GraphJCL: A Dual-Perspective Graph-Based Framework for Urban Region Representation via Joint Contrastive Learning

Yaya Zhao[1], Kaiqi Zhao[2], Zixuan Tang[1], Xiaoling Lu[1]✉, Yuanyuan Zhang[3], and Yalei Du[3]

[1] Center for Applied Statistics, School of Statistics, Renmin University of China {zhaoyaya, 2021201741, xiaolinglu}@ruc.edu.cn
[2] The University of Auckland kaiqi.zhao@auckland.ac.nz
[3] Beijing Baixingkefu Network Technology Co., Ltd. zhang.huanzhiyuan@gmail.com yaleidu@163.com

Abstract. Graph learning for urban region modeling has gained significant attention for leveraging multi-modal data to generate region representations for downstream task prediction. However, existing models face two key limitations: (1) they primarily adopt a global perspective, overlooking the joint modeling of both local and global aspects, and (2) they rely on redundant, low-information nodes, leading to suboptimal region representations. To address these challenges, we propose GraphJCL, a dual-perspective framework that models both local and global perspectives. Specifically, GraphJCL first constructs local graphs for individual regions and a global graph encompassing all regions, integrating POI, taxi flow, remote sensing, street view, and road network data. Additionally, GraphJCL employs specialized message-passing mechanisms to efficiently capture both local and global graph node representations. Furthermore, GraphJCL incorporates entropy-optimized graph node pruning, retaining only the most informative nodes to enhance final region representations. To ensure the effectiveness of the designed dual-perspective graph framework, GraphJCL introduces a joint contrastive learning approach, optimizing region representations through geography-driven, entropy-optimized, and mutual information-based optimization techniques. Extensive experiments on two real-world datasets across five modalities demonstrate that GraphJCL consistently outperforms state-of-the-art methods on three tasks, validating its flexibility and effectiveness.

Keywords: Urban region representation · Graph neural networks · Joint contrastive learning.

1 Introduction

Graph learning [7,18,32] for urban region representation leverages multi-modal data, including Points of Interest (POI), taxi flow, remote sensing imagery, street view imagery, road network data, and socioeconomic indicators, to generate embeddings that effectively capture cross-modal relationships and semantic structures. These embeddings facilitate various downstream tasks, such as check-in prediction [9], crime fore-

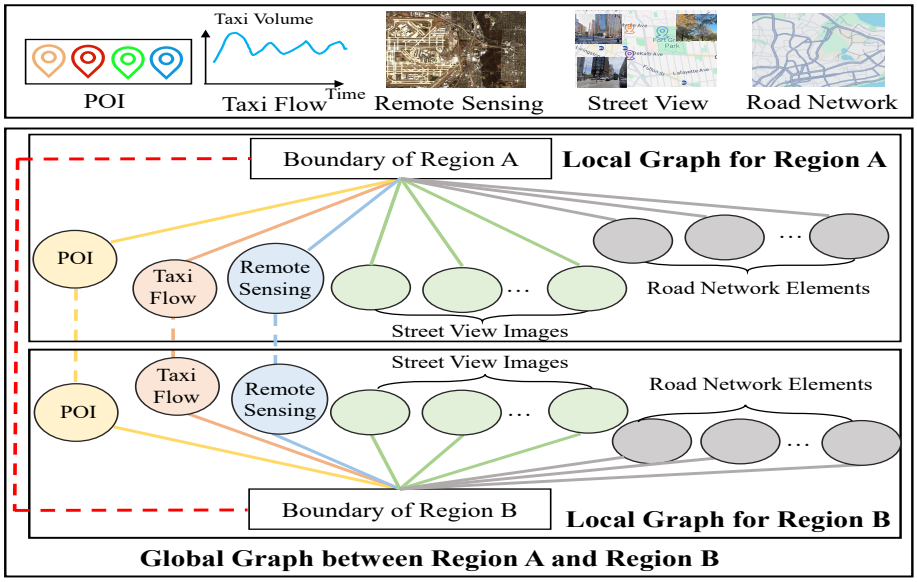

Fig. 1. An illustration of local and global graph construction using two regions. The local graph for each region is modeled independently, containing only nodes and intra-region edges within itself. In contrast, the global graph spans both regions, incorporating nodes from both, intra-region edges (solid lines), and inter-region coarse-grained modality connections edges (dashed lines).

casting [19], and traffic crash prediction [5], thereby supporting smart urban optimization [11]. Despite the effectiveness of existing graph learning methods, two key limitations persist. First, **neglecting joint region modeling of local and global aspects.** Local region modeling integrates multi-modal data within a specific region to capture localized information. However, global modeling simultaneously incorporates both intra-region information and inter-region interactions but does not explicitly model individual regions in isolation. Ideally, local and global modeling should work in tandem to produce a comprehensive urban region representation. However, existing methods predominantly focus on global modeling while often neglecting the independent local modeling of individual regions. For instance, methods such as [2,29,30,32] construct a global heterogeneous graph where region representations are optimized jointly, without an independent process for learning region-specific embeddings. Second, **reliance on redundant, low-information nodes.** Many existing methods rely on excessively redundant, low-information nodes to generate final region representations, which negatively impacts representation quality. For example, [28] directly averages the representations of all modality nodes to obtain the final region representation, while [7] averages the representations of all modality nodes and modality-type nodes to produce the final representation. These methods fail to prune low-information nodes, resulting in graphs that include numerous redundant nodes. This lack of refinement leads to inefficient representation learning and ultimately hampers the overall performance of final representations.

To address these limitations, we propose GraphJCL, a dual-perspective graph-based framework that models both local and global perspectives. It integrates five modalities: POI, taxi flow, remote sensing imagery, street view imagery, and road network data, categorized into coarse-grained and fine-grained types based on their characteristics. (1) Coarse-grained modalities capture region-level urban functions by aggregating raw data into representative vectors. POI data is clustered by category to reflect commercial activity, taxi flow is aggregated over time to represent mobility patterns, and remote sensing imagery provides a macro-level view of land use and urban structure. (2) Fine-grained modalities capture spatial topology and local environmental structures. Street view imagery consists of diverse location-specific images, while road network data represents individual road segments with distinct positions and attributes.

Based on this modality classification, GraphJCL constructs both local graphs for individual regions and a global graph encompassing all regions, as illustrated in Fig. 1. Specifically: (1) Local graphs connect each region node to its aggregated POI, taxi flow, and remote sensing imagery vector nodes, along with multiple street view images nodes and road network elements nodes, capturing both functional and spatial characteristics. (2) The global graph establishes inter-region edges only between coarse-grained region nodes (POI, taxi flow, remote sensing) and region boundary nodes, while fine-grained nodes remain unconnected to prevent edge explosion and unnecessary computational overhead. This structure ensures effective inter-region interactions while keeping the graph compact and efficient. Additionally, GraphJCL introduces specialized message-passing mechanisms to capture local and global graph node representations effectively. Furthermore, GraphJCL incorporates entropy-optimized graph node pruning to retain high-information nodes while eliminating redundant ones, ensuring the generation of effective region representations. Finally, to ensure that our designed dual-perspective graph framework functions effectively and draws inspiration from contrastive learning [14,25], GraphJCL employs a joint contrastive learning approach to optimize local region representations from three views, ultimately generating the final region representation. Specifically, it refines region representations using geography-driven and entropy-optimized techniques and integrates global region representations through mutual information-based optimization. These three aspects work together to collectively enhance region representations. In summary, our key contributions are as follows:

- We propose GraphJCL, a dual-perspective graph-based framework, as the first to jointly model both local and global perspectives for urban region representation.
- GraphJCL constructs local graphs for individual regions and a global graph encompassing all regions. It employs tailored message-passing mechanisms to effectively capture both local and global node representations, enabling joint modeling of regional structures. Additionally, it integrates graph node pruning and attention mechanisms to derive more efficient and informative region representations.
- GraphJCL introduces a joint contrastive learning approach, incorporating geography-driven and entropy-optimized contrastive learning techniques, as well as mutual information-based optimization, to refine and enhance region representations.
- Extensive experiments on two real-world datasets spanning five modalities demonstrate that GraphJCL outperforms state-of-the-art methods across three downstream tasks, highlighting its flexibility and effectiveness.

