# UNCERTAINTY-AWARE MULTIMODAL FUSION FOR ROBUST URBAN REPRESENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Urban region embedding increasingly combines multiple data modalities, such as points of interest, human mobility, and street-view imagery, to capture the functional, behavioral, and physical characteristics of urban regions. Existing methods integrate these modalities through cross-modal alignment or adaptive fusion, but encode each modality as a single deterministic embedding, implicitly treating every observed modality as a reliable description of the region. This assumption is fragile in real urban sensing: because urban data are collected through heterogeneous processes, the reliability of the same modality varies substantially across regions, even when the available modalities are identical. We introduce URGE, an uncertainty-aware framework that models each modality as a probabilistic distribution over regional semantics, capturing both its semantic information and associated uncertainty. Without downstream labels or explicit reliability annotations, URGE learns observation uncertainty from cross-modal structure and carries it throughout representation learning. The learned uncertainty modulates both cross-modal semantic learning and multimodal fusion, reducing the influence of unreliable observations and adapting the fused representation to changes in observation completeness. Across real-world urban benchmarks covering multiple cities and downstream tasks, URGE consistently improves region representation quality and robustness under unreliable and incomplete multimodal observations.

## 1 INTRODUCTION

Urban region embedding encodes each geographic region into a compact vector that supports diverse downstream tasks, such as population estimation, socioeconomic prediction, and land-use inference (Zhou et al., 2023; Yan et al., 2024; Xiao et al., 2024). Recent methods learn such embeddings from multiple urban data modalities: points of interest (POIs) reflect regional functions, human mobility reflects interactions between regions, road networks reflect connectivity, and satellite and street-view imagery reflect the built environment. Because each modality captures a different aspect of a region, combining them is central to learning general-purpose region representations.

Existing multimodal methods improve this combination along two lines. Alignment-based methods encourage the embeddings of different modalities to agree on the same region (Zhang et al., 2022; Li et al., 2024b; Liu et al., 2023b), while fusion-based methods combine modality embeddings through attention or graph-based interactions (Zhang et al., 2021; Zhou et al., 2023; Sun et al., 2025). Both lines encode each modality as a single deterministic embedding, which carries an implicit assumption: *if a modality is observed in a region, its embedding is a reliable description of that region, regardless of how well the modality was actually observed there.*

This assumption often fails in practice. Urban data are collected through heterogeneous processes, such as commercial listings, app usage, and street-view collection vehicles, whose coverage varies substantially across space (Fan et al., 2025; Zhang & Zhu, 2018). Consequently, the *reliability* of a modality, i.e., how faithfully its observations in a region reflect that region's actual characteristics, varies from region to region. POI records may be dense and up to date downtown but capture only a fraction of the facilities in a newly developed neighborhood; street-view imagery is abundant along major roads but sparse in peripheral areas (Fan et al., 2025); and mobility data depend on platform penetration and sampling behavior. Crucially, this variation arises even when two regions



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

have exactly the same set of modalities, and therefore differs from the missing-modality setting considered by methods that adapt to varying modality availability (Ma et al., 2021).

Addressing this variation is difficult because reliability cannot be observed directly. Coverage indicators capture data volume rather than representativeness: a region may contain many outdated POI records, or hundreds of street-view images concentrated along a single arterial road. Nor is there ground truth indicating how faithfully a modality describes a region. Existing uncertainty-aware multimodal methods typically estimate predictive uncertainty or use uncertainty to guide fusion under task supervision (Han et al., 2022; Zhang et al., 2023). Region embedding, however, is learned without downstream labels or explicit reliability annotations, providing no direct supervision for estimating the reliability of each modality for each region.

We address this limitation by representing each modality as a probabilistic distribution over regional semantics, allowing the model to capture both the semantics suggested by an observation and the uncertainty associated with that estimate. Building on this idea, we propose **URGE**, an **U**ncertainty-aware multimodal **R**e**G**ion **E**mbedding framework. URGE comprises two main components. First, **probabilistic semantic modeling** learns cross-modal semantic agreement while accounting for the uncertainty of the participating observations. This prevents uncertain evidence from being treated as equally reliable during semantic learning and uses cross-modal disagreement as a signal for uncertainty estimation. Second, **uncertainty-aware multimodal fusion** carries the learned uncertainty into multimodal integration, allowing the contribution of each observation to adapt to its estimated reliability. We further introduce a relative robustness objective that encourages uncertainty to adapt when observation completeness changes and promotes stable fused representations. Together, these components carry uncertainty from semantic modeling into multimodal fusion, enabling robust urban representation learning under varying observation reliability.

Our contributions are summarized as follows:

- We develop a label-free framework for reliability-aware multimodal urban representation learning, which learns observation reliability directly from multimodal observations without downstream labels or explicit reliability annotations, and incorporates it into representation learning to obtain robust region embeddings.
- We propose **URGE**, which models each modality as a probabilistic distribution over regional semantics and carries the resulting uncertainty throughout representation learning. This allows cross-modal agreement and multimodal integration to adapt to the reliability of individual observations, reducing the influence of unreliable evidence on the learned region representation.
- We conduct comprehensive experiments on real-world urban datasets across multiple cities and downstream tasks. The results show that URGE consistently improves urban representation quality and maintains robust performance under partial and unreliable multimodal observations. Our code is publicly available at `https://anonymous.4open.science/r/URGE`.

## 2 Related Work

### 2.1 Multimodal Urban Representation Learning

Urban region representation learning aims to encode geographic regions into compact embeddings that support downstream urban analysis. With increasingly heterogeneous urban sensing data, recent studies have shifted from individual data sources toward multimodal modeling (Zhang et al., 2021; Zhou et al., 2023; Li et al., 2024b). Different views characterize complementary aspects of regions: mobility captures human interaction patterns, POIs and map features reflect regional functions and spatial organization, while satellite and street-view imagery describe the physical environment.

