	66.1 ± 2.7	84.4 ± 2.3	0.674 ± 0.009	88.3 ± 6.4	114.4 ± 5.5	0.578 ± 0.041
FlexiReg †	53.2 ± 0.4	73.6 ± 1.6	0.751 ± 0.009	61.7 ± 0.2	85.1 ± 2.0	0.766 ± 0.011
RAMEN	**49.3** ± 0.9 *	**67.7** ± 2.1 *	**0.798** ± 0.015 *	**57.1** ± 0.3 *	**74.1** ± 2.3 *	**0.823** ± 0.010 *
Improv.	**7.3%**	**8.0%**	**6.3%**	**7.4%**	**12.9%**	**7.4%**
Check-in	**NYC**			**CHI**		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MVURE	285.1 ± 6.2	461.0 ± 8.4	0.682 ± 0.015	1693 ± 74	3171 ± 128	0.656 ± 0.029
MGFN	345.0 ± 13.3	503.5 ± 20.8	0.621 ± 0.032	1281 ± 41	2276 ± 86	0.817 ± 0.011
ReCP	233.8 ± 3.6	392.7 ± 19.1	0.763 ± 0.027	1272 ± 92	2341 ± 267	0.804 ± 0.045
HAFusion	202.8 ± 7.2	322.8 ± 12.6	0.844 ± 0.012	929 ± 62	1947 ± 75	0.870 ± 0.010
UrbanCLIP	393.6 ± 5.9	602.4 ± 3.1	0.458 ± 0.005	2612 ± 29	4885 ± 73	0.186 ± 0.024
ComSRE	251.7 ± 17.1	434.8 ± 44.4	0.716 ± 0.057	1609 ± 286	3927 ± 327	0.453 ± 0.088
HREP	274.1 ± 8.3	417.7 ± 14.9	0.739 ± 0.008	1679 ± 71	3135 ± 79	0.664 ± 0.017
FlexiReg †	198.2 ± 3.6	309.0 ± 17.2	0.850 ± 0.011	922 ± 76	1775 ± 199	0.891 ± 0.022
RAMEN	**173.1** ± 4.5 *	**251.4** ± 26.8 *	**0.901** ± 0.013 *	**597** ± 84 *	**1207** ± 216 *	**0.949** ± 0.014 *
Improv.	**12.7%**	**18.6%**	**6.0%**	**35.3%**	**32.0%**	**6.5%**
Service Call	**NYC**			**CHI**		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MVURE	1402 ± 27	2128 ± 36	0.398 ± 0.022	190.3 ± 9.8	266.9 ± 12.1	0.441 ± 0.050
MGFN	1653 ± 70	2250 ± 108	0.327 ± 0.060	208.2 ± 11.3	293.4 ± 16.6	0.329 ± 0.077
ReCP	1478 ± 59	2136 ± 41	0.366 ± 0.017	206.7 ± 11.1	303.4 ± 16.1	0.284 ± 0.076
HAFusion	1273 ± 20	1951 ± 27	0.493 ± 0.014	159.3 ± 13.9	222.0 ± 18.9	0.613 ± 0.067
UrbanCLIP	1409 ± 7	2401 ± 16	0.232 ± 0.005	183.2 ± 0.9	256.3 ± 1.8	0.491 ± 0.003
ComSRE	1459 ± 69	2254 ± 98	0.324 ± 0.058	213.0 ± 11.7	286.2 ± 18.8	0.362 ± 0.085
HREP	1396 ± 20	2011 ± 36	0.462 ± 0.021	185.7 ± 6.1	262.2 ± 10.8	0.468 ± 0.022
FlexiReg †	1303 ± 16	1989 ± 43	0.497 ± 0.012	121.1 ± 2.3	178.2 ± 5.1	0.753 ± 0.014
RAMEN	**1187** ± 21 *	**1785** ± 52 *	**0.591** ± 0.010 *	**110.3** ± 3.6 *	**177.6** ± 3.2	**0.755** ± 0.008
Improv.	**6.8%**	**8.5%**	**18.9%**	**9.0%**	**0.3%**	**0.3%**
Population	**NYC**			**CHI**		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MVURE	2899 ± 53	3708 ± 70	0.508 ± 0.007	13717 ± 322	17174 ± 552	0.313 ± 0.043
MGFN	3222 ± 47	4207 ± 119	0.367 ± 0.026	13071 ± 505	16578 ± 707	0.359 ± 0.054
ReCP	3527 ± 101	4434 ± 147	0.329 ± 0.043	12085 ± 400	17029 ± 561	0.325 ± 0.044
HAFusion	2497 ± 50	3277 ± 82	0.616 ± 0.019	10678 ± 390	13988 ± 548	0.544 ± 0.035
UrbanCLIP	3338 ± 11	4499 ± 16	0.276 ± 0.002	13328 ± 69	17498 ± 74	0.288 ± 0.006
ComSRE	2964 ± 95	3916 ± 100	0.452 ± 0.028	12063 ± 688	15308 ± 841	0.475 ± 0.057
HREP	3118 ± 67	4023 ± 121	0.421 ± 0.037	12063 ± 539	15397 ± 832	0.447 ± 0.061
FlexiReg †	2231 ± 24	2974 ± 60	0.701 ± 0.002	8126 ± 320	11395 ± 508	0.698 ± 0.028
RAMEN	**2167** ± 21 *	**2849** ± 59 *	**0.717** ± 0.008 *	**7881** ± 103	**10669** ± 336 *	**0.745** ± 0.023 *
Improv.	**2.9%**	**4.2%**	**2.3%**	**3.0%**	**6.4%**	**6.7%**