2 Related Work

Graph Representation for Multi-Modal Data Graph embedding for multi-modal data aims to learn low-dimensional vector representations of graph nodes from diverse data sources. Recent advances in graph neural networks (GNNs) have significantly improved these representations. For instance, HetCAN [31] enhances heterogeneous graph representation by incorporating both type-aware and dimension-aware encoders. These GNN-based approaches have gained substantial attention for their effectiveness in multi-modal data representation learning [10,21]. Early methods employed taxi flow patterns to define graph edges [18,23], whereas more recent studies integrate spatial and socio-environmental attributes into heterogeneous graphs for comprehensive urban modeling [7,32,33]. Additionally, graph contrastive learning has shown great potential in urban representation tasks [29]. These approaches highlight the effectiveness of graph-based multi-modal learning for urban region modeling.

Urban Region Representation Learning Urban region representation learning models can be categorized based on the number of modalities they utilize. Some approaches focus on single-modal data. For instance, [18] leverages taxi data to model urban region embeddings, capturing vehicle movement patterns to reflect the semantic characteristics of urban areas. Similarly, [6] primarily utilizes Points of Interest (POI) features for region representation learning, while [22] employs satellite imagery, leveraging large-scale models to enhance final region representations. Other studies adopt multi-modal data to construct richer and more comprehensive urban region representations [9,24,26,27,30]. For example, [27] introduces a multi-view joint learning framework that integrates taxi data, POI data, and check-in records, effectively capturing cross-modal correlations to model urban regions from multiple perspectives. From a methodological perspective, some studies employ attention mechanisms for modality fusion [8,15,20], while others utilize graph-based approaches for urban region representation learning [7,32]. Our framework utilizes five modalities in a graph-based structure with contrastive learning, effectively capturing effective urban region representations.

3 Preliminaries & Problem Statement

Definition 1 (Urban Regions ($\mathcal{U}$)). *An urban area is partitioned into N non-overlapping regions, denoted as $\mathcal{U} = \{U_1, U_2, \ldots, U_N\}$. Each U_i has a geographical boundary.*

Definition 2 (Coarse-Grained Modalities ($\mathcal{CM}$)). *A modality is considered coarse-grained if it captures region-level urban functional attributes, allowing its data within a region to be aggregated into a single vector representing overall urban functionality.* ***The coarse-grained modalities $\mathcal{CM}$ used are as follows:***

- ***Point-of-Interest (POI) ($\mathcal{P}$):*** *POI data reflects the commercial activity of a region. Each region U_i is associated with a vector $P_i = \{P_{i1}, P_{i2}, \ldots, P_{iK}\}$, where P_{ij} denotes the count of POIs in category j within region U_i. The K represents the top K POI categories with the highest occurrence frequency across all regions.*

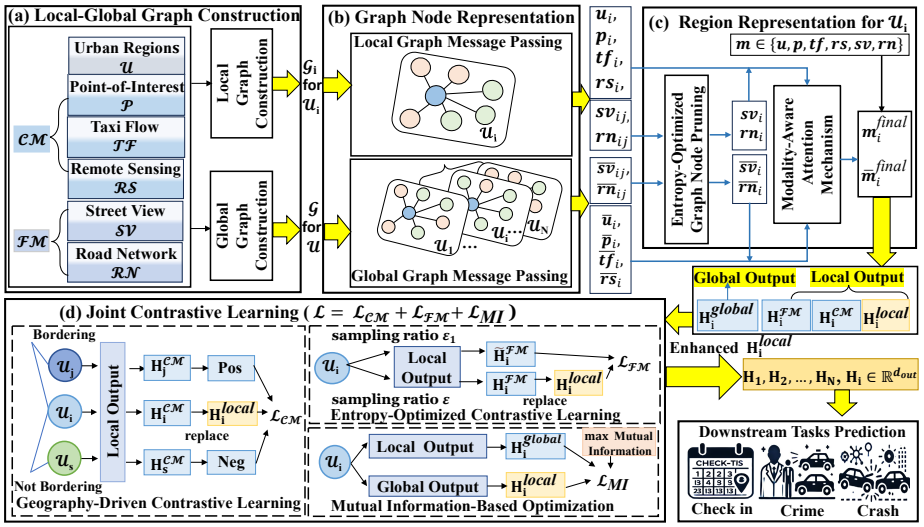

Fig. 2. The framework: (a) GraphJCL constructs local graphs for individual regions and a global graph for all regions. (b) GraphJCL introduces tailored message-passing mechanisms to capture local and global graph node representations. (c) GraphJCL employs entropy-optimized graph node pruning and modality-aware attention mechanisms to derive effective region representations. (d) Geography-driven, entropy-optimized, and mutual information-based contrastive techniques jointly enhance local region representations to generate the final region representations.

- ***Taxi Flow*** *($\mathcal{TF}$): Taxi flow captures mobility patterns within a region. Each region U_i is associated with a vector $TF_i = \{TF_{i1}, TF_{i2}, \ldots, TF_{iT}\}$, where TF_{ij} represents the taxi flow count in region U_i during the j-th time period. The T represents the total number of periods (e.g., hours in a month), depending on the dataset.*
- ***Remote Sensing Imagery*** *($\mathcal{RS}$): Remote sensing imagery provides a macro-level view of land use and urban structure within a region. Each region U_i is associated with a remote sensing image RS_i.*

Definition 3 (Fine-Grained Modalities ($\mathcal{FM}$)). *A modality is considered fine-grained if it captures spatial topology and local environmental structures within a region, meaning that its data points exhibit significant variability and cannot be effectively aggregated into a single vector representation.*
The fine-grained modalities $\mathcal{FM}$ used are as follows:

- ***Street View Imagery*** *($\mathcal{SV}$): Street view imagery captures local visual characteristics of urban regions through numerous spatially distributed images. Each region U_i contains a collection of street view images, denoted as $\{SV_{i1}, SV_{i2}, \ldots, SV_{i|\mathcal{SV}_i|}\}$, where SV_{ij} represents the street view image with index j within region U_i.*
- ***Road Network*** *($\mathcal{RN}$): The road network captures the topological structure of urban regions. Each region U_i contains a collection of road network elements, denoted as $\{RN_{i1}, RN_{i2}, \ldots, RN_{i|\mathcal{RN}_i|}\}$, where RN_{ij} represents a road segment*

or junction with index j within region U_i. The category index RNS_{ij} specifies the type of road element (e.g., trunk road or motorway).

Definition 4 (Downstream Tasks (Y)). *Downstream tasks refer to socio-economic and environmental indicators in urban contexts. For urban area $\mathcal{U}$ with$N$ regions , the L task targets are represented as $Y \in \mathbb{R}^{N \times L}$. This paper focuses on three downstream tasks, check-in counts, crime incidents, and traffic crash counts, where $L = 3$.*

Definition 5 (Problem Statement: Urban Region Representation). *Given urban regions $\mathcal{U}$ along with their associated coarse-grained modalities $\mathcal{CM}$ and fine-grained modalities $\mathcal{FM}$, the objective is to learn a low-dimensional representation $H_i \in \mathbb{R}^{d_{out}}$ for each region U_i. These embeddings $\mathcal{H} = \{H_1, H_2, \ldots, H_N\}$ are used to predict downstream tasks Y.*

4 Methodology

In this section, we introduce GraphJCL, a dual-perspective graph-based framework for urban region representation. Its key components are illustrated in Fig. 2.