Recent work has moved beyond feature combination toward explicitly modeling how information is shared and integrated across views. One line strengthens cross-view semantic coordination. Zhang et al. (2022) promotes intra- and inter-view consistency through contrastive learning, Li et al. (2024b) encourages consistency between POI and mobility representations through contrastive prediction, and Li et al. (2026b) disentangles cross-view commonality from view-specific semantics. Another line develops adaptive multimodal integration. Zhou et al. (2023) models heterogeneous

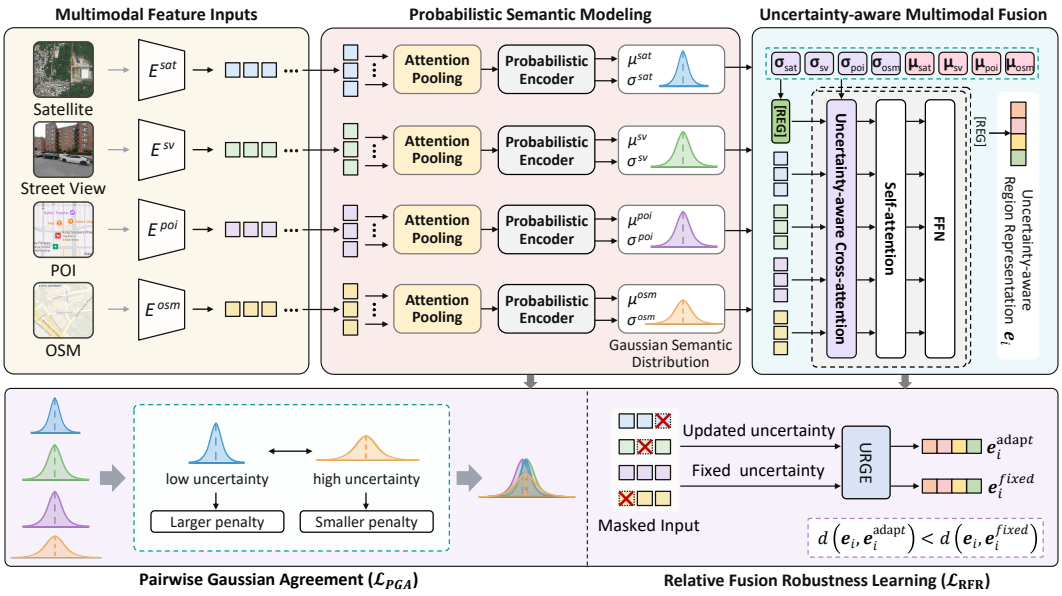



Figure 1: Overview of URGE.

regional relations with attention-based fusion, Sun et al. (2024) captures dependencies across region features through attentive fusion, and Jin et al. (2025) combines transferable multi-view pre-training with task-specific adaptation. Despite more structured modeling of shared and complementary urban semantics, these methods generally leave evidence reliability implicit in deterministic modality representations.

### 2.2 Robust Multimodal Learning

Robust multimodal learning studies robustness to incomplete, corrupted, or uneven-quality observations. Missing-modality methods recover absent information, align complete and incomplete representations, or transfer knowledge across modality subsets Ma et al. (2021); Wang et al. (2023); Lin & Hu (2023); Li et al. (2024a). Other work targets low-quality or conflicting inputs through adaptive fusion and reliability-aware evidence aggregation Zhang et al. (2023); Xu et al. (2024). These studies mainly address explicit missingness or corruption, yet reliability can vary even when every modality is present: a view may provide weaker or less representative evidence for some instances.

Uncertainty offers a principled representation of this variation. Prior work estimates view-specific uncertainty or confidence to support reliable multi-view prediction and fusion Han et al. (2022); Liu et al. (2023a). Uncertainty has also been used for test-time multimodal fusion under varying modality availability Zhang et al. (2024). These approaches mainly connect uncertainty to supervised prediction or fusion decisions. Label-free representation learning is different because no task labels are available to indicate which modality is reliable for each instance. URGE instead models each view as an uncertain estimate of shared semantics and infers reliability from cross-view structure.

## 3 Method

In this section, we present **URGE**, an uncertainty-aware multimodal framework for robust urban representation learning. As illustrated in Fig. 1, URGE takes heterogeneous urban sensing modalities as input and learns region representations through two main components: (1) *Probabilistic Semantic Modeling*, which models each modality as a probabilistic distribution over regional semantics and learns cross-modal agreement under varying observation uncertainty; and (2) *Uncertainty-Aware Multimodal Fusion*, which incorporates the learned uncertainty into fine-grained multimodal integration to reduce the influence of unreliable observations and support robust region representations.

### 3.1 PROBLEM FORMULATION

We consider multimodal representation learning over a collection of $N$ geographic regions. Let $\mathcal{V} = \{1, \ldots, V\}$ denote the set of sensing modalities. For the $i$-th region, the observation from modality $v$ is represented as a variable-length sequence of feature tokens $X_i^v = [x_{i,1}^v, \ldots, x_{i,N_i^v}^v]$, where $x_{i,j}^v \in \mathbb{R}^{d_v}$, $N_i^v$ denotes the number of tokens, and $d_v$ is the feature dimension of modality $v$. The multimodal observations of region $i$ are denoted by $\mathcal{X}_i = \{X_i^v\}_{v \in \mathcal{V}}$.

