NARA: ANCHOR-CONDITIONED REPRESENTATION LEARNING FOR HETEROGENEOUS VECTOR GEOENTITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vector geospatial data represent the world as discrete geoentities, such as roads, buildings, and points of interest (POIs), each with semantic attributes, geometry, and spatial relations to other geoentities, including metric proximity and topology. Existing methods for learning geoentity representations typically support a single geometry type or model only a subset of these relations, limiting their ability to capture spatial context across heterogeneous geoentities and support diverse downstream tasks. We propose **NARA** (**N**eural **A**nchor-conditioned **R**elation-**A**ware representation learning), a novel self-supervised representation framework for heterogeneous vector geoentities. NARA contextualizes geoentities through spatial-context-aware attention that models spatial autocorrelation using geometry distance modulated by topological relations across surrounding points, polylines, and polygons. NARA introduces masked geoentity semantic modeling and geometry-aware spatial relation modeling, as well as relation-conditioned regularization that encourages similar representations for geoentities sharing the same spatial relation to a common reference entity, while accounting for spatial autocorrelation. NARA's frozen, task-agnostic encoder outperforms state-of-the-art methods, each with an architecture tailored to its respective task, across traffic-speed prediction for polylines, building-function classification for polygons, and next point-of-interest prediction for points.

1 INTRODUCTION

Vector maps such as OpenStreetMap (OSM) (OpenStreetMap contributors, 2017) and Overture Maps (OpenStreetMap contributors, Overture Maps Foundation, 2026) provide large-scale, human-curated abstractions of the world as collections of discrete *geoentities*, including roads, buildings, and points of interest (POIs). Unlike sensed raster data (e.g., satellite imagery), vector data reflect intentional human curation (Goodchild, 1992), in which important entities are selected, preserved, and represented at different geometric granularities (e.g., polygons vs. points) according to their functional roles (Longley et al., 2015). Geoentities are further organized through spatial relations, both metric and topological, that characterize their spatial configurations within the surrounding environment. Vector data thus describe what geoentities are, what shapes they have, and how they are spatially related, providing a rich foundation for representation learning that supports diverse geospatial tasks (Mai et al., 2024; Siampou et al., 2025).

Learning task-agnostic representations across geometry types requires spatial contextualization (Balsebre et al., 2024), in which a target geoentity is represented through its semantics, geometry, and spatial relations to surrounding geoentities (Mai et al., 2020). For example, a Starbucks *inside a mall* may exhibit visiting patterns unlike one *in an airport* or *along Fifth Avenue*, and a Starbucks mapped as a polygon rather than a point may denote a larger venue. Existing methods on contextual geoentity representation learning, however, often support only a single geometry type, learning representations separately for points (Mai et al., 2020; Li et al., 2022; Cheng et al., 2025), polylines (Zhang & Long, 2023; Zhou et al., 2024), or polygons (Li et al., 2023b; Mai et al., 2023) with type-specific architectures that do not transfer across geometry types. Recent methods supporting multiple geometry types model spatial context only partially, through point-based proximity (Tempelmeier et al., 2021) or predefined relational schemas that depend on sparse auxiliary data such as mobility (Yang et al., 2025) or transit routes (Balsebre et al., 2024). A distinct line of work learns unified geometry representations across points, polylines, and polygons (Siampou et al., 2025; Chu & Shahabi, 2026). These

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.

methods encode geometry effectively but represent each geoentity in isolation, without modeling their semantics or spatial context. To our knowledge, no existing approach jointly contextualizes semantics, complete geometry, and spatial relations across heterogeneous geoentities, the capability needed for task-agnostic representations that transfer across geometry types and downstream tasks.

We propose **NARA** (**N**eural **A**nchor-conditioned **R**elation-**A**ware representation learning), a novel self-supervised framework for learning task-agnostic contextual representations of heterogeneous vector geoentities. NARA contextualizes each geoentity from its semantics, complete geometry, and spatial relations to surrounding points, polylines, and polygons through spatial-context-aware attention, whose topology-modulated distance bias modulates the effect of complete-geometry distance by topological relation. NARA introduces (i) *masked geoentity semantic modeling*, which reconstructs a target's masked semantics from its retained geometry and spatial context, (ii) *geometry-aware spatial relation modeling*, which requires the contextual representations to encode spatial relations, and (iii) *relation-conditioned regularization*, which encourages geoentities sharing the same spatial relation to a common reference entity, termed an *anchor*, to learn similar representations beyond what their proximity explains. We pretrain NARA separately on New York City (NYC) and Singapore with the same architecture and keep its task-agnostic encoder frozen for downstream tasks. NARA outperforms state-of-the-art approaches with task-specific architectures and task-aligned pretraining objectives on all three tasks spanning every geometry type.

2 Related Work

Existing representation learning methods for geospatial vector data are typically designed for a single geometry type (e.g., points) (Klemmer et al., 2025; Liu et al., 2026), rely on specific data sources or attributes (e.g., place categories) (Balsebre et al., 2024), and define context using fixed similarity notions such as spatial proximity or predefined functional relations (Chen et al., 2025). This limits their ability to capture spatial interactions and generalize across heterogeneous geoentities. Point-based methods encode semantic attributes with coordinate embeddings (Mai et al., 2020; Li et al., 2022), but lack spatial extent and topology. Line-based methods, largely focused on road networks, learn from graph connectivity (Grover & Leskovec, 2016; Wang et al., 2019; Jepsen et al., 2022; Zhang & Long, 2023; Zhou et al., 2024) but ignore geometric shape and spatial relations with surrounding entities, and are difficult to generalize beyond road networks. Polygon-based methods either rasterize shapes for visual encoders (Mai et al., 2023; Balsebre et al., 2024; Li et al., 2023b), losing geometric precision and scale, or encode coordinates only (e.g., Mai et al. (2023)), and model entities in isolation without inter-entity context. Geometry-type-agnostic encoders (Siampou et al., 2025; Chu & Shahabi, 2026) unify representations across geometry types but focus solely on geometry and require downstream adaptation to capture semantic, metric, or topological relationships. More general frameworks in urban computing (Tempelmeier et al., 2021; Balsebre et al., 2024; Yang et al., 2025) contextualize heterogeneous entities, yet rely on proximity or predefined functional dependencies from available data schemas (e.g., OSM relations), rather than learning how topological relations shape functional similarity. As a result, semantics and spatial structure are treated separately instead of jointly as complementary information sources. NARA addresses these limitations by jointly encoding semantic attributes and geometry across heterogeneous entity types. In addition, prior work has incorporated spatial autocorrelation, one of the critical spatial principles, through strategies such as training data sampling (Yan et al., 2017), negative sampling (Mai et al., 2019; Huang et al., 2022), and specialized loss functions (Jean et al., 2019; Wang et al., 2024). However, these approaches focus solely on metric relations and are limited to satellite imagery (Jean et al., 2019; Li et al., 2023a) or point-based entities (Mai et al., 2019; Huang et al., 2022; Wang et al., 2024). In contrast, NARA models spatial autocorrelation across geoentities with diverse geometry types and explicitly captures how metric and topological relations jointly shape contextual representations of vector geoentities.

