TRAJGANR: TRAJECTORY-CENTRIC URBAN MULTIMODAL LEARNING VIA GEOSPATIALLY ALIGNED NEURAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many urban prediction tasks depend not only on the static attributes of a location, such as the visual appearance of its built environment, but also on its dynamic *function*: how people navigate and use it over time. Predicting congestion, travel demand, or road safety risks, for instance, requires mobility information that imagery and geographic coordinates alone cannot provide. Trajectories carry this signal, yet current trajectory representations are difficult to align with static geospatial observations: sparse, irregular GPS samples rarely coincide with street-view image coordinates, and existing encoders produce either a single embedding per trajectory or road segment, or embeddings only at the observed GPS samples. We introduce **TRAJGANR**, a trajectory-centric multimodal self-supervised learning framework that represents each trajectory as a *continuous, path-conditioned neural field*. Reconstructed from the observed GPS samples, this representation can be queried at arbitrary coordinates (including the off-path locations where street-view images were captured), yielding localized mobility embeddings that align with street-view image and location representations at the same coordinate. Across four fine-grained road-level tasks in San Francisco and Porto, TRAJGANR consistently outperforms recent geospatial and trajectory foundation models, with ablations attributing the gains to both the mobility signal and the fine granularity at which it is aligned. TRAJGANR's location encoder further generalizes beyond the road network to areas where neither street-view imagery nor trajectories are available.

1 INTRODUCTION

Cities are jointly shaped both by the layout and appearance of their built environment (e.g., streets, buildings, and public spaces) as well as by their patterns of human mobility and activity (Batty, 2013). Many urban properties depend on the latter: road safety risk reflects localized travel behavior Li et al. (2026), congestion and travel demand reflect collective movement (Zheng et al., 2014), and an area's land use is partly revealed by when, how often, and from where people visit it (Yuan et al., 2012). Yet, visual appearance and movement often carry different information: two nearby road segments can look similar in street-view imagery while supporting very different traffic volumes, speeds, or driving behaviors. Static observations, including imagery and coordinates, therefore describe the *form* of urban space, whereas trajectories provide the dynamic signal needed to characterize its *function*.

Learning general-purpose representations of urban space requires integrating these complementary signals (Mai et al., 2022; 2024). Multimodal self-supervised learning (MSSL) offers a natural framework for this integration and has become the main pretraining paradigm in geospatial representation learning (Xiao et al., 2025; Janowicz et al., 2025). Existing urban MSSL frameworks, however, pretrain almost exclusively on *static* modalities, such as street-view images, satellite imagery, and geographic coordinates, relying on two primary objective families: contrastive approaches that align observations recorded at the same or nearby locations (Cepeda et al., 2023; Klemmer et al., 2025; Mai et al., 2023; Mühlematter et al., 2025; Liu et al., 2026; Balsebre et al., 2023), and reconstruction-based approaches that mask and reconstruct multi-source observations represented on a shared spatial grid (Fuller et al., 2024; Guo et al., 2024; Choudhury et al., 2025; Bai et al., 2025). Both families thus presuppose that every modality is observed at a common set of locations, a condition trajectories do not satisfy. Although the value of mobility trajectories for geospatial pretraining has been

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.

articulated explicitly (Choudhury et al., 2024; Siampou et al., 2025a), no existing MSSL framework operationalizes this by aligning GPS trajectories with static urban modalities.

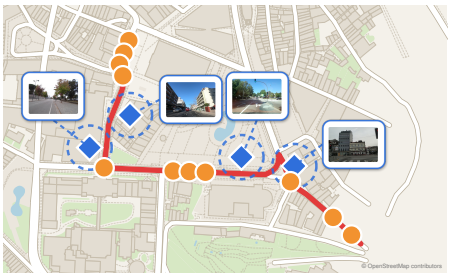

Figure 1: A vehicle's continuous path (red) is observed at irregularly spaced GPS sample waypoints (orange), while street-view images (blue) are captured at locations that typically do not co-locate with waypoints.

Integrating trajectories into the pretraining of urban MSSL frameworks is not straightforward. Static modalities such as street-view images (SVIs) are anchored to individual geographic locations, whereas trajectories describe movement along a continuous path through space and time. Figure 1 illustrates this mismatch: a vehicle's movement is observed only at irregularly spaced GPS samples, and nearby street-view images are captured at locations that rarely coincide with them. Existing trajectory encoders cannot directly supply the localized movement signal at an image's coordinate, as they typically produce either a single embedding summarizing the entire trajectory or one embedding per road segment/observed waypoint (Lin et al., 2024b; Zhao et al., 2025; Zhu et al., 2025; Lin et al., 2024a). None of these representations can directly answer the location-centered question needed for multimodal learning: *What movement context does this trajectory provide at the coordinate of this image?* Discretizing trajectories into grid cells or map-matched road segments (Li et al., 2018; Ma et al., 2024; Wang et al., 2025) creates common alignment units, but makes the representation constant within each unit, erasing variation between nearby locations (Zhou et al., 2025). Thus, incorporating trajectories into multimodal pretraining **requires a representation that preserves continuous variation along a path and can return a localized trajectory embedding at any queried coordinate along or near that path.**

We introduce **TRAJGANR**, a trajectory-centric geospatial MSSL framework designed around this requirement. Rather than forcing a static observation to match the nearest recorded GPS point or a pooled road-segment representation, TRAJGANR uses the observation's coordinate as a query into the trajectory, which it represents as a continuous, path-conditioned neural field. Given the observed GPS samples together with query coordinates supplied by nearby static modalities (SVI locations in our setup), the model jointly contextualizes the observed and queried positions and infers the trajectory's localized movement context at every query. This produces a localized trajectory embedding at the exact coordinate of each image, even when no GPS point was recorded there. TRAJGANR then aligns that embedding with the corresponding street-view and location representations through a multimodal self-supervised objective. Across many trajectories, this training encourages the resulting location representations to capture both the *form* of urban space and the *function* it supports.

We evaluate TRAJGANR on four road-segment-level downstream tasks in San Francisco and Porto: average traffic speed, road popularity, area-of-interest (AOI) function, and hard-braking event prediction. Together, these tasks isolate the two effects our design targets: the contribution of mobility signals to multimodal geospatial pretraining, and the granularity at which trajectories and static observations are aligned. TRAJGANR outperforms recent geospatial MSSL frameworks that combine imagery with locations, as well as a unimodal trajectory foundation model, on every task. In San Francisco, it reduces traffic-speed error by 44% and hard-braking error by 82% relative to the strongest baseline. Critically, when trajectories are instead aligned using a pooled road-segment representation, this advantage disappears on traffic speed and hard braking prediction tasks. This result supports the role of fine-grained alignment precisely for outcomes that vary within a segment. Finally, we demonstrate that TRAJGANR's location encoder generalizes beyond the road network: at off-network locations, where neither street-view imagery nor trajectories are available, it improves F1 by 28% over the strongest baseline for the POI popularity prediction task, demonstrating that mobility-grounded supervision transfers to regions unseen during model pretraining.

In summary, our contributions are:

- We introduce **TRAJGANR**, a trajectory-centric MSSL framework that aligns human mobility trajectories with street-view images and their geographic locations.
- We develop a path-conditioned neural implicit function that uses sparse observed GPS points to infer localized movement context at arbitrary query coordinates supplied by static modalities.

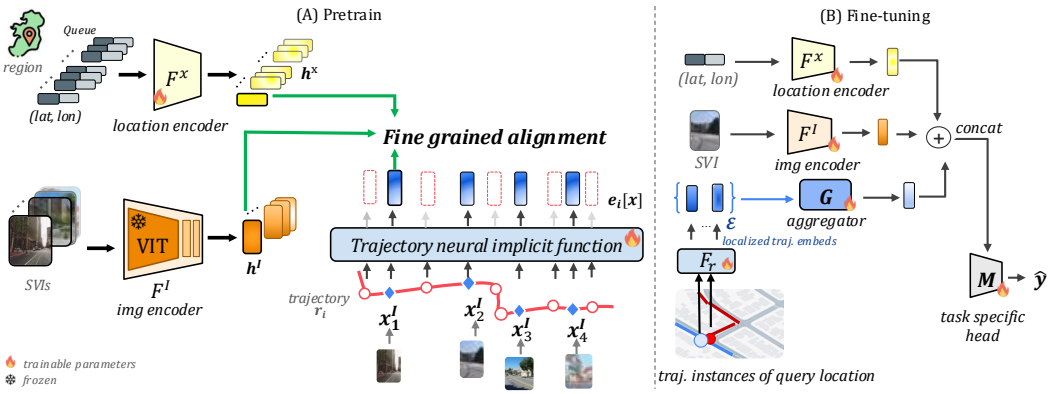

Figure 2: **Overview of TRAJGANR. (a) Pretraining:** Fine-grained geospatial alignment jointly trains trajectories, street-view images, and coordinates. White circles denote observed points of $\mathbf{r}_i$, while blue diamonds indicate SVI locations $\{\mathbf{x}_k^{(I)}\}$. **(b) Fine-Tuning:** For a query location $\mathbf{x}_k^{(I)}$, localized trajectory embeddings $\mathcal{E}_r(\mathbf{x}_k^{(I)})$ are queried, aggregated, and concatenated with image ($\mathbf{e}_k^{(I)}$) and location ($\mathbf{e}_k^{(x)}$) representations for downstream prediction.