Different modalities provide complementary observations of the same region, while the reliability of their observed evidence can vary substantially across modalities and regions. Importantly, such variation can arise even when the same set of modalities is available. Our goal is to learn, without downstream task supervision, a compact region representation

$$e_i = f\left(X_i^1, X_i^2, \ldots, X_i^V\right) \in \mathbb{R}^d, \tag{1}$$

while explicitly accounting for such variation in evidence reliability during multimodal representation learning and fusion. The learned embedding $e_i$ can then serve as a representation for diverse downstream urban prediction tasks.

### 3.2 PROBABILISTIC SEMANTIC MODELING

Different modalities provide complementary observations of the same region, yet the certainty with which they support the underlying regional semantics can vary substantially across modalities and regions. A deterministic point estimate does not express the uncertainty associated with the inferred semantics. Consequently, cross-modal consistency is often modeled without regard to evidence reliability, allowing uncertain observations to distort the learned cross-modal semantics. To address this, URGE models each modality-specific semantic estimate as a Gaussian distribution, whose mean captures the inferred region semantics and whose variance characterizes the associated uncertainty. Cross-modal agreement is then learned under a probabilistic formulation that jointly aligns modality-specific semantic estimates and learns their uncertainty from cross-modal discrepancies.

**Modality Probabilistic Encoding.** For each modality, we model its estimate of the underlying region semantics as a diagonal Gaussian distribution:

$$q_i^v(\mathbf{s}_i) = \mathcal{N}\left(\mathbf{s}_i; \boldsymbol{\mu}_i^v, \operatorname{diag}\left((\boldsymbol{\sigma}_i^v)^2\right)\right), \tag{2}$$

where $\boldsymbol{\mu}_i^v \in \mathbb{R}^{d_z}$ represents the semantic estimate and $\boldsymbol{\sigma}_i^v \in \mathbb{R}_+^{d_z}$ characterizes its associated uncertainty. Specifically, each feature token is first projected into a common $d$-dimensional latent space:

$$\mathbf{z}_{i,j}^v = \mathrm{Proj}^v\left(\mathbf{x}_{i,j}^v\right), \tag{3}$$

where $\mathrm{Proj}^v$ is a modality-specific projection network. Let $\mathbf{Z}_i^v = [\mathbf{z}_{i,1}^v, \ldots, \mathbf{z}_{i,N_i^v}^v]^\top \in \mathbb{R}^{N_i^v \times d}$. We then aggregate the variable-length token sequence through cross-attention with a learnable pooling query:

$$\mathbf{h}_i^v = \mathrm{CrossAttn}^v\left(\mathbf{q}^v, \mathbf{Z}_i^v, \mathbf{Z}_i^v\right) \in \mathbb{R}^d, \tag{4}$$

where $\mathbf{q}^v \in \mathbb{R}^d$ denotes the learnable query. The pooled representation is then used to jointly estimate the semantic mean and uncertainty:

$$\begin{bmatrix} \boldsymbol{\mu}_i^v \\ \log\left((\boldsymbol{\sigma}_i^v)^2\right) \end{bmatrix} = g^v\left(\mathbf{h}_i^v\right), \tag{5}$$

where $g^v : \mathbb{R}^d \to \mathbb{R}^{2d_z}$ is a prediction network that jointly outputs the semantic mean and log-variance. The corresponding variance is obtained by exponentiation, ensuring positivity.

**Pairwise Gaussian Agreement.** Under this probabilistic formulation, each modality provides an imperfect observation of the same underlying region semantics, with uncertainty characterizing the reliability of its semantic estimate. For two modalities observing the same region, their semantic discrepancy provides a natural supervisory signal for estimating the uncertainty of their semantic predictions. Based on this principle, we introduce *Pairwise Gaussian Agreement* (PGA), which

jointly promotes distributional agreement between semantic estimates from different modalities and learns their associated uncertainty from cross-modal disagreement:

$$\mathcal{L}_{\mathrm{PGA}}^{i,u,v} = -\frac{1}{d_z} \log \int_{\mathbb{R}^{d_z}} q_i^u(\mathbf{s})\, q_i^v(\mathbf{s})\, \mathrm{d}\mathbf{s}. \tag{6}$$

where $\mathbf{s} \in \mathbb{R}^{d_z}$ denotes a latent semantic vector in the shared semantic space. Since $q_i^u$ and $q_i^v$ are diagonal Gaussian distributions, the objective admits a closed-form solution. Omitting constants independent of the model parameters, we obtain

$$\mathcal{L}_{\mathrm{PGA}}^{i,u,v} = \frac{1}{2d_z} \sum_{j=1}^{d_z} \left[ \frac{\left(\mu_{i,j}^u - \mu_{i,j}^v\right)^2}{\left(\sigma_{i,j}^u\right)^2 + \left(\sigma_{i,j}^v\right)^2} + \log\left( \left(\sigma_{i,j}^u\right)^2 + \left(\sigma_{i,j}^v\right)^2 \right) \right]. \tag{7}$$

The pairwise variance $\left(\sigma_{i,j}^u\right)^2 + \left(\sigma_{i,j}^v\right)^2$ captures the combined uncertainty of the two semantic estimates. Accordingly, the first term imposes stronger alignment when both estimates are reliable and weaker alignment when their uncertainty is high, while the logarithmic term penalizes excessively large uncertainty. PGA therefore allows different modalities to express different levels of uncertainty while learning a shared semantic space.

### 3.3 Uncertainty-Aware Multimodal Fusion

The probabilistic semantic modeling above provides each modality with fine-grained latent tokens and an uncertainty-aware semantic estimate. We use this uncertainty to guide how multimodal evidence is aggregated into the region representation. Specifically, uncertainty affects both the initialization of the region latent and its subsequent interaction with modality tokens, allowing less reliable observations to exert weaker influence during fusion.