3 Spatial Contextualization Principles

Preliminaries Given a set of geoentities $\mathcal{G} = \{v_1, \ldots, v_n\}$, each entity v_i has semantic attributes s_i (e.g., a category or textual description), geometry g_i, and geometry type $\eta(v_i) \in \{\text{point}, \text{polyline}, \text{polygon}\}$. Our goal is to learn a function $f : \mathcal{G} \to \mathbb{R}^{d_{\text{model}}}$ that maps these geoentities to a shared space while retaining their semantics, geometry, and spatial structure, including both metric proximity and topological relations.

Spatial Dependencies Geographic systems exhibit structured spatial dependencies. Tobler's First Law states that nearby entities tend to be more related than distant ones (Tobler, 1970; 2004), motivating *spatial autocorrelation*, where similarity decays with distance, often characterized by the

semivariogram (Bachmaier & Backes, 2011). Lin et al. (2020); Wang et al. (2024) further generalized the principle of semivariograms into the latent embedding space:

$$\gamma(b) \;=\; \frac{1}{2}\,\mathbb{E}\big[\|Z(v_i) - Z(v_j)\|^2 \,\big|\, d_{ij} = b\big]\,, \tag{1}$$

where $Z(v_i)$ denotes a latent representation associated with geoentity v_i, and d_{ij} is the geographic distance between v_i and v_j. Smaller $\gamma(b)$ indicates stronger similarity at distance separation b, and $\gamma(b)$ typically increases with distance in a nonlinear, saturating manner. For vector geoentities, this suggests that nearby entities provide stronger contextual information and that contextual influence should decay with distance in a data-dependent way.

Beyond metric proximity, spatial reasoning also depends on *topological* structure (e.g., containment and adjacency) (Egenhofer, 1990; Randell et al., 1992; Regalia et al., 2019). Such relations can override proximity: entities sharing meaningful topology (e.g., stores within the same mall or businesses along the same avenue) may exhibit similar functions despite being geographically separated. Let a denote a reference (anchor) entity and $\mathrm{rel}(v_i, a) \in \mathcal{R}$ the topological relation between v_i and a. We expect entities sharing the same relation to a common anchor to be more similar than entities at the same distance in general:

$$\mathbb{E}\big[\mathrm{sim}\big(Z(v_i), Z(v_j)\big) \,\big|\, \mathrm{rel}(v_i, a) = \mathrm{rel}(v_j, a) \neq \varnothing,\; d_{ij} = b\big] > \mathbb{E}\big[\mathrm{sim}\big(Z(v_i), Z(v_j)\big) \,\big|\, d_{ij} = b\big]\,. \tag{2}$$

Here, Eq. (2) should be viewed as a modeling assumption motivated by geographic spatial organization rather than a theoretical guarantee. Therefore, motivated by Eq. (1) and Eq. (2), we model functional similarity as jointly governed by metric distance and relational agreement:

$$\mathrm{sim}(Z(v_i), Z(v_j)) \propto k(d_{ij}) \cdot \mathbb{E}_{a \sim p(a \mid v_i, v_j)}\big[\psi\big(\mathrm{rel}(v_i, a), \mathrm{rel}(v_j, a)\big)\big]\,, \tag{3}$$

where $k(\cdot)$ models distance decay and $\psi(\cdot)$ captures topological modulation. The expectation is taken over anchors relevant to the pair (v_i, v_j), allowing entities to participate in multiple relational contexts. Eq. (3) is not intended as a probabilistic generative model, but rather as a conceptual formulation motivating the representation objectives introduced in §4. This perspective suggests two desirable properties for contextual representation learning over heterogeneous geoentities. **Distance-decay:** contextual similarity decreases with geographic distance in a nonlinear manner. **Topological modulation:** entities sharing similar topological relations to a common anchor should exhibit enhanced similarity beyond metric proximity alone.

4 NARA

§3 motivates the design of NARA (Fig. 1), a self-supervised framework for learning contextual geoentity representations, in which pairwise metric and topological relations directly modulate contextualization through topology-modulated distance bias, and relation-conditioned regularization regularizes anchor relation-dependent similarity beyond pairwise proximity.

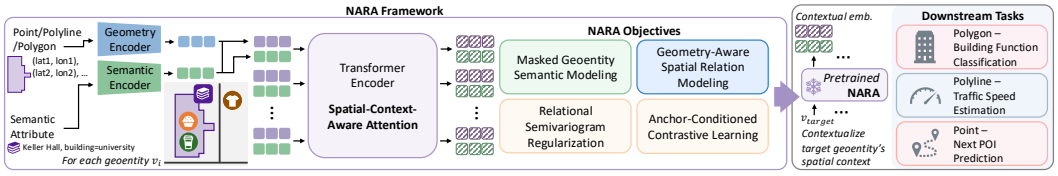

Figure 1: Overview of NARA, the proposed self-supervised learning framework.

4.1 Semantic and Geometric Representations of Geoentity

For each geoentity $v_i \in \mathcal{G}$, NARA encodes its semantic attributes s_i, represented as a sentence (e.g., composed of name, category, or other textual descriptors), with a language encoder (e.g., BERT (Devlin et al., 2019)), and its complete geometry g_i with a geometry encoder (e.g., Poly2Vec (Siampou et al., 2025)), to preserve shape, spatial extent, and spatial relations across geometry types:

$$\mathbf{e}_i^{\mathrm{sem}} = \mathrm{Enc}_{\mathrm{sem}}(s_i), \qquad \mathbf{e}_i^{\mathrm{geom}} = \mathrm{Enc}_{\mathrm{geom}}(g_i). \tag{4}$$

NARA projects both embeddings into $\mathbb{R}^{d_{\mathrm{model}}}$, yielding $\mathbf{z}_i^{\mathrm{sem}}$ and $\mathbf{z}_i^{\mathrm{geom}}$, and fuses them adaptively:

$$\rho_i = \sigma\big(f_{\mathrm{gate}}\big([\mathbf{z}_i^{\mathrm{sem}}; \mathbf{z}_i^{\mathrm{geom}}]\big)\big)\,, \qquad \mathbf{z}_i^{\mathrm{fuse}} = (1 - \rho_i)\mathbf{z}_i^{\mathrm{sem}} + \rho_i \mathbf{z}_i^{\mathrm{geom}}. \tag{5}$$

The scalar $\rho_i \in [0, 1]$ modulates the contribution of geometric information. We initialize the gating bias negatively so that training prioritizes semantic information and gradually incorporates geometry.

