LocationBind: Coordinate-Pivot Alignment of Heterogeneous Urban Modalities in a Unified Latent Space

Naoki Tamura
tam@ucl.nuee.nagoya-u.ac.jp
Nagoya University
Nagoya, Aichi, Japan

Haru Terashima
haru@ucl.nuee.nagoya-u.ac.jp
Nagoya University
Nagoya, Aichi, Japan

Kazuyuki Shoji
shoji@ucl.nuee.nagoya-u.ac.jp
Nagoya University
Nagoya, Aichi, Japan

Kenta Urano
urano@nagoya-u.jp
Nagoya University
Nagoya, Aichi, Japan

Takuro Yonezawa
takuro@nagoya-u.jp
Nagoya University
Nagoya, Aichi, Japan

Nobuo Kawaguchi
kawaguti@nagoya-u.jp
Nagoya University
Nagoya, Aichi, Japan

Abstract

Urban environments continuously generate and collect data from diverse modalities such as points of interest (POIs), satellite imagery, and human mobility traces, each anchored to geographic coordinates. There is growing demand for frameworks that integrate these heterogeneous sources to capture both the multifaceted characteristics of locations and the latent correspondences across modalities. However, existing approaches either align coordinates with a single modality, thereby failing to capture multiple aspects of a location, or require explicit cross-modal pairs (e.g., image–text) that are difficult to construct when modalities differ in spatial resolution and observation density. We propose LocationBind, a framework that uses geographic coordinates as a common pivot to jointly align multiple urban modalities in a shared latent space, requiring only coordinate–modality pairs. Experiments on Osaka and Nagoya show that LocationBind outperforms baselines on most metrics across missing-modality completion, cross-modal retrieval, and urban-indicator prediction (e.g., land use and population density), demonstrating the effectiveness of coordinate-pivot alignment. We further illustrate applications such as cross-modal pattern discovery and compositional cross-modal retrieval, suggesting that LocationBind can serve as a unified framework for multifaceted urban data analysis.

CCS Concepts

• **Information systems** → **Geographic information systems**; **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Learning latent representations**.

Keywords

Urban Computing, Representation Learning, Multimodal

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGSPATIAL '26, Riverside, CA, USA
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Naoki Tamura, Haru Terashima, Kazuyuki Shoji, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2026. LocationBind: Coordinate-Pivot Alignment of Heterogeneous Urban Modalities in a Unified Latent Space. In *Proceedings of 34th ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '26).* ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Cities are complex socio-technical systems in which economic, infrastructural, and environmental processes interact across multiple spatial and temporal scales [1]. Urban computing increasingly uses large-scale spatiotemporal data to monitor, model, and predict these dynamics, enabling evidence-based urban planning and management [2]. At the same time, urban data are rapidly diversifying. Multimodal data such as Points of Interest (POIs), review and survey texts, satellite imagery, street view images, and human mobility logs are increasingly collected at scale together with their geographic coordinates, and many such datasets are also being released as open resources [3, 4]. These data make it possible to capture the characteristics of the same urban location from multiple perspectives. However, frameworks that integrate such multimodal data and support cross-modal understanding, retrieval, and inference are still lacking.

A representative line of work in urban data modeling focuses on geospatial embedding. These methods embed modality data tied to geographic space, such as satellite imagery and human mobility data, as vector representations in a latent space through self-supervised learning [5–11]. In particular, recent methods such as GeoCLIP [12], SatCLIP [13], and CaLLiPer [14] jointly learn aligned representations of a modality and geographic coordinates, enabling the acquisition of embeddings defined over continuous geographic coordinates. However, most existing methods rely on unimodal observation sources. Even coordinate-to-modality alignment methods such as GeoCLIP, SatCLIP, and CaLLiPer align coordinates with only one modality at a time. A naïve extension that trains such methods separately for each modality and combines them fails because each modality occupies its own latent space, preventing cross-modal operations. As a result, it becomes difficult to leverage these representations for cross-modal retrieval, completion, or pattern discovery. In addition, designs tied to a specific modality cannot capture the multifaceted characteristics of each location,

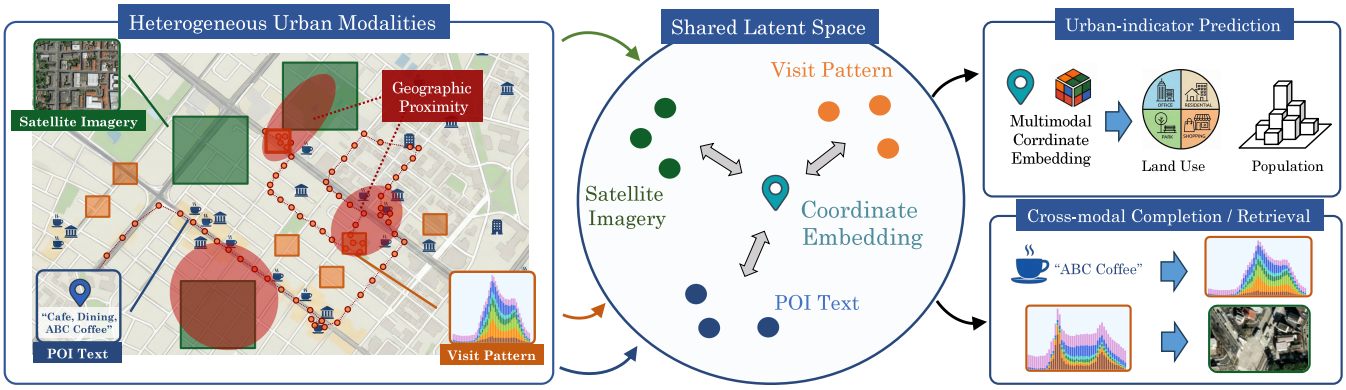


Figure 1: Alignment of multimodal urban data via LocationBind.

and they can also be sensitive to temporal and spatial sampling bias and missing observations.

In contrast, multimodal representation learning methods such as ImageBind [15] and KnowCL [11] provide frameworks that align multiple modalities into a shared latent space. However, to learn cross-modal correspondences, these methods require explicit cross-modal pairs or the aggregation of observations into fixed spatial units such as regional nodes or mesh cells. In actual urban data, spatial resolution and observation density differ across modalities, and data collection conditions vary by region and over time, making it difficult to construct stable large-scale pairs for arbitrary combinations of modalities.

We hypothesize that heterogeneous modality data observed at the same location (or its vicinity) reflect the urban characteristics of that location from different perspectives, and therefore share latent correspondences. Under this assumption, the correspondence between modalities can be learned indirectly by aligning them in a shared latent space through geographic coordinates as a common pivot, even in the absence of explicit cross-modal pairs. Building on this idea, we propose LocationBind, a framework for unified modeling of multimodal urban data. Figure 1 presents an overview of LocationBind, and Figure 2 shows its training architecture. LocationBind comprises a LocationEncoder that maps geographic coordinates into the shared latent space, modality-specific ModalityProjectors that project each modality embedding into the same space, and ModalityDecoders that reconstruct modality embeddings from shared latent representations. By using geographic coordinates as the pivot for alignment, the framework integrates modalities with different spatial resolutions and observation densities into a shared latent space, without requiring explicit cross-modal pairs or pre-aggregation into fixed spatial units. This design enables cross-modal capabilities that are difficult to support with existing methods, including missing-modality completion, cross-modal retrieval, and learning unified location representations that integrate multiple modalities. Furthermore, through case studies and qualitative analyses, we show that LocationBind can serve as a versatile foundation for multifaceted urban understanding.

The main contributions of this study are as follows:

- We propose LocationBind, a framework that uses geographic coordinates as a pivot to embed multiple urban modalities into a shared latent space without requiring explicit cross-modal pairs.
- We show that LocationBind achieves strong and balanced performance across missing-modality completion, cross-modal retrieval, and urban-indicator prediction.
- We demonstrate the utility of the learned latent space through applications, including cross-modal pattern discovery and compositional cross-modal retrieval.

2 Related Work

2.1 Unimodal Embeddings of Urban Data

As urban data become increasingly available, many self-supervised representation learning methods have been proposed to uncover the latent patterns embedded in such data. Most of these methods associate or aggregate observations with discrete spatial units such as POIs or grid cells, and embed them into a latent space. Existing work has covered modalities including POIs [5, 6, 16], satellite imagery [17–20], street view images [7, 21], and mobility logs [4, 8–10, 22–24]. These unimodal embeddings have been used for visualizing geographic space and as transferable features for a range of downstream tasks.