**Uncertainty-Guided Multimodal Aggregation.** Observation reliability varies across regions and modalities, and should therefore be reflected during multimodal aggregation. We use the learned uncertainty to guide both the initialization of the region state and its subsequent fusion with multimodal tokens.

We first construct a region-specific [REG] token from the probabilistic semantic estimates:

$$\begin{aligned} \mathbf{u}_{i,[\mathrm{REG}]}^{(0)} &= g_{\mathrm{REG}}\left([\overline{\boldsymbol{\mu}}_i; -\log \boldsymbol{\Lambda}_i]\right), \\ \boldsymbol{\Lambda}_i &= \sum_{v \in \mathcal{V}} (\boldsymbol{\sigma}_i^v)^{-2}, \qquad \overline{\boldsymbol{\mu}}_i = \boldsymbol{\Lambda}_i^{-1} \odot \sum_{v \in \mathcal{V}} (\boldsymbol{\sigma}_i^v)^{-2} \odot \boldsymbol{\mu}_i^v. \end{aligned} \tag{8}$$

Here, $\boldsymbol{\Lambda}_i$ is the aggregated precision and $\overline{\boldsymbol{\mu}}_i$ is the corresponding precision-weighted semantic estimate. Their concatenation allows the initial [REG] token to encode both the semantic consensus and its reliability.

Together with learnable latent tokens, the [REG] token attends to fine-grained tokens from all modalities. Omitting layer and head indices for clarity, the cross-attention weight is

$$\alpha_{i,r,(v,j)} = \mathrm{softmax}_{v,j}\left( \frac{\mathbf{q}_{i,r}^\top \mathbf{k}_{i,j}^v}{\sqrt{d_h}} + \gamma \rho_i^v - \log N_i^v \right), \tag{9}$$

where $N_i^v$ is the number of valid tokens, and $\gamma \geq 0$ is learnable. The modality reliability is summarized from its dimension-wise uncertainty as $\rho_i^v = -\frac{1}{2d_z} \sum_{k=1}^{d_z} \log\left( (\sigma_{i,k}^v)^2 \right)$. Thus, $\gamma \rho_i^v$ modulates attention according to observation reliability, while $-\log N_i^v$ corrects for differences in token counts. Standard self-attention and feed-forward layers are then applied to the fusion latents. After $L$ layers, the final [REG] token forms the region representation: $\mathbf{e}_i = \mathbf{u}_{i,[\mathrm{REG}]}^{(L)} \in \mathbb{R}^d$.

**Relative Fusion Robustness Learning.** The uncertainty used for multimodal fusion should reflect changes in observation completeness and induce a corresponding adjustment of the fusion process. We therefore introduce *Relative Fusion Robustness* (RFR), which encourages the learned uncertainty to support stable multimodal representations under partial observations.

During training, we construct a partial observation by random masking a subset of valid tokens from one modality and recompute its semantic estimate and uncertainty. Let $\widetilde{\mathbf{Z}}_i$, $\widetilde{\boldsymbol{\mu}}_i$, and $\widetilde{\boldsymbol{\sigma}}_i$ denote the resulting token representations, semantic estimates, and uncertainty. We obtain three fused representations:

$$\begin{aligned} \mathbf{e}_i &= F(\mathbf{Z}_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i), \\ \mathbf{e}_i^{\text{adapt}} &= F(\widetilde{\mathbf{Z}}_i, \widetilde{\boldsymbol{\mu}}_i, \widetilde{\boldsymbol{\sigma}}_i), \\ \mathbf{e}_i^{\text{fixed}} &= F(\widetilde{\mathbf{Z}}_i, \widetilde{\boldsymbol{\mu}}_i, \boldsymbol{\sigma}_i), \end{aligned} \tag{10}$$

where $F$ denotes the multimodal fusion module. $\mathbf{e}_i^{\text{adapt}}$ uses uncertainty recomputed from the partial observation, whereas $\mathbf{e}_i^{\text{fixed}}$ retains the uncertainty estimated from the complete observation while using the same partial semantic evidence.

We encourage $\mathbf{e}_i^{\text{adapt}}$ to remain closer to $\mathbf{e}_i$ than $\mathbf{e}_i^{\text{fixed}}$:

$$\mathcal{L}_{\text{RFR}} = \left[ d\left(\text{sg}(\mathbf{e}_i), \mathbf{e}_i^{\text{adapt}}\right) - d\left(\text{sg}(\mathbf{e}_i), \text{sg}(\mathbf{e}_i^{\text{fixed}})\right) + m \right]_+, \tag{11}$$

where $d(\cdot, \cdot)$ denotes cosine distance, $m$ is a margin, and $\text{sg}(\cdot)$ denotes stop-gradient. RFR therefore encourages uncertainty adaptation to improve representation stability when observation completeness changes.

### 3.4 Training Objective

To prevent representation collapse, we encourage each semantic dimension to maintain sufficient variation across regions:

$$\mathcal{L}_{\text{var}} = \frac{1}{V d_z} \sum_{v=1}^{V} \sum_{j=1}^{d_z} \left[ \tau - \text{Std}_i\left(\mu_{i,j}^v\right) \right]_+, \tag{12}$$

where $\tau$ specifies the minimum desired standard deviation across regions.