† The Chicago results of FlexiReg are taken from the original publication [23], since its pipeline requires a street-view image dataset that has not been publicly released. All other results in this table are obtained from our own reproduction.

(1) Region-level fine-grained adaptation outperforms global coarse-grained fusion. Compared with the first category of baselines (i.e., coarse-grained fusion and alignment methods), RAMEN shows consistent performance gains. While models such as HAFusion attain competitive results via dual-attention mechanisms, they still rely on globally shared fusion weights and, thus, have limited capacity to customize the semantic integration for distinct micro-environments. UrbanCLIP also yields relatively lower performance on most regression tasks, suggesting that cross-modal semantic alignment alone (via LLMs) may be insufficient without explicitly modeling fine-grained spatial topologies. In contrast, by aggregating information within the bounded receptive field of ego-networks, our REST module restricts message passing to locally relevant neighbors, which contributes to more accurate region representations.

(2) Region-level fine-grained adaptation outperforms task-Level prompts. Within the second category of baselines, FlexiReg is the strongest competitor and ranks second in most cases, indicating that adapting universal representations toward specific downstream objectives is, indeed, beneficial. Nevertheless, RAMEN still consistently outperforms FlexiReg, with the largest margins observed on the check-in and service-call tasks. This gap supports our core hypothesis that task-level tuning alone is insufficient in the presence of pronounced spatial heterogeneity and suggests that even within the same task, different regions may rely on different data views. By introducing the region-adaptive gating mechanism, RAMEN allocates region-specific fusion weights that are more consistent with the underlying multi-view urban characteristics.

4.3. Ablation Study (RQ2)

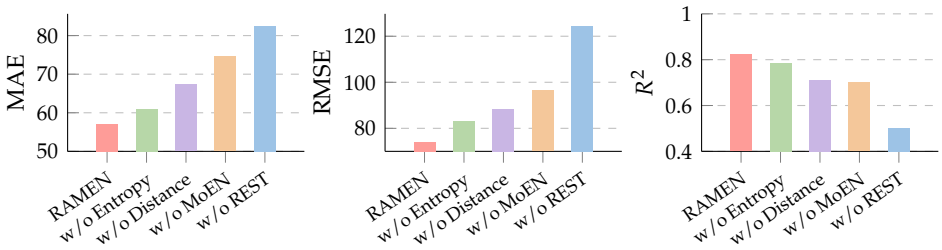
To thoroughly validate the necessity and effectiveness of our core architectural designs, we conduct a comprehensive ablation study on the Chicago dataset across all four downstream tasks. Specifically, we compare the complete RAMEN framework against four critical degraded variants:

- **w/o MoEN:** We remove the entire Mixture of Ego-Networks module, reverting the model to coarse-grained fusion by applying a globally shared attention weight across all regions.
- **w/o REST:** We further replace our tailored Region Ego-Aware Spatial Transformer with a standard Graph Convolutional Network (GCN) [38] to evaluate the specific impact of our fine-grained ego-aware aggregation.
- **w/o Entropy:** We remove the entropy-based regularization term from the routing/fusion objective to evaluate whether this constraint stabilizes fine-grained region-adaptive fusion.
- **w/o Distance:** We remove the spatial distance encoding from the REST module to evaluate the contribution of explicit spatial priors in ego-aware aggregation.