A notable recent development is the emergence of location encoders [25–27], which map continuous geographic coordinates into an embedding space. Whereas earlier work typically modeled geographic information at discrete and coarse spatial scales ranging from hundreds of meters to several kilometers, location encoders can represent locations at a finer resolution from a few meters to tens of meters while explicitly modeling continuous spatial proximity. A key component of many location encoders is multi-scale positional encoding, which makes it possible to model global and local geographic relationships simultaneously. Building on this, several methods perform contrastive learning between coordinate-based geographic embeddings and modality-specific embeddings. GeoCLIP [12] aligns geotagged image embeddings with geographic coordinate embeddings, enabling geolocalization. SatCLIP [13] performs contrastive learning between satellite imagery and a global location encoder, and transfers the resulting

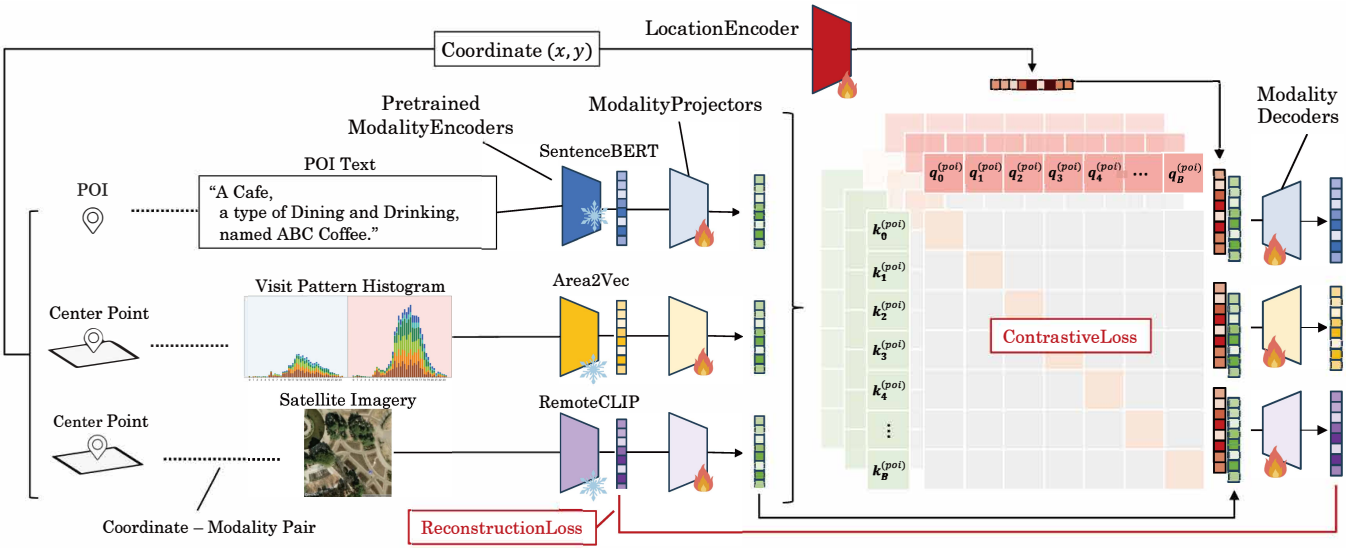


Figure 2: The architecture of LocationBind.

coordinate embeddings to location-dependent tasks such as temperature prediction, species recognition, and population density estimation. CaLLiPer [14] aligns POI text embeddings with geographic coordinate embeddings to obtain semantically informed coordinate embeddings, and applies them to downstream tasks such as land use classification and socioeconomic attribute prediction. Because these methods align geographic coordinate representations with only a single non-coordinate modality, we refer to them as single-modality coordinate alignment methods in this paper, distinguishing them from approaches that align multiple non-coordinate modalities in a shared latent space.

2.2 Multimodal Representation Learning

Representation learning that relies on a single modality has inherent limitations, particularly when observations are sparse or missing, or when the multifaceted nature of a location must be captured through representations that integrate multiple modalities. Motivated by this, extensive research has been conducted on multimodal representation learning. Outside the urban domain, contrastive learning with image–text pairs has emerged as a leading approach, exemplified by CLIP [28], which learns a shared latent space from large-scale image–text pairs and enables cross-modal retrieval as well as zero-shot classification. Multimodal learning has since been extended beyond two modalities, making the alignment of many modalities into a single shared latent space increasingly feasible. ImageBind [15] treats images as an anchor modality and maps diverse observations paired with images, including audio, depth, thermal signals, and IMU measurements, into the image embedding space. Such alignment via an anchor modality is promising in realistic settings where complete pairings for all modality combinations are unavailable, and it enables cross-modal retrieval and transfer even under such conditions.

Similar multimodal representation learning has also been explored in urban computing [11, 29–32]. For instance, KnowCL [11] constructs an urban graph representing the spatial proximity and similarity of urban regions and POIs, and aligns this graph representation with satellite imagery and street view images in the same latent space. MoRA [33] uses graph representations derived from human mobility data as an anchor modality and aligns other modalities with this anchor in a shared latent space. These studies suggest that multimodal integration improves downstream task performance in many settings and provides a more comprehensive understanding of locations than single-modality approaches.

3 Method

LocationBind is trained on pairs of a geographic coordinate and a single-modality observation, as illustrated in Figure 2. The model jointly learns a LocationEncoder, modality-specific ModalityProjectors, and ModalityDecoders, all operating in a shared latent space. Raw modality data are first converted into fixed embeddings using pretrained ModalityEncoders, whose weights are kept frozen. This keeps large backbone encoders out of the optimization and stabilizes alignment based on coordinate–modality pairs. Training jointly minimizes a contrastive loss and a reconstruction loss.

3.1 Paired Data and Modality-Specific Embeddings

We construct the set of coordinate–embedding pairs independently for each modality. We denote the pair set for modality m as $\mathcal{D}^{(m)} = \{(\ell_i^{(m)}, z_i^{(m)})\}_{i=1}^{N_m}$, where $N_m = |\mathcal{D}^{(m)}|$, $\ell_i^{(m)} = (x_i, y_i)$ is a planar geographic coordinate, and $z_i^{(m)}$ is a fixed embedding produced by the pretrained ModalityEncoder for modality m. We consider three modalities, $\mathcal{M} = \{\text{poi}, \text{visit}, \text{sat}\}$: POI text, visit pattern histograms, and satellite imagery. The construction of $z_i^{(m)}$ for each modality is described below.

3.1.1 POI text. We represent POIs, such as buildings, facilities, and landmarks, using text constructed from their categories and names.

Following CaLLiPer [14], we format a two-level POI category and the POI name into a single Japanese sentence (e.g., a sentence conveying "a cafe named ABC Coffee in the food-and-drink category"). The sentence is encoded by a Sentence-BERT [34] model pretrained on Japanese text to obtain $z^{(\mathrm{poi})}$, with ℓ set to the coordinate of the corresponding POI. This modality captures the semantic identity of facilities and landmarks present at a location.

3.1.2 Visit pattern. Following Area2Vec [10], we build a $2 \times 24 \times 7$ frequency vector for each grid cell, summarizing visits along three axes: weekday/holiday (2 classes), hour of day (24 classes), and length of stay (7 classes). Area2Vec is trained on these vectors, and its output is used as the fixed visit pattern embedding $z^{(\mathrm{visit})}$, with ℓ set to the area centroid. This modality captures behavioral features such as stay tendencies and temporal usage patterns around each location.

3.1.3 Satellite imagery. We use satellite images centered on each location and encode them with RemoteCLIP [35], a model pretrained on a large-scale satellite image corpus, to obtain $z^{(\mathrm{sat})}$. The coordinate ℓ is set to the image center. This modality captures visual features of the urban surface, such as building shapes, building density, and road patterns.

3.2 Multimodal Alignment Learning

3.2.1 LocationEncoder. The LocationEncoder f must handle differences in spatial resolution and observation density across modalities while modeling multi-scale neighborhood relationships among coordinates. Following CaLLiPer [14], we combine a grid positional encoding (PE) [27] with a residual MLP:

$$\begin{aligned} f(x,y) &= NN(PE(x,y)) \\ PE(x,y) &= \Big[\cos\tfrac{x}{\alpha_0}, \sin\tfrac{x}{\alpha_0}, \cos\tfrac{y}{\alpha_0}, \sin\tfrac{y}{\alpha_0}, \cdots, \\ &\qquad \cos\tfrac{x}{\alpha_{S-1}}, \sin\tfrac{x}{\alpha_{S-1}}, \cos\tfrac{y}{\alpha_{S-1}}, \sin\tfrac{y}{\alpha_{S-1}}\Big] \\ \alpha_s &= r_{\min} \cdot \left(\tfrac{r_{\max}}{r_{\min}}\right)^{\frac{s}{S-1}}. \end{aligned} \tag{1}$$

Here, (x, y) is a planar coordinate in meters, and NN is a fully connected residual network (FC-Net). The hyperparameters $r_{\min}$, $r_{\max}$, and S control the spatial frequency bands the encoder can capture. The output is a d-dimensional representation that jointly reflects global position and local neighborhood structure.

3.2.2 ModalityProjectors and ModalityDecoders. Each ModalityProjector $g^{(m)}$ is an independent MLP with one hidden layer that maps a d_m-dimensional modality embedding into the d-dimensional shared latent space; the hidden dimension is $2d_m$. Each ModalityDecoder $h^{(m)}$ is structurally symmetric: it maps the d-dimensional shared representation back to d_m via a hidden layer of size $2d_m$. We apply L2 normalization to the decoder output.