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{PGA}} + \lambda_1 \mathcal{L}_{\text{RFR}} + \lambda_2 \mathcal{L}_{\text{var}}, \tag{13}$$

where $\mathcal{L}_{\text{PGA}}$ learns cross-modal semantic agreement and the associated uncertainty, $\mathcal{L}_{\text{RFR}}$ encourages uncertainty adaptation to support robust multimodal fusion under partial observations, and $\mathcal{L}_{\text{var}}$ maintains representation diversity across regions. The coefficients $\lambda_1$ and $\lambda_2$ control the relative contributions of the auxiliary objectives.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.** We conduct experiments on two real-world urban datasets from New York City (NYC) and London, introduced in Li et al. (2026a). The NYC dataset contains 2,324 Census Tracts, while the London dataset contains 1,000 Middle Layer Super Output Areas (MSOAs). Each region is characterized by four heterogeneous sensing modalities: satellite imagery, street-view imagery, points of interest (POIs), and OpenStreetMap (OSM) map tiles. The datasets provide 15 and 14 region-level prediction tasks for NYC and London, respectively, spanning socioeconomic, health, environmental, and transportation indicators. Detailed data sources and preprocessing procedures are provided in Appendix A.

**Evaluation protocol.** We evaluate the learned region representations on the 29 downstream tasks. Following the standard protocol for urban region representation learning(Li et al., 2024b; Yan et al., 2024; Li et al., 2023), we first learn region representations without downstream supervision and then evaluate the frozen representations using downstream prediction tasks. In the main paper, we report four representative tasks covering distinct urban domains: *education* (bachelor's degree rate), *health* (obesity rate for NYC and life expectancy for London), *income*, and *transport* (carpool commuting ratio). For each downstream task, we fit a Ridge regression model on the frozen region representations and assess prediction performance using 5-fold cross-validation. We report mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination ($R^2$).

Table 1: Performance comparison on representative downstream tasks. Best results are in **bold** and second-best are underlined.

| NYC | Education | | | Health | | | Income | | | Transport | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | MAE↓ | RMSE↓ | $R^2$↑ | MAE↓ | RMSE↓ | $R^2$↑ | MAE↓ | RMSE↓ | $R^2$↑ | MAE↓ | RMSE↓ | $R^2$↑ |
| Urban2Vec | 10.66 | 13.90 | 0.59 | 3.81 | 4.91 | 0.44 | 22460 | 30410 | 0.48 | 3.16 | 4.48 | 0.16 |
| RegionDCL | 10.71 | 14.08 | 0.58 | 3.88 | 5.00 | 0.42 | 21712 | 29732 | 0.51 | 3.16 | 4.51 | 0.15 |
| UrbanCLIP | 11.24 | 14.62 | 0.54 | 3.80 | 4.84 | 0.46 | 23533 | 31526 | 0.44 | 3.23 | 4.57 | 0.15 |
| HREP | 10.80 | 14.02 | 0.58 | 3.42 | 4.43 | 0.54 | 25914 | 34338 | 0.34 | 3.27 | 4.60 | 0.12 |
| ReCP | 10.55 | 13.89 | 0.59 | 3.28 | 4.22 | 0.59 | 21640 | 29241 | 0.52 | 3.21 | 4.59 | 0.12 |
| HAFusion | 12.46 | 16.08 | 0.45 | 4.40 | 5.50 | 0.30 | 24674 | 33505 | 0.37 | 3.15 | 4.49 | 0.16 |
| FlexiReg | 13.79 | 17.86 | 0.32 | 4.30 | 5.29 | 0.34 | 26551 | 36085 | 0.27 | 3.29 | 4.63 | 0.11 |
| GURPP | 10.60 | 13.79 | 0.59 | 3.18 | 4.11 | 0.61 | 22221 | 29712 | 0.51 | 3.07 | 4.40 | 0.21 |
| ComSRE | 11.06 | 14.35 | 0.56 | 3.85 | 4.89 | 0.45 | 23755 | 31529 | 0.44 | 3.11 | 4.45 | 0.19 |
| **URGE** | **9.09** | **12.17** | **0.68** | **2.82** | **3.69** | **0.68** | **20066** | **27374** | **0.58** | **3.02** | **4.34** | **0.23** |
| **Improvement** | **13.84** | **11.75** | **15.25** | **11.32** | **10.22** | **11.48** | **7.27** | **6.38** | **11.54** | **1.63** | **1.36** | **9.52** |

| London | Education | | | Health | | | Income | | | Transport | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | MAE↓ | RMSE↓ | $R^2$↑ | MAE↓ | RMSE↓ | $R^2$↑ | MAE↓ | RMSE↓ | $R^2$↑ | MAE↓ | RMSE↓ | $R^2$↑ |
| Urban2Vec | 5.68 | 7.26 | 0.66 | 1.68 | 2.16 | 0.41 | 2187 | 2785 | 0.57 | 0.38 | 0.50 | 0.70 |
| RegionDCL | 5.67 | 7.33 | 0.65 | 1.70 | 2.20 | 0.39 | 2203 | 2839 | 0.54 | 0.50 | 0.64 | 0.51 |
| UrbanCLIP | 6.42 | 8.05 | 0.58 | 1.85 | 2.37 | 0.29 | 2372 | 2999 | 0.49 | 0.44 | 0.56 | 0.63 |
| HREP | 6.52 | 8.37 | 0.55 | 2.15 | 2.70 | 0.08 | 2558 | 3217 | 0.42 | 0.42 | 0.54 | 0.65 |
| ReCP | 7.75 | 9.60 | 0.41 | 2.05 | 2.60 | 0.14 | 2056 | 2659 | 0.60 | 0.33 | 0.44 | 0.77 |
| HAFusion | 6.60 | 8.24 | 0.56 | 1.81 | 2.30 | 0.33 | 2544 | 3219 | 0.42 | 0.43 | 0.55 | 0.64 |
| FlexiReg | 9.63 | 11.59 | 0.14 | 2.16 | 2.71 | 0.07 | 3502 | 4215 | 0.01 | 0.59 | 0.75 | 0.33 |
| GURPP | 7.49 | 9.30 | 0.44 | 2.06 | 2.61 | 0.13 | 2881 | 3594 | 0.28 | 0.43 | 0.54 | 0.65 |
| ComSRE | 6.23 | 7.84 | 0.60 | 1.85 | 2.34 | 0.31 | 2446 | 3107 | 0.46 | 0.37 | 0.50 | 0.71 |
| **URGE** | **4.72** | **6.02** | **0.77** | **1.65** | **2.12** | **0.43** | **1826** | **2357** | **0.69** | **0.32** | **0.42** | **0.79** |
| **Improvement** | **16.75** | **17.08** | **16.67** | **1.79** | **1.85** | **4.88** | **11.19** | **11.36** | **15.00** | **3.03** | **4.55** | **2.60** |