4.1 Local-Global Graph Construction

In this subsection, to achieve joint region modeling of local and global aspects, we construct a local graph for each individual region and a global graph for all regions.

Local Graph Construction For each region U_i, we construct a local heterogeneous graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$ to capture its intrinsic features without influence from other regions. Specifically, the node set $\mathcal{V}_i$ consists of six types of nodes:

- **Region**: A node U_i representing the geographic boundary of the region.
- **POI**: A node P_i representing aggregated POI data.
- **Taxi Flow**: A node TF_i representing aggregated taxi mobility patterns.
- **Remote Sensing**: A node RS_i capturing macro-level land use and urban structure.
- **Street View Imagery**: A set of nodes $\{SV_{i1}, SV_{i2}, \ldots, SV_{i|\mathcal{SV}_i|}\}$ representing localized street-level visual environmental features.
- **Road Network**: A set of nodes $\{RN_{i1}, RN_{i2}, \ldots, RN_{i|\mathcal{RN}_i|}\}$ representing the topological structure of road segments and junctions.

The edge set $\mathcal{E}_i$ consists of five types of edges, defining relationships between the region node U_i and other modality nodes:

- $\mathcal{U}\_has\_\mathcal{P}$: An edge (U_i, P_i) that connects the region node to the POI node.
- $\mathcal{U}\_has\_\mathcal{TF}$: An edge (U_i, TF_i) that connects the region node to the taxi node.
- $\mathcal{U}\_has\_\mathcal{RS}$: An edge (U_i, RS_i) that connects the region node to the remote sensing imagery node.
- $\mathcal{U}\_has\_\mathcal{SV}$: A set of edges (U_i, SV_{ij}), where $j = 1, 2, \ldots, |\mathcal{SV}_i|$, that connect the region node to all street view images nodes.
- $\mathcal{U}\_has\_\mathcal{RN}$: A set of edges (U_i, RN_{ij}), where $j = 1, 2, \ldots, |\mathcal{RN}_i|$, that connect the region node to all road network elements nodes.

Based on the local graph construction method above, N local graphs $\mathcal{G}_i$, where $i = 1, 2, \ldots, N$, can be constructed for urban regions $U_1, U_2, \ldots, U_N$, respectively.

Global Graph Construction For all regions $U_i \in \mathcal{U}, i = 1, 2, \ldots, N$, we construct a global graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which includes all nodes and edges from the local graphs $\mathcal{G}_1, \ldots, \mathcal{G}_N$. Additionally, it introduces four extra types of edges that capture inter-region relationships, based on region nodes $\mathcal{U}$ and coarse-grained modalities: POI ($\mathcal{P}$), Taxi Flow ($\mathcal{TF}$), and Remote Sensing ($\mathcal{RS}$). Fine-grained modalities Street View ($\mathcal{SV}$) and Road Network ($\mathcal{RN}$), do not establish inter-region edges, as they primarily capture local environmental and topological features within a region. The four types are:

- $\mathcal{U}\_sim\_\mathcal{U}$: A set of edges (U_i, U_j) connecting region nodes whose geographical boundaries are adjacent, where $i, j = 1, 2, \ldots, N$ and $i \neq j$.
- $\mathcal{P}\_sim\_\mathcal{P}$: A set of edges (P_i, P_j) connecting POI nodes based on cosine semantic similarity across regions, where $i, j = 1, 2, \ldots, N$ and $i \neq j$.
- $\mathcal{TF}\_sim\_\mathcal{TF}$: A set of edges (TF_i, TF_j) connecting taxi flow nodes based on cosine mobility similarity across regions, where $i, j = 1, 2, \ldots, N$ and $i \neq j$.
- $\mathcal{RS}\_sim\_\mathcal{RS}$: A set of edges (RS_i, RS_j) connecting remote sensing nodes based on cosine satellite similarity across regions, where $i, j = 1, 2, \ldots, N$ and $i \neq j$.

4.2 Graph Node Representation

This subsection introduces the initialization of node representations for different modalities, followed by our designed message passing mechanisms for iterative updates.

Node initialization for each modality

- **Region ($\mathcal{U}$)**: An undirected graph is constructed with all region geographic boundary U_i, i = 1,2,...N and edges connecting adjacent regions. The initialization vector $\boldsymbol{u}_i^{(0)} \in \mathbb{R}^{d_{in}}$ for U_i is derived using the Node2Vec algorithm [3].
- **Remote Sensing Imagery ($\mathcal{RS}$)**: The initialization vector $\boldsymbol{rs}_i^{(0)} \in \mathbb{R}^{d_{in}}$ for each RS_i is obtained by applying the EfficientNet-B4 model [16].
- **Point-of-Interest (POI) ($\mathcal{P}$)**: The initialization vector $\boldsymbol{p}_i^{(0)} \in \mathbb{R}^{d_{in}}$ for each P_i is generated by setting $K = d_{in}$.
- **Taxi Flow ($\mathcal{TF}$)**: The initialization vector $\boldsymbol{tf}_i^{(0)} \in \mathbb{R}^{d_{in}}$ for each TF_i is computed by setting $T = d_{in}$.
- **Street View Imagery ($\mathcal{SV}$)**: The initialization vector $\boldsymbol{sv}_{i,j}^{(0)} \in \mathbb{R}^{d_{in}}$ is obtained for each street view image $SV_{i,j}$, where $j = 1, 2, \ldots, |\mathcal{SV}_i|$, by leveraging the CLIP-ViT-B/32 model [13].
- **Road Network ($\mathcal{RN}$)**: The initialization vector $\boldsymbol{rn}_{i,j}^{(0)} \in \mathbb{R}^{d_{in}}$ is generated for each road network element $RN_{i,j}$, where $j = 1, 2, \ldots, |\mathcal{RN}_i|$, by employing an embedding technique [12].

Local Graph Message Passing Mechanism For each local graph $\mathcal{G}_i$ in region U_i, the nodes include the region node U_i, POI node P_i, taxi flow node TF_i, remote sensing node RS_i, street view node SV_{ij}, and road network node RN_{ij}. Their node representations at the l-th layer are denoted as $\boldsymbol{u}_i^{(l)}$, $\boldsymbol{p}_i^{(l)}$, $\boldsymbol{tf}_i^{(l)}$, $\boldsymbol{rs}_i^{(l)}$, $\boldsymbol{sv}_{ij}^{(l)}$, and $\boldsymbol{rn}_{ij}^{(l)}$, respectively, with initial embeddings defined as previously described. To capture region-specific features without interference from other regions, message passing in the local

graph occurs only between nodes within the same region. Specifically, the node representations are updated at the l-th layer according to the following GNN update rule:

$$\begin{aligned}
\boldsymbol{u}_i^{(l)} &= \sigma\left(\sum_{m\in\{p,tf,rs\}} W_m \boldsymbol{m}_i^{(l-1)} + \sum_{m\in\{sv,rn\}} \sum_{j=1}^{|\mathcal{M}_i|} W_m \boldsymbol{m}_{ij}^{(l-1)}\right), \\
\boldsymbol{m}_i^{(l)} &= \sigma\left(W_u \boldsymbol{u}_i^{(l-1)}\right), \quad m \in \{p, tf, rs\}, \\
\boldsymbol{m}_{ij}^{(l)} &= \sigma\left(W_u \boldsymbol{u}_i^{(l-1)}\right), \quad m \in \{sv, rn\}, \quad j = 1, 2, \ldots, |\mathcal{M}_i|,
\end{aligned} \tag{1}$$

where σ is the activation function, and W_m and W_u are learnable. The final layer outputs are denoted as $\boldsymbol{u}_i$, $\boldsymbol{p}_i$, $\boldsymbol{tf}_i$, $\boldsymbol{rs}_i$, $\boldsymbol{sv}_{ij}$, and $\boldsymbol{rn}_{ij}$, all of which belong to $\mathbb{R}^{d_{in}}$.