3.2.3 Contrastive loss. For each pair $(\ell_i^{(m)}, z_i^{(m)})$, we obtain the L2-normalized representations $\boldsymbol{q}_i^{(m)} = \mathrm{normalize}(f(\ell_i^{(m)}))$ and $\boldsymbol{k}_i^{(m)} = \mathrm{normalize}(g^{(m)}(z_i^{(m)}))$. Using the other samples of the same modality in the mini-batch as negatives, we define the location-to-modality InfoNCE loss [38]:

$$\mathcal{L}_{l\to m}^{(m)} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\boldsymbol{q}_i^{(m)\top}\boldsymbol{k}_i^{(m)}/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{q}_i^{(m)\top}\boldsymbol{k}_j^{(m)}/\tau)}, \tag{2}$$

where $\tau = 0.07$ is the temperature and B is the per-modality mini-batch size. The reverse-direction loss $\mathcal{L}_{m\to l}^{(m)}$ is defined symmetrically.

3.2.4 Reconstruction loss. To encourage the shared latent space to retain modality-specific information and to support modality completion, we train each ModalityDecoder on a stochastically sampled input that is either the coordinate representation, the modality representation, or their mean:

$$\boldsymbol{u}_i^{(m)} = \mathrm{normalize}\left(\begin{cases} (\boldsymbol{k}_i^{(m)} + \boldsymbol{q}_i^{(m)})/2 & \text{w.p. } p_{\mathrm{both}} \\ \boldsymbol{q}_i^{(m)} & \text{w.p. } p_{\mathrm{loc}} \\ \boldsymbol{k}_i^{(m)} & \text{w.p. } p_{\mathrm{mod}} \end{cases}\right), \tag{3}$$

with $p_{\mathrm{both}} = 0.50$ and $p_{\mathrm{loc}} = p_{\mathrm{mod}} = 0.25$. This mixing trains the decoder to recover the original embedding from either input alone or from their mean. Using $\bar{z}_i^{(m)} = z_i^{(m)}/\|z_i^{(m)}\|_2$ as the target, the reconstruction loss is the cosine distance

$$\mathcal{L}_{\mathrm{decode}}^{(m)} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - h^{(m)}(\boldsymbol{u}_i^{(m)})^{\top}\bar{z}_i^{(m)}\right). \tag{4}$$

3.2.5 Total objective. The final loss averages the bidirectional contrastive losses and the reconstruction loss across modalities:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\left[\frac{1}{2}\left(\mathcal{L}_{l\to m}^{(m)} + \mathcal{L}_{m\to l}^{(m)}\right) + \lambda_{\mathrm{decode}}\,\mathcal{L}_{\mathrm{decode}}^{(m)}\right], \tag{5}$$

where $\lambda_{\mathrm{decode}}$ balances the two terms. Because the contrastive loss aligns each modality with the same coordinate representation, representations of different modalities are aligned indirectly through this common pivot, while the reconstruction loss preserves modality-specific information within the shared space.

4 Experiments

We conduct experiments on two major urban areas in Japan, Osaka and Nagoya. Both cities are core cities of their respective metropolitan areas and contain diverse spatial configurations, ranging from dense city centers to suburban areas. They also provide suitable testbeds because relevant urban indicator datasets are available for both cities. All geographic coordinates are projected into a planar metric coordinate system using the Japan Plane Rectangular Coordinate System (EPSG:6674 for Osaka and EPSG:6675 for Nagoya). For each city, we collect three modalities and train a separate LocationBind model. We then evaluate the trained models on three downstream tasks: missing-modality completion, cross-modal retrieval, and urban-indicator prediction.

4.1 Modality Datasets

For each city, we collect three urban modalities: POI text, visit patterns, and satellite imagery. Table 1 summarizes the statistics of the modality datasets. All modalities are associated with geographic coordinates, but they differ in spatial unit and observation density: POIs are point observations, visit patterns are aggregated on a 50 m

Table 1: Statistics of each modality dataset and the input embeddings.

Modality	Source and period	Input embedding (dim.)	Spatial unit	Data volume (Osaka)	Data volume (Nagoya)
POI text	Foursquare [3], 2023	Sentence-BERT (768)	Point	188,390	78,224
Visit pattern	Blogwatcher [36], 2023	Area2Vec (64)	50 m grid	64,880	74,202
Satellite imagery	Mapbox [37], 2025	RemoteCLIP (512)	250 m grid	4,320	6,096

grid, and satellite images are collected on a 250 m grid. This setting reflects the heterogeneous nature of urban observations.

4.1.1 POI text. POI data are obtained from Foursquare [3]. Each POI entry contains the POI name and a hierarchical category label with up to five levels. We use the level-1 category as the major category and the finest available category label as the subcategory, and combine them with the POI name to construct a Japanese text description for each POI. The resulting text is encoded by the Japanese Sentence-BERT model [39], yielding a 768-dimensional embedding for each POI. The geographic coordinate of each POI is used as the corresponding location.

4.1.2 Visit pattern. Visit patterns are constructed from GPS-based human stay logs provided by Blogwatcher [36]. Mobility logs from consenting users are abstracted into stay records, each containing latitude, longitude, stay start time, and length of stay. We aggregate these records on a 50 m grid and represent each grid cell by temporal stay features. Area2Vec is then used to obtain a 64-dimensional visit-pattern embedding for each grid cell. To protect user privacy and avoid sparsely observed cells, grid cells with fewer than 100 visits are excluded when pretraining Area2Vec and LocationBind.

4.1.3 Satellite imagery. Satellite images are acquired using the Mapbox API [37]. Each city is partitioned into a 250 m grid, and a 512×512 pixel image at zoom level 17 is downloaded for each grid cell, using the grid centroid as the image center. Because historical imagery matching the POI and mobility periods was not available, we used 2025 imagery, the most recent available at the time of collection. Following the RemoteCLIP preprocessing protocol [35], images are resized to 224×224 pixels and encoded by RemoteCLIP ViT-B/32, yielding a 512-dimensional image embedding.

4.2 Evaluation Tasks and Metrics

We evaluate LocationBind from three perspectives. The first two tasks directly examine the quality of the shared cross-modal latent space: missing-modality completion and cross-modal retrieval. The third task, urban-indicator prediction, evaluates whether the learned coordinate representations are also useful as general-purpose features for standard urban analysis.

4.2.1 Missing-modality completion. Given a target modality $m \in \{\text{poi}, \text{visit}, \text{sat}\}$ designated as missing, we evaluate whether its true embedding $z^{(m)}$ can be reconstructed from the other observed modalities. We use two splits. *Random* assigns samples uniformly to five folds. *Region* applies K-means to the coordinates (x, y) to divide each city into five contiguous regions and holds out one region per fold. For each fold, we exclude only the held-out target-modality samples, allowing the model to learn that modality's projector and decoder from the remaining samples without seeing the evaluation embeddings. To complete the missing modality, for each target location we retrieve the geographically nearest sample of each non-missing modality, project these into the shared latent space via the ModalityProjectors, average the projections, and pass the mean to the ModalityDecoder $h^{(m)}$ to obtain $\hat{z}^{(m)}$. Although LocationBind can in principle complete a modality from the geographic coordinate alone, we restrict this experiment to the observed modalities, for fair comparison with ImageBind and MoRA, which lack a coordinate encoder. The evaluation metric is the cosine similarity between $\hat{z}_i^{(m)}$ and $z_i^{(m)}$, averaged over five folds.

4.2.2 Cross-modal retrieval. Given a query in one modality, we retrieve the corresponding instance in another modality within the shared latent space. Since our training data contain only coordinate–modality pairs and no explicit cross-modal pairs, we construct evaluation pairs from geographic correspondence. For POI–Visit pairs, we use isolated POIs to reduce ambiguous correspondences. Specifically, we retain POIs with at least 100 stays within a 30 m radius and no other POI within the same radius, aggregate the stays into a temporal visit pattern, and encode it in the same Area2Vec space used during training. For POI–Sat pairs, the POIs falling within the 250 m bounding box of each satellite image tile are taken as positives. For Visit–Sat pairs, the 50 m grid visit pattern embeddings within the same bounding box are used as positives. When a satellite tile contains multiple POIs or 50 m grid cells, the POI text or visit-pattern embeddings are averaged to form the paired representation. Both the query and candidate embeddings are mapped into the shared latent space via the corresponding ModalityProjector. For each positive pair, we sample 1,000 negatives and report R@1 and R@10 based on cosine similarity.