**Baselines.** We compare URGE with representative urban region representation approaches, including HREP (Zhou et al., 2023), ReCP (Li et al., 2024b), HAFusion (Sun et al., 2024), FlexiReg (Sun et al., 2025), RegionDCL (Li et al., 2023), Urban2Vec (Wang et al., 2020), UrbanCLIP (Yan et al., 2024), ComSRE (Li et al., 2026b), and GURPP (Jin et al., 2025). RegionDCL and Urban2Vec are evaluated using their supported modality subsets, namely OSM with POIs and street-view imagery with POIs, respectively. All other baselines use the same four modality inputs as URGE. Detailed baseline descriptions are provided in Appendix C. Implementation details are provided in Appendix B.

### 4.2 Performance Comparison

We compare different methods on representative downstream tasks in NYC and London. The main results are reported in Table 1 using MAE, RMSE, and $R^2$. We highlight three key observations from the results:

- Existing methods exhibit pronounced task- and city-specific performance variations. The strongest baseline changes considerably across downstream targets and urban contexts. GURPP is particularly competitive on Health and Transport in NYC, whereas ReCP performs strongly on Income in both cities and Transport in London. Such shifts suggest that different downstream tasks may emphasize different aspects of multimodal urban features, thereby placing different demands on how heterogeneous modalities should be integrated.
- More modalities do not necessarily lead to better region representations. RegionDCL and Urban2Vec remain highly competitive despite using only a subset of the available modalities. In particular, both methods outperform four-view baselines such as HAFusion and FlexiReg on several tasks, while Urban2Vec achieves the strongest baseline performance on Education and Health in London. These results indicate that simply increasing modality coverage does not necessarily improve representation quality; the benefit of additional modalities depends critically on how heterogeneous and potentially uneven evidence is integrated.

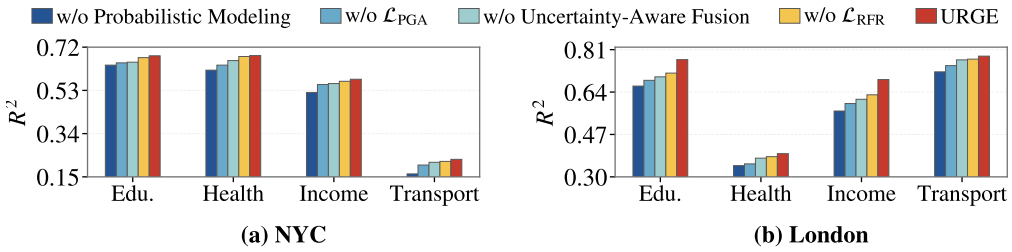



Figure 2: Performance comparison of different variants.

- URGE achieves consistent improvements through uncertainty-aware semantic alignment and fusion. URGE achieves the strongest performance across all reported tasks in both NYC and London. Averaged over the four representative tasks, URGE improves $R^2$ over the strongest baseline by 11.95% in NYC and 9.79% in London. Unlike existing approaches that primarily focus on either cross-view representation learning or multimodal aggregation, URGE jointly aligns view-specific semantics while explicitly modeling their uncertainty during fusion. The consistent gains across diverse prediction targets support the effectiveness of using uncertainty as a region-specific signal for integrating heterogeneous urban evidence.

### 4.3 Ablation Study

To evaluate the importance of each module to the overall performance, we conduct an ablation study measured by $R^2$ (see Figure 2) by testing the following model variants:

**w/o Probabilistic Modeling** replaces the Gaussian semantic representation $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ with deterministic embeddings, replaces Pairwise Gaussian Agreement with cosine alignment between view semantics, and removes uncertainty conditioning and Relative Fusion Robustness from fusion. This variant exhibits the largest and most consistent performance degradation across the four tasks in both NYC and London. The results indicate that explicitly modeling semantic uncertainty is fundamental to learning reliable cross-view semantics and multimodal region representations.

**w/o** $\mathcal{L}_{\text{PGA}}$ removes the Pairwise Gaussian Agreement objective $\mathcal{L}_{\text{PGA}}$, eliminating explicit distributional agreement across views. Its performance consistently decreases across both datasets compared with the full model. This suggests that explicitly encouraging uncertainty-aware agreement between view-specific semantic distributions improves the consistency of heterogeneous multimodal representations.

**w/o Uncertainty-Aware Fusion** replaces the uncertainty-conditioned token interaction with standard multimodal self-attention and disables $\mathcal{L}_{\text{RFR}}$ accordingly, preventing learned uncertainty from conditioning cross-modal interactions. The resulting model performs consistently worse than URGE across all reported tasks in both cities. This indicates that incorporating uncertainty into fine-grained cross-modal interaction helps the fusion process account for differences in evidence reliability across modalities and regions.