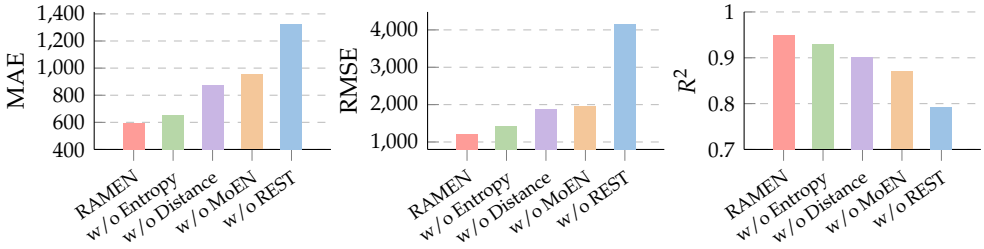
As illustrated in Figure 3, the Chicago ablation results present an overall hierarchy where RAMEN remains the strongest model; w/o Entropy is the best variant; w/o Distance is competitive with or stronger than w/o MoEN on crime, check-in, and population tasks though slightly weaker on service calls; and w/o REST is consistently the weakest. This specific performance hierarchy explicitly answers RQ2 and reveals critical insights into urban spatial modeling:

(1) Comparing the complete RAMEN with w/o MoEN, we observe a massive performance drop when the entire module is removed. For instance, in the check-in prediction task, removing the MoEN module causes the RMSE to sharply inflate from 1207 to 1952. This degradation directly corroborates our core hypothesis: coarse-grained universal representations fail to capture the nuanced, task-specific spatial heterogeneities of different

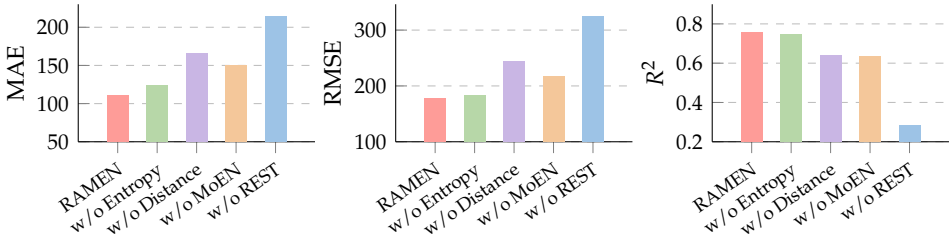
regions. Extracting fine-grained, region-adaptive ego-networks is essential for accurate urban profiling.



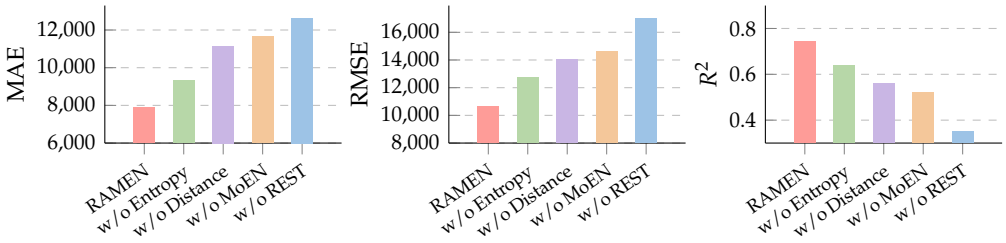
(**a**) Crime prediction



(**b**) Check-in prediction



(**c**) Service-call estimation



(**d**) Population prediction

Figure 3. Ablation study of core components on the Chicago dataset across four downstream tasks.

(2) The most striking observation is that the w/o REST variant (which applies gating but uses a GCN for aggregation) performs significantly worse than the purely coarse-grained w/o MoEN variant. For example, in crime prediction, the R^2 plunges from 0.701 (w/o MoEN) down to a mere 0.502 (w/o REST). This severe collapse highlights a profound structural defect in traditional message-passing GNNs: when confronted with highly heterogeneous, dynamically weighted multi-view ego-networks, standard GCNs treat local topologies homogeneously, leading to catastrophic over-smoothing and structural noise.

(3) The newly added w/o Entropy variant isolates the effect of the entropy-based constraint. On the Chicago dataset, removing this term consistently weakens RAMEN: the RMSE increases from 74.1 to 83.1 for crime prediction, from 1207 to 1421 for check-in predic-

tion, from 177.6 to 182.5 for service-call estimation, and from 10,669 to 12,738 for population prediction. The corresponding R^2 values also decrease from 0.823/0.949/0.755/0.745 to 0.787/0.929/0.747/0.637 across the four tasks. These results indicate that entropy-guided routing provides an additional regularizing effect that improves the robustness of fine-grained ego-network fusion.

(4) The w/o Distance variant further substantiates the importance of explicit spatial priors in REST. When distance encoding is removed, RMSE consistently increases across all tasks: from 74.1 to 88.5 for crime prediction, from 1207 to 1871 for check-in prediction, from 177.6 to 244.8 for service-call estimation, and from 10,669 to 14,040 for population prediction. These results indicate that spatial distance encoding provides essential structural guidance, enabling REST to effectively distinguish interactions among proximate and distant ego-networks instead of relying solely on data-driven attention weights.