4.2 MASKED GEOENTITY SEMANTIC MODELING WITH SPATIAL-CONTEXT ATTENTION

Following §3, NARA contextualizes each geoentity through its semantics, geometry, and metric and topological relations to surrounding geoentities. We propose **Masked Geoentity Semantic Modeling (MGSM) with Spatial-Context-Aware attention**. Given a target whose semantic representation is masked, MGSM reconstructs its semantics from its retained geometry and the surrounding heterogeneous geoentities. For example, a point geoentity with masked semantics is likely a "shop" if it lies within a "commercial building" polygon, but more likely a "parking lot" if it lies just outside the same polygon. NARA models these interactions using a Transformer encoder over local geoentities with spatial-context-aware attention, which determines how much information a target geoentity receives from each neighbor based on the neighbor's semantics, geometry, and spatial relation to the target.

We define local spatial contexts by covering the target area (e.g., NYC) with overlapping windows $\{\mathcal{W}_\ell\}_{\ell=1}^{L}$ (e.g., $250\,\mathrm{m} \times 250\,\mathrm{m}$) placed at a fixed stride. Each window yields a context $\mathcal{N}(\ell)$, the set of heterogeneous geoentities intersecting $\mathcal{W}_\ell$, and the overlap preserves relational neighborhoods across window boundaries. Because geoentities near a window boundary may have incomplete context, NARA samples masked targets only from a central region of each window, allowing all entities in $\mathcal{N}(\ell)$ to provide contextual information. Let $\mathcal{M}_\ell \subseteq \mathcal{N}(\ell)$ denote the set of geoentities sampled for masking. We replace each target's semantic representation with a learned mask vector:

$$\mathbf{e}_{i,\ell}^{\mathrm{sem}} = \mathbf{e}^{\mathrm{mask}} \text{ if } v_i \in \mathcal{M}_\ell, \quad \mathbf{e}_{i,\ell}^{\mathrm{sem}} = \mathbf{e}_i^{\mathrm{sem}} \text{ otherwise}, \tag{6}$$

NARA projects the masked semantic representation and recomputes the adaptive fusion in Eq. (5), yielding the semantic and fused embeddings $\mathbf{z}_{i,\ell}^{\mathrm{sem}}$ and $\mathbf{z}_{i,\ell}^{\mathrm{fuse}}$. Prior approaches to contextual geoentity representation learning represent geographic position using point coordinates, often by adding location embeddings to token representations, primarily capturing point-based spatial proximity (Li et al., 2022; 2023c). In contrast, NARA represents heterogeneous vector geoentities by their complete geometry (e.g., location, shape, and spatial extent).

To reconstruct masked semantics, spatial-context-aware attention learns how these spatial relations among neighboring entities modulate the strength of contextualization. For clarity, we describe one attention head of the first layer and omit head and layer indices; subsequent layers operate on the preceding layer's outputs. Queries and keys are computed from the fused geoentity embeddings:$\mathbf{q}_{i,\ell} = \mathbf{W}_Q \mathbf{z}_{i,\ell}^{\mathrm{fuse}}, \mathbf{k}_{j,\ell} = \mathbf{W}_K \mathbf{z}_{j,\ell}^{\mathrm{fuse}}$. Standard dot-product attention scores a pair (v_i, v_j) using only their query and key representations, which do not explicitly capture the pairwise spatial relation between them. To let the attention score reflect this relation, we introduce a learned **topology-modulated distance bias** b_{ij}:

$$a_{ij,\ell} = \underbrace{\frac{\mathbf{q}_{i,\ell}^\top \mathbf{k}_{j,\ell}}{\sqrt{d_{\mathrm{head}}}}}_{\text{from fused geoentity representations}} + \underbrace{b_{ij}}_{\text{from their spatial relation}}, \qquad A_{ij,\ell} = \frac{\exp(a_{ij,\ell})}{\sum_{v_k \in \mathcal{N}(\ell)} \exp(a_{ik,\ell})}, \tag{7}$$

where d_{head} is the dimension of one attention head. A larger b_{ij} increases the contribution of v_j relative to other entities in the context, and a smaller value decreases its contribution. Adding pairwise structure to attention scores as a bias has proven effective in other domains, such as shortest-path distances between nodes in the graph domain (Ying et al., 2021) and 3D Euclidean distances between atoms conditioned on atom-pair types in the molecular domain (Zhou et al., 2023). For heterogeneous geoentities, we design a bias that jointly captures their distance and topological relation.

The proposed bias is derived directly from geoentity geometry. For each pair (v_i, v_j), $\bar{d}_{ij}$ denotes the normalized minimum distance between their complete geometry and $t_{ij} \in \mathcal{T} = \{\text{DISJOINT}, \text{CONTACT}, \text{CONTAINMENT}\}$ denotes their symmetric topological relation (see Appdx. A.1). These quantities are fixed geometric properties of the pair and require neither manual annotations nor downstream labels. We model topology as a symmetric predicate, since "a restaurant within a mall" and "a mall contains a restaurant" indicate the same shared functional context, and collapsing such inverse relations yields a simpler and denser relational representation across heterogeneous geoentities. Following §3, we instantiate the topology-modulated distance bias so that topology conditions the effect of distance:

$$b_{ij} = \underbrace{\beta_{t_{ij}}}_{\text{topology-specific offset}} + \underbrace{\mathbf{w}_{t_{ij}}^\top \boldsymbol{\phi}(\bar{d}_{ij})}_{\text{distance effect under topology } t_{ij}}, \tag{8}$$