Global Graph Message Passing Mechanism For the global graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the nodes in region U_i include the region node U_i, POI node P_i, taxi flow node TF_i, remote sensing node RS_i, street view node SV_{ij}, and road network node RN_{ij}. The node representations at the l-th layer are $\overline{\boldsymbol{u}}_i^{(l)}$, $\overline{\boldsymbol{p}}_i^{(l)}$, $\overline{\boldsymbol{tf}}_i^{(l)}$, $\overline{\boldsymbol{rs}}_i^{(l)}$, $\overline{\boldsymbol{sv}}_{ij}^{(l)}$, and $\overline{\boldsymbol{rn}}_{ij}^{(l)}$. Each vector lies in $\mathbb{R}^{d_{\text{in}}}$ and is initialized as previously described. The global graph captures region representations through interactions across regions. Message passing between regions occurs only for coarse-grained modalities, while fine-grained modalities focusing on local topology do not exchange information across regions. The sets $\mathcal{N}_i^U$ and $\mathcal{N}_i^m$ denote neighboring region nodes and same-modality neighboring nodes across regions, respectively, and σ is the activation function. Specifically, the node representations are updated at the l-th layer according to the following GNN update rule:

$$\begin{aligned}
\overline{\boldsymbol{u}}_i^{(l)} &= \sigma\left(\sum_{m\in\{p,tf,rs\}} \overline{W}_m \overline{\boldsymbol{m}}_i^{(l-1)} + \sum_{m\in\{sv,rn\}} \sum_{j=1}^{|\mathcal{M}_i|} \overline{W}_m \overline{\boldsymbol{m}}_{ij}^{(l-1)} + \sum_{j\in\mathcal{N}_i^U} \overline{W}_{uu} \overline{\boldsymbol{u}}_j^{(l-1)}\right), \\
\overline{\boldsymbol{m}}_i^{(l)} &= \sigma\left(\overline{W}_u \overline{\boldsymbol{u}}_i^{(l-1)} + \sum_{j\in\mathcal{N}_i^m} \overline{W}_{mm} \overline{\boldsymbol{m}}_j^{(l-1)}\right), \quad m \in \{p, tf, rs\}, \\
\overline{\boldsymbol{m}}_{ij}^{(l)} &= \sigma\left(\overline{W}_u \overline{\boldsymbol{u}}_i^{(l-1)}\right), \quad m \in \{sv, rn\}, \quad j = 1, 2, \ldots, |\mathcal{M}_i|,
\end{aligned} \tag{2}$$

where $\overline{W}_m$, $\overline{W}_{uu}$, $\overline{W}_u$, and $\overline{W}_{mm}$ are learnable weight matrices. The final layer outputs are denoted as $\overline{\boldsymbol{u}}_i$, $\overline{\boldsymbol{p}}_i$, $\overline{\boldsymbol{tf}}_i$, $\overline{\boldsymbol{rs}}_i$, $\overline{\boldsymbol{sv}}_{ij}$, and $\overline{\boldsymbol{rn}}_{ij}$, all of which belong to $\mathbb{R}^{d_{in}}$.

4.3 Entropy-Optimized and Multi-Modal Region Representation

To eliminate redundant low-information nodes, we propose an entropy-optimized graph node pruning mechanism. Specifically, we compute each node's entropy based on its graph node representation. High-entropy nodes, which carry more informative content, are preserved, while low-entropy redundant nodes are discarded. To obtain region representations, we design a modality-aware attention mechanism to dynamically adjust each modality's contribution, ensuring an effective multi-modal region representation.

Entropy-Optimized Graph Node Pruning Mechanism To enhance efficiency and reduce redundancy, we perform entropy-optimized graph node pruning on both the local and global graphs. **(1)** In the local graph, $\mathcal{G}_i$, a large number of street view images nodes SV_{ij} are present, with each node represented as $\boldsymbol{sv}_{ij} \in \mathbb{R}^{din}$. Here, $j = 1, 2, \ldots, |\mathcal{M}_i|$. To assess the informativeness of each node SV_{ij}, we first normalize its feature vector as $\boldsymbol{sv}_{ij,f}^{\text{norm}} = \frac{|\boldsymbol{sv}_{ij,f}|}{\sum_{f=1}^{d_{in}} |\boldsymbol{sv}_{ij,f}|}$ to ensure proportional scaling. We then compute its entropy $H(SV_{ij})$ and sampling probability $P(SV_{ij})$ as:

$$H(SV_{ij}) = -\sum_{f=1}^{d_{in}} \boldsymbol{sv}_{ij,f}^{\text{norm}} \log(\boldsymbol{sv}_{ij,f}^{\text{norm}}), \quad P(SV_{ij}) = H(SV_{ij}) / \sum_{j=1}^{|\mathcal{SV}_i|} H(SV_{ij}). \tag{3}$$

Then existence of node SV_{ij} follows a Bernoulli distribution, $SV_{ij} \sim \text{Bern}(P(SV_{ij}))$. Nodes with higher entropy values, and consequently higher $P(SV_{ij})$, are more likely to be retained, while low-entropy nodes are pruned. The retained nodes are selected based on the *sampling ratio* $\boldsymbol{\varepsilon}$. The sampled street view node representations, where each is $\boldsymbol{sv}_{ij} \in \mathbb{R}^{d_{in}}$, are averaged to derive an embedding $\boldsymbol{sv}_i \in \mathbb{R}^{d_{in}}$ for region U_i. Similarly, an embedding for the road network is obtained as $\boldsymbol{rn}_i \in \mathbb{R}^{d_{in}}$.

(2) In the global graph, $\mathcal{G}$, a large number of fine-grained street view images nodes and road network elements nodes are present in the region U_i. Applying the same pruning strategy as in the local graph, we retain high-entropy nodes and average their representations to obtain the final embeddings $\overline{\boldsymbol{sv}}_i, \overline{\boldsymbol{rn}}_i \in \mathbb{R}^{d_{in}}$ for region U_i.

Modality-Aware Attention Mechanism To effectively capture the varying contributions of different modalities, we apply a modality-aware attention mechanism in both the local and global graphs. **(1)** In the local graph, $\mathcal{G}_i$ of region U_i, modality representations are denoted as $\boldsymbol{m}_i \in \mathbb{R}^{d_{in}}$, where $\boldsymbol{m} \in \{\boldsymbol{u}, \boldsymbol{p}, \boldsymbol{tf}, \boldsymbol{rs}, \boldsymbol{sv}, \boldsymbol{rn}\}$. Since different modalities contribute unequally to the final region representation, an attention weight vector $\boldsymbol{w}_{\boldsymbol{m}} \in \mathbb{R}^{d_{in}}$ is learned for each modality to capture its importance. The modality-specific representation is computed as $\boldsymbol{m}_i^{attn} = \sigma(\boldsymbol{w}_{\boldsymbol{m}} \odot \boldsymbol{m}_i)$, where $\odot$ denotes element-wise multiplication and σ is an activation function. To refine the information, a linear transformation [17] is applied, followed by an autoencoder [1] with a *hidden dimension* of d_{hid}, producing the final local modality representation $\boldsymbol{m}_i^{final} \in \mathbb{R}^{d_{out}}$.