These results demonstrate that entropy-guided routing, spatial encoding, fine-grained fusion and region ego-aware aggregation are strictly co-dependent. Attempting to perform fine-grained region-adaptive fusion without an advanced spatial aggregator (like REST) does more harm than good. Our REST module, by explicitly injecting distance and connectivity priors into the attention mechanism, helps the model better characterize the complex micro-topologies within ego-networks, thereby complementing the MoEN framework.

4.4. Robustness Analysis Under Missing Modalities (RQ3)

To further evaluate the robustness of RAMEN in data-scarce scenarios, we conduct an experiment where one specific data modality (POI, satellite, or mobility) is deliberately removed during the fine-grained adaptation stage. We compare these variants against the SOTA baseline, FlexiReg, which utilizes a full multi-modal profile. The results are summarized in Figure 4.

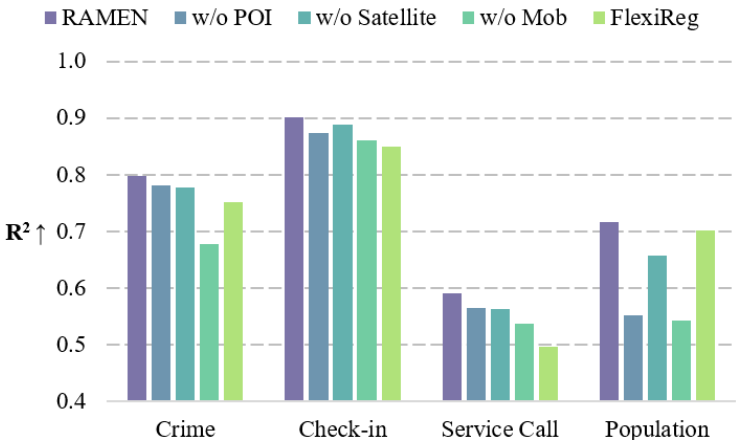


Figure 4. Model robustness analysis under missing modalities. ↑ indicates higher is better.

A key observation is that the full version of RAMEN consistently provides the performance upper bound across all tasks. More importantly, in the crime, check-in, and service-call tasks, even when one view is missing, RAMEN variants generally maintain superior or competitive performance compared to the full FlexiReg. This resilience can be attributed to our MoEN module's ability to autonomously re-allocate weights: when one modality becomes unavailable, the gating mechanism adaptively increases the reliance on the remaining informative views to compensate for the lost semantics. For instance, in the service-call task, all three missing-modality variants significantly surpass FlexiReg, demonstrating that our region-level ego-network modeling captures more robust spatial dependencies than coarse-grained paradigms, even with incomplete data.

The impact of missing modalities varies significantly across tasks, revealing the distinct "dominant views" inherent in different urban phenomena. In crime prediction, the removal of the mobility view leads to a more pronounced performance drop compared to other views, confirming that human movement patterns are a critical precursor for criminal activities. In check-in prediction, the model exhibits high stability, regardless of which single view is missing, suggesting a degree of information redundancy where POI attractions and mobility flows provide overlapping signals for social activity.

While RAMEN demonstrates strong overall robustness, the population estimation task presents a more challenging case. Although the full RAMEN achieves the best R^2 score, the performance drops notably when the POI or mobility views are removed. Specifically, the w/o Mob variant performs slightly worse than FlexiReg in this task. This indicates that for socio-demographic indicators like population density, POIs (representing residential/commercial land use) and mobility (reflecting population distribution) provide non-redundant and essential signals. In such cases, the MoEN's capacity to compensate is limited because each modality captures a unique, indispensable dimension of the region's demographic profile. This observation underscores the necessity of maintaining high-quality multi-source data for complex demographic modeling.

4.5. Parameter Sensitivity Analysis (RQ4)

To evaluate the robustness of our framework and explicitly answer RQ4, we conduct a parameter sensitivity analysis on the New York City dataset. We focus on two critical hyperparameters that govern our spatial modeling: the size of the ego-network (determined by the top-K neighbors) and the spatial inductive bias weight (β) in the REST module. The fluctuations in the R^2 metric across the four downstream tasks are illustrated in Figure 5.