**w/o** $\mathcal{L}_{\text{RFR}}$ removes the Relative Fusion Robustness objective $\mathcal{L}_{\text{RFR}}$, eliminating explicit robustness supervision for uncertainty adaptation under evidence degradation. This variant remains close to the full model but shows consistent performance degradation across the evaluated tasks. The results suggest that uncertainty-conditioned fusion already provides useful reliability information, while $\mathcal{L}_{\text{RFR}}$ further encourages the learned uncertainty to adapt effectively to changes in evidence quality.

### 4.4 Robustness to Unreliable Multimodal Observations

To assess the robustness of URGE to unreliable multimodal observations, we construct two families of perturbed evaluation sets that simulate reduced evidence availability and semantic unreliability, respectively, and compare URGE with strong baselines. Results are shown in Figure 3.

**Robustness to Sparse Evidence.** We independently mask each evidence token of every region and modality with probability $r \in \{0, 0.1, 0.2, 0.3, 0.4\}$, resulting in progressively sparser and heteroge-

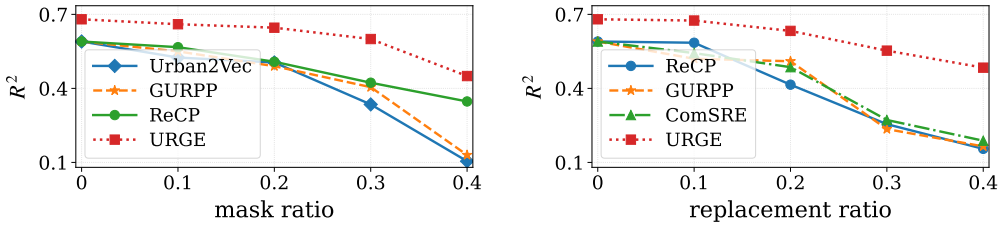



Figure 3: Robustness comparison under unreliable multimodal evidence.

neous observations. All models are trained only on the original dataset $\mathcal{D}^{(0)}$ and directly evaluated on the perturbed sets $\mathcal{D}^{(r)}$ without retraining or adaptation. As shown in Figure 3, URGE consistently maintains stronger performance than the competing baselines as the masking ratio increases. This indicates that URGE is more robust when the amount of available multimodal evidence becomes increasingly limited.

**Robustness to Semantically Perturbed Evidence.** We further consider a complementary setting in which evidence quantity is preserved but its semantic reliability is perturbed. Specifically, each token is replaced with a real same-modality token sampled from a randomly selected region in the same city with probability $q \in \{0, 0.1, 0.2, 0.3, 0.4\}$, without changing token counts or modality availability. As shown in Figure 3, URGE consistently outperforms the competing baselines across different replacement ratios, with a smaller performance decline as more tokens are replaced. These results further demonstrate the robustness of URGE, even when multimodal observations remain structurally complete but part of its semantic content becomes less reliable.

## 5 Conclusion

In this paper, we introduce URGE, an uncertainty-aware multimodal framework for robust urban region embedding. URGE models each modality as a probabilistic distribution over regional semantics and learns observation uncertainty directly from multimodal observations in a self-supervised setting. Based on this formulation, URGE incorporates uncertainty into both cross-modal semantic learning and fine-grained multimodal fusion, reducing the influence of unreliable observations throughout representation learning. A relative robustness objective further encourages the learned uncertainty to adapt when observation completeness changes, supporting stable region representations under partial observations. Extensive experiments across multiple cities and downstream tasks demonstrate that URGE consistently improves region representation quality and maintains strong robustness under unreliable multimodal observations.

### AI use statement

In this work, we used generative AI tools for text polishing and language editing. We have not used generative AI tools for research idea generation, methodology design, data analysis, or result interpretation, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for improving grammar, wording, and readability. We have reviewed all AI-assisted work. All AI-assisted language revisions were manually checked by the authors to ensure that the technical meaning and scientific content remained unchanged. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## References

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

Zicheng Fan, Chen-Chieh Feng, and Filip Biljecki. Coverage and bias of street view imagery in mapping the urban environment. *Computers, Environment and Urban Systems*, 117:102253, 2025.

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

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

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

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12458–12468. IEEE, 2024a.

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

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

Zechen Li, Hongwei Jia, Weiming Huang, Kai Zhao, and Meng Chen. Urbanexpert: Task-conditioned multi-modal fusion via semantic expert routing for urban socioeconomic prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2909–2920, 2026a.

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

Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023.

Wei Liu, Yufei Chen, Xiaodong Yue, Changqing Zhang, and Shaorong Xie. Safe multi-view deep classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8870–8878, 2023a.

Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*, pp. 4150–4160, 2023b.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2302–2310, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

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

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

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21968–21977. IEEE, 2023.

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

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

Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16129–16137, 2024.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Guiming Zhang and A-Xing Zhu. The representativeness and spatial bias of volunteered geographic information: A review. *Annals of GIS*, 24(3):151–162, 2018.

Liang Zhang, Cheng Long, and Gao Cong. Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9031–9036, 2022.

Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 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.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27446–27456. IEEE, 2024.

Silin Zhou, Dan He, Lisi Chen, Shuo Shang, and Peng Han. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4981–4989, 2023.

# APPENDIX

## A DATASET AND PREPROCESSING DETAILS

**Downstream Tasks and Data Sources.** Table 2 lists the downstream prediction tasks used for evaluation and their corresponding data sources. We retain the target construction provided by Li et al. (2026a).