(2) In the global graph $\mathcal{G}$, modality representations in region $\mathcal{U}_i$ are given by $\overline{\boldsymbol{m}_i} \in \mathbb{R}^{d_{in}}$, where $\boldsymbol{m} \in \{\boldsymbol{u}, \boldsymbol{p}, \boldsymbol{tf}, \boldsymbol{rs}, \boldsymbol{sv}, \boldsymbol{rn}\}$. Applying the same modality-aware attention mechanism, we derive the final global modality representation as $\overline{\boldsymbol{m}}_i^{final} \in \mathbb{R}^{d_{out}}$.

Local-Global Region Representation We compute coarse-grained, fine-grained, and overall local region representations $\boldsymbol{H}_i^{\mathcal{CM}}, \boldsymbol{H}_i^{\mathcal{FM}}, \boldsymbol{H}_i^{local} \in \mathbb{R}^{d_{out}}$ and global representation $\boldsymbol{H}_i^{global} \in \mathbb{R}^{d_{out}}$ as described below:

$$\begin{aligned}
\boldsymbol{H}_i^{\mathcal{X}} &= \text{Mean}\left(\{\boldsymbol{m}_i^{final} \mid \boldsymbol{m} \in \mathcal{X}\}\right), \quad \mathcal{X} \in \{\mathcal{CM}, \mathcal{FM}, local\}, \\
\mathcal{CM} &= \{\boldsymbol{p}, \boldsymbol{tf}, \boldsymbol{rs}\}, \quad \mathcal{FM} = \{\boldsymbol{sv}, \boldsymbol{rn}\}, \quad local = \{\boldsymbol{u}, \boldsymbol{p}, \boldsymbol{tf}, \boldsymbol{rs}, \boldsymbol{sv}, \boldsymbol{rn}\}, \\
\boldsymbol{H}_i^{global} &= \text{Mean}\left(\left\{\overline{\boldsymbol{m}}_i^{final} \mid \boldsymbol{m} \in \{\boldsymbol{u}, \boldsymbol{p}, \boldsymbol{tf}, \boldsymbol{rs}, \boldsymbol{sv}, \boldsymbol{rn}\}\right\}\right).
\end{aligned} \tag{4}$$

4.4 Joint Contrastive Learning

To effectively learn multi-modal region representations within our dual-perspective local and global graph framework, we introduce a joint contrastive learning approach that optimizes $\boldsymbol{H}_i^{local}$ from three views.

Geography-Driven Contrastive Learning with $\boldsymbol{H}_i^{\mathcal{CM}}$ The coarse-grained local region representation $\boldsymbol{H}_i^{\mathcal{CM}}$ captures functional attributes with strong geographical continuity. Since neighboring regions exhibit similar coarse-grained representations, we apply contrastive learning by treating geographically adjacent regions, such as U_j, as positive samples and non-adjacent regions, such as U_s, as negative samples. To refine the final region representation, we replace $\boldsymbol{H}_i^{\mathcal{CM}}$ with $\boldsymbol{H}_i^{local}$,γ is a margin hyperparameter and $\|\cdot\|_2$ represents the L2 norm, the contrastive loss is defined as follows:

$$\mathcal{L}_{\mathcal{CM}} = \max\left(\|\boldsymbol{H}_i^{local} - \boldsymbol{H}_j^{\mathcal{CM}}\|_2 - \|\boldsymbol{H}_i^{local} - \boldsymbol{H}_s^{\mathcal{CM}}\|_2 + \gamma, 0\right). \tag{5}$$

Entropy-Optimized Contrastive Learning with $\boldsymbol{H}_i^{\mathcal{FM}}$ The fine-grained local region representation $\boldsymbol{H}_i^{\mathcal{FM}}$ captures spatial topology and local environmental structures. Due to high variability across regions, all regions except the target are treated as negative samples, with positive samples drawn from the region's own fine-grained data. During graph node pruning, a different *sampling ratio* ε_1. is used to obtain $\tilde{\boldsymbol{H}}_i^{\mathcal{FM}}$, which serves as a positive sample. To refine the final representation, $\boldsymbol{H}_i^{local}$ replaces $\boldsymbol{H}_i^{\mathcal{FM}}$. τ is the temperature hyperparameter. The entropy-optimized contrastive loss is:

$$\begin{aligned}\mathcal{L}_{\mathcal{FM}} = \sum_{i=1}^{N}\Bigg[&- \log\exp\left(\frac{\boldsymbol{H}_i^{local}\cdot\tilde{\boldsymbol{H}}_i^{\mathcal{FM}}}{\tau}\right) \\ &+ \log\left(\exp\left(\frac{\boldsymbol{H}_i^{local}\cdot\tilde{\boldsymbol{H}}_i^{\mathcal{FM}}}{\tau}\right) + \sum_{j\neq i}\exp\left(\frac{\boldsymbol{H}_i^{local}\cdot\boldsymbol{H}_j^{local}}{\tau}\right)\right)\Bigg],\end{aligned} \tag{6}$$

Mutual Information-Based Optimization with $\boldsymbol{H}_i^{global}$ The global region representation $\boldsymbol{H}_i^{global}$ and the local region representation $\boldsymbol{H}_i^{local}$ capture complementary perspectives from the same regional data. To integrate global information into $\boldsymbol{H}_i^{local}$, we maximize the mutual information $\mathcal{I}(\boldsymbol{H}_i^{local}, \boldsymbol{H}_i^{global}) = \mathbb{E}\left[\log\frac{p_i^{l,g}}{p_i^l p_i^g}\right]$, where $p_i^{l,g}$, p_i^l, and p_i^g represent the joint and marginal distributions of the local and global representations, respectively. The mutual information is optimized by minimizing the following contrastive loss, with λ as a hyperparameter:

$$\mathcal{L}_{MI} = -\sum_i p_i^{l,g}\left[\log p_i^{l,g} - \lambda\left(\log p_i^l + \log p_i^g\right)\right]. \tag{7}$$

Table 1. Dataset statistics. * denotes model training data, and # denotes downstream task data.

	$\mathcal{U}^*$	$\mathcal{P}^*$	$\mathcal{TF}^*$	$\mathcal{RS}^*$	$\mathcal{SV}^*$	$\mathcal{RN}^*$	Crime#	Crash#	Check-in#
NYC	180	20872	848006	180	20698	13200	77701	14564	329153
CHI	77	36963	922746	77	36179	56005	91252	109645	167222

Training. The final objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{CM}} + \mathcal{L}_{\mathcal{FM}} + \mathcal{L}_{MI}. \tag{8}$$

The enhanced local region representation, $\boldsymbol{H}_i^{local}$, learned jointly through $\mathcal{L}$, serve as the final region embeddings $\boldsymbol{H}_i \in \mathbb{R}^{dout}$. $\mathcal{H} = \boldsymbol{H}_1, \boldsymbol{H}_2, \ldots, \boldsymbol{H}_N$ corresponds to the urban regions $\mathcal{U} = U_1, U_2, \ldots, U_N$ and is used for predicting downstream tasks Y.

5 Evaluation

5.1 Experimental Setup

Datasets and Metrics. We conduct experiments on two real-world urban datasets from New York City (NYC) and Chicago (CHI) [7,24,28], which provide region-level information $\mathcal{U}$ along with five heterogeneous modalities: Points of Interest (POI) $\mathcal{P}$, taxi flow $\mathcal{TF}$, remote sensing imagery $\mathcal{RS}$, street view data $\mathcal{SV}$, and road network topology $\mathcal{RN}$. These multimodal datasets support three representative downstream tasks: check-in prediction, crime forecasting, and traffic crash prediction. The detailed dataset statistics are summarized in Table 1. To comprehensively evaluate prediction performance, we adopt three widely used metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These metrics jointly assess both the accuracy and robustness of the model across different predictive tasks.