- We develop a fine-grained MSSL objective that aligns each queried trajectory embedding with its street-view and location representations at the same coordinate.
- We introduce four road-segment-level tasks across two metropolitan areas, on which TRAJGANR outperforms recent geospatial MSSL frameworks and a trajectory foundation model on every task.
- Through alignment and modality ablations, we show that the gains arise from both the trajectory signal and the ability to query it at fine spatial granularity.
- We further show that TRAJGANR's location encoder generalizes to off-road network locations.

2 PRELIMINARIES

Definition 1 (Street-view Image Dataset)**.** Let $\mathcal{S} \subseteq \mathbb{R}^2$ denote the continuous spatial extent of interest, and let $\mathcal{D}^{(I)} = \{\mathbf{o}_k\}_{k=1}^{N^{(I)}}$ be a dataset of $N^{(I)}$ street-view images georeferenced within $\mathcal{S}$. Each image is represented as a tuple $\mathbf{o}_k = (\mathbf{I}_k, \mathbf{x}_k^{(I)})$, where $\mathbf{I}_k \in \mathbb{R}^{C\times H\times W}$ denotes the high-dimensional visual feature tensor and $\mathbf{x}_k^{(I)} \in \mathcal{S}$ is its corresponding geographic coordinate. We denote the set of image locations by $\mathcal{D}^{(x)} = \{\mathbf{x}_k^{(I)}\}_{k=1}^{N^{(I)}} \subset \mathcal{S}$. In our study, $\mathcal{S}$ corresponds to a bounded region.

Definition 2 (GPS Trajectory Dataset)**.** Let $\mathcal{D}^{(r)} = \{\mathbf{r}_i\}_{i=1}^{N^{(r)}}$ be a collection of $N^{(r)}$ GPS trajectories observed within $\mathcal{S}$ over a temporal horizon $\mathcal{T} \subseteq \mathbb{R}$. Each trajectory $\mathbf{r}_i$ is an ordered sequence of $L_i^{(r)}$ sampled GPS points, $\mathbf{r}_i = (\mathbf{p}_{i1}, \mathbf{p}_{i2}, \ldots, \mathbf{p}_{iL_i^{(r)}})$, where each $\mathbf{p}_{ij} = (\mathbf{x}_{ij}^{(r)}, t_{ij})$ consists of a location $\mathbf{x}_{ij}^{(r)} \in \mathcal{S}$ and a strictly increasing timestamp $t_{ij} \in \mathcal{T}$ such that $t_{i1} < t_{i2} < \cdots < t_{iL_i^{(r)}}$. Although observed as a finite sequence of discrete points, $\mathbf{r}_i$ samples an underlying continuous movement trajectory through space and time. We denote this continuous spatial footprint and temporal span by $\mathcal{X}(\mathbf{r}_i) \subset \mathcal{S}$ and $\mathcal{T}(\mathbf{r}_i) \subseteq \mathcal{T}$, respectively, such that $\mathbf{x}_{ij}^{(r)} \in \mathcal{X}(\mathbf{r}_i)$ and $t_{ij} \in \mathcal{T}(\mathbf{r}_i) \forall j \in \{1, \ldots, L_i^{(r)}\}$.

Problem Formulation. Let $\mathbb{D} = (\mathcal{D}^{(I)}, \mathcal{D}^{(r)})$ denote a multimodal geospatial dataset defined over the continuous spatial domain $\mathcal{S}$. Our objective is to pretrain a parameterized representation function $\mathcal{F}_\theta$ on $\mathbb{D}$. For any geographic query coordinate $\mathbf{x} \in \mathcal{S}$, the function produces a continuous localized neural embedding $\mathbf{u}(\mathbf{x}) = \mathcal{F}_\theta(\mathbf{x}; \mathbb{D}) \in \mathbb{R}^d$, where $\mathbf{u}(\mathbf{x})$ encodes localized multimodal context by integrating *visual features* from $\mathcal{D}^{(I)}$ and *mobility patterns* from $\mathcal{D}^{(r)}$ within the neighborhood of $\mathbf{x}$.

3 METHODOLOGY

In this section, we present the components of our TRAJGANR framework, as illustrated in Fig. 2.

3.1 STATIC MODALITY ENCODERS

We first describe the encoding of the **static** modalities of $\mathbb{D}$, which include the street-view image dataset $\mathcal{D}^{(I)}$ and their geographic coordinate set $\mathcal{D}^{(x)}$. Following standard practices in geospatial multimodal representation learning (Cepeda et al., 2023; Liu et al., 2026), we encode them using specialized, modality-specific encoders to project them into a shared d-dimensional embedding space.

Street-view Image Encoder. Let $\mathcal{F}^{(I)}$ denote the street-view image encoder. Given the image tensor $\mathbf{I}_k \in \mathbb{R}^{C\times H\times W}$ of tuple $\mathbf{o}_k = (\mathbf{I}_k, \mathbf{x}_k^{(I)})$, the encoder produces a visual representation $\mathbf{e}_k^{(I)} \in \mathbb{R}^d$:

$$\mathbf{e}_k^{(I)} = \mathcal{F}^{(I)}(\mathbf{I}_k), \tag{1}$$

where $\mathcal{F}^{(I)} = g_{\mathrm{I}} \circ \phi_{\mathrm{I}}$. Here, ϕ_{I} is a frozen, pretrained Vision Transformer (ViT) backbone (Dosovitskiy et al., 2020) while g_{I} is a learnable projection head that adapts the visual features to a shared embedding space. We keep the ϕ_{I} frozen to prevent its generalized representations from overfitting to our localized spatial distribution, and minimize computational overhead during cross-modal training.

Street-view Location Encoder. Let $\mathcal{F}_{\mathbf{o}}^{(x)}$ denote the geographic location encoder. Given the image coordinates $\mathbf{x}_k^{(I)} \in \mathcal{S}$ of street-view image $\mathbf{o}_k$, the encoder produces a location embedding $\mathbf{e}_k^{(x)} \in \mathbb{R}^d$:

$$\mathbf{e}_k^{(x)} = \mathcal{F}_{\mathbf{o}}^{(x)}(\mathbf{x}_k^{(I)}), \tag{2}$$

where $\mathcal{F}_{\mathbf{o}}^{(x)}$ is parameterized using Space2Vec (Mai et al., 2020). We choose Space2Vec as it is specifically designed to preserve multi-scale spatial frequency relationships within a local study area.[1]

3.2 TRAJECTORY NEURAL IMPLICIT FUNCTION

We now describe our modeling approach for the **dynamic** modality $\mathcal{D}^{(r)}$ of $\mathbb{D}$. Aligning trajectories with static modalities in real-world scenarios is fundamentally challenged by a spatial mismatch: trajectories represent continuous movement, while their sparse and irregularly sampled GPS points rarely co-locate with street-view camera locations. We resolve this challenge by formulating trajectories **as continuous neural implicit representations** that can be queried at image's location coordinates. This allows TRAJGANR to infer a localized trajectory embedding at the exact capture site of any nearby street-view image, enabling cross-modal alignment at the most granular level.

In the following, we detail the encoding of discrete GPS observations (Section 3.2.1), formalize the coordinate-conditioned query mechanism (Section 3.2.2), and present the joint spatiotemporal contextualization that reconstructs the continuous coordinate-conditioned trajectory field (Section 3.2.3).

3.2.1 TRAJECTORY OBSERVED POINTS ENCODING

Each trajectory $r_i \in \mathcal{D}^{(r)}$ is represented as a sequence of discretely sampled GPS points $\mathbf{p}_{ij}$, which we refer to as observed points. Each such point is characterized by a pair of coordinates $\mathbf{x}_{ij}^{(r)} \in \mathcal{S}$ and a timestamp $t_{ij} \in \mathcal{T}$, which we project into a shared d-dimensional space via two separate factorized encoders. We then combine them additively to construct our initial trajectory tokens:

$$\mathbf{h}_{ij}^0 = \mathcal{F}_r^{(x)}(\mathbf{x}_{ij}^{(r)}) + \mathcal{F}_r^{(t)}(t_{ij}). \tag{3}$$

Here, $\mathcal{F}_r^{(x)} : \mathcal{S} \to \mathbb{R}^d$ is a trajectory-specific Space2Vec location encoder (Mai et al., 2020) and $\mathcal{F}_r^{(t)} : \mathbb{R} \to \mathbb{R}^d$ is a Time2Vec temporal encoder (Kazemi et al., 2019). By summing these two representations, the temporal encoding $\mathcal{F}_r^{(t)}(t_{ij})$ served as a positional encoding, structurally preserving the sequential order of the points along the trajectory.