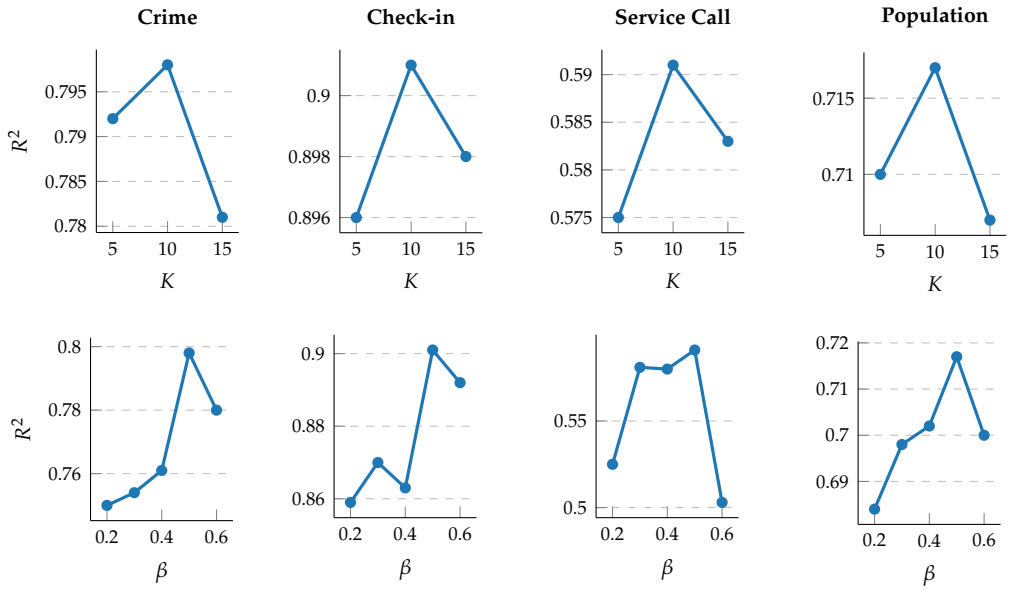


Figure 5. Parameter sensitivity analysis on the New York City dataset. The **first row** demonstrates the impact of varying the ego-network size (top-K neighbors), while the **second row** illustrates the influence of the spatial inductive bias weight (β) in the REST module.

Impact of Ego-Network Size (K): The first row of Figure 5 demonstrates the model's performance as we vary the bounded receptive field of the ego-networks ($K \in \{5, 10, 15\}$). Strikingly, across all four diverse urban tasks, the predictive performance consistently peaks at exactly $K = 10$. When K is restricted to 5, the model misses crucial local context, leading to slight performance drops. Conversely, expanding the ego-network too broadly (e.g., $K = 15$) causes a noticeable decline. For instance, in population prediction, the R^2 drops

from 0.717 at $K = 10$ to 0.707 at $K = 15$. This consistent trend yields a profound insight: the defining functional attributes of an urban region are fundamentally dictated by its immediate, highly localized micro-environment. Forcing the ego-network to include distant, functionally irrelevant regions introduces structural noise, which effectively dilutes the fine-grained spatial heterogeneity. This observation powerfully validates our architectural choice of utilizing bounded ego-networks rather than global city graphs.

Impact of Spatial Inductive Bias (β): The second row depicts the sensitivity to β, which balances the influence of the physical distance prior against the semantic attention mechanism within the REST module. Overall, the curves improve as β approaches 0.5 and universally peak at $\beta = 0.5$ across all tasks. Setting β too low (e.g., 0.2) forces the model to rely almost entirely on semantic correlations while neglecting geographic reality, leading to sub-optimal results (e.g., crime prediction R^2 drops from 0.798 at $\beta = 0.5$ to 0.750 at $\beta = 0.2$). On the other hand, an excessively large β (e.g., 0.6) forces the model to over-emphasize physical proximity, thereby overshadowing critical latent functional dependencies. The universal optimal sweet spot at $\beta = 0.5$ confirms that robust urban region embedding necessitates a synergistic, carefully calibrated fusion of both global functional semantics and local spatial–physical laws.

4.6. Case Study (RQ5)

To intuitively illustrate whether our proposed MoEN module effectively captures spatial heterogeneity, we visualize the view-fusion weights learned by the region-adaptive gating mechanism across Chicago. As illustrated in the top row of Figure 6a, the importance assigned to each data modality is far from uniform, exhibiting distinct spatial patterns that appear consistent with commonly observed urban structures.