4.2.3 Urban-indicator prediction. This task evaluates whether the pretrained representations capture features useful for predicting urban indicators at each location. Following prior work on geographic embedding evaluation [13, 14, 33, 35], we obtain a fixed representation for each 50 m grid cell and train a linear probe to predict the indicators of that cell. For LocationBind, SatCLIP, and CaLLiPer, we feed the centroid coordinate of each grid cell into the LocationEncoder. For GeoCLIP, we use its publicly released LocationEncoder. For methods that cannot embed continuous coordinates [10, 33, 35], we use the embedding for each discrete spatial unit, obtained from a model pretrained on the same data as LocationBind, as input. We evaluate three indicators, land use classification (LUC), population regression (PR), and building height regression (BHR). LUC is an 11-class classification problem evaluated by the Macro-F1 score, where we remap land use survey data [40, 41] from each local government into 11 unified classes on the 50 m grid. PR and BHR are regression tasks evaluated by the root mean squared error (RMSE). The targets are aggregated to the 50 m grid from the national census [42] for

Table 2: Results of missing-modality completion. Each value is the mean ± standard deviation of cosine similarity (↑) over five folds. POI, Visit, and Sat indicate the missing modality.

		POI		Visit		Sat	
City	Model	*Random*	*Region*	*Random*	*Region*	*Random*	*Region*
Osaka	ImageBind	0.445 ± 0.006	0.423 ± 0.016	0.711 ± 0.008	0.706 ± 0.024	0.753 ± 0.004	0.740 ± 0.025
	MoRA	0.466 ± 0.009	0.455 ± 0.030	0.714 ± 0.010	0.709 ± 0.029	0.742 ± 0.008	0.737 ± 0.044
	LocationBind	**0.642 ± 0.000**	**0.618 ± 0.021**	**0.814 ± 0.001**	**0.798 ± 0.015**	**0.818 ± 0.003**	**0.809 ± 0.039**
Nagoya	ImageBind	0.423 ± 0.003	0.389 ± 0.007	0.634 ± 0.007	0.624 ± 0.040	0.746 ± 0.001	0.723 ± 0.019
	MoRA	0.448 ± 0.004	0.428 ± 0.018	0.667 ± 0.011	0.652 ± 0.014	0.747 ± 0.004	0.735 ± 0.017
	LocationBind	**0.599 ± 0.001**	**0.562 ± 0.008**	**0.756 ± 0.001**	**0.738 ± 0.026**	**0.825 ± 0.002**	**0.812 ± 0.020**

Table 3: Results of cross-modal retrieval. Each value is the Recall@k (↑) of retrieval.

	Osaka						Nagoya					
	POI–Visit		POI–Sat		Visit–Sat		POI–Visit		POI–Sat		Visit–Sat	
Model	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
ImageBind	0.006	0.029	0.093	0.265	0.060	0.210	0.004	0.031	0.207	0.450	0.037	0.154
MoRA	0.007	0.052	0.021	0.107	0.017	0.098	0.010	0.068	0.017	0.081	0.009	0.061
LocationBind	**0.013**	**0.075**	**0.331**	**0.619**	**0.614**	**0.866**	**0.019**	**0.119**	**0.392**	**0.674**	**0.450**	**0.750**

population density and from PLATEAU [43] for the mean building height. Because of public-data availability constraints, some indicator targets are not perfectly aligned in observation period with the modality data used for pretraining; the exact years of each source are listed in Table 9. PR and BHR targets are standardized per city. For the linear probe, we use a linear classifier for LUC and a linear regressor for PR and BHR. We evaluate two five-fold splits. *Random* partitions 50 m grid cells uniformly at random, and *Region* holds out one of five spatially contiguous regions per fold. The same splits are applied to all methods, and we report the mean and standard deviation over the five folds.

4.3 Baselines

For missing-modality completion and cross-modal retrieval, we use ImageBind-style [15] and MoRA-style [33] baselines, both of which embed multiple modalities into a shared latent space through an anchor modality. ImageBind-style baseline uses satellite imagery as the anchor and independently aligns, via contrastive learning, the POIs falling within each satellite tile and the visit-pattern embeddings of the 50 m grid cells contained in that tile with the corresponding satellite-image representation. MoRA-style baseline uses human mobility as the anchor: it first pretrains LINE [44] embeddings on a 50 m grid graph built from GPS data, and in the contrastive phase feeds these LINE embeddings into a LightGCN encoder whose output is aligned with the embedding of each modality associated with the same 50 m grid cell.

For urban-indicator prediction, we compare LocationBind against representative geographic embedding methods. GeoCLIP [12] aligns a location encoder with geotagged images such as street-level photos; we use its publicly released pretrained weights. SatCLIP [13] learns a location encoder jointly with satellite image representations. CaLLiPer [14] contrastively aligns a location encoder with POI text representations. We train SatCLIP and CaLLiPer on the same modality data and use the same Grid+FC-Net location encoder and positional encoding hyperparameters as LocationBind. We additionally include per-grid embeddings from Area2Vec [10] and RemoteCLIP [35]. For MoRA, we map the LightGCN graph embeddings learned in the contrastive phase onto the 50 m grid and use them as input. All baselines use the same one-hidden-layer projector as LocationBind, except for anchor modalities, which are used directly.

4.4 Results

Table 2 reports missing-modality completion results. LocationBind achieves the highest cosine similarity for both cities, all target modalities, and both splits. The improvement is most pronounced when the target is POI, and consistent gains are observed for Visit and Sat. This indicates that the embedding of a missing modality can be reconstructed with high fidelity from observations of the other modalities through the aligned shared latent space. The *Region* split yields lower scores and larger variances than *Random* across all methods, confirming that completing unobserved regions is harder than *Random* missingness; even so, LocationBind retains a clear advantage over the baselines.

Next, Table 3 reports cross-modal retrieval results. LocationBind achieves the best recall for both cities, all modality pairs, and both metrics. The improvement is most pronounced for POI–Sat and Visit–Sat. The POI–Visit pair shows lower recall overall than the pairs involving satellite imagery, likely because the correspondence between POI semantics and visit patterns is less direct and is further degraded by fine-grained GPS noise in the association between facility-level POIs and nearby stays. Nevertheless, LocationBind consistently outperforms ImageBind and MoRA across all modality pairs, showing that coordinate-pivot alignment induces meaningful

Table 4: Results of urban-indicator prediction. Each value is the mean ± standard deviation over five folds. LUC is evaluated by Macro-F1 (↑), while PR and BHR are evaluated by RMSE (↓).

		LUC (↑)		PR (↓)		BHR (↓)	
City	Model	*Random*	*Region*	*Random*	*Region*	*Random*	*Region*
Osaka	Area2Vec	0.326 ± 0.006	0.283 ± 0.020	0.855 ± 0.018	0.839 ± 0.132	0.873 ± 0.021	0.870 ± 0.212
	RemoteCLIP	0.269 ± 0.004	0.212 ± 0.017	0.874 ± 0.021	0.840 ± 0.193	**0.856 ± 0.023**	**0.858 ± 0.223**
	GeoCLIP	0.145 ± 0.001	0.073 ± 0.024	0.942 ± 0.021	0.926 ± 0.192	0.925 ± 0.021	0.956 ± 0.253
	SatCLIP	0.265 ± 0.001	0.182 ± 0.010	0.897 ± 0.020	0.893 ± 0.147	0.888 ± 0.021	0.906 ± 0.235
	CaLLiPer	0.280 ± 0.005	0.215 ± 0.017	0.900 ± 0.020	0.869 ± 0.184	0.910 ± 0.019	0.919 ± 0.241
	MoRA	0.270 ± 0.003	0.192 ± 0.023	0.849 ± 0.018	0.860 ± 0.175	0.875 ± 0.021	0.902 ± 0.227
	LocationBind	**0.392 ± 0.007**	**0.342 ± 0.015**	**0.789 ± 0.018**	**0.762 ± 0.149**	0.865 ± 0.021	0.890 ± 0.236
Nagoya	Area2Vec	0.325 ± 0.004	0.293 ± 0.014	0.894 ± 0.011	0.870 ± 0.165	0.869 ± 0.024	0.855 ± 0.246
	RemoteCLIP	0.269 ± 0.004	0.204 ± 0.008	0.865 ± 0.009	0.841 ± 0.167	**0.838 ± 0.024**	**0.844 ± 0.252**
	GeoCLIP	0.134 ± 0.001	0.055 ± 0.011	0.943 ± 0.010	0.940 ± 0.168	0.913 ± 0.026	0.926 ± 0.268
	SatCLIP	0.246 ± 0.002	0.168 ± 0.012	0.903 ± 0.011	0.922 ± 0.152	0.883 ± 0.026	0.913 ± 0.257
	CaLLiPer	0.237 ± 0.002	0.171 ± 0.015	0.932 ± 0.009	0.923 ± 0.164	0.929 ± 0.023	0.931 ± 0.256
	MoRA	0.214 ± 0.003	0.124 ± 0.026	0.887 ± 0.010	0.932 ± 0.174	0.898 ± 0.025	0.983 ± 0.300
	LocationBind	**0.384 ± 0.005**	**0.319 ± 0.020**	**0.838 ± 0.010**	**0.822 ± 0.157**	0.880 ± 0.026	0.888 ± 0.256

cross-modal correspondences among semantic, behavioral, and visual urban modalities.