where β_t is a learned offset and $\mathbf{w}_t$ is a learned weight vector for each topological relation $t \in \mathcal{T}$, and $\boldsymbol{\phi}(\cdot)$ denotes Gaussian radial basis features (see Appdx. A.2). The distance term models distance decay, and conditioning β_t and $\mathbf{w}_t$ on t_{ij} yields topological modulation, so the same pairwise distance affects attention differently for DISJOINT, CONTACT, and CONTAINMENT pairs. Both CONTACT and CONTAINMENT pairs have $\bar{d}_{ij} = 0$, but their learned distance effects under topology t_{ij} can still differ since the weights $\mathbf{w}_{t_{ij}}$ are topology-specific. Thus, both the learned topology-specific offset $\beta_{t_{ij}}$ and the learned distance effect $\mathbf{w}_{t_{ij}}^{\top}\phi(0)$ can distinguish these relations. For DISJOINT pairs, the distance effect under topology t_{ij} can additionally vary with $\bar{d}_{ij}$.These parameters are optimized jointly with the encoder rather than following a predetermined distance-decay function. They are learned separately for each attention head and shared across layers, so different heads can capture distinct topology–distance effects. Within each head, the same spatially guided attention weights aggregate the semantic and fused geoentity embeddings through separate value projections: $\tilde{\mathbf{h}}^{s}_{i,\ell} = \sum_{v_j \in \mathcal{N}(\ell)} A_{ij,\ell} \mathbf{W}_{V,s} \mathbf{z}^{s}_{j,\ell}$ for $s \in \{\mathrm{sem}, \mathrm{fuse}\}$. Sharing the attention map ties both embeddings to the same neighbor selection, while the separate value projections $\mathbf{W}_{V,\mathrm{sem}}$ and $\mathbf{W}_{V,\mathrm{fuse}}$ let each specialize. After multi-head aggregation and the remaining Transformer layers, each with separate semantic and fused feed-forward networks (Appdx. B.1), we denote the final outputs by the semantic contextual embedding $\mathbf{h}^{\mathrm{sem}}_{i,\ell}$ and the fused contextual embedding $\mathbf{h}^{\mathrm{fuse}}_{i,\ell}$. The semantic contextual embedding carries spatially contextualized semantics used to reconstruct masked attributes, described below, while the fused contextual embedding additionally carries geometry, from which spatial relations are modeled (§4.3).

MGSM reconstructs the original semantic embedding of each masked target as $\hat{\mathbf{e}}^{\mathrm{sem}}_{i,\ell} = f_{\mathrm{rec}}(\mathbf{h}^{\mathrm{sem}}_{i,\ell})$ for $v_i \in \mathcal{M}_\ell$. Reconstruction uses a contrastive objective in which each predicted embedding $\hat{\mathbf{e}}^{\mathrm{sem}}_{i,\ell}$ is pulled toward the target's original embedding $\mathbf{e}^{\mathrm{sem}}_{i}$ and pushed away from the original semantic embeddings of other geoentities in the same local context:

$$\mathcal{L}_{\text{MGSM}} = -\frac{1}{\sum_\ell |\mathcal{M}_\ell|} \sum_{\ell} \sum_{i \in \mathcal{M}_\ell} \log \frac{\exp\Big(\mathrm{sim}(\hat{\mathbf{e}}^{\mathrm{sem}}_{i,\ell}, \mathbf{e}^{\mathrm{sem}}_{i})/\tau_{\text{MGSM}}\Big)}{\sum_{j \in \mathcal{P}_\ell(i)} \exp\big(\mathrm{sim}(\hat{\mathbf{e}}^{\mathrm{sem}}_{i,\ell}, \mathbf{e}^{\mathrm{sem}}_{j})/\tau_{\text{MGSM}}\big)}, \tag{9}$$

where $\mathcal{P}_\ell(i)$ is the candidate set for v_i. It contains the target as the positive and the other geoentities in $\mathcal{N}(\ell)$ as negatives, with any geoentity whose semantic embedding is identical to the target's removed to avoid false negatives. $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and $\tau_{\text{MGSM}} > 0$ is the temperature.

4.3 Geometry-aware Spatial Relation Modeling

The spatial relation bias uses pairwise distance and topology to shape attention, yet these relations are not explicitly encoded in the representations. We introduce **Geometry-Aware Spatial Relation Modeling (GSRM)**, a self-supervised objective that serves two purposes. First, GSRM guides the fused contextual embeddings to encode pairwise distance and topology. Second, the gradient propagates back through the encoder layers and fused geoentity embeddings to the geometry encoder, grounding the geometry embeddings in spatial relations. The idea is that semantic context allows NARA to model these relations jointly with functional context rather than from geometry alone, such as containment versus adjacency, as in a shop inside a commercial building versus a parking lot beside it. GSRM therefore predicts these relations from the fused contextual embeddings, so that each embedding jointly encodes a geoentity's semantic context and its spatial relations to neighbors.

NARA samples two sets of unmasked geoentity pairs within each local context, for topology classification and distance regression, respectively. Topology pairs are stratified jointly by topology class and geometry-type pair, and distance pairs by distance interval and geometry-type pair, to mitigate class imbalance, as pairs are dominated by DISJOINT and common geometry types (see Appdx. B.1). For each sampled pair (v_i, v_j) in context ℓ, we construct the pair representation $\mathbf{p}_{ij,\ell} = \Big[\mathbf{h}^{\mathrm{fuse}}_{i,\ell} \odot \mathbf{h}^{\mathrm{fuse}}_{j,\ell}; \Big|\mathbf{h}^{\mathrm{fuse}}_{i,\ell} - \mathbf{h}^{\mathrm{fuse}}_{j,\ell}\Big|\Big]$, which is order-invariant, matching the symmetry of t_{ij} and $\bar{d}_{ij}$. For each sampled topology pair and distance pair, GSRM minimizes, respectively,

$$\mathrm{CE}(f_{\mathrm{topo}}(\mathbf{p}_{ij,\ell}),\, t_{ij}) \qquad \text{and} \qquad \big(f_{\mathrm{dist}}(\mathbf{p}_{ij,\ell}) - \bar{d}_{ij}\big)^2. \tag{10}$$

$\mathcal{L}_{\mathrm{topo}}$ and $\mathcal{L}_{\mathrm{dist}}$ average these terms over the sampled topology and distance pairs, and are combined as $\mathcal{L}_{\text{GSRM}} = \alpha_{\mathrm{topo}}\mathcal{L}_{\mathrm{topo}} + \alpha_{\mathrm{dist}}\mathcal{L}_{\mathrm{dist}}$. MGSM trains the semantic contextual embeddings through reconstruction, and GSRM grounds the fused contextual embeddings and the geometry encoder in

spatial relations. Because both objectives share one attention map, NARA learns which neighbors inform each target while keeping their distance and topology recoverable from the fused contextual embeddings for downstream tasks.