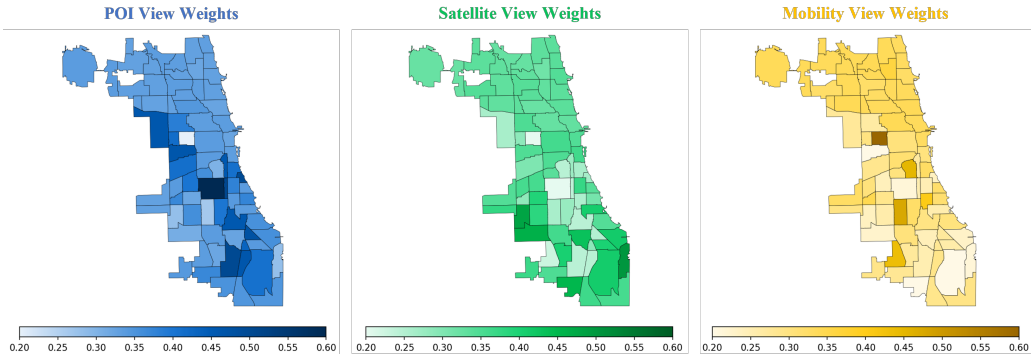
Specifically, the POI view weights tend to be higher in the commercial corridors and residential clusters of the central and southern areas, which is consistent with the intuition that POI density can be informative for characterizing regional function. In contrast, the satellite view weights are prominently highlighted in the peripheral and coastal regions, where physical land-cover textures (e.g., water bodies and vegetation) tend to be visually more distinguishable than through sparse digital records. The mobility view weights show a concentrated dominance along major transit arteries and transportation hubs, suggesting a possible association with human flow patterns in these regions.

To further investigate the model's reasoning, we categorize each region by its dominant view (the view with the highest weight) and highlight two representative cases, as shown in Figure 6b:

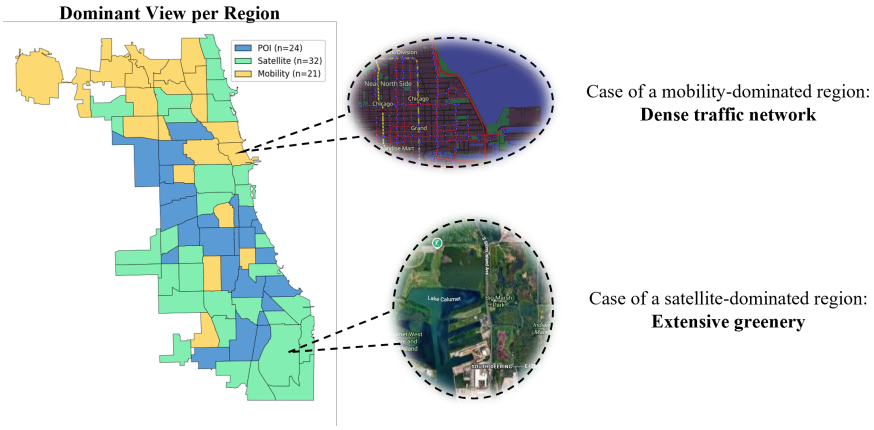
- Mobility-Dominated Region: For the downtown core, characterized by a dense traffic network, our model automatically assigns the highest weight to the mobility view. This appears consistent with the common intuition that transit-intensive hubs are often characterized by human flows and transport connectivity, in addition to static POI categories.
- Satellite-Dominated Region: In contrast, for regions featuring extensive greenery and open landscapes, the model shifts its focus to the satellite view. In these suburban or ecological zones, where POI density and mobility flows are relatively low, the visual morphological features from remote sensing imagery may offer complementary cues for characterizing the region's profile.

These qualitative observations are consistent with our core motivation: a "one-size-fits-all" fusion weight, as used in existing coarse-grained methods, may be insufficient to capture such fine-grained spatial variations. By dynamically adjusting view importance at the region level, RAMEN's gating mechanism provides an interpretable lens into how different modalities contribute across regions, complementing its quantitative performance

gains. We note that the visualizations here are intended to be illustrative rather than a rigorous quantitative validation of interpretability, which we leave as an important direction for future work.



(**a**) Weight distribution of each view



(**b**) Dominant view of each region and two representative cases

Figure 6. Case study on the check-in task in Chicago.

4.7. Model Running-Time Analysis

Table 3 reports the running time of downstream task adaptation on the NYC and CHI datasets. RAMEN achieves comparable efficiency to the task-adaptive baselines, requiring 129 s on NYC and 127 s on CHI. Although the proposed region-adaptive fusion and REST-based ego-network aggregation introduce additional spatial modeling operations, the overall running time remains within the same scale as HREP and FlexiReg. These results indicate that RAMEN improves prediction performance while maintaining practical computational efficiency for downstream urban computing tasks.

Table 3. Running time (in seconds) of downstream task adaptation.