3.2.2 COORDINATE-CONDITIONED QUERY CONSTRUCTION

To align trajectories with static modalities at the most granular geographic level, we treat each street-view image's coordinate as a spatial query that probes the continuous trajectory footprint.

[1] For global scale modeling, alternative encoding approaches such as Random Fourier Features (Cepeda et al., 2023) could be utilized. Similarly, to encode more complex, non-point geographic geometries, methods like Poly2Vec (Siampou et al., 2025b) represent orthogonal approaches that can be integrated into this framework.

Formally, for a given trajectory r_i, let $\mathcal{Q}_i = \{\mathbf{x} \in \mathcal{S} \mid d(\mathbf{x}, \mathcal{X}(\mathbf{r}_i)) \leq \epsilon\}$[2] denote a set of queries representing street-view image locations $\mathbf{x}_k^{(I)}$ within a distance threshold ϵ of the trajectory's continuous spatial footprint $\mathcal{X}(\mathbf{r}_i)$. For each query coordinate $\mathbf{x} \in \mathcal{Q}_i$, we construct an initial query token:

$$\mathbf{q}_i^0[\mathbf{x}] = \mathcal{F}_r^{(x)}(\mathbf{x}) + \mathbf{t}_i[\mathbf{x}], \tag{4}$$

where $\mathcal{F}_r^{(x)}(\mathbf{x})$ provides spatial grounding using the trajectory-specific location encoder, and $\mathbf{t}_i[\mathbf{x}] \in \mathbb{R}^d$ represents the trajectory-conditioned temporal embedding assigned to $\mathbf{x}$. We clarify that $\mathcal{F}_r^{(x)}(\mathbf{x})$ is distinct from the street-view location encoder $\mathcal{F}_{\mathbf{o}}^{(x)}(\mathbf{x})$ defined in Section 3.1, as the latter is intended to provide general regional geographic context, whereas $\mathcal{F}_r^{(x)}(\mathbf{x})$ aims to project coordinates into a subspace that aligns with sequential and directional dynamics of trajectories.

We now describe how the trajectory-conditioned temporal embedding $\mathbf{t}_i[\mathbf{x}] \in \mathbb{R}^d$ is computed. Given that street-view images are captured independently of the trajectory dataset, our query set $\mathcal{Q}_i$ lacks intrinsic timestamps. While image capture times could be used, we choose to not constrain our framework, since our goal is not to match transient, real-time events (e.g., co-locating a vehicle with a simultaneous GPS ping) but to learn the **long-term visual and behavioral characteristics of a location**. Since the physical infrastructure of a neighborhood (e.g., road layouts, neighborhood aesthetics) evolves slowly, asynchronous images still accurately reflect the environment that dictates current trajectory patterns. Under this assumption, we investigate two query-time encoding strategies:

Synthetic-Time Query Encoding. The first strategy assigns a synthetic timestamp to the query location. Since $\mathbf{r}_i$ represents continuous movement over the spatial footprint $\mathcal{X}(\mathbf{r}_i)$ and temporal span $\mathcal{T}(\mathbf{r}_i)$, a query location $\mathbf{x} \in \mathcal{X}(\mathbf{r}_i)$ can be associated with an estimated time $\hat{t}_i[\mathbf{x}] \in \mathcal{T}(\mathbf{r}_i)$. Let $\mathbf{p}_{ia}$ and $\mathbf{p}_{ib}$ denote two neighboring observed GPS samples along $\mathbf{r}_i$ such that $\mathbf{x}$ lies between their locations along the trajectory footprint. We estimate the query time by linear interpolation:

$$\hat{t}_i[\mathbf{x}] = (1 - \lambda)t_{ia} + \lambda t_{ib}, \tag{5}$$

where $\lambda \in [0, 1]$ denotes the relative position of $\mathbf{x}$ between $\mathbf{x}_{ia}^{(r)}$ and $\mathbf{x}_{ib}^{(r)}$ along the trajectory, with $\lambda = \frac{\|\mathbf{x} - \mathbf{x}_{ia}^{(r)}\|_2}{\|\mathbf{x}_{ib}^{(r)} - \mathbf{x}_{ia}^{(r)}\|_2}$. The temporal embedding is then obtained using the temporal encoder:

$$\mathbf{t}_i[\mathbf{x}] = \mathcal{F}_r^{(t)}\left(\hat{t}_i[\mathbf{x}]\right). \tag{6}$$

This variant treats the query location as a point that the trajectory traverses at an estimated time.

Masked-Time Query Encoding. The second strategy represents time as a latent property by assigning a sequence-level learnable mask to each query. Specifically, for a query set of size k, we assign a sequence of learnable temporal masks $\{\mathbf{m}_{\text{time}}^{(j)}\}_{j=1}^k$ such that:

$$\mathbf{t}_i[\mathbf{x}_j] = \mathbf{m}_{\text{time}}^{(j)}, \tag{7}$$

where $\mathbf{m}_{\text{time}}^{(j)} \in \mathbb{R}^d$ is optimized end-to-end. This approach treats each query's temporal feature as a mask, allowing the self-attention mechanism to dynamically reconstruct the missing timestamp of $\mathbf{x}_j$ by contextualizing its unique spatial encoding $\mathcal{F}_r^{(x)}(\mathbf{x}_j)$ against the observed trajectory points.

3.2.3 Joint Spatiotemporal Contextualization

To obtain localized trajectory embeddings at the query locations, we jointly process the observed trajectory sequence and the queries. Recall from Section 3.2.2 that the query set $\mathcal{Q}_i$ contains the coordinates of all street-view images located in the neighborhood of trajectory's $\mathbf{r}_i$ continuous spatial footprint $\mathcal{X}(\mathbf{r}_i)$. To preserve this relationship while modeling the path's dynamics, we concatenate the sequence of initial trajectory tokens $\mathbf{H}_i^0 = [\mathbf{h}_{i1}^0, \ldots, \mathbf{h}_{iL_i^{(r)}}^0]$ of $\mathcal{X}(\mathbf{r}_i)$ with its set of spatial query tokens $\{\mathbf{q}_i^0[\mathbf{x}]\}_{\mathbf{x} \in \mathcal{Q}_i}$. This yields an augmented spatiotemporal sequence:

$$\widetilde{\mathbf{H}}_i^0 = \left[\mathbf{h}_{i1}^0, \ldots, \mathbf{h}_{iL_i^{(r)}}^0, \{\mathbf{q}_i^0[\mathbf{x}]\}_{\mathbf{x} \in \mathcal{Q}_i}\right]. \tag{8}$$

[2] In practice, ϵ is set to a 20-50 m buffer on the order of typical GPS error to make sure that the street-view images associated with query locations were taken along the road traversed by trajectories while we filter out SVIs that are taken in completely off-road areas (e.g., parking lots).

We pass the augmented sequence to an L-layer Transformer encoder backbone (Vaswani et al., 2017):

$$\widetilde{\mathbf{H}}_i^L = \text{Transformer}\left(\widetilde{\mathbf{H}}_i^0\right). \tag{9}$$

Through self-attention, each query token is processed in parallel with the observed sequence, allowing the Transformer to implicitly interpolate the trajectory's movement patterns at each continuous query location $\mathbf{x}$. The final-layer representation of each $\mathbf{q}_i^0[\mathbf{x}]$ serves as the localized trajectory embedding:

$$\mathbf{e}_i[\mathbf{x}] = \widetilde{\mathbf{H}}_{i,\text{idx}(\mathbf{x})}^L \in \mathbb{R}^d, \tag{10}$$

where $\text{idx}(\mathbf{x})$ maps the query coordinate to its corresponding positional index in the augmented sequence $\widetilde{\mathbf{H}}_i^L$. The representation $\mathbf{e}_i[\mathbf{x}]$ captures the localized movement context of trajectory r_i at coordinate $\mathbf{x}$, and is subsequently aligned with the visual and location representations of the corresponding street-view image. We describe this alignment process in the next section.