4.4 Relation-Conditioned Regularization

We introduce two complementary objectives that promote relation-conditioned contextualization (Eq. (3)). The first, **Anchor-Conditioned Contrastive Learning (ACC)**, focuses on *anchor-conditioned relations* by encouraging NARA to pull sibling entities sharing the same relation to an anchor closer in the contextual semantic embedding space, while pushing non-siblings apart, with attraction modulated by geographic distance. The second, **Relational Semivariogram Regularization (RSR)**, inspired by Kriging for spatial prediction (Lin et al., 2020), takes a broader *neighborhood-level* view by comparing anchor-conditioned sibling pairs with nearby, geometry-type-matched non-sibling pairs within the same geographic distance bins. It encourages the average similarity of sibling pairs to exceed the non-sibling baseline by a margin in each valid distance bin, reinforcing relation-conditioned similarity beyond metric proximity alone.

For an anchor a and a member v_i, $\mathrm{rel}(v_i, a) \in \mathcal{R}$ denotes their relation, where $\mathcal{R}$ consists of polygons along a polyline, points along a polyline, and points inside a polygon (e.g., buildings or POIs near a road, and POIs inside a building; see Appdx. A.3). This anchor relation is directed, since it specifies which geoentity plays the role of the shared anchor. Following Eq. (2), the members that hold the same relation to the same anchor form a *sibling group*, denoted as $\mathcal{S}_{a,r} = \{v_i \in \mathcal{N}(\ell) \setminus \mathcal{M}_\ell \mid \mathrm{rel}(v_i, a) = r\}$. The siblings $\mathcal{S}_i$ of v_i are the other members of every sibling group that contains v_i. A geoentity can therefore participate in several relational contexts, and, by Eq. (2), siblings are expected to be more similar than other geoentities at the same distance, motivating the two objectives below.

Anchor-Conditioned Contrastive Learning. ACC enforces relation-conditioned similarity by pulling sibling pairs together and pushing non-siblings apart, with attraction modulated by distance. Let c_{ij} denote the cosine similarity between $\mathbf{h}^{\mathrm{sem}}_{i,\ell}$ and $\mathbf{h}^{\mathrm{sem}}_{j,\ell}$, and let $\mathcal{E}_i$ denote the unmasked geoentities in $\mathcal{N}(\ell)$ sharing the geometry type of v_i, excluding v_i. The positives are the v_i's siblings, and the negatives are the remaining geoentities in $\mathcal{E}_i$. Because nearby siblings are expected to be more similar than distant ones (Eq. (1)), each sibling receives a weight $\omega_{ij} \propto \exp(-d_{ij}/\lambda)$ with scale $\lambda > 0$, normalized to sum to one over $\mathcal{S}_i$. For each geoentity v_i with at least one sibling, ACC minimizes

$$-\sum_{v_j \in \mathcal{S}_i} \omega_{ij} \log \frac{\exp(c_{ij}/\tau_{\mathrm{ACC}})}{\sum_{v_k \in \mathcal{E}_i} \exp(c_{ik}/\tau_{\mathrm{ACC}})}, \tag{11}$$

where d_{ij} denotes complete-geometry distance and $\tau_{\mathrm{ACC}} > 0$ is the temperature. $\mathcal{L}_{\mathrm{ACC}}$ averages this term over such geoentities within each context and then over contexts. Restricting negatives to the same geometry type keeps the contrast from being dominated by differences between points and polygons. This formulation allows entities to belong to multiple relational contexts, with distance-aware weighting ensuring proximate siblings exhibit stronger similarity than distant siblings or non-siblings at the same distance.

Relational Semivariogram Regularization. Building on Eq. (1) and Kriging for spatial prediction (Lin et al., 2020), RSR employs semivariogram-based regularization over ℓ_2-normalized contextual semantic embeddings $\mathbf{h}^{\mathrm{sem}}_{i,\ell}$. RSR compares each sibling group's trend with that of a relation-agnostic baseline at a matched distance, requiring siblings to be more similar than the baseline at every distance, as expected by Eq. (2). NARA uses cosine distance as a proxy for squared Euclidean distance (see Appdx. A.4), the empirical semivariance over a pair set Π is: $\hat{\gamma}(\Pi) = \frac{1}{|\Pi|} \sum_{(v_i, v_j) \in \Pi} (1 - c_{ij})$. NARA computes two semivariance estimates conditioned on a bin b of a geographic distance d_{ij}. For a sibling group $\mathcal{S}_{a,r}$ and bin b, the relational semivariance $\gamma^{(a,r)}_{\mathrm{rel}}(b)$ is $\hat{\gamma}$ over the sibling pairs of $\mathcal{S}_{a,r}$ in bin b. The neighborhood semivariance $\gamma^{(\eta_{a,r})}_{\mathrm{glob}}(b)$ serves as a relation-agnostic baseline, computed as $\hat{\gamma}$ over nearby non-sibling pairs in the same context and bin whose geometry type $\eta_{a,r}$ matches the group's members, and captures distance decay without relational structure. A gap between the two at the same distance therefore reflects the shared anchor relation. For each sibling group $\mathcal{S}_{a,r}$, RSR penalizes

$$\sum_b \pi^{(a,r)}_b \ \max\Big(0,\ \gamma^{(a,r)}_{\mathrm{rel}}(b) - \gamma^{(\eta_{a,r})}_{\mathrm{glob}}(b) + \delta\Big), \tag{12}$$

where the sum runs over valid distance bins containing sufficient sibling and non-sibling pairs, and $\delta > 0$ is a margin. The weight $\pi_b^{(a,r)}$ reflects the relative number of sibling pairs assigned to bin b. $\mathcal{L}_{\text{RSR}}$ aggregates these group-level penalties across sibling groups and spatial contexts.

Overall, NARA pretraining loss combines all four objectives:

$$\mathcal{L} = \mathcal{L}_{\text{MGSM}} + \mathcal{L}_{\text{GSRM}} + \alpha_{\text{ACC}}\,\mathcal{L}_{\text{ACC}} + \alpha_{\text{RSR}}\,\mathcal{L}_{\text{RSR}}, \tag{13}$$

where $\alpha_{\text{ACC}}, \alpha_{\text{RSR}} \geq 0$ weight the objectives, and $\mathcal{L}_{\text{GSRM}}$ carries its own internal weights $\alpha_{\text{topo}}, \alpha_{\text{dist}}$.

5 Experiments

We evaluate NARA on three standard tasks spanning all geometry types: traffic speed estimation on polylines (§5.1), building function classification on polygons (§5.2), and next POI prediction on points (§5.3). In all three tasks, the state-of-the-art method relies on task-specific feature engineering or geometry-type-specific architectures. In contrast, NARA uses a single task-agnostic architecture with shared hyperparameters, pretrained separately on OSM data from New York City (NYC) and Singapore (SG). No downstream labels are used during NARA pretraining. For downstream evaluation, the pretrained encoder is frozen, and each target geoentity is contextualized with the point, polyline, and polygon geoentities intersecting its 100-meter buffer. The target's own semantic representation is replaced with the mask vector, as in MGSM, so NARA infers the target from its geometry and spatial context rather than from its own attributes. Its contextual representation $[\mathbf{h}_i^{\text{sem}}; \mathbf{h}_i^{\text{fuse}}]$ is passed to a lightweight task-specific head. Experimental details are in Appdx. B.1–B.4.