Table 2: Downstream prediction tasks and their data sources in NYC and London. "–" indicates that the corresponding task is not available in that city.

| Domain | Task | NYC Source | London Source |
|---|---|---|---|
| Socioeconomic | Education | ACS 2023 | Census 2021 |
| | Income | ACS 2023 | ONS Income Estimates |
| | Rent | ACS 2023 | – |
| | Unemployment | ACS 2023 | Census 2021 |
| | Poverty | ACS 2023 | IMD 2019 |
| Health | Obesity | PLACES | – |
| | Life Expectancy | – | OHID Fingertips |
| Environment | NDVI | Sentinel-2 | Sentinel-2 |
| Population | Population Density | WorldPop | WorldPop |
| Built Environment | Building Height | GHS-BUILT-H | GHS-BUILT-H |
| Safety | Crime Rate | NYC OpenData | UK Police |
| Transportation | Public Transport | ACS 2023 | Census 2021 |
| | Drive | ACS 2023 | Census 2021 |
| | Walk/Bike | ACS 2023 | Census 2021 |
| | Work from Home | ACS 2023 | Census 2021 |
| | Carpool | ACS 2023 | Census 2021 |

**Multi-Modal Data and Preprocessing.** We follow the multi-modal data construction and preprocessing pipeline of Li et al. (2026a). Each region is associated with four heterogeneous sensing modalities. **Satellite imagery** is represented using pretrained AlphaEarth Foundations embeddings (Brown et al., 2025). The native 10 m embeddings are aggregated to a 100 m grid, with each grid cell treated as one token, resulting in $Z_i^{\mathrm{sat}} \in \mathbb{R}^{N_i^{\mathrm{sat}} \times 64}$. **Street-view imagery** is collected through the Google Street View Static API, with up to 10 images per region. Each image is encoded using CLIP ViT-L/14 (Radford et al., 2021), and its `[CLS]` representation is retained as a 768-dimensional token, yielding $Z_i^{\mathrm{sv}} \in \mathbb{R}^{N_i^{\mathrm{sv}} \times 768}$. **POI records** are obtained from OpenStreetMap (Haklay & Weber, 2008) and summarized into structured textual descriptions of regional category distributions and functional characteristics. These descriptions are encoded using Qwen3-8B (Yang et al., 2025), producing $Z_i^{\mathrm{poi}} \in \mathbb{R}^{N_i^{\mathrm{poi}} \times 1024}$. Finally, **OSM map tiles** are rendered from OpenStreetMap (Haklay & Weber, 2008) to capture building footprints, road networks, and land-cover information, and are encoded using DINOv2 ViT-B/14 (Oquab et al., 2024) into $Z_i^{\mathrm{osm}} \in \mathbb{R}^{N_i^{\mathrm{osm}} \times 768}$.

## B MODEL CONFIGURATION

For probabilistic semantic modeling, features from each modality are projected into a common latent space with dimension $d = 512$. Each modality is aggregated using cross-attention with a learnable pooling query, and the pooled representation is mapped to a diagonal Gaussian distribution with semantic dimension $d_z = 64$.

For uncertainty-aware multimodal fusion, the fusion module contains $L = 2$ layers with 4 attention heads and a per-head dimension of $d_h = 128$. Each layer applies uncertainty-aware cross-attention, followed by self-attention and a feed-forward network with hidden dimension $d_{\mathrm{ff}} = 1024$.

For Relative Fusion Robustness learning, we randomly select a masking fraction $p_{\text{mask}} \in \{0.5\}$ with equal probability and mask valid tokens from one modality during training, then recompute its semantic distribution from the partial observation. The overall objective combines Pairwise Gaussian Agreement, Relative Fusion Robustness, and variance regularization with $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$, respectively.

For optimization, we use Adam with a learning rate of $10^{-3}$ for the modality encoders and $3 \times 10^{-4}$ for the fusion module, a weight decay of $10^{-7}$, and a batch size of 128.

## C Baseline Descriptions

- **Urban2Vec** Wang et al. (2020): An unsupervised multi-modal framework that jointly encodes street-view images via CNN and POI text via bag-of-words embeddings to learn neighborhood representations preserving geospatial similarity.
- **RegionDCL** Li et al. (2023): Leverages OSM building footprints and POIs with a Transformer encoder using distance-biased self-attention. Dual contrastive learning operates at both building-group and region levels to handle data-sparse areas.
- **UrbanCLIP** Yan et al. (2024) generates textual descriptions of satellite images and learns text-enhanced region representations through image–text contrastive pretraining and language modeling.
- **HREP** Zhou et al. (2023) constructs a heterogeneous region graph from mobility, POI, and geographic proximity, and applies relation-aware GCN with attention-based fusion. Prefix-tuning prompts enable downstream task adaptation.
- **ReCP** Li et al. (2024b) learns consistent region representations from POI and mobility views via intra-view contrastive learning and inter-view consistency constraints.
- **HAFusion** Sun et al. (2024) employs a dual-feature attentive fusion module to capture inter-region correlations across views, with hybrid attentive learning over intra- and inter-view interactions.
- **FlexiReg** Sun et al. (2025) learns grid embeddings from public data (POI, land use, satellite, street view), aggregates them into region embeddings, and applies a prompt enhancer for task-specific adaptation.
- **GURPP** Jin et al. (2025) constructs a heterogeneous urban region graph from POIs, roads, junctions, and typed spatial relations, learns subgraph-centric region embeddings via pre-training, and adapts them with graph-based prompts for downstream tasks.
- **ComSRE** Li et al. (2026b) disentangles shared and view-specific information in POI and mobility data through view-to-commonality contrastive alignment and differential orthogonality constraints.