3.3 TRAJGANR MULTIMODAL SELF-SUPERVISED PRETRAINING

For each street-view coordinate $\mathbf{x}_k^{(I)} \in \mathcal{Q}_i$, the framework yields three representations co-anchored within the same local neighborhood: the visual feature $\mathbf{e}_k^{(I)}$, the global location prior $\mathbf{e}_k^{(x)}$, and the localized trajectory embedding $\mathbf{e}_i[\mathbf{x}_k^{(I)}]$. We denote by $\mathcal{L}_{\text{info}}(\mathbf{u}, \mathbf{v}^+; \mathcal{N})$ the standard InfoNCE loss (Oord et al., 2018) between an anchor $\mathbf{u}$ and a positive target $\mathbf{v}^+$, contrasted against a negative set $\mathcal{N}$. The overall pretraining objective is the sum of all cross-modal pairs:

$$\mathcal{L}_{\text{pre}} = \underbrace{\mathcal{L}_{\text{info}}(\mathbf{e}_k^{(I)}, \mathbf{e}_k^{(x)}; \mathcal{N}_{\text{loc}})}_{\text{Image–Location}} + \underbrace{\mathcal{L}_{\text{info}}(\mathbf{e}_i[\mathbf{x}_k^{(I)}], \mathbf{e}_k^{(x)}; \mathcal{N}_{\text{loc}})}_{\text{Trajectory–Location}} + \underbrace{\mathcal{L}_{\text{info}}(\mathbf{e}_i[\mathbf{x}_k^{(I)}], \mathbf{e}_k^{(I)}; \mathcal{N}_{\text{img}})}_{\text{Trajectory–Image}}. \tag{11}$$

Here, the negative set $\mathcal{N}_{\text{loc}}$ consists of coordinates sampled from a MoCo-style memory queue (He et al., 2020), while $\mathcal{N}_{\text{img}}$ draws in-batch negatives to contrast the positive street-view against images of different locations. Furthermore, by design, a single street-view coordinate $\mathbf{x}_k^{(I)}$ can have multiple trajectories passing through its local neighborhood. Each of these trajectories generates a unique localized embedding $\mathbf{e}_i[\mathbf{x}_k^{(I)}]$, providing multiple positive trajectory views for the same static anchor $\mathbf{x}_k^{(I)}$. This multi-view contrastive setup incentivizes TRAJGANR to extract generalized, location-consistent mobility patterns rather than overfitting to or memorizing individual trajectory instances.

3.4 TRAJGANR FINE-TUNING

To adapt our pretrained encoders to downstream prediction tasks, we construct a single multimodal final embedding $\mathbf{e}_k$ for each target location $\mathbf{x}_k^{(I)}$. Each place is usually associated with a single image $\mathbf{e}_k^{(I)}$ and location embeddings $\mathbf{e}_k^{(x)}$, but a variable-sized set of localized trajectory embeddings $\mathcal{E}_r(\mathbf{x}_k^{(I)}) = \{\mathbf{e}_i[\mathbf{x}_k^{(I)}] \mid \mathbf{x}_k^{(I)} \in \mathcal{Q}_i\}$. To aggregate this set into a single vector of fixed dimensionality, we feed the set into a permutation-invariant multi-head self-attention pooling block $\mathcal{G}_\theta$ prepended with a learnable `[CLS]` token. The final-layer `[CLS]` embedding is the collective mobility profile:

$$\mathbf{e}_k^{(r)} = \mathcal{G}_\theta\left(\mathcal{E}_r(\mathbf{x}_k^{(I)})\right). \tag{12}$$

We then concatenate these three modalities to get the final location representation: $\mathbf{e}_k = \left[\mathbf{e}_k^{(I)} \parallel \mathbf{e}_k^{(x)} \parallel \mathbf{e}_k^{(r)}\right]$, where $\parallel$ denotes concatenation. Finally, for each downstream task k, we attach a task-specific prediction head $M_\psi^{(k)}$ to generate the target prediction, which is optimized end-to-end using the corresponding task's supervised loss:

$$\hat{\mathbf{y}}_k^{(k)} = M_\psi^{(k)}(\mathbf{e}_k). \tag{13}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate TRAJGANR using two publicly available and widely adopted trajectory datasets: the Porto Taxi dataset (O'Connell et al., 2015) and the San Francisco Cabspotting dataset [3].

[3] https://stamen.com/work/cabspotting/

Table 1: Main experimental results on San Francisco. **Best** and second best values are highlighted. Results are averaged over 3 runs with different train/test splits. "$\pm x$" denotes the standard deviation.

Method	*traffic speed*		*road popularity*		*AOI function*		*HBE*	
	(↓) MAE	(↓) RMSE	(↑) F1	(↑) AUPRC	(↓) L1	(↑) Cosine	(↓) MAE	(↓) RMSE
UNITRAJ	$2.353_{\pm.06}$	$5.303_{\pm.08}$	$0.181_{\pm.01}$	$0.354_{\pm.05}$	$1.194_{\pm.01}$	$0.602_{\pm.00}$	$2.373_{\pm.01}$	$5.975_{\pm.01}$
GEOCLIP	$2.607_{\pm.04}$	$4.848_{\pm.12}$	$0.431_{\pm.01}$	$0.615_{\pm.20}$	$1.144_{\pm.05}$	$0.634_{\pm.03}$	$5.853_{\pm.27}$	$7.297_{\pm.35}$
SATCLIP	$2.316_{\pm.24}$	$5.239_{\pm.09}$	$0.181_{\pm.01}$	$0.393_{\pm.06}$	$1.202_{\pm.00}$	$0.600_{\pm.00}$	$10.95_{\pm.46}$	$13.11_{\pm.09}$
GAIR $_{\text{SVI}}$	$2.203_{\pm.09}$	$4.384_{\pm.04}$	$0.582_{\pm.06}$	$0.894_{\pm.02}$	$1.114_{\pm.01}$	$0.649_{\pm.00}$	$1.744_{\pm.04}$	$5.061_{\pm.11}$
GAIR $_{\text{LOC}}$	$1.420_{\pm.18}$	$2.981_{\pm.09}$	$0.748_{\pm.03}$	$0.832_{\pm.02}$	$0.570_{\pm.03}$	$0.879_{\pm.01}$	$1.185_{\pm.24}$	$3.586_{\pm.56}$
GAIR $_{\text{LOC + SVI}}$	$\underline{0.725}_{\pm.03}$	$\underline{1.694}_{\pm.01}$	$\underline{0.930}_{\pm.02}$	$\underline{0.981}_{\pm.00}$	$\underline{0.262}_{\pm.01}$	$\underline{0.958}_{\pm.00}$	$\underline{0.715}_{\pm.07}$	$\underline{1.742}_{\pm.08}$
TrajGANR	$\mathbf{0.403}_{\pm.12}$	$\mathbf{1.187}_{\pm.22}$	$\mathbf{0.960}_{\pm.02}$	$\mathbf{0.992}_{\pm.00}$	$\mathbf{0.121}_{\pm.03}$	$\mathbf{0.981}_{\pm.01}$	$\mathbf{0.130}_{\pm.02}$	$\mathbf{0.926}_{\pm.19}$

Table 2: Main experimental results on Porto. **Best** and second best values are highlighted.

Method	*road popularity*		*AOI function*		*HBE*	
	(↑) F1	(↑) AUPRC	(↓) L1	(↑) Cosine	(↓) MAE	(↓) RMSE
UNITRAJ	$0.383_{\pm.05}$	$0.459_{\pm.08}$	$0.865_{\pm.01}$	$0.728_{\pm.00}$	$1.121_{\pm.13}$	$4.378_{\pm.31}$
GEOCLIP	$0.632_{\pm.04}$	$0.685_{\pm.01}$	$0.777_{\pm.02}$	$0.778_{\pm.01}$	$9.807_{\pm1.2}$	$14.39_{\pm1.2}$
SATCLIP	$0.153_{\pm.01}$	$0.373_{\pm.08}$	$0.856_{\pm.03}$	$0.737_{\pm.01}$	$9.971_{\pm1.3}$	$14.80_{\pm1.4}$
GAIR $_{\text{SVI}}$	$0.449_{\pm.26}$	$0.557_{\pm.24}$	$0.758_{\pm.02}$	$0.784_{\pm.01}$	$6.180_{\pm.10}$	$8.842_{\pm.08}$
GAIR $_{\text{LOC}}$	$0.795_{\pm.02}$	$0.823_{\pm.02}$	$0.644_{\pm.01}$	$0.837_{\pm.00}$	$1.081_{\pm.13}$	$3.923_{\pm.38}$
GAIR $_{\text{LOC + SVI}}$	$\underline{0.894}_{\pm.04}$	$\underline{0.910}_{\pm.06}$	$\underline{0.496}_{\pm.12}$	$\underline{0.891}_{\pm.03}$	$\underline{0.937}_{\pm.04}$	$\underline{2.652}_{\pm.16}$
TrajGANR	$\mathbf{0.908}_{\pm.04}$	$\mathbf{0.939}_{\pm.04}$	$\mathbf{0.279}_{\pm.01}$	$\mathbf{0.951}_{\pm.00}$	$\mathbf{0.401}_{\pm.28}$	$\mathbf{1.358}_{\pm.04}$

Following previous works, we download street-view images and their capture locations from the Mapillary API [4] (Liu et al., 2026). Additional dataset details can be found in Appendix A.1.1.