Finally, Table 4 reports the urban-indicator prediction results. LocationBind achieves the best performance on LUC and PR across both cities and both split settings. The gains are especially clear for LUC, suggesting that the coordinate-pivot representation effectively captures semantic information related to urban functions. LocationBind also performs strongly on PR, indicating that the shared representation retains socioeconomic signals associated with population distribution. For BHR, RemoteCLIP achieves the best performance in both cities and both splits, as building height is a visually grounded indicator that RemoteCLIP encodes directly from satellite-image features. Nevertheless, LocationBind remains competitive on BHR while using a shared coordinate-pivot representation that balances semantic, mobility, and visual information. These results indicate that LocationBind is not specialized to a single indicator, but provides a general-purpose coordinate representation that supports both cross-modal reasoning and standard urban prediction tasks.

4.5 Additional Analysis

4.5.1 Coordinate-only Modality Completion. The missing-modality completion deliberately restricts the input to the observed modalities, for fair comparison with ImageBind and MoRA, which have no coordinate encoder. Here, we instead evaluate the capability that is specific to coordinate-pivot models, recovering a modality embedding from the geographic coordinate alone. For a target modality m, we feed only the coordinate representation $\boldsymbol{q}^{(m)} = \text{normalize}(f(\ell))$ into the ModalityDecoder $h^{(m)}$, i.e., the location-only branch (p_{loc}) of the reconstruction objective in Eq. (3). We compare against two types of coordinate-only baselines. Spatial NN returns the embedding of the geographically nearest observed sample of the target modality, and thus reflects the raw geographic-proximity prior. Since CaLLiPer and SatCLIP are single-modality coordinate-alignment methods that lack a decoder, we encode the query coordinate with their location encoder and retrieve the target-modality embedding of the nearest training sample in the aligned latent space. Because each aligns coordinates with a single modality, CaLLiPer applies only to POI and SatCLIP only to satellite imagery.

Table 5 shows that LocationBind outperforms every baseline on all three target modalities under both splits. Two observations stand out. First, LocationBind consistently exceeds Spatial NN, confirming that the learned coordinate-pivot representation captures modality-specific urban structure beyond what geographic proximity alone provides. Second, this advantage widens under the harder *Region* split: because it holds out entire contiguous areas, a proximity-based estimate degrades sharply, whereas LocationBind degrades far less, indicating that its gains cannot be explained solely by spatial interpolation. The improvement is largest for POI and Visit. For satellite imagery the margin is smaller, as RemoteCLIP embeddings vary slowly over space and are therefore already well approximated by spatial nearest neighbors.

4.5.2 Ablation Study. The training objective of LocationBind comprises a contrastive loss and a reconstruction loss. To examine the contribution of each, we evaluate two ablations. **w/o Decoder** removes the ModalityDecoders and the reconstruction loss, retaining only the contrastive alignment between coordinates and each modality. Since this variant has no decoder, completion is performed by averaging the observed modality representations in the shared latent space and returning the training sample of the target modality nearest to that mean. **w/o CL** removes the contrastive loss and trains only the reconstruction loss, so each ModalityDecoder is trained independently to reconstruct its modality embedding from the geographic coordinate without any inter-modality alignment.

Table 6 reports the results. For missing-modality completion, the Full Model achieves the best performance on all three target modalities, indicating that completion requires both aligned shared representations and modality-specific reconstruction. For cross-modal retrieval, w/o Decoder slightly outperforms the Full Model, especially on POI–Sat and Visit–Sat, whereas w/o CL performs poorly across all modality pairs. This confirms that the contrastive

Table 5: Results of coordinate-only modality completion. Each value is the cosine similarity (↑) averaged over the five folds and over Osaka and Nagoya.

	POI		Visit		Sat	
Model	*Random*	*Region*	*Random*	*Region*	*Random*	*Region*
Spatial NN	0.500	0.374	0.742	0.659	0.823	0.748
CaLLiPer	0.481	0.361	–	–	–	–
SatCLIP	–	–	–	–	0.825	0.751
LocationBind	**0.615**	**0.567**	**0.795**	**0.754**	**0.849**	**0.814**

Table 6: Results of the ablation study. Each value is the average over Osaka and Nagoya. Missing-modality completion reports cosine similarity in the *Region* setting. Cross-modal retrieval reports R@10. For Urban-indicator prediction, LUC is Macro-F1, while PR and BHR are RMSE, using the *Region* setting.

	Missing-modality completion			Cross-modal retrieval			Urban-indicator prediction		
Model	POI (↑)	Visit (↑)	Sat (↑)	POI–Visit (↑)	POI–Sat (↑)	Visit–Sat (↑)	LUC (↑)	PR (↓)	BHR (↓)
w/o Decoder	0.417	0.688	0.745	**0.097**	**0.678**	**0.817**	0.315	0.794	0.899
w/o CL	0.492	0.562	0.749	0.013	0.032	0.020	0.214	0.832	**0.855**
Full Model	**0.590**	**0.768**	**0.810**	**0.097**	0.646	0.808	**0.330**	**0.791**	0.889

Table 7: Results of the zero-shot cross-city transfer experiment. Each value is the average over both transfer directions. Missing-modality completion reports cosine similarity in the *Region* setting. Cross-modal retrieval reports R@10.

	Missing-modality completion			Cross-modal retrieval		
Model	POI	Visit	Sat	POI–Visit	POI–Sat	Visit–Sat
ImageBind	0.404	0.672	0.731	0.016	0.029	**0.066**
MoRA	0.432	0.681	0.735	0.041	0.024	0.042
LocationBind	**0.583**	**0.769**	**0.803**	**0.059**	**0.049**	0.057

loss is essential for cross-modal alignment. For urban-indicator prediction, the Full Model achieves the best performance on LUC and PR, while w/o CL performs best on BHR. Overall, the two losses play complementary roles. The contrastive loss is crucial for organizing modalities into a shared latent space, while the reconstruction loss enables modality completion and helps preserve information useful for coordinate-based prediction. Although the Full Model is not the single best on every metric, it provides the most balanced performance across completion, retrieval, and urban-indicator prediction.

4.5.3 Cross-City Zero-Shot Transfer. The previous experiments evaluate models on the same city used for training. To examine whether the shared latent space generalizes beyond a single city, we conduct a cross-city zero-shot transfer experiment: we apply a model pretrained on all data from one city directly to all data from the other, and evaluate on missing-modality completion and cross-modal retrieval. Urban-indicator prediction is excluded because the LocationEncoder is city-specific. We report results averaged over the two transfer directions.

Table 7 shows the results. For missing-modality completion, LocationBind is the best on all three modalities and maintains performance close to the within-city setting, with the advantage over the baselines preserved. This indicates that aligning modalities via geographic coordinates captures partially transferable cross-modal correspondences that do not rely on city-specific details, supporting the cross-city generalization of the learned shared latent space. In contrast, the performance drop under transfer is noticeable for cross-modal retrieval, especially on pairs that include Sat. Unlike completion, which only requires local correspondence from the observed modalities to the missing one, retrieval must identify the positive among 1,000 negatives and is therefore sensitive to the global structure of the latent space, making it vulnerable to distributional differences between cities. Satellite imagery in particular differs substantially between cities in building style, density, and color tone, so the pretrained encoder's output may exhibit city-specific bias, weakening the ModalityProjector's alignment in the target city. As a result, ImageBind surpasses LocationBind on the Visit–Sat pair, as it anchors all modalities on the RemoteCLIP satellite representation, whose large-scale pretraining yields city-independent features. LocationBind, by contrast, optimizes all modality representations without an explicit cross-city constraint, leaving its shared space more susceptible to city-specific bias.