5.1 Task A: Traffic Speed Estimation

This task predicts the average observed traffic speed on road segments, each represented as a polyline geoentity, using Uber Movement data mapped to OSM road segments. Traffic speed depends on the spatial context, surrounding land use and topological relations with nearby geoentities. NARA uses the same frozen, task-agnostic encoder as in the other two tasks, without road-specific architectures or pretraining objectives. It encodes each segment of a road and average-pools the segment embeddings into a road-level representation, handling arbitrary road lengths and segment granularities without architectural changes. In contrast, the state-of-the-art CityFM (Balsebre et al., 2024) is tailored to this task. It fine-tunes BERT (Devlin et al., 2019) on OSM tags of point and polygon entities, represents polylines with engineered features (e.g., segment length, vertex count, and one-hot road type), and uses the traffic speeds of neighboring roads from the training set as input. We also compare against road-network graph methods (Node2Vec (Grover & Leskovec, 2016), GCWC (Hu et al., 2019), RFN (Jepsen et al., 2022)), geometry encoders (Poly2Vec (Siampou et al., 2025), Geo2Vec (Chu & Shahabi, 2026)), and spatial representation methods (IRN2Vec (Wang et al., 2019), GeoVectors (Tempelmeier et al., 2021)). For a fair comparison with CityFM, NARA's downstream head receives the frozen NARA contextual representation concatenated with engineered road features and training-set neighboring speeds, both from CityFM Balsebre et al. (2024).

Results. NARA achieves the best results on every metric, reducing RMSE by 6.62%, MAE by 8.44%, and MAPE by 8.25% relative to the strongest baseline, CityFM, and raising R^2 from 0.77 to 0.80 (Tab. 1). All improvements are significant under Holm-corrected two-sided paired t-tests ($p < 0.01$). NARA improves on it solely by replacing CityFM's polyline-specific pipeline with its frozen, task-agnostic representation. The baseline categories show how each source of spatial context contributes: topology-only models (MAE 4.83–5.31) capture connectivity but lack semantics and geometry; geometry encoders (4.20–4.24) add shape; and semantic methods (3.78–3.92) add road attributes. CityFM (MAE 3.20, MAPE 19.27%) shows high performance among baselines with a polyline-specific contextualization pipeline. Across road types (Appdx. C.1), NARA lowers MAE relative to CityFM in 13 of 16 types, by 5.4–9.9% on the four most frequent (residential, primary, secondary, and tertiary; 83% of test roads) and by 11.9% and 15.9% on motorways and trunk roads, respectively. This suggests that modeling heterogeneous spatial context improves robustness across common urban road environments.

5.2 Task B: Building Function Classification

This task predicts land-use categories for untagged OSM building polygons using the standard 8-class Singapore government land-use taxonomy, following the CityFM protocol (Balsebre et al., 2024). A building's function depends not only on its footprint but on its spatial relations with surrounding POIs, roads, and neighboring buildings. NARA uses the frozen, task-agnostic encoder and classifies

Table 1: Traffic speed prediction: results averaged over 10 independent runs (SD in parentheses). Gains are computed relative to the second-best method. **Bold** = best, underline = second-best.

Category	Model	RMSE ↓	MAE ↓	R^2 ↑	MAPE ↓
Topology (Road Network)	Node2Vec	6.82 (± 0.12)	5.31 (± 0.05)	0.38 (± 0.04)	32.22% (± 0.6%)
	GCWC	6.74 (± 0.04)	5.20 (± 0.04)	0.41 (± 0.04)	32.75% (± 1.2%)
	RFN	6.45 (± 0.09)	4.83 (± 0.02)	0.46 (± 0.02)	30.10% (± 0.0%)
Geometry	Poly2Vec	5.66 (± 0.01)	4.24 (± 0.01)	0.56 (± 0.00)	25.75% (± 0.1%)
	Geo2Vec	5.64 (± 0.01)	4.20 (± 0.01)	0.56 (± 0.00)	25.42% (± 0.2%)
Semantics + Metric	IRN2Vec	5.02 (± 0.04)	3.78 (± 0.02)	0.66 (± 0.01)	24.30% (± 0.0%)
	GeoVectors	5.21 (± 0.07)	3.92 (± 0.06)	0.64 (± 0.06)	24.03% (± 0.3%)
Semantics + Geometry	CityFM	4.08 (± 0.01)	3.20 (± 0.01)	0.77 (± 0.02)	19.27% (± 0.8%)
	NARA	**3.81** (± 0.01) +6.62%	**2.93** (± 0.01) +8.44%	**0.80** (± 0.00) +3.90%	**17.68%** (± 0.1%) +8.25%

each building from its contextual representation alone. As for every downstream target, the building's semantic representation is masked, so the prediction relies only on its footprint and surrounding context. In contrast, the state-of-the-art CityFM is tailored to this task, adding a building-type-aligned pretraining objective that contrasts rasterized building footprints with OSM building-type labels and relying on engineered features such as precomputed building area. We also compare against non-spatial semantic models (BERT (Devlin et al., 2019)), geometry encoders (Poly2Vec (Siampou et al., 2025), Geo2Vec (Chu & Shahabi, 2026)), and spatial methods with neighborhood context (SpaBERT (Li et al., 2022), GeoVectors (Tempelmeier et al., 2021)).

Results. NARA achieves the best results on every metric, improving Macro-F1 by 7.15% (69.34 → 74.30), Weighted-F1 by 1.99%, and accuracy by 1.44% relative to the strongest baseline, CityFM (Tab. 2). All improvements are significant under Holm-corrected two-sided paired t-tests ($p < 0.001$). Unlike CityFM, NARA uses neither a building-specific pretraining objective nor engineered building features, relying only on its task-agnostic contextual representation. Since the target's semantics are masked, these gains come from its geometry and spatial context, which MGSM trains the encoder to exploit. The baseline categories show how each source of spatial context contributes: BERT (Macro-F1 37.56) shows semantic attributes alone are insufficient; geometry methods (45.12–58.35) show footprint shape is highly informative; and point-based spatial methods (GeoVectors, SpaBERT) underperform them, as point-level proximity fails to capture building extent and shape. Per class (Appdx. C.2), NARA's Macro-F1 gain comes mainly from underrepresented classes whose function cannot be inferred from geometry alone: Civic & Community Institutions (+27.49 F1), Commercial & Residential (+20.61), and Educational (+16.74).