Baselines. To evaluate the performance of GraphJCL, we compare it with seven baseline models: **i)** GraphST [29], a spatiotemporal graph learning model designed for self-supervised learning; **ii)** ReCP [9], a pipeline for consistent representation learning across diverse views; **iii)** HREP [32], a framework leveraging heterogeneous region embedding with prompt learning; **iv)** HAFusion [15], which applies a dual-feature attentive fusion module to capture higher-order correlations within and across region features; **v)** UrbanVLP [4], integrating multi-granularity macro (satellite) and micro (street-view) information; **vi)** MuseCL [24], a multi-semantic contrastive learning framework for fine-grained urban region profiling; and **vii)** GURPP [7], a graph-based urban region pretraining and prompting framework for improved representation learning.

Parameter Settings For experiments in NYC and CHI, the input and output dimensionality (d_{in} and d_{out}) of all modalities is consistently set to 168. Models for both cities are optimized using Adam optimizer with a learning rate of 0.001 and weight decay of 0.01, and trained for a maximum of 50 epochs. The batch size is configured as 180 for

NYC and 77 for CHI. The hyperparameters for contrastive learning are as follows: the margin hyperparameter γ and the temperature hyperparameter τ are both set to 2 and 0.1, respectively, for both cities. The parameter λ is set to 8 for NYC and 7 for CHI.

5.2 Overall Performance

Table 2 presents a performance comparison of baseline models across three tasks in two cities. Since different region representation models utilize different modality sets $\mathcal{UM}$, a total of ten modality sets are employed in the regional representation models reviewed in this paper, as detailed in the header of Table 2. Among these, $\mathcal{UM}_1$, $\mathcal{UM}_2$, $\mathcal{UM}_3$, and $\mathcal{UM}_4$ correspond to modality sets used by existing models, while the remaining six are novel modality sets introduced by our model. **(1) High Flexibility and Effectiveness:** On existing modality sets, GraphJCL outperforms baseline models across all tasks and cities, highlighting its effectiveness in capturing regional representations from both local and global perspectives. Moreover, its support for ten modality sets demonstrates superior flexibility. **(2) Adaptability to Novel Modality Sets:** GraphJCL consistently performs well across all ten modality sets, proving its adaptability to diverse data distributions and robustness with novel modality sets. **(3) Effective Modality Information Utilization:** The modality sets ($\mathcal{UM}_i$, $i = 3, 4, 5, 6, 7$) are derived by sequentially removing $\mathcal{RN}$, $\mathcal{SV}$, $\mathcal{P}$, $\mathcal{TF}$, or $\mathcal{RS}$ from the full set $(\mathcal{P}, \mathcal{TF}, \mathcal{RS}, \mathcal{SV}, \mathcal{RN})$. The performance decline of GraphJCL when excluding any modality underscores its capability to effectively leverage each modality and integrate their distinct information features.

5.3 Ablation Study

To assess the contributions of different modules in our model, we evaluate five variants of **GraphJCL**: i) **w/o NP**: Removes entropy-optimized graph node pruning. ii) **w/o MA**: Removes the modality-aware attention mechanism. iii) **w/o** $\mathcal{L}_{\mathcal{CM}}$: Removes geography-driven contrastive learning. iv) **w/o** $\mathcal{L}_{\mathcal{FM}}$: Removes entropy-optimized contrastive learning. v) **w/o** $\mathcal{L}_{MI}$: Removes mutual information-based optimization. Experimental results are shown in Table 3. The results show that **GraphJCL** outperforms all variants, highlighting the importance of its key modules. Notably, removing entropy-optimized local graph node pruning (**w/o NP**) causes the largest performance drop, emphasizing the importance of selecting high mutual-information nodes. Omitting the modality-aware attention mechanism (**w/o MA**) also reduces performance, underlining the need for modality weighting in final region representations. Finally, removing any of the contrastive learning components (**w/o** $\mathcal{L}_{\mathcal{CM}}$, **w/o** $\mathcal{L}_{\mathcal{FM}}$, or **w/o** $\mathcal{L}_{MI}$) leads to performance degradation, highlighting the significance of joint contrastive learning.

5.4 Hyper-parameter Study

We conduct a detailed analysis of two key hyper-parameters that significantly influence the model's performance: the *sampling ratio* ε used in the entropy-optimized graph node pruning module, and the *hidden dimension* d_{hid} employed in the modality-aware attention mechanism, both introduced in subsection 4.3. The effects of these hyper-parameters are illustrated in Fig. 3, where model performance is evaluated using the

Table 2. Performance comparison of baseline models across three tasks in two cities. $\mathcal{UM}$ represents the corresponding modality set used by each model. A total of 10 modality sets are utilized across all models, as listed below: $\mathcal{TM} = (\mathcal{P}, \mathcal{TF}, \mathcal{RS}, \mathcal{SV}, \mathcal{RN})$, $\mathcal{UM}_1 = (\mathcal{P}, \mathcal{TF})$, $\mathcal{UM}_2 = (\mathcal{RS}, \mathcal{SV})$, $\mathcal{UM}_3 = \mathcal{TM} \setminus \mathcal{RN}$ (i.e., $\mathcal{TM}$ excluding $\mathcal{RN}$), $\mathcal{UM}_4 = \mathcal{TM} \setminus \mathcal{SV}$ (i.e., $\mathcal{TM}$ excluding $\mathcal{SV}$), $\mathcal{UM}_5 = \mathcal{TM} \setminus \mathcal{P}$ (i.e., $\mathcal{TM}$ excluding $\mathcal{P}$), $\mathcal{UM}_6 = \mathcal{TM} \setminus \mathcal{TF}$ (i.e., $\mathcal{TM}$ excluding $\mathcal{TF}$), $\mathcal{UM}_7 = \mathcal{TM} \setminus \mathcal{RS}$ (i.e., $\mathcal{TM}$ excluding $\mathcal{RS}$), $\mathcal{UM}_8 = (\mathcal{P}, \mathcal{TF}, \mathcal{RN})$, $\mathcal{UM}_9 = (\mathcal{RS}, \mathcal{SV}, \mathcal{RN})$.

Model	$\mathcal{UM}$	NYC								
		Check-in			Crime			Crash		
		MAE↓	RMSE↓	R2↑	MAE↓	RMSE↓	R2↑	MAE↓	RMSE↓	R2↑
GraphST[29]	$\mathcal{UM}_1$	1501.3	2838.2	0.284	316.04	496.41	0.103	30.855	42.633	0.162
ReCP[9]	$\mathcal{UM}_1$	823.4	1404.6	0.719	183.74	280.56	0.498	33.518	43.110	0.226
HREP[32]	$\mathcal{UM}_1$	1122.6	1745.4	0.566	201.69	308.86	0.391	26.859	34.483	0.505
HAFusion[15]	$\mathcal{UM}_1$	1051.7	1709.0	0.584	188.15	299.60	0.427	25.177	32.566	0.558
UrbanVLP[4]	$\mathcal{UM}_2$	1508.5	2652.6	0.286	195.30	404.53	0.395	38.421	50.695	0.183
MuseCL[24]	$\mathcal{UM}_3$	1479.3	2781.6	0.313	278.28	394.83	0.433	33.977	44.077	0.104
GURPP[7]	$\mathcal{UM}_4$	900.3	1356.0	0.738	194.64	280.50	0.498	30.194	41.764	0.274
GraphJCL	$\mathcal{UM}_1$	747.5	1269.6	0.771	165.33	251.80	0.595	26.987	34.771	0.497
	$\mathcal{UM}_2$	**610.1**	1072.8	0.836	167.19	270.28	0.534	27.672	35.732	0.468
	$\mathcal{UM}_3$	613.0	1117.1	0.822	162.73	257.18	0.578	26.106	34.869	0.495
	$\mathcal{UM}_4$	691.5	1248.9	0.778	158.35	246.95	0.611	26.861	36.399	0.448
	$\mathcal{UM}_5$	684.4	1142.5	0.814	164.67	258.01	0.575	23.199	30.542	0.612
	$\mathcal{UM}_6$	683.8	1135.7	0.816	170.37	275.84	0.514	22.498	29.430	0.639
	$\mathcal{UM}_7$	683.7	1103.7	0.827	176.39	276.10	0.513	22.891	29.979	**0.651**
	$\mathcal{UM}_8$	770.5	1320.5	0.752	172.18	266.48	0.546	23.027	29.908	0.628
	$\mathcal{UM}_9$	676.7	1142.9	0.814	177.99	264.27	0.554	24.481	31.482	0.587
	$\mathcal{TM}$	676.2	**1027.0**	**0.850**	**158.20**	**235.26**	**0.647**	**22.829**	**29.733**	0.632