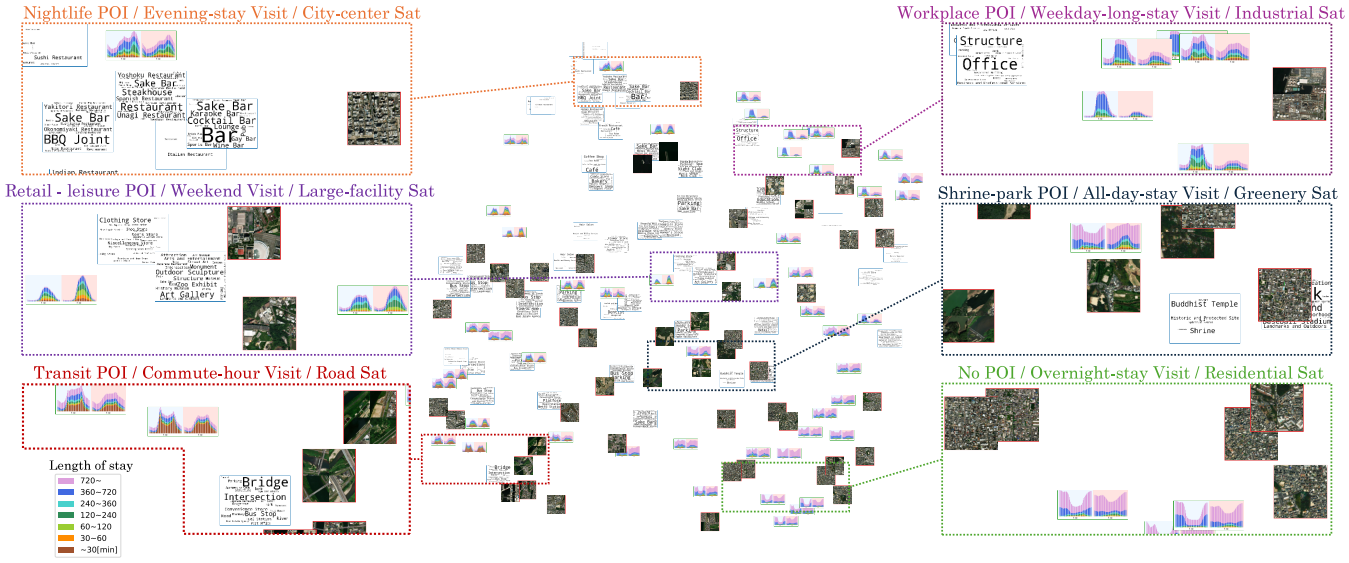
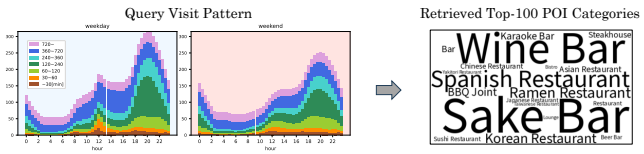
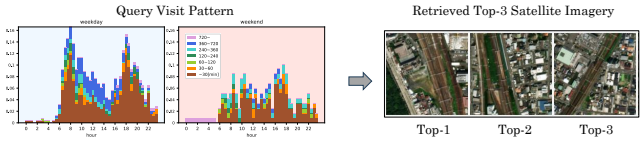


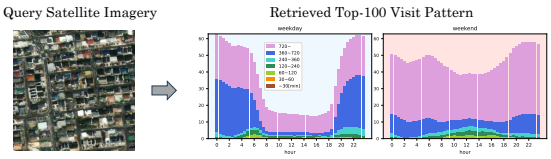
Figure 3: Cross-modal pattern discovery. UMAP projection of 150 cluster centroids (50 per modality) in the shared latent space.



(a) Visit Pattern to POI Text.



(b) Visit Pattern to Satellite Imagery.



(c) Satellite Imagery to Visit Pattern.

Figure 4: Cross-modal retrieval case studies.

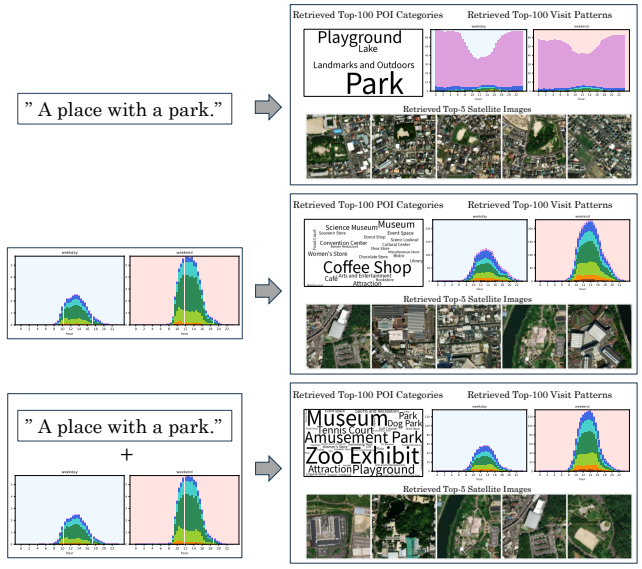


Figure 5: Compositional cross-modal retrieval.

5 Applications

5.1 Cross-modal Pattern Discovery

To examine whether the shared latent space encodes coherent cross-modal structure, we cluster each modality independently into 50 clusters across POI text, visit patterns, and satellite imagery. We compute the centroid of each cluster in the shared latent space and project the 150 centroids jointly into two dimensions via UMAP. At each projected position, we place a glyph summarizing the cluster: a word cloud of representative POI categories, a 2×24 heatmap of the mean visit pattern, or a medoid satellite image. The result is shown in Figure 3. Clusters from different modalities that describe the same kind of urban environment consistently lie in nearby regions of the projected space. For example, the *Nightlife / Evening-stay / City-center* region combines bar and izakaya POIs, evening-peaking visit patterns, and densely built-up cores; Figure 3 annotates six such patterns in total, ranging from *Workplace / Weekday-long-stay / Industrial* to *Shrine-park / All-day-stay / Greenery*. Because LocationBind is trained only on coordinate–modality pairs, these co-locations emerge without any cross-modal supervision, indicating

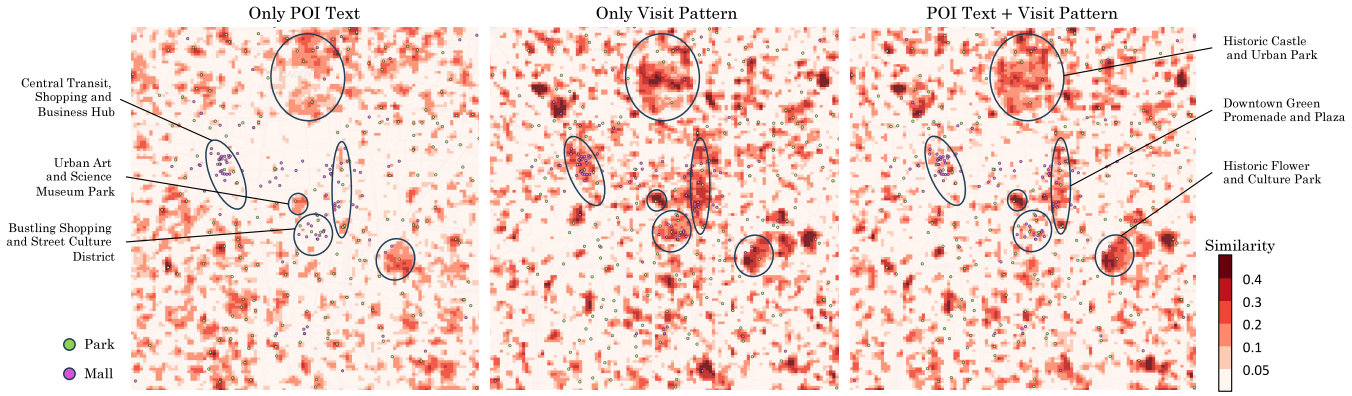


Figure 6: Compositional multimodal queries over locations in Nagoya.

that coordinate-pivot alignment alone organizes heterogeneous modalities into semantically coherent patterns.

5.2 Cross-modal Retrieval Case Studies

Beyond aggregated patterns, the shared latent space supports retrieval from any modality to any other modality at the instance level. Figure 4 shows three representative cases. In case (a), the query is a visit pattern with sharp evening peaks. The top-100 retrieved POIs form a word cloud dominated by Wine Bar, Sake Bar, Spanish Restaurant, and similar nightlife categories, indicating that the temporal signature is associated with food-and-drink establishments. In case (b), the query is a visit pattern with short stays concentrated in the morning and evening commute hours. The top-3 retrieved satellite images show bridges, railway lines, and arterial roads, consistent with commuting infrastructure. In case (c), the query is a satellite image of a low-rise residential neighborhood, and the top-100 retrieved visit patterns concentrate on long overnight stays, matching the expected daily rhythm of residential areas. These cases show that less directly interpretable modalities, such as visit-pattern histograms or satellite tiles, become interpretable through retrieval from a more human-readable modality such as POI categories.

5.3 Compositional Cross-modal Queries

Because all modalities share a common latent space, queries can be composed across modalities using simple vector operations. We encode two queries via their ModalityProjectors: (i) the POI text "A place with a park" and (ii) a visit pattern peaking on holiday afternoons, and we take their normalized mean as a composed query. For each of the two single-modality queries and their composition, we retrieve the nearest POI categories, visit patterns, and satellite images (Figure 5), as well as the most similar 50 m grid locations in Nagoya via the LocationEncoder (Figure 6). Each single-modality query returns results consistent with its own semantics. The POI-text query retrieves small neighborhood parks and outdoor-landmark categories, concentrated on residential parks scattered across the city, whereas the visit-pattern query retrieves weekend-leisure venues such as coffee shops and large retail facilities, concentrated on commercial centers and transit hubs. The composed query combines both aspects: it retrieves categories such as Zoo, Museum, and Amusement Park and satellite images of urban parks and cultural complexes, and its geographic heatmap (Figure 6) concentrates on landmark green spaces such as the central park-and-promenade district, the science-museum park, and the historic castle grounds. The composed query thus localizes to places that are simultaneously park-like and active on weekends, a combination neither single-modality query can express alone, demonstrating that compositional cross-modal queries arise naturally from the shared latent space without any additional training objective.