Baselines. We compare TRAJGANR against representative baselines across two complementary categories: (i) **static geospatial foundation models**, including GEOCLIP (Cepeda et al., 2023), SATCLIP (Klemmer et al., 2025), and GAIR (Liu et al., 2026), and (ii) **trajectory-only foundation model**, including the state-of-the-art UNITRAJ (Zhu et al., 2025) to assess whether mobility dynamics alone suffice for downstream tasks. During fine-tuning, we freeze all pretrained encoders and train a lightweight, task-specific prediction head. For UNITRAJ, we mean-pool the point embeddings along the road segment corresponding to each target location. More details are provided in Appendix A.1.7.

Evaluation Tasks. We evaluate TRAJGANR on four downstream fine-grained road-level tasks derived from Google Maps: (i) **traffic speed** prediction, which estimates the historical average driving speed along each road segment, (ii) **road popularity**, which categorizes segments based on aggregated user navigation demand, (iii) **AOI function** prediction, which predicts the distribution of Point-of-Interest (POI) categories within the Area of Interest (AOI) surrounding each segment, and (iv) **hard braking events (HBE)** prediction, which estimates the frequency of abrupt deceleration events as a proxy for driving difficulty and safety risk (Li et al., 2026). Task ground-truth labels are anchored to street-view image coordinates, which serve as the target locations. Due to sparse coverage in Porto, the traffic speed estimation tasks is skipped. We provide more details in Appendix A.1.2.

4.2 MAIN RESULTS

Tables 1 and 2 report downstream evaluation results across San Francisco and Porto, respectively. Across both regions and all tasks, TRAJGANR consistently outperforms all baseline models. The gains are most pronounced on tasks that depend on driving behavior, namely traffic speed estimation and hard-braking event (HBE) prediction. On these tasks, TRAJGANR reduces MAE over the strongest baseline by 44.4% and 81.8% in San Francisco, and by 57.2% on HBEs in Porto. Among static baselines, $\text{GAIR}_{\text{Loc+SVI}}$ is the most competitive, indicating that geographic coordinates combined with street-level visual features provide a strong structural prior. In contrast, while satellite-trained models like SATCLIP can capture broad regional land-use patterns, their overhead perspective misses the localized ground-level details essential for our fine-grained tasks. Finally, we observe that UNITRAJ struggles in performance despite being pretrained on large-scale trajectory data. We

[4] https://www.mapillary.com/

attribute this to two factors: First, UNITRAJ outputs embeddings strictly at the observed waypoints, lacking a continuous mechanism to infer representations at unseen target coordinates. Thus, it requires segment-level pooling, which averages out fine-grained movement. Second, as a unimodal model, it lacks cross-modal alignment with other modalities, preventing it from grounding its representations in the physical built environment. Overall, these findings demonstrate that TRAJGANR's advantage comes from integrating complementary static visual features with dynamic movement patterns, as well as its continuous implicit function mechanism that queries trajectory embeddings directly at any location, bypassing the need for lossy segment pooling.

4.3 ABLATION STUDIES

The Effect of Alignment Granularity. We examine how spatial alignment granularity affects representation quality across three pretraining variants: (1) NO ALIGN, which skips multimodal contrastive alignment entirely, (2) SEGMENT-BASED, which aligns modalities at the road-segment level by mean-pooling observed trajectory point embeddings, and (3) FINE-GRAIN (our full model), which aligns modalities directly at street-view coordinates using our continuous neural implicit function. As shown in Table 3, NO ALIGN yields the weakest performance, while FINE-GRAIN achieves the strongest overall results, with large gains over SEGMENT-BASED on tasks driven by driving behavior, such as traffic speed (MAE reduced by 43%) and HBE prediction (MAE reduced by 60%). Segment pooling assigns the same trajectory representation to every SVI on a segment, averaging away how driving behavior varies along it, whereas fine-grained alignment learns its own trajectory-conditioned signal, yielding a richer description of how the road is driven. In contrast, SEGMENT-BASED performs comparably on AOI function and road popularity, which depend less on driving behavior. Fine-grained alignment is thus preferable in practice: it matches segment pooling where driving behavior matters less, and substantially outperforms it where it matters most.

Table 3: Ablation of alignment granularity on San Francisco. **Best** values are highlighted.

Method	*Traffic Speed*		*Road Popularity*		*AOI Function*		*HBE*	
	(↓) MAE	(↓) RMSE	(↑) F1	(↑) AUPRC	(↓) L1	(↑) Cosine	(↓) MAE	(↓) RMSE
NO ALIGN	$0.907_{\pm.13}$	$1.591_{\pm.85}$	$0.805_{\pm.07}$	$0.882_{\pm.08}$	$0.322_{\pm.00}$	$0.940_{\pm.00}$	$0.385_{\pm.01}$	$1.283_{\pm.09}$
SEGMENT-BASED	$0.709_{\pm.06}$	$1.464_{\pm.09}$	$0.958_{\pm.04}$	$0.985_{\pm.01}$	$\mathbf{0.121}_{\pm.01}$	$\mathbf{0.982}_{\pm.00}$	$0.328_{\pm.05}$	$1.106_{\pm.15}$
FINE-GRAIN	$\mathbf{0.403}_{\pm.12}$	$\mathbf{1.187}_{\pm.22}$	$\mathbf{0.960}_{\pm.02}$	$\mathbf{0.992}_{\pm.00}$	$0.121_{\pm.03}$	$0.981_{\pm.01}$	$\mathbf{0.130}_{\pm.02}$	$\mathbf{0.926}_{\pm.19}$

The Effect of Modality Combination. Table 4 evaluates how combining each modality impacts downstream performance. Relying solely on image features yields the weakest results across all benchmarks, showing that visual appearance alone is insufficient for road-level tasks without geographic or behavioral grounding. In comparison, geographic coordinates alone establish a strong spatial prior for road popularity and HBE due to spatial autocorrelation across neighboring areas, but are insufficient for dynamic tasks like traffic speed estimation. Introducing trajectory dynamics directly addresses this gap: pairing trajectories with coordinates (LOC + TRAJ) substantially outperforms the static pair (LOC + SVI) on traffic speed and reduces HBE error fivefold. Ultimately, integrating all three data sources in TRAJGANR achieves the best overall results across every benchmark, demonstrating that they provide complementary signals, essential for learning general-purpose location representations.

Generalization to off-network locations. A natural question is whether our framework can represent arbitrary locations that do not lie along observed trajectories. To evaluate this, we test our pretrained location encoder against baseline location backbones on an additional Point-of-Interest (POI) popularity prediction task in San Francisco, predicting venue popularity strictly from raw coordinates that lie off the drivable road network and away from observed trajectories. As shown in Table 5, TRAJGANR's location-only encoder outperforms the strongest baseline, with relative improvements of 10.6% in accuracy, 28.3% in macro-F1, and 3.9% in AUROC. Absolute performance is modest on this task indicating that predicting venue popularity from coordinates alone is difficult for all models. The consistent ordering nonetheless suggests that TRAJGANR's location encoder to effectively generalizes to off-network coordinates rather than memorizing locations seen during pretraining.

5 RELATED WORK

Trajectory Representation Learning. Trajectory representation learning uses mobility traces as a primary signal to capture human travel dynamics across urban spaces. While existing trajectory

Table 4: Modality combination ablation on San Francisco. **Best** and second best values highlighted.

Modalities			*Traffic Speed*		*Road Popularity*		*AOI Function*		*HBE*	
Loc	**SVI**	**Traj**	(↓) MAE	(↓) RMSE	(↑) F1	(↑) AUPRC	(↓) L1	(↑) Cosine	(↓) MAE	(↓) RMSE
✓	–	–	1.117±.03	2.101±.07	0.928±.03	0.982±.00	0.188±.06	0.956±.01	0.237±.05	1.277±.12
–	✓	–	2.203±.09	4.384±.04	0.582±.06	0.894±.02	1.114±.01	0.648±.00	1.744±.04	5.061±.11
✓	✓	–	1.051±.08	1.724±.17	0.921±.04	0.972±.02	0.262±.01	0.962±.00	1.076±.13	1.653±.10
✓	–	✓	0.907±.10	1.774±.20	0.952±.00	0.984±.00	0.204±.02	0.967±.01	0.191±.00	1.128±.02
✓	✓	✓	**0.403**±.12	**1.187**±.22	**0.960**±.02	**0.992**±.00	**0.121**±.03	**0.981**±.01	**0.130**±.02	**0.926**±.19

Table 5: Off-network location-only performance on the POI popularity prediction task in San Francisco. **Best** values highlighted.