Table 2: Building function classification: results averaged over 10 folds (SD in parentheses). Gains are computed relative to the second-best method. **Bold** = best, underline = second-best.

Category	Model	Macro-F1 ↑	Weighted-F1 ↑	Accuracy ↑
Non-Spatial Semantics	BERT (fine-tuned)	37.56 (± 0.3)	44.19 (± 0.7)	51.77 (± 0.7)
Geometry	Poly2Vec	45.12 (± 2.9)	84.25 (± 1.0)	86.65 (± 0.8)
	Geo2Vec	58.35 (± 0.7)	86.90 (± 0.3)	88.05 (± 0.3)
Semantics + Metric	GeoVectors	47.24 (± 1.4)	64.18 (± 2.1)	69.49 (± 1.5)
	SpaBERT	45.06 (± 1.1)	78.47 (± 2.4)	75.95 (± 2.7)
Semantics + Geometry	CityFM	69.34 (± 2.1)	90.95 (± 0.4)	91.81 (± 0.3)
	NARA	**74.30** (± 1.2) + 7.15%	**92.76** (± 0.3) + 1.99%	**93.13** (± 0.3) + 1.44%

5.3 Task C: Next Point-of-Interest Prediction

This task predicts the next POI a user will visit from a historical check-in sequence, using Foursquare data from NYC and SG (Yang et al., 2014) and following the POI-Enhancer protocol (Cheng et al., 2025). POI visit dynamics depend on surrounding context: a cafe in a commercial district attracts different visits from one in a residential area despite sharing a category. NARA's frozen, task-agnostic encoder jointly encodes each Foursquare POI with the OSM geoentities within its 100-meter buffer, without check-in data or other task-specific inputs. Because NARA is pretrained only on OSM, this setting also tests transfer to a vector data source unseen during pretraining (i.e., Foursquare). In contrast, the state-of-the-art POI-Enhancer Cheng et al. (2025) is tailored to this task: it enriches POI embeddings by prompting Llama-2-7B (Touvron et al., 2023) with task-specific inputs, including check-in statistics, OSM addresses (e.g., street name, postal code), and nearby-POI categories. Following POI-Enhancer, we fuse the resulting embeddings into two trajectory models, TALE (Wan

et al., 2022) and CTLE (Lin et al., 2021), via cross-attention, and compare each trajectory model alone, augmented with POI-Enhancer or NARA.

Results. NARA achieves the best Hit@1 and Hit@5 in all eight city–model–metric combinations, improving over POI-Enhancer by 3.25–17.38% and over the base trajectory models by up to 43.98% (NY, TALE, Hit@1) (Tab. 3). The contrast is sharpest in Singapore, where POI-Enhancer yields little or negative benefit (-2.60% to $+1.54\%$), whereas NARA improves every setting ($+2.57\%$ to $+13.89\%$). NARA thus outperforms an LLM-based method that uses task-specific inputs such as check-in statistics, while relying only on pretrained OSM spatial context. Per venue type in NYC (Appdx. C.3.1), NARA improves Hit@1 in 13 of the 15 most frequent categories (e.g., Subway, Bar) and ties in the remaining two, and improves Hit@5 in all 15.

Table 3: Next POI prediction: results averaged over 3 independent runs. Parentheses indicate relative improvement over the base model (TALE/CTLE); Δ denotes gain over POI-Enhancer. **Bold** = best.

Model	NY		SG	
	Hit@1 ↑	Hit@5 ↑	Hit@1 ↑	Hit@5 ↑
TALE	5.680 ± 0.184	12.170 ± 0.350	9.669 ± 0.073	21.801 ± 0.073
+ POI-Enhancer	7.728 ± 0.265 (+36.06%)	15.928 ± 0.051 (+30.87%)	9.690 ± 0.069 (+0.22%)	21.235 ± 0.246 (-2.60%)
+ **NARA**	**8.178** ± 0.087 (+43.98%) (Δ5.82%)	**17.121** ± 0.124 (+40.67%) (Δ7.49%)	**11.012** ± 0.275 (+13.89%) (Δ13.64%)	**24.460** ± 0.283 (+12.20%) (Δ15.19%)
CTLE	6.782 ± 0.300	13.735 ± 0.678	10.607 ± 0.064	22.922 ± 0.194
+ POI-Enhancer	7.076 ± 0.031 (+4.34%)	14.324 ± 0.154 (+4.29%)	10.529 ± 0.066 (-0.74%)	23.275 ± 0.055 (+1.54%)
+ **NARA**	**8.072** ± 0.085 (+19.02%) (Δ14.06%)	**16.814** ± 0.176 (+22.43%) (Δ17.38%)	**10.880** ± 0.103 (+2.57%) (Δ3.33%)	**24.033** ± 0.128 (+4.85%) (Δ3.25%)

5.4 Ablation Studies

We ablate each component on building function classification (Tab. 4), since inferring a building's function requires both its geometry and spatial relations to surrounding geoentities, and thus exercises every part of NARA. The topology-modulated distance bias improves Macro-F1 from 68.46 to 69.94, and GSRM yields the largest additional gain to 72.93; ACC and RSR further improve performance to 73.98 and 74.30. These results show that NARA's gains come from contextualizing each geoentity with the semantics and geometry of its neighbors through their spatial relations, rather than from richer geometry alone.

Table 4: NARA ablations on building function classification. Each row adds one component to the row above. T/D bias denotes the topology-modulated distance bias.

Configuration	Macro-F1 ↑	Weighted-F1 ↑	Accuracy ↑
MGSM	68.463 ± 2.135	90.617 ± 0.608	91.218 ± 0.456
+ T/D bias	69.943 ± 1.482	91.677 ± 0.360	92.189 ± 0.315
+ GSRM	72.934 ± 1.243	92.371 ± 0.421	92.793 ± 0.340
+ ACC	73.982 ± 1.625	92.757 ± 0.335	**93.142 ± 0.280**
+ RSR (**Full NARA**)	**74.298 ± 1.153**	**92.764 ± 0.306**	93.134 ± 0.296

5.5 Robustness Across Prediction Conditions

Performance on next POI prediction varies with spatial context and user history (Tab. 9 in Appdx. C.3.2). To assess whether NARA's gains depend on particular conditions, we partition test events by neighborhood density, sibling share, target popularity, history length (number of prior check-ins), and whether the target POI is new or revisited. NARA outperforms POI-Enhancer (Cheng et al., 2025) in nearly all slices across both cities and trajectory models, with statistically significant gains in most slices. Gains increase with neighborhood density: in Singapore, the Hit@1 gain rises from sparse to dense neighborhoods, reaching $+27.4\%$ for TALE. The only slices without improvement occur under SG-CTLE, and none of these differences is statistically significant.