6 Conclusion

We presented LocationBind, a coordinate-pivot framework that aligns heterogeneous urban modalities into a shared latent space from coordinate–modality pairs alone, combining contrastive alignment with modality reconstruction. Experiments showed that LocationBind outperforms baselines on most tasks and remains competitive on the others across missing-modality completion, cross-modal retrieval, and urban-indicator prediction. Case studies further illustrated that the learned space supports cross-modal pattern discovery, instance-level retrieval, and compositional queries.

Several limitations remain. Cross-city retrieval, especially for pairs involving satellite imagery, still suffers from distribution shifts between cities. In addition, the current LocationEncoder represents areal units only by their centroid coordinates. Future work will address domain adaptation across cities, extend the model to additional modalities such as street-view images and transportation networks, and develop encoders that jointly handle continuous coordinates and discrete spatial units.

Acknowledgments

This work was supported by JST CREST (JPMJCR22M4), JST ACT-X (JPMJAX24M8), and JST RISTEX (JPMJRS23K). We would like to thank Blogwatcher Inc. for providing the data used in this paper. OpenAI's ChatGPT was used to assist with translation and language editing of the manuscript. The authors reviewed and edited all AI-assisted output incorporated into the manuscript and take full responsibility for the final content of this paper.

References

[1] Luís M. A. Bettencourt. The origins of scaling in cities. Science, 340(6139):1438–1441, 2013.

[2] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology, 5(3):38:1–38:55, 2014.

[3] Foursquare. Foursquare os places, 2025. Accessed: 2025-09-01.

[4] Naoki Tamura, Kazuyuki Shoji, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. Openuas: Embeddings of cities in japan with anchor data for cross-city analysis of area usage patterns, 2024.

[5] Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang, and Ke Mai. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. International Journal of Geographical Information Science, 31(4):825–848, 2017.

[6] Weiming Huang, Lizhen Cui, Meng Chen, Daokun Zhang, and Yao Yao. Estimating urban functional distributions with semantics preserved poi embedding. International Journal of Geographical Information Science, 36(10):1905–1930, 2022.

[7] Xiucheng Liang, Tianhong Zhao, and Filip Biljecki. Revealing spatio-temporal evolution of urban visual environments with street view imagery. Landscape and Urban Planning, 237:104802, 2023.

[8] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. Representing urban functions through zone embedding with human mobility patterns. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 3919–3925. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[9] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, and Xiaolin Li. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):906–913, Jul. 2019.

[10] Kazuyuki Shoji, Haru Terashima, Naoki Tamura, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. Life pattern based human attribute estimation using only raw gps data. IEEE Access, 7 2025.

[11] 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, WWW '23, page 4150–4160, New York, NY, USA, 2023. Association for Computing Machinery.

[12] Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In Advances in Neural Information Processing Systems, 2023.

[13] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. Proceedings of the AAAI Conference on Artificial Intelligence, 39(4):4347–4355, Apr. 2025.

[14] Xinglei Wang, Tao Cheng, Stephen Law, Zichao Zeng, Lu Yin, and Junyuan Liu. Multi-modal contrastive learning of urban space representations from poi data. Computers, Environment and Urban Systems, 120:102299, 2025.

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In CVPR, 2023.

[16] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '17, New York, NY, USA, 2017. Association for Computing Machinery.

[17] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 10161–10170, 2021.

[18] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, 2022.

[19] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 4065–4076, 2023.

[20] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9394–9403, 2021.

[21] Steven Stalder, Michele Volpi, Nicolas Büttner, Stephen Law, Kenneth Harttgen, and Esra Suel. Self-supervised learning unveils urban change from street-level images. Computers, Environment and Urban Systems, 112:102156, 2024.

[22] Hongjian Wang and Zhenhui Li. Region representation learning via mobility flow. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, page 237–246, New York, NY, USA, 2017. Association for Computing Machinery.

[23] Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1), Feb. 2017.

[24] Toru Shimizu, Takahiro Yabe, and Kota Tsubouchi. Enabling finer grained place embeddings using spatial hierarchy from human mobility trajectories. In Proceedings of the 28th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '20, page 187–190, New York, NY, USA, 2020. Association for Computing Machinery.

[25] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-Only Geographical Priors for Fine-Grained Image Classification. In ICCV, 2019.

[26] Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In The Eighth International Conference on Learning Representations. openreview, 2020.

[27] Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. ISPRS Journal of Photogrammetry and Remote Sensing, 202:439–462, 2023.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In ICML, 2021.

[29] 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, pages 1013–1020, 2020.

[30] Tianyuan Huang, Zhecheng Wang, Hao Sheng, Andrew Y. Ng, and Ram Rajagopal. Learning neighborhood representation from multi-modal multi-graph: Image, text, mobility graph and beyond. ArXiv, abs/2105.02489, 2021.

[31] Lubin Bai, Weiming Huang, Xiuyuan Zhang, Shihong Du, Gao Cong, Haoyu Wang, and Bo Liu. Geographic mapping with unsupervised multi-modal representation learning from vhr images and pois. ISPRS Journal of Photogrammetry and Remote Sensing, 201:193–208, 2023.

[32] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In Proceedings of the ACM Web Conference 2024, WWW '24, page 4006–4017, New York, NY, USA, 2024. Association for Computing Machinery.

[33] Ya Wen, Jixuan Cai, Qiyao Ma, Linyan Li, Xinhuan Chen, Chris Webster, and Yulun Zhou. MoRA: Mobility as the backbone for geospatial representation learning at scale. In The Fourteenth International Conference on Learning Representations, 2026.

[34] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv [cs.CL], August 2019.

[35] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62:1–16, 2024.

[36] Blogwatcher, Inc. Blogwatcher mobility dataset, 2026.

[37] Mapbox, Inc. Mapbox satellite imagery. https://www.mapbox.com/imagery, 2025. Accessed: 2025-11-20.

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In NeurIPS, 2018.

[39] Isao Sonobe. sonoisa/sentence-bert-base-ja-mean-tokens-v2: Japanese sentence-bert model. Hugging Face model repository, 2021. Accessed: 2025-09-01.

[40] City of Osaka, Urban Planning Division. Basic urban development data (mesh data etc.), 2025.

[41] Aichi Prefecture, Urban Planning Division. Urban planning survey (toshikeikaku no chosa), 2022.

[42] Statistics Bureau of Japan. 2020 population census of japan, small area tabulation (population by sex and number of households by basic unit block). https://www.e-stat.go.jp/stat-search/files?toukei=00200521&tstat=000001136464&tclass1=000001136472, 2022. Accessed: 2026-05-08.

[43] Ministry of Land, Infrastructure, Transport and Tourism (MLIT), Japan. Plateau open data, 2025. Accessed: 2026-01-17.

[44] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In Proceedings of the 24th International Conference on World Wide Web, WWW '15, page 1067–1077, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.

Table 8: Main hyperparameters used for LocationBind pretraining. Values are shared by Osaka and Nagoya unless otherwise noted.

Item	Value
Shared latent dimension d	128
Coordinate system	Planar coordinates in meters
LocationEncoder	Grid PE + FC-Net
PE scales $(r_{\min}, r_{\max}, S)$	(50m, 20000m, 2048)
Projector hidden dimension	$2d_m$
Decoder hidden dimension	$2d_m$
Contrastive temperature τ	0.07
Decoder loss weight λ_{decode}	1.0
Fusion probabilities	$p_{\text{both}} = 0.50$, $p_{\text{loc}} = p_{\text{mod}} = 0.25$
Optimizer	AdamW [45]
LocationEncoder learning rate	10^{-5}
Projector/Decoder learning rate	10^{-4}
LocationEncoder/Decoder weight decay	0.1
Projector weight decay	0.5
Batch size per modality	128
Training epochs	100

A Training and Implementation Details of LocationBind

We train LocationBind independently for Osaka and Nagoya using identical configurations. As described in Section 3, the pretrained ModalityEncoders are frozen, and only the LocationEncoder, ModalityProjectors, and ModalityDecoders are optimized. The LocationEncoder is a grid positional encoding followed by an FC-Net; each ModalityProjector maps a d_m-dimensional input into the shared space ($d = 128$) through one hidden layer of size $2d_m$, and each ModalityDecoder is its symmetric counterpart with L2-normalized output. During training, each mini-batch contains an equal number of samples from all three modalities, drawn independently and uniformly at random from each modality so that every modality contributes equally to each optimization step.

We optimize with AdamW, using linear warmup followed by cosine decay and separate learning rates for the LocationEncoder and the projector/decoder modules. The full hyperparameter configuration, shared by both cities, is listed in Table 8. By default, all coordinate–modality pairs of a city are used for training; in the missing-modality completion experiments the held-out fold of the target modality is excluded, while the other modalities remain fully observed (Appendix B).