Model *(location encoder only)*	↑ Accuracy	↑ F1	↑ AUROC
SATCLIP	0.251 ± .00	0.101 ± .00	0.500 ± .02
GEOCLIP	0.298 ± .01	0.209 ± .00	0.545 ± .02
GAIR	0.310 ± .02	0.247 ± .03	0.568 ± .01
TRAJGANR	**0.343** ± .01	**0.317** ± .02	**0.590** ± .01

foundation models successfully power downstream applications such as next-location prediction, classification, and trajectory generation (Rao et al., 2020; Chen et al., 2021; Zhu et al., 2023; 2025; Li et al., 2017; Anastasiou et al., 2022; Zhou et al., 2026; Yang et al., 2025), they almost universally formulate trajectories as discrete sequences of GPS points or spatial grid cells. Consequently, learned representations remain restricted to global sequence embeddings or discrete point- and cell-level tokens. This discrete formulation cannot be directly applied in our multimodal setting, where continuous trajectories must be aligned with static observations at locations that lie within or near a transit path without matching any GPS point. In such scenarios, whole-trajectory embeddings are too coarse, while discrete point tokens provide incomplete alignment anchors. In contrast, TRAJGANR's continuous trajectory representation via its neural implicit function enables querying localized embeddings at arbitrary target locations, unlocking fine-grained geospatial alignment.

Geospatial Multimodal SSL. Multimodal self-supervised learning has emerged as a prominent paradigm for pretraining geospatial foundation models, primarily aligning street-view, satellite images, and text with geographic coordinates (Cepeda et al., 2023; Klemmer et al., 2025; Mai et al., 2023; Liu et al., 2024; Dhakal et al., 2025; Sastry et al., 2025). While straightforward to co-locate, these static modalities only capture the visual appearance and semantics of locations, overlooking their dynamic urban function. Temporal dynamics have been explored in weather forecasting (Nguyen et al., 2023; Bi et al., 2023) and multi-temporal Earth observation embedding fields (Brown et al., 2025). However, these approaches model how the physical environment evolves on fixed grids, rather than how humans navigate through it. Recently, ME-POIs (Siampou et al., 2026) introduced the concept of enriching static POI representations with human mobility signals, however, it models discrete check-in sequences rather than continuous trajectories. In contrast, TRAJGANR is the first framework to incorporate continious trajectories into geospatial multimodal pretraining. By modeling trajectories via a continuous neural implicit function, TRAJGANR extracts movement representations at arbitrary locations, enabling fine-grained alignment between street-view images and trajectories.

6 CONCLUSION

We introduced TRAJGANR, a trajectory-centric MSSL framework that aligns human mobility with street-view images and geographic locations at the finest granularity. By representing each trajectory as a path-conditioned neural implicit function, TRAJGANR infers localized trajectory embeddings directly at street-view image coordinates and aligns them with the corresponding visual and location representations. Experiments showed that TRAJGANR outperforms static multimodal baseslines and a trajectory foundation model on every task, with ablations attributing these gains the addition of trajectory modality and the alignment granularity. Future work will extend TRAJGANR to additional modalities (e.g., road networks, building footprints, remote sensing imagery), scale pretraining from regional to global coverage, and explore zero- and few-shot transfer techniques.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance (revising sections to improve clarity), and for coding assistance during the implementation of our method. We have not used generative AI tools for research ideation or experimental design. All AI-assisted text and technical claim was read, edited, and approved by the authors. AI-assisted code was reviewed by the authors and validated through the experiments reported in this paper. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here. Our multimodal, trajectory-centric geospatial framework could enable better location based services and urban planning tools. Our experiments rely on anonymized trajectory data and aggregated, anonymized tasks labels minimizing potential privacy concerns.

REPRODUCIBILITY STATEMENT

We have taken several steps to support the reproducibility of our work. The architecture of TRAJ-GANR, is described in detail, and all hyperparameters, model sizes, and training details are listed in Appendix A.1. We also describe the downstream tasks, label definitions, discretization procedures, data splits, evaluation metrics, and the baseline fine-tuning protocol. The trajectory datasets and street-view images used for pretraining are publicly available and widely adopted by the geospatial research community. The downstream labels are derived from aggregated and anonymized proprietary data and cannot be released. We therefore document their construction in detail so that the evaluation can be replicated with comparable data sources. All results are averaged over three splits with standard deviations reported. Our codebase will be publicly available upon acceptance.

REFERENCES

Chrysovalantis Anastasiou, Seon Ho Kim, and Cyrus Shahabi. Generation of synthetic urban vehicle trajectories. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 359–366. IEEE, 2022.

Lubin Bai, Xiuyuan Zhang, Haoyu Wang, and Shihong Du. Integrating remote sensing with open-streetmap data for comprehensive scene understanding through multi-modal self-supervised learning. *Remote Sensing of Environment*, 318:114573, 2025.

Pasquale Balsebre, Weiming Huang, Gao Cong, and Yi Li. Cityfm: City foundation models to solve urban challenges. *arXiv preprint arXiv:2310.00583*, 2023.

Michael Batty. *The new science of cities*. The MIT press, 2013.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

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.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinyu Chen, Jiajie Xu, Rui Zhou, Wei Chen, Junhua Fang, and Chengfei Liu. Trajvae: A variational autoencoder model for trajectory generation. *Neurocomputing*, 428:332–339, 2021.

Shushman Choudhury, Abdul Rahman Kreidieh, Ivan Kuznetsov, and Neha Arora. Towards a trajectory-powered foundation model of mobility. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications*, pp. 1–4, 2024.

Shushman Choudhury, Chandrakumari Suvarna, Iveel Tsogsuren, Abdul Rahman Kreidieh, Elad Aharoni, Chun-Ta Lu, and Neha Arora. S2vec: Self-supervised geospatial embeddings for the built environment. *ACM Transactions on Spatial Algorithms and Systems*, 2025.

Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. Range: Retrieval augmented neural fields for multi-resolution geo-embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24680–24689, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27672–27683, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Yujun Hou, Matias Quintana, Maxim Khomiakov, Winston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, Tianhong Zhao, and Filip Biljecki. Global streetscapes—a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215:216–238, 2024.

Weiming Huang, Daokun Zhang, Gengchen Mai, Xu Guo, and Lizhen Cui. Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:134–145, 2023.

Krzysztof Janowicz, Gengchen Mai, Weiming Huang, Rui Zhu, Ni Lao, and Ling Cai. Geofm: how will geo-foundation models reshape spatial data science and geoai? *International Journal of Geographical Information Science*, 39(9):1849–1865, 2025.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4347–4355, 2025.

Xiucheng Li, Kaiqi Zhao, Gao Cong, Christian S Jensen, and Wei Wei. Deep representation learning for trajectory similarity computation. In *2018 IEEE 34th international conference on data engineering (ICDE)*, pp. 617–628. IEEE, 2018.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

Yechen Li, Shantanu Shahane, Shoshana Vasserman, Carolina Osorio, Yi-fan Chen, Ivan Kuznetsov, Kristin White, Justyna Swiatkowska, Neha Arora, and Feng Guo. From lagging to leading: Validating hard braking events as high-density indicators of segment crash risk. *arXiv preprint arXiv:2601.06327*, 2026.

Yan Lin, Tonglong Wei, Zeyu Zhou, Haomin Wen, Jilin Hu, Shengnan Guo, Youfang Lin, and Huaiyu Wan. Trajfm: A vehicle trajectory foundation model for region and task transferability. *arXiv preprint arXiv:2408.15251*, 2024a.

Yan Lin, Zeyu Zhou, Yicheng Liu, Haochen Lv, Haomin Wen, Tianyi Li, Yushuai Li, Christian S Jensen, Shengnan Guo, Youfang Lin, et al. Unite: a survey and unified pipeline for pre-training spatiotemporal trajectory embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 37(3):1475–1494, 2024b.

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*, 2024.

Kang Liu, Ling Yin, Feng Lu, and Naixia Mou. Visualizing and exploring poi configurations of urban regions on poi-type semantic space. *Cities*, 99:102610, 2020.

Zeping Liu, Lao Ni, Zhangyu Wang, Junfeng Jiao, and Gengchen Mai. Gair: Location-aware self-supervised contrastive pre-training with geo-aligned implicit representations. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2026.

Zhipeng Ma, Zheyan Tu, Xinhai Chen, Yan Zhang, Deguo Xia, Guyue Zhou, Yilun Chen, Yu Zheng, and Jiangtao Gong. More than routing: Joint gps and route modeling for refine trajectory representation learning. In *Proceedings of the ACM Web Conference 2024*, pp. 3064–3075, 2024.

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 *ICLR 2020*. openreview, 2020.

Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*. PMLR, 2023.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Transactions on Spatial Algorithms and Systems*, 2024.

Gengchen Mai et al. A review of location encoding for geoai: methods and applications. *IJGIS*, 36 (4):639–673, 2022.

Dominik J Mühlematter, Lin Che, Ye Hong, Martin Raubal, and Nina Wiedemann. Urbanfusion: Stochastic multimodal fusion for contrastive learning of robust spatial representations. In *Forty-third International Conference on Machine Learning*, 2025.