Method	NYC	CHI
HREP	92	146
FlexiReg	137	103
RAMEN	129	127

5. Conclusions

In this paper, we proposed RAMEN, a hierarchical framework designed to capture the fine-grained spatial heterogeneity in urban region representation learning. By shifting the multi-view fusion paradigm from a coarse-grained global level to a region-adaptive

ego-network level, our model successfully addresses the limitation of uniform view dependencies across diverse urban landscapes. The core innovations—the MoEN module for dynamic weight allocation and the REST module for spatial-prior-guided aggregation—provide a robust mechanism for balancing universal urban knowledge with task-specific micro-structures. Experimental results across multiple metropolitan cities and diverse downstream tasks confirm the superiority and generalizability of RAMEN. Furthermore, the interpretability analysis suggests a qualitative alignment between the learned gating weights and actual urban land-cover/mobility patterns, offering valuable insights for urban planners.

Author Contributions: Conceptualization, G.D.; Methodology, G.D. and Z.G.; Software, Z.G.; Writing—original draft, Z.G.; Writing—review & editing, G.D., H.H., J.C., L.J. and B.Z.; Supervision, H.H., J.C. and B.Z.; Funding acquisition, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2025YFE0103100), the Guangdong Basic and Applied Basic Research Foundation (2026A1515011875), the Shenzhen Science and Technology Program (JCYJ20240813113300001 and 20231127180406001), and the Natural Science Foundation for Top Talents of SZTU (Nos. GDRC202518 and GDRC202320).

Data Availability Statement: The data presented in this study are available in NYC Open Data at https://opendata.cityofnewyork.us/. These data were derived from the following resources available in the public domain: https://opendata.cityofnewyork.us/.

Acknowledgments: The code and data are available at https://github.com/townSeven/RAMEN.git (accessed on 20 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cao, J.; Chen, J.; Wang, X.; Huang, W.; Chen, D.; Zhao, T.; Tu, W.; Li, Q. UrbanMMCL: Urban Region Representations via Multi-Modal and Multi-Graph Self-Supervised Contrastive Learning. *ISPRS J. Photogramm. Remote Sens.* **2026**, *232*, 75–93. [CrossRef]
2. Wang, Z.; Li, H.; Rajagopal, R. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2020; Volume 34, pp. 1013–1020. [CrossRef]
3. Li, Y.; Huang, W.; Cong, G.; Wang, H.; Wang, Z. Urban region representation learning with openstreetmap building footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2023; pp. 1363–1373. [CrossRef]
4. Wang, H.; Li, Z. Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*; ACM: New York, NY, USA, 2017; pp. 237–246. [CrossRef]
5. Zhang, M.; Li, T.; Li, Y.; Hui, P. Multi-view joint graph representation learning for urban region embedding. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 4431–4437. [CrossRef] [PubMed]
6. Chan, W.; Ren, Q. Region-wise attentive multi-view representation learning for urban region embedding. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*; ACM: New York, NY, USA, 2023; pp. 3763–3767. [CrossRef]
7. Sun, F.; Qi, J.; Chang, Y.; Fan, X.; Karunasekera, S.; Tanin, E. Urban region representation learning with attentive fusion. In *Proceedings of the 2024 IEEE 40th International Conference on Data Engineering (ICDE)*; IEEE: New York, NY, USA, 2024; pp. 4409–4421. [CrossRef]
8. Luo, Y.; Chung, F.L.; Chen, K. Urban region profiling via multi-graph representation learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*; ACM: New York, NY, USA, 2022; pp. 4294–4298. [CrossRef]
9. Zhang, L.; Long, C.; Cong, G. Region embedding with intra and inter-view contrastive learning. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 9031–9036. [CrossRef]