B Details of the Missing-Modality Completion Experiment

We use two 5-fold splits. *Random* partitions each modality's samples uniformly with a seed shared across methods. *Region* (Figure 7) applies K-Means on the coordinates at the 250 m mesh level to form five contiguous regions; POIs and 50 m (Visit) cells inherit the region of their containing mesh, and one region is held out per fold.

For each combination of a missing target modality m and a fold k, the model is trained with the samples of m that belong to fold k excluded from the pretraining data. The non-target modalities are used as observed modalities with their data from all regions. Therefore, the ground truth embeddings of the missing modality used for evaluation are not used during training of the corresponding fold. These fold splits and training data exclusion settings are also applied identically to the baseline methods ImageBind and MoRA.

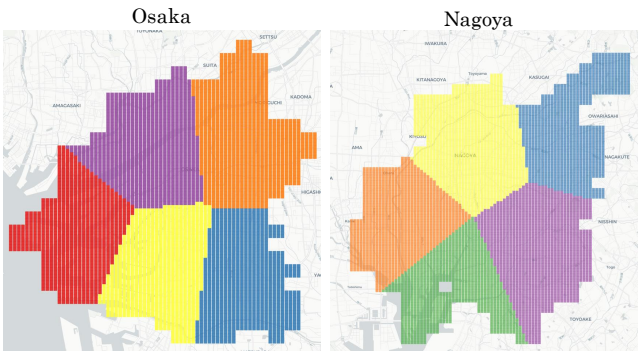


Figure 7: Mapping of region splits.

The evaluation set and the candidate set for nearest-neighbor-based completion in fold k are defined as follows. Here, $\text{split}(i, m) \in \{1, \ldots, 5\}$ is the function that indicates the fold to which sample i belongs in the 5-fold split of modality m.

$$\begin{aligned} I^{(m)} &= \{1, 2, \ldots, N_m\}, \\ T_{m,k} &= \{i \in I^{(m)} \mid \text{split}(i, m) = k\}, \\ C_{m,k} &= I^{(m)} \setminus T_{m,k}. \end{aligned} \tag{6}$$

Here, $I^{(m)}$ is the index set of all samples for modality m. The candidate set $C_{m,k}$ corresponds to the index set of the remaining samples used for training. In nearest neighbor based completion, retrieval candidates are restricted to the samples corresponding to $C_{m,k}$, and the ground truth samples in the evaluation fold are excluded from the candidates.

Let $\ell_i^{(m)}$ denote the coordinate of an evaluation sample $i \in T_{m,k}$. For each observed modality $m' \in \mathcal{M} \setminus \{m\}$, we retrieve the sample of m' that is closest to $\ell_i^{(m)}$.

$$j^*(i, m') = \arg\min_{j \in I^{(m')}} ||\ell_i^{(m)} - \ell_j^{(m')}||_2. \tag{7}$$

In this experiment, no maximum search radius is set, and the single nearest neighbor is used from each observed modality. The retrieved original embeddings are projected into the shared latent space by the Projector of each method, L2 normalized, and then averaged.

$$\bar{s}_i = \text{normalize}\left(\frac{1}{|\mathcal{M}| - 1} \sum_{m' \in \mathcal{M} \setminus \{m\}} g^{(m')}\left(z_{j^*(i,m')}^{(m')} \right) \right)$$

This $\bar{s}_i$ is used as the query representation for estimating the missing target modality m. For fair comparison, the geographic coordinate representation obtained from the LocationEncoder is not used as input for completion.

Since LocationBind has a ModalityDecoder, $\bar{s}_i$, which is composed only of the observed modalities, is fed into the decoder $h^{(m)}$

of the missing target modality.

$$\hat{z}_i^{(m)} = h^{(m)}(\bar{s}_i)$$

In contrast, since ImageBind and MoRA do not have a decoder in this experimental setting, the training candidate set $C_{m,k}$ of the missing target modality m is projected into the shared latent space, and the original embedding of the candidate most similar to $\bar{s}_i$ is used as the completion result.

$$j^\dagger = \arg\max_{j \in C_{m,k}} \bar{s}_i^\top \text{normalize}\left(g^{(m)}(z_j^{(m)})\right), \qquad \hat{z}_i^{(m)} = z_{j^\dagger}^{(m)}$$

For modalities without a Projector, such as the satellite image anchor in ImageBind, the L2 normalized original embedding is treated as the representation in the shared latent space. We report the mean and standard deviation, over the five folds, of the cosine similarity between $\hat{z}_i^{(m)}$ and $z_i^{(m)}$.

C Details of the Cross-Modal Retrieval Experiment

In this section, we describe the details of the cross-modal retrieval experiment. Unlike the missing-modality completion experiment, in cross-modal retrieval, no per-fold exclusion is performed, and a single model trained on all data from each city is used for evaluation. The evaluation targets are the three pairs of POI–Visit, POI–Sat, and Visit–Sat.

Let the set of positive pairs be $\mathcal{P}_{a,b} = \{(\tilde{z}_i^{(a)}, \tilde{z}_i^{(b)})\}_{i=1}^{N_{a,b}}$. For each pair, bidirectional retrieval is performed. The original embeddings are projected into the shared latent space by the Projector of each method and L2 normalized.

$$u_i^{(a)} = \text{normalize}\left(g^{(a)}(\tilde{z}_i^{(a)})\right),$$
$$v_i^{(b)} = \text{normalize}\left(g^{(b)}(\tilde{z}_i^{(b)})\right)$$

For modalities without a Projector, such as the satellite image anchor in ImageBind, the L2 normalized original embedding is used as the representation in the shared latent space.

For each query, a candidate set of 1,001 instances is constructed, consisting of one positive and 1,000 randomly sampled negatives. Negatives are sampled from the candidate side of the same pair set after excluding the positive, and all methods use the same positive-pair construction and evaluation settings. For each modality pair, bidirectional retrieval is performed in both $a \to b$ and $b \to a$, and Recall@1 and Recall@10 aggregated at the pair level are reported.

LocationBind, ImageBind, and MoRA are compared under the same retrieval pairs, candidate set, and negative sampling conditions. For LocationBind, the output of the ModalityProjector of each modality is used, and the Decoder is not used at retrieval time. For MoRA, the shared space learned with the mobility graph as the anchor is used, and at retrieval time, only the representations obtained by projecting each of the POI, Visit, and Sat modalities with the Projector are used.

D Details of the Urban-Indicator Prediction Experiment

In this section, we describe the details of the urban-indicator prediction experiment. This experiment uses the 50 m grid representation of each model as a fixed feature to predict three indicators. The task datasets are summarized in Table 9, and the 11 LUC classes are summarized in Table 10. PR and BHR targets are standardized per city before computing RMSE.

Table 9: Datasets used in the urban-indicator prediction experiment.

City	Task	Source	Year	Spatial unit	# Instances
Osaka	**LUC**	[40]	2022	50 m grid	81,998
	PR	[42]	2020	50 m grid	81,998
	BHR	[43]	2024	50 m grid	81,998
Nagoya	**LUC**	[41]	2021	50 m grid	125,741
	PR	[42]	2020	50 m grid	125,741
	BHR	[43]	2022	50 m grid	125,741

Table 10: The 11 land use classes used in the LUC task.

Land use class	# Meshes (Nagoya)	# Meshes (Osaka)
Business	3,353	8,743
Industrial	10,948	7,213
Educational facility	5,213	3,714
Cultural and religious	3,054	1,041
Medical	1,693	839
Commercial	5,268	4,199
Residential	44,694	26,225
Park	8,460	5,336
Road and railway	21,977	16,476
Farmland, forest, and river	14,335	377
Other	6,746	7,835
Total	125,741	81,998

For input features, for LocationBind, SatCLIP, and CaLLiPer, the representative coordinate of each 50 m grid is fed into the LocationEncoder to obtain the geographic coordinate representation. For GeoCLIP, latitude and longitude are fed into the publicly available LocationEncoder. Area2Vec, RemoteCLIP, and MoRA use precomputed mesh representations. For Area2Vec, we use the visit pattern embedding of the 50 m grid. For RemoteCLIP, we use the image embedding of the 250 m satellite image tile assigned to the corresponding 50 m grid. For MoRA, we use the region representations learned with the mobility graph as the anchor.

For the downstream predictor, all models use a linear probe. For LUC, the probe is a linear classifier with output dimension 11, trained with cross-entropy loss, and evaluated by Macro-F1. For PR and BHR, the probe is a linear regressor with output dimension 1, trained with mean squared error loss, and evaluated by RMSE. We use AdamW for optimization, with a learning rate of 1.0×10^{-4}, a weight decay of 1.0×10^{-1}, a batch size of 128, and 100 epochs.

We use two five-fold splits. In the *Random* split, 50 m grid cells are partitioned uniformly at random. In the *Region* split, we use the five-region partition shown in Figure 7; each fold uses four regions for training and the remaining region for testing. The same split assignment is applied to all models, and we report the mean and standard deviation over the five folds.