Model	$\mathcal{UM}$	CHI								
		Check-in			Crime			Crash		
		MAE↓	RMSE↓	R2↑	MAE↓	RMSE↓	R2↑	MAE↓	RMSE↓	R2↑
GraphST[29]	$\mathcal{UM}_1$	2066.7	5281.3	0.545	566.79	721.12	0.381	548.625	761.611	0.464
ReCP[9]	$\mathcal{UM}_1$	1508.2	3448.3	0.581	322.41	445.39	0.763	384.472	555.310	0.701
HREP[32]	$\mathcal{UM}_1$	2139.5	4174.6	0.385	518.71	679.53	0.448	623.910	811.459	0.362
HAFusion[15]	$\mathcal{UM}_1$	1280.7	3424.5	0.586	578.57	758.15	0.312	655.469	899.352	0.217
UrbanVLP[4]	$\mathcal{UM}_2$	2524.5	5786.6	0.388	570.67	859.28	0.388	613.180	883.680	0.510
MuseCL[24]	$\mathcal{UM}_3$	2264.2	5944.2	0.424	624.35	858.21	0.123	473.063	596.230	0.672
GURPP[7]	$\mathcal{UM}_4$	1251.5	2850.9	0.713	346.84	455.69	0.752	386.788	586.449	0.667
GraphJCL	$\mathcal{UM}_1$	1145.9	2783.3	0.727	302.54	400.827	0.797	367.110	529.178	0.729
	$\mathcal{UM}_2$	1555.0	3526.5	0.561	381.56	498.66	0.703	330.951	469.629	0.786
	$\mathcal{UM}_3$	1154.8	2812.7	0.721	300.32	427.16	0.782	342.628	505.661	0.752
	$\mathcal{UM}_4$	1194.2	2695.4	0.744	317.21	464.22	0.742	335.786	466.867	0.789
	$\mathcal{UM}_5$	1267.7	2860.0	0.712	436.94	592.05	0.581	319.905	469.139	0.789
	$\mathcal{UM}_6$	1139.5	2692.2	0.744	349.06	529.48	0.665	**306.288**	496.011	0.762
	$\mathcal{UM}_7$	1217.8	3011.6	0.680	338.89	463.84	0.743	326.074	479.446	0.777
	$\mathcal{UM}_8$	1001.0	2326.3	0.809	305.38	417.81	0.791	322.057	528.207	0.730
	$\mathcal{UM}_9$	1361.3	2985.7	0.685	437.97	575.19	0.604	313.112	469.137	0.787
	$\mathcal{TM}$	**894.9**	**2115.2**	**0.842**	**293.24**	**391.67**	**0.817**	316.244	**465.734**	**0.790**

Table 3. Performance evaluation of ablation experiments across three tasks in two cities.

Method	NYC								
	Check-in			Crime			Crash		
	MAE↓	RMSE↓	R2↑	MAE↓	RMSE↓	R2↑	MAE↓	RMSE↓	R2↑
w/o NP	730.8	1364.7	0.735	165.68	238.62	0.637	23.968	30.809	0.605
w/o MA	681.4	1035.5	0.847	160.06	247.06	0.610	23.626	31.183	0.595
w/o $\mathcal{L}_{\mathcal{CM}}$	687.3	1135.8	0.816	170.02	277.19	0.510	24.555	32.049	0.572
w/o $\mathcal{L}_{\mathcal{FM}}$	758.2	1238.7	0.782	174.63	283.53	0.487	23.945	31.593	0.584
w/o $\mathcal{L}_{MI}$	701.4	1053.0	0.842	160.50	241.36	0.628	23.760	31.177	0.595
GraphJCL	**676.2**	**1027.0**	**0.850**	**158.20**	**235.26**	**0.647**	**22.829**	**29.733**	**0.632**
Method	**CHI**								
	Check-in			**Crime**			**Crash**		
	MAE↓	**RMSE↓**	**R2↑**	**MAE↓**	**RMSE↓**	**R2↑**	**MAE↓**	**RMSE↓**	**R2↑**
w/o NP	1080.3	2446.7	0.789	366.08	497.45	0.704	333.024	468.868	0.787
w/o MA	928.0	2297.8	0.813	311.24	435.34	0.773	317.946	484.319	0.772
w/o $\mathcal{L}_{\mathcal{CM}}$	1098.7	2412.6	0.795	315.98	438.36	0.770	351.831	545.538	0.712
w/o $\mathcal{L}_{\mathcal{FM}}$	1103.8	2516.8	0.777	317.49	426.89	0.782	314.124	456.12	0.783
w/o $\mathcal{L}_{MI}$	895.9	2232.3	0.824	294.44	393.11	0.809	323.476	488.833	0.769
GraphJCL	**894.9**	**2115.2**	**0.842**	**293.24**	**391.67**	**0.817**	**288.56**	**434.242**	**0.812**

average coefficient of determination (R^2) across three representative urban prediction tasks: check-in prediction, crime forecasting, and traffic crash prediction. For the NYC dataset, setting the sampling ratio ε to 0.5 and hidden dimension d_{hid} to 144 consistently yields the highest average R^2 across tasks. Similar trends are observed for the CHI dataset, where we also set $\varepsilon = 0.5$ and $d_{hid} = 144$ as optimal. Consequently, we adopt $\varepsilon = 0.5$ and $d_{hid} = 144$ as the default settings for both cities.

5.5 Model Efficiency Study

We evaluate the efficiency of GraphJCL in comparison with state-of-the-art region representation models using the complete modality set $(\mathcal{P}, \mathcal{TF}, \mathcal{RS}, \mathcal{SV}, \mathcal{RN})$. The evaluation procedure involves loading the same raw datasets, applying model-specific preprocessing pipelines, and training each model for one epoch, as illustrated in Fig.4. All experiments were conducted on an Intel® Xeon® Gold 6148 CPU (80 cores, 2.40 GHz) and a 24 GB NVIDIA RTX 4090 GPU to ensure consistent and fair comparisons. GraphJCL demonstrates superior efficiency by achieving higher predictive accuracy while maintaining competitive processing time when compared to HAFusion, UrbanVLP, and MuseCL. Although GraphST, ReCP, HREP, and GURPP exhibit faster training speeds, their overall performance is limited due to reduced modality usage and oversimplified modeling strategies. By integrating dual-perspective global and local modeling with five different modalities, GraphJCL slightly increases training time

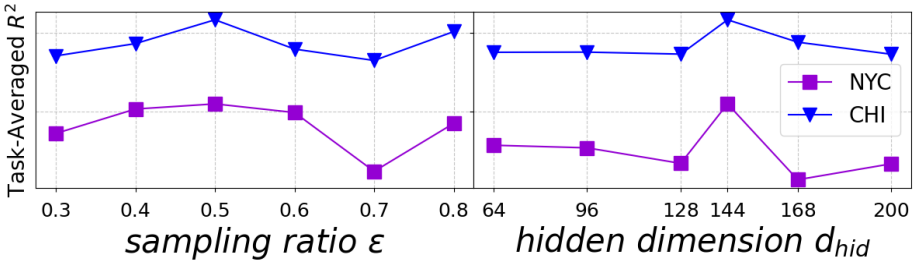

Fig. 3. Hyperparameter study of GraphJCL.