10. Li, Z.; Huang, W.; Zhao, K.; Yang, M.; Gong, Y.; Chen, M. Urban region embedding via multi-view contrastive prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2024; Volume 38, pp. 8724–8732. [CrossRef]
11. Yan, Y.; Wen, H.; Zhong, S.; Chen, W.; Chen, H.; Wen, Q.; Zimmermann, R.; Liang, Y. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference*; Association for Computing Machinery: New York, NY, USA, 2024; Volume 2024, pp. 4006–4017. [CrossRef]
12. Jin, J.; Song, Y.; Kan, D.; Zhu, H.; Sun, X.; Li, Z.; Sun, X.; Zhang, J. Urban Region Pre-training and Prompting: A Graph-based Approach. In *KDD'25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, New York, NY, USA*; ACM: New York, NY, USA, 2025; pp. 1071–1082. [CrossRef]
13. Hao, X.; Chen, W.; Yan, Y.; Zhong, S.; Wang, K.; Wen, Q.; Liang, Y. Urbanvlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2025; Volume 39, pp. 28061–28069. [CrossRef]
14. Yao, Z.; Fu, Y.; Liu, B.; Hu, W.; Xiong, H. Representing urban functions through zone embedding with human mobility patterns. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018. [CrossRef] [PubMed]
15. Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; Ermon, S. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2019; Volume 33, pp. 3967–3974. [CrossRef]
16. Zhao, W.; Li, M.; Wu, C.; Zhou, W.; Chu, G. Identifying urban functional regions from high-resolution satellite images using a context-aware segmentation network. *Remote Sens.* **2022**, *14*, 3996. [CrossRef]
17. Huang, W.; Zhang, D.; Mai, G.; Guo, X.; Cui, L. Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 134–145. [CrossRef]
18. Pastorino, M.; Gallo, F.; Di Febbraro, A.; Moser, G.; Sacco, N.; Serpico, S.B. Multimodal fusion of mobility demand data and remote sensing imagery for urban land-use and land-cover mapping. *Remote Sens.* **2022**, *14*, 3370. [CrossRef]
19. Wang, J.; Feng, C.C.; Guo, Z. A novel graph-based framework for classifying urban functional zones with multisource data and human mobility patterns. *Remote Sens.* **2023**, *15*, 730. [CrossRef]
20. Fu, Y.; Wang, P.; Du, J.; Wu, L.; Li, X. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2019; Volume 33, pp. 906–913. [CrossRef]
21. Zhang, Y.; Fu, Y.; Wang, P.; Li, X.; Zheng, Y. Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM: New York, NY, USA, 2019; pp. 1700–1708. [CrossRef]
22. Zhou, S.; He, D.; Chen, L.; Shang, S.; Han, P. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2023; Volume 37, pp. 4981–4989. [CrossRef]
23. Sun, F.; Chang, Y.; Tanin, E.; Karunasekera, S.; Qi, J. FlexiReg: Flexible Urban Region Representation Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*; ACM: New York, NY, USA, 2025; pp. 2702–2713. [CrossRef]
24. Wang, Y.; Han, T.; Rui, L.; Ma, J.; Jin, Q. Ego-centric multiple-correlation and temporal graph neural networks based residential load forecasting. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111933. [CrossRef]
25. Yu, Z.; Li, W.; Yin, Z.; Hong, X.; Xiao, S.; Lu, S. Contextual structure knowledge transfer for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2025; Volume 39, pp. 22263–22271.
26. He, Y.; Sui, Y.; He, X.; Liu, Y.; Sun, Y.; Hooi, B. Unigraph2: Learning a unified embedding space to bind multimodal graphs. In *Proceedings of the ACM on Web Conference*; Association for Computing Machinery: New York, NY, USA, 2025; Volume 2025, pp. 1759–1770.
27. Xin, J.; Yun, S.; Peng, J.; Choi, I.; Ballard, J.L.; Chen, T.; Long, Q. I2MoE: Interpretable Multimodal Interaction-aware Mixture-of-Experts. *arXiv* **2025**, arXiv:2505.19190.
28. Wang, Y.; Yang, Z.; Che, X. A hierarchical mixture-of-experts framework for few labeled node classification. *Neural Netw.* **2025**, *188*, 107285. [CrossRef] [PubMed]
29. Zhu, J.; Zou, X.; Sun, J.; Luo, C.; Liu, L.; Zeng, L.; Zhang, N.; Wu, B.; Tang, C.; Dai, L. MoEGCL: Mixture of Ego-Graphs Contrastive Representation Learning for Multi-View Clustering. *arXiv* **2025**, arXiv:2511.05876.
30. NYC Government. NYC Open Data. Available online: https://opendata.cityofnewyork.us/ (accessed on 28 June 2025).
31. Chicago Government. Chicago Data Portal. Available online: https://data.cityofchicago.org/ (accessed on 28 June 2025).
32. OpenStreetMap. Available online: https://www.openstreetmap.org/ (accessed on 28 June 2025).
33. Google Maps. Available online: https://www.google.com/maps (accessed on 28 June 2025).
34. Foursquare. Available online: https://foursquare.com/ (accessed on 28 June 2025).

35. WorldPop. Available online: https://www.worldpop.org/ (accessed on 28 June 2025).
36. Wu, S.; Yan, X.; Fan, X.; Pan, S.; Zhu, S.; Zheng, C.; Cheng, M.; Wang, C. Multi-graph fusion networks for urban region embedding. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), Vienna, Austria, 23–29 July 2022; pp. 321–327. [CrossRef] [PubMed]
37. Li, Z.; Jia, H.; Zhao, K.; Huang, W.; Chen, M. 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*; ACM: New York, NY, USA, 2026; pp. 748–758.
38. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.