Dominik J. Mühlematter, Lin Che, Ye Hong, Martin Raubal, and Nina Wiedemann. Urbanfusion: Stochastic multimodal fusion for contrastive learning of robust spatial representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=C0NCLB6ZGY.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning*, pp. 25904–25938. PMLR, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Meghan O'Connell, L Moreira-Matias, and Wendy Kan. Ecml/pkdd 15: Taxi trajectory prediction (i). *Kaggle. Retrieved April*, 11:2025, 2015.

Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. Crawdad data set epfl/mobility (v. 2009-02-24), 2009.

Jinmeng Rao, Song Gao, Yuhao Kang, and Qunying Huang. Lstm-trajgan: A deep learning approach to trajectory privacy protection. In *11th International Conference on Geographic Information Science (GIScience 2021)-Part I (2020)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2020.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1765–1774. IEEE, 2025.

Maria Despoina Siampou, Shang-Ling Hsu, Shushman Choudhury, Neha Arora, and Cyrus Shahabi. Toward foundation models for mobility enriched geospatially embedded objects. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 774–779, 2025a.

Maria Despoina Siampou, Jialiang Li, John Krumm, Cyrus Shahabi, and Hua Lu. Poly2vec: Polymorphic fourier-based encoding of geospatial objects for geoai applications. *Proceedings of Machine Learning Research*, 267:55511–55532, 2025b.

Maria Despoina Siampou, Shushman Choudhury, Shang-Ling Hsu, Neha Arora, and Cyrus Shahabi. Mobility-embedded pois: Learning what a place is and how it is used from human movement. *arXiv preprint arXiv:2601.21149*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jue Wang and Anoop Cherian. Learning discriminative video representations using adversarial perturbations. *arXiv preprint arXiv:1807.09380*, 2018.

Xiangheng Wang, Ziquan Fang, Chenglong Huang, Danlei Hu, Lu Chen, and Yunjun Gao. Gtr: A general, multi-view, and dynamic framework for trajectory representation learning. In *Forty-second International Conference on Machine Learning*, 2025.

Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

Can Yang and Gyozo Gidofalvi. Fast map matching, an algorithm integrating hidden markov model with precomputation. *International Journal of Geographical Information Science*, 32(3):547–570, 2018.

Hao Yang, Angela Yao, Christopher C Whalen, and Gengchen Mai. Bert4traj: Transformer-based trajectory reconstruction for sparse mobility data. In *13th International Conference on Geographic Information Science (GIScience 2025)*, pp. 8–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2025.

Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 186–194, 2012.

Jie Zhao, Chao Chen, Yuanshao Zhu, Mingyu Deng, and Yuxuan Liang. Unitr: A unified framework for joint representation learning of trajectories and road networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13348–13356, 2025.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–55, 2014.

Haicang Zhou, Weiming Huang, Yile Chen, Tiantian He, Gao Cong, and Yew-Soon Ong. Road network representation learning with the third law of geography. *Advances in Neural Information Processing Systems*, 37:11789–11813, 2024.

Silin Zhou, Yao Chen, Shuo Shang, Lisi Chen, Bingsheng He, and Ryosuke Shibasaki. Blurred encoding for trajectory representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4132–4143, 2025.

Silin Zhou, Chenhao Wang, Yuntao Wen, Shuo Shang, Lisi Chen, and Panos Kalnis. From gps points to travel patterns: Flexible and semantic trajectory generation with llms. *arXiv preprint arXiv:2605.30014*, 2026.

Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. Difftraj: Generating gps trajectory with diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36:65168–65188, 2023.

Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Xun Zhou, Liang Han, Xuetao Wei, and Yuxuan Liang. Unitraj: Learning a universal trajectory foundation model from billion-scale worldwide traces. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A Appendix

A.1 Additional Details on Experimental Setup

A.1.1 Dataset Statistics and Preprocessing

Street-View Images. We download street-view images via the Mapillary API. To ensure visual fidelity, we filter out low-visibility and obstructed frames by excluding images with a `quality_score` below 0.8, using the automated quality metrics provided in the Mapillary metadata. Each retained image is associated with its camera capture coordinates and the identifier of its snapped OpenStreetMap (OSM) [5] road segment.

Trajectory Preprocessing and Map Matching. We spatially filter the raw trajectory datasets to retain only trips that fall within the geographic bounding box established by the street-view image dataset. Following standard practices in trajectory modeling (Zhu et al., 2025; Wang & Cherian, 2018), we extract the drivable road network for each study region from OSM using the `OSMnx` library. To correct for GPS noise, each raw trajectory point is map-matched to the road network graph (Yang & Gidofalvi, 2018). We filter out short trips with less than 10 points to ensure meaningful sequential context. Because both trajectories and street-view images are aligned to the same underlying OSM network, we use these shared segment identifiers to determine the query set $\mathcal{Q}_i$ for each trajectory $\mathbf{r}_i$.

Summary Statistics. Table 6 summarizes the statistics of our multimodal datasets for both regions.

Table 6: Summary statistics of the preprocessed multimodal datasets.

Metric	Porto	San Francisco
Bounding Box (`lat, lon`)	$[41.13, 41.19] \times [-8.69, -8.55]$	$[37.74, 37.77] \times [-122.45, -122.43]$
Time range	05/17/2008 - 06/10/2008	07/01/2013 - 06/30/2014
# Trajectories	674,740	5,381
# Street-view images	51,729	68,040

A.1.2 Downstream Tasks

We evaluate on four road-level tasks that are representative of fine-grained urban understanding. These tasks cover complementary aspects of a road segment: its traffic conditions (traffic speed), destination demand (road popularity), surrounding land use (AOI function), and safety risk (HBEs). To assess generalization beyond the road network, we also evaluate on the POI popularity task. All labels are defined per road segment. Each example is represented by the coordinate of a street-view image on the segment, which the model embeds directly.

- **Traffic speed**: A regression task that predicts the average observed traffic speed on each road segment (km/h), derived from aggregated and anonymized Google Maps traffic data that are independent of the input trajectories. Traffic speed can differ from the posted speed limit, as it reflects real-world driving behavior. There are 40,338 labels in San Francisco.
- **Road popularity**:We define road popularity as a proxy for user interest in reaching destinations along a road segment, measured by the average number of Google Maps direction queries to the segment. Since we do not observe actual visits, this provides an indirect signal of the segment's popularity as a destination. We discretize the continuous query values into four ordinal classes from low to high intent. This task evaluates the model's ability to predict destination demand, i.e., where people intend to go. There are 2,084 labels in San Francisco, and 1,752 in Porto.
- **Area-of-interest (AOI) function**: The AOI function prediction task aims to predict the distribution of POI categories associated with a specific road segment. To construct the ground truth, we define an AOI by generating a 50-meter buffer around each road segment and retrieving all enclosed POIs. We then compute and normalize the frequency of each POI category to create a probability distribution, which serves as a proxy for the segment's underlying function (Liu et al., 2020; Huang et al., 2023; Zhou et al., 2024). This distribution is assigned as the ground truth label for all

[5] https://openstreetmap.org

streetview images located along that segment, and is tested against the predicted distribution of our fine-tuned model (Liu et al., 2020). There are 56,549 labels in San Francisco, and 20,199 in Porto.

- **Hard breaking events (HBE)**: A regression task that predicts the number of hard-braking events on each road segment, derived from aggregated and anonymized data from the Google Android Auto platform. Hard-braking events indicate evasive maneuvers and have been validated as a segment-level surrogate for crash risk (Li et al., 2026). There are 2,974 labels in San Francisco, and 1,282 in Porto.
- **POI popularity**: Similar to road-segment popularity, we define visit intent as a proxy for user interest in visiting a POI, measured by the average number of Google Maps direction queries to the location. Since we do not observe actual visits, this provides an indirect signal of interest. We discretize the continuous query values into four ordinal classes from low to high intent. In San Fransciso, we evaluate on 1,629 POIs.

A.1.3 Details on Baselines

We compare TrajGANR against the latest geospatial MSSL models spanning the modalities it integrates: location encoders pretrained with visual supervision (GeoCLIP, SatCLIP), a static multimodal model aligning street-view imagery, remote sensing imagery, and coordinates (GAIR), and a trajectory foundation model (UniTraj). Each covers a subset of TrajGANR's modalities.