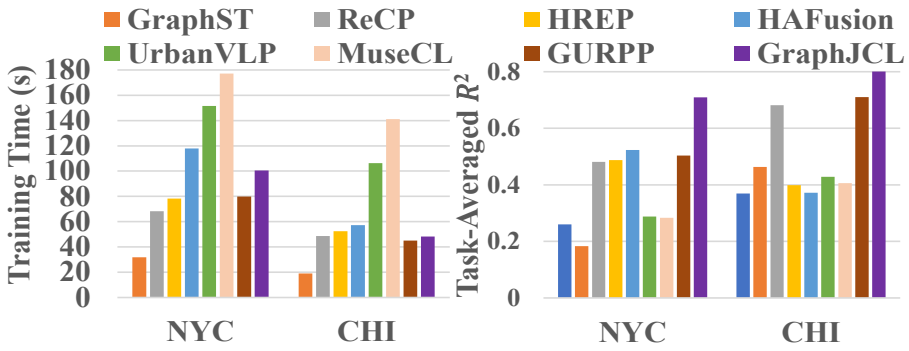

Fig. 4. Efficiency study: The time required for each model to load all data and train for one epoch.

but significantly enhances predictive performance, making the additional computational cost a highly worthwhile trade-off for practical applications.

6 Discussion and Conclusion

In this paper, we propose **GraphJCL**, a novel dual-perspective graph framework for urban region representation that models both local and global perspectives. It employs joint contrastive learning to enhance region representations. Experimental results demonstrate the model's flexibility and effectiveness. Future research will explore alternative strategies for integrating local and global learning beyond contrastive optimization, as well as extending the framework to incorporate dynamic data, temporal variations, and contextual information to enhance real-time prediction accuracy and robustness.

Acknowledgments. This work is supported by the MOE Project of Key Research Institute of Humanities and Social Sciences (No.22JJD110001), the National Natural Science Foundation of China (No.72171229) and the Outstanding Innovative Talents Cultivation Funded Programs 2023 of Renmin University of China.

References

1. Berahmand, K., Daneshfar, F., Salehi, E.S., Li, Y., Xu, Y.: Autoencoders and their applications in machine learning: a survey. Artificial Intelligence Review **57**(2), 28 (2024)
2. Chan, W., Ren, Q., Li, J.: Enhanced urban region profiling with adversarial self-supervised learning. arXiv preprint arXiv:2402.01163 (2024)
3. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864 (2016)
4. Hao, X., Chen, W., Yan, Y., Zhong, S., Wang, K., Wen, Q., Liang, Y.: Urbanvlp: Multi-granularity vision-language pretraining for urban region profiling. arXiv preprint arXiv:2403.16831 (2024)
5. Hu, J., Bai, J., Yang, J., Lee, J.J.: Crash risk prediction using sparse collision data: Granger causal inference and graph convolutional network approaches. Expert Systems with Applications **259**, 125315 (2025)
6. Huang, W., Zhang, D., Mai, G., Guo, X., Cui, L.: Learning urban region representations with pois and hierarchical graph infomax. ISPRS Journal of Photogrammetry and Remote Sensing **196**, 134–145 (2023)
7. Jin, J., Song, Y., Kan, D., Zhu, H., Sun, X., Li, Z., Sun, X., Zhang, J.: Urban region pretraining and prompting: A graph-based approach. arXiv preprint arXiv:2408.05920 (2024)
8. 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. pp. 1363–1373 (2023)
9. 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. vol. 38, pp. 8724–8732 (2024)
10. 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. pp. 4294–4298 (2022)
11. Mandal, S., O'Connor, N.E.: Llmasmmkg: Llm assisted synthetic multi-modal knowledge graph creation for smart city cognitive digital twins. Proceedings of the AAAI Symposium Series **4**, 210–221 (2024)
12. Mikolov, T., Le, Q.V., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (ICLR) (2013)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
14. Shui, C., Li, X., Qi, J., Jiang, G., Yu, Y.: Hierarchical graph contrastive learning for review-enhanced recommendation. In: Bifet, A., Davis, J., Krilavičius, T., Kull, M., Ntoutsi, E., Žliobaitė, I. (eds.) Machine Learning and Knowledge Discovery in Databases. Research Track. pp. 423–440. Springer Nature Switzerland, Cham (2024)
15. Sun, F., Qi, J., Chang, Y., Fan, X., Karunasekera, S., Tanin, E.: Urban region representation learning with attentive fusion. In: 2024 IEEE 40th International Conference on Data Engineering (ICDE). pp. 4409–4421. IEEE (2024)
16. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

18. 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. pp. 2312–2318 (2022)
19. Xia, L., Huang, C., Xu, Y., Dai, P., Bo, L., Zhang, X., Chen, T.: Spatial-temporal sequential hypergraph network for crime prediction with dynamic multiplex relation learning. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 1631–1637 (2021)
20. Xiao, C., Zhou, J., Xiao, Y., Huang, J., Xiong, H.: Refound: Crafting a foundation model for urban region understanding upon language and visual foundations. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2024)
21. Xu, Z., Zhou, X.: Cgap: Urban region representation learning with coarsened graph attention pooling. In: Larson, K. (ed.) Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. pp. 7518–7526 (2024)
22. 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 Web Conference 2024. p. 4006–4017. WWW '24 (2024)
23. 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) (2018)
24. Yong, X., Zhou, X.: Musecl: Predicting urban socioeconomic indicators via multi-semantic contrastive learning. In: Larson, K. (ed.) Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. pp. 7536–7544 (2024)
25. You, Y., Chen, T., Wang, Z., Shen, Y.: Bringing your own view: Graph contrastive learning without prefabricated data augmentations. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. p. 1300–1309. WSDM '22 (2022)
26. Zhang, L., Long, C., Cong, G.: Region embedding with intra and inter-view contrastive learning. IEEE Transactions on Knowledge and Data Engineering **35**(9), 9031–9036 (2022)
27. 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 conference on international joint conferences on artificial intelligence. pp. 4431–4437 (2021)
28. Zhang, Q., Huang, C., Xia, L., Wang, Z., Li, Z., Yiu, S.: Automated spatio-temporal graph contrastive learning. In: Proceedings of the ACM Web Conference 2023. pp. 295–305 (2023)
29. Zhang, Q., Huang, C., Xia, L., Wang, Z., Yiu, S.M., Han, R.: Spatial-temporal graph learning with adversarial contrastive adaptation. In: International Conference on Machine Learning. pp. 41151–41163. PMLR (2023)
30. 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. pp. 1700–1708 (2019)
31. Zhao, Z., Ge, Q., Cheng, A., Liu, Y., Li, X., Wang, S.: Hetcan: A heterogeneous graph cascade attention network with dual-level awareness. In: Bifet, A., Davis, J., Krilavičius, T., Kull, M., Ntoutsi, E., Žliobaitė, I. (eds.) Machine Learning and Knowledge Discovery in Databases. Research Track. pp. 57–73. Springer Nature Switzerland, Cham (2024)
32. 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. vol. 37, pp. 4981–4989 (2023)
33. Zou, X., Huang, J., Hao, X., Yang, Y., Wen, H., Yan, Y., Huang, C., Liang, Y.: Learning geospatial region embedding with heterogeneous graph. arXiv preprint arXiv:2405.14135 (2024)