- **GeoCLIP** (Cepeda et al., 2023) is a CLIP-inspired geospatial model that contrastively aligns images with their GPS coordinates. GeoCLIP's image encoder builds on a frozen CLIP backbone, while its location encoder models the Earth as a continuous function, combining Random Fourier Feature (RFF) positional encodings at multiple resolutions into a hierarchical location representation. Although developed for worldwide image geolocalization, its pretrained location encoder has since been adopted as a general-purpose location representation for downstream geospatial prediction tasks (Klemmer et al., 2025; Mühlematter et al., 2026). We use its pretrained location encoder to embed each query coordinate.
- **SatCLIP** (Klemmer et al., 2025) is a global, general-purpose location encoder pretrained by contrastively matching openly available satellite imagery with its geographic coordinates. Its location encoder combines spherical harmonic basis functions with a sinusoidal representation network, yielding embeddings that summarize the environmental characteristics of any location on Earth. Similar to GeoCLIP, we use its pretrained location encoder to embed each query coordinate.
- **GAIR** (Liu et al., 2026) is a location-aware self-supervised framework that jointly pretrains factorized encoders for overhead remote sensing imagery, street-view imagery, and geographic coordinates with contrastive objectives. To align street-view images with overhead imagery, it interpolates ViT patch features within each remote sensing image to obtain a localized overhead feature at the street-view location. Since GAIR has highlighted the predictive value of each of its encoders, we evaluate three variants: its location encoder GAIR $_{\text{LOC}}$, its street-view encoder GAIR $_{\text{SVI}}$, and their combination GAIR $_{\text{LOC+SVI}}$. Our TrajGANR adopts GAIR's street-view encoder as its visual backbone, so GAIR $_{\text{SVI}}$ corresponds to TrajGANR without the trajectory and location modality. GAIR is pretrained on Streetscapes1M, which samples one million street-view images across 688 cities from the Global Streetscapes dataset (Hou et al., 2024). The street-view images used to pretrain TrajGANR are drawn from the same Global Streetscapes source.
- **UniTraj** (Zhu et al., 2025) is a universal trajectory foundation model pretrained on WorldTrace [6], a corpus of 2.45 million trajectories spanning 70 countries. It employs an encoder-decoder Transformer with rotary position encodings, trained with self-supervised trajectory masking and adaptive resampling to handle heterogeneous sampling rates. Notably, UniTraj has been evaluated on the Porto taxi dataset in both zero-shot and fine-tuning settings, so one of our two study regions is not foreign to it. UniTraj produces embeddings only at observed GPS points and, like other trajectory encoders, cannot be queried at coordinates where no point was recorded. We therefore obtain a representation for each road segment by mean-pooling the embeddings of all GPS points assigned to that segment, and fine-tune the model with the same task heads as TrajGANR.

[6] https://huggingface.co/datasets/OpenTrace/WorldTrace

A.1.4 IMPLEMENTATION DETAILS

Data Splits. We partition the data using a spatial split over the road network. Road segments are divided into disjoint training and test sets such that approximately 60% of trajectories fall in the training set and 40% in the test set, with each trajectory assigned entirely to one side. Similarly, each street-view image is assigned to the same split as the road segment it is associated with. Consequently, evaluation query locations, along with their street-view images, labels, and intersecting trajectories, are unseen during model fine-tuning. Our TRAJGANR and all baselines are fine-tuned on the same training split and evaluated on the test split. All reported results are averaged over three independent spatial splits, and we report mean and standard deviation across runs.

Coordinate Normalization. Given our focus on specific urban regions across two datasets, we normalize all geographic coordinates into the range $[0, 1]$ relative to each area's bounding box. This local normalization ensures the model remains sensitive to fine-grained regional structures.

Sampling and Augmentation. During pretraining, we employ a trajectory-centric alignment strategy, where each sample consists of a single trajectory and 32 street-view images whose geolocations lie along the trajectory's continuous spatial footprint. We enforce a fixed trajectory window size of $L_i^{(r)} = 240$, applying random downsampling for longer trajectories and zero-padding for shorter ones. This stochastic dropping of waypoints serves as data augmentation and forces the model to learn trajectory representations invariant to varying sampling frequencies and diverse temporal gaps. Street-view images undergo standard augmentations following prior work (Cepeda et al., 2023; Chen et al., 2020; Liu et al., 2026). During fine-tuning, we shift to a location-centric sampling strategy, where each sample consists of a street-view image, its associated location, and 16 sampled intersecting trajectories.

Architecture and Optimization. We use a shared projection dimension of 128 and a location embedding dimension of 128. For the Space2Vec encoders, we set $\lambda_{\min} = 1$, $\lambda_{\max} = \sqrt{2}$, and use 64 spatial scales. The trajectory encoder is a 4-layer Transformer with 8 attention heads and feed-forward dimension 512. We pretrain for 300 epochs using Adam with batch size 256, learning rate 10^{-4}, and weight decay 10^{-4}, and maintain a location negative queue of 2048 embeddings. For downstream fine-tuning, we train for 100 epochs using a two-layer MLP prediction head with hidden dimension 1024 and ReLU activation.

Baselines. We use the publicly released checkpoints for GeoCLIP and SatCLIP, and follow the hyperparameters provided in the official implementations of GAIR and UniTraj. All baselines are initialized from their released pretrained weights and fine-tuned end-to-end with the same task heads, folds, and training protocol as TRAJGANR. Because the methods require different inputs, each is evaluated on the locations where its inputs are available.

A.1.5 PRETRAINING OBJECTIVE

We utilize InfoNCE loss (Oord et al., 2018) to pretrain our TRAJGANR. In more detail, we define the pairwise InfoNCE loss between anchor $\mathbf{u}$ and positive target $\mathbf{v}^+$ over negative set $\mathcal{N}$ as:

$$\mathcal{L}_{\text{info}}(\mathbf{u}, \mathbf{v}^+; \mathcal{N}) = -\log \frac{\exp\left(\text{sim}(\mathbf{u}, \mathbf{v}^+)/\tau\right)}{\exp\left(\text{sim}(\mathbf{u}, \mathbf{v}^+)/\tau\right) + \sum_{\mathbf{v}^- \in \mathcal{N}} \exp\left(\text{sim}(\mathbf{u}, \mathbf{v}^-)/\tau\right)},$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity scaled by temperature τ.

A.1.6 COMPUTATIONAL ENVIRONMENT

Our experiments are performed on a cluster node equipped with 8 CPU cores, 128 GB of memory, and one NVIDIA H100 GPU. Furthermore, all neural network models are implemented in Python 3.10 using PyTorch 2.3.0 and CUDA 12.1.

A.1.7 LICENSES AND TERMS OF USE

We use publicly available trajectory datasets and street-view imagery sources. The Porto taxi trajectory dataset from the ECML/PKDD 2015 Taxi Trajectory Prediction Challenge is distributed under the Creative Commons Attribution 4.0 International license (CC BY 4.0), and we cite the original dataset source. The Cabspotting dataset contains taxi mobility traces from San Francisco and was originally

provided by the Exploratorium Piorkowski et al. (2009). Street-view images are retrieved from Mapillary, whose images are shared under a Creative Commons Attribution-ShareAlike license (CC BY-SA). We follow Mapillary's terms of use. All datasets are used explicitly for research purposes, with attribution to the original creators and providers.

A.2 Evaluation of Query-Time Encoding Strategies

Table 7: Ablation study on query-time temporal encoding strategies for the trajectory neural implicit function: synthetic-time query encoding and masked-time query encoding – as we described in Section 3.2. The results of each setting on the SF dataset are averaged over 3 runs, and "$\pm x$" denotes the standard deviation. **Best** and second best values are highlighted.

Method	*Traffic Speed*		*Road Popularity*		*AOI Function*		*HBE*	
	(↓) MAE	(↓) RMSE	(↑) F1	(↑) AUPRC	(↓) L1	(↑) Cosine	(↓) MAE	(↓) RMSE
SYNTHETIC-TIME	**0.403**$_{\pm.12}$	**1.187**$_{\pm.22}$	0.960$_{\pm.02}$	**0.992**$_{\pm.00}$	**0.121**$_{\pm.03}$	**0.981**$_{\pm.01}$	**0.130**$_{\pm.02}$	0.926$_{\pm.19}$
MASKED-TIME	0.451$_{\pm.04}$	1.230$_{\pm.07}$	**0.970**$_{\pm.08}$	0.987$_{\pm.01}$	0.179$_{\pm.04}$	0.972$_{\pm.01}$	0.147$_{\pm.02}$	**0.873**$_{\pm.20}$

We empirically compare the two proposed query-time encoding strategies, to study their impact on downstream performance. Table 7 presents the results. Overall, the synthetic-time variant achieves stronger performance across most metrics. This suggests that although street-view image locations are not observed GPS samples, the interpolated timestamp allows the model to place each query point within the continuous temporal path of the trajectory, which is especially useful for tasks that depend on localized mobility context. In contrast, the masked-time variant avoids the explicit assumption about the query timestamp, but requires the Transformer to infer the temporal position of the query token from the observed trajectory samples. This inference becomes more difficult when trajectories are sparse or irregularly sampled, which may explain its weaker overall performance. Nevertheless, the masked-time variant remains competitive and achieves the best result on a few metrics, suggesting that explicit temporal interpolation is not always necessary and may occasionally introduce noise. Taken together, these results suggest that synthetic-time encoding provides a strong overall design choice despite its simplicity, while masked-time encoding remains a reasonable alternative when interpolation is unreliable.