6 Conclusion

We present NARA, a self-supervised framework for contextual representation learning over heterogeneous vector geoentities. NARA contextualizes geoentities through spatial-context-aware attention with a topology-modulated distance bias, and learns semantic–geometric representations through masked geoentity semantic modeling, geometry-aware spatial relation modeling, and relation-conditioned regularization grounded in spatial autocorrelation. With a single frozen, task-agnostic encoder, NARA outperforms task-specialized state-of-the-art methods on traffic speed estimation, building function classification, and next POI prediction, spanning all three geometry types. These results suggest that jointly modeling semantics, complete geometry, and spatial relations provides a strong foundation for task-agnostic contextual geoentity representation learning.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity and readability of the manuscript, including polishing transitions, refining the presentation and framing of tables, and assisting with polishing the implementation pipeline, including the presentation of arguments. Generative AI was not used to formulate scientific claims, develop theoretical arguments, design the overall methodology, or draw research conclusions. AI-assisted code was reviewed, adapted, and tested by the authors for correctness and consistency with the intended implementation. We have reviewed all AI-assisted work, including the manuscript text, presentation, and code, and verified that it accurately reflects the authors' intended content and implementation. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All data used in this study are publicly available and were not collected by the authors: OpenStreetMap geoentities, released under the Open Database License; Uber Movement traffic speeds, which are aggregated over trips and road segments and contain no individual trip records; the Singapore land use labels used in the CityFM benchmark (Balsebre et al., 2024); and the Foursquare check-in datasets of Yang et al. (2014), which are distributed for research use with anonymized user identifiers. No new human-subject data were collected, and we do not attempt to identify individuals or link check-in records to external sources.

REPRODUCIBILITY STATEMENT

We release an anonymous implementation, including data preprocessing, pretraining, and downstream evaluation code, at `https://anonymous.4open.science/r/nara-7881/README.md`. Method details are in §A, and implementation details, hyperparameters, data statistics, downstream protocols, and baselines are in §B. All datasets are publicly available, and all downstream results are reported as means and standard deviations over multiple seeds or folds.

REFERENCES

Martin Bachmaier and Matthias Backes. Variogram or semivariogram? variance or semivariance? allan variance or introducing a new term? *Mathematical Geosciences*, 43(6):735–740, 2011.

Pasquale Balsebre, Weiming Huang, Gao Cong, and Yi Li. City foundation models for learning general purpose representations from openstreetmap. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 87–97, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679662. URL `https://doi.org/10.1145/3627673.3679662`.

Yile Chen, Weiming Huang, Kaiqi Zhao, Yue Jiang, and Gao Cong. Self-supervised representation learning for geospatial objects: A survey. *Information Fusion*, 123:103265, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103265. URL `https://www.sciencedirect.com/science/article/pii/S1566253525003380`.

Jiawei Cheng, Jingyuan Wang, Yichuan Zhang, Jiahao Ji, Yuanshao Zhu, Zhibo Zhang, and Xiangyu Zhao. Poi-enhancer: An llm-based semantic enhancement framework for poi representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11509–11517, Apr. 2025. doi: 10.1609/aaai.v39i11.33252. URL `https://ojs.aaai.org/index.php/AAAI/article/view/33252`.

Chen Chu and Cyrus Shahabi. Geo2vec: Shape-and distance-aware neural representation of geospatial entities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18985–18993, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

Max Egenhofer. A mathematical framework for the definition of topological relations. In *Proc. the fourth international symposium on spatial data handing*, pp. 803–813, 1990.

Michael F Goodchild. Geographical data modeling. *Computers & geosciences*, 18(4):401–408, 1992.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 855–864, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939754. URL https://doi.org/10.1145/2939672.2939754.

Jilin Hu, Chenjuan Guo, Bin Yang, and Christian S. Jensen. Stochastic weight completion for road networks using graph convolutional networks. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1274–1285, 2019. doi: 10.1109/ICDE.2019.00116.

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.

Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3967–3974, 2019.

Tobias Skovgaard Jepsen, Christian S. Jensen, and Thomas Dyhre Nielsen. Relational fusion networks: Graph convolutional networks for road networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):418–429, 2022. doi: 10.1109/TITS.2020.3011799.

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.

Hao Li, Jiapan Wang, Johann Maximilian Zollner, Gengchen Mai, Ni Lao, and Martin Werner. Rethink geographical generalizability with unsupervised self-attention model ensemble: A case study of openstreetmap missing building detection in africa. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pp. 1–9, 2023a.

Yi Li, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. Urban region representation learning with openstreetmap building footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 1363–1373, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599538. URL https://doi.org/10.1145/3580305.3599538.

Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. SpaBERT: A pretrained language model from geographic data for geo-entity representation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2757–2769, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.200. URL https://aclanthology.org/2022.findings-emnlp.200/.

Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. GeoLM: Empowering language models for geospatially grounded language understanding. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5227–5240, Singapore, December 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.317. URL https://aclanthology.org/2023.emnlp-main.317/.

Yan Lin, Huaiyu Wan, Shengnan Guo, and Youfang Lin. Pre-training context and time aware location embeddings from spatial-temporal trajectories for user next location prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4241–4248, May 2021. doi: 10.1609/aaai.v35i5.16548. URL https://ojs.aaai.org/index.php/AAAI/article/view/16548.

Yijun Lin, Yao-Yi Chiang, Meredith Franklin, Sandrah P. Eckel, and José Luis Ambite. Building autocorrelation-aware representations for fine-scale spatiotemporal prediction. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 352–361, 2020. doi: 10.1109/ICDM50108.2020.00044.

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.

Paul A Longley, Michael F Goodchild, David J Maguire, and David W Rhind. *Geographic information science and systems*. John Wiley & Sons, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Gengchen Mai, Bo Yan, Krzysztof Janowicz, and Rui Zhu. Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. In *International conference on geographic information science*, pp. 21–39. Springer, 2019.

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 *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJljdh4KDH.

Gengchen Mai, Chiyu Jiang, Weiwei Sun, Rui Zhu, Yao Xuan, Ling Cai, Krzysztof Janowicz, Stefano Ermon, and Ni Lao. Towards general-purpose representation learning of polygonal geometries. *GeoInformatica*, 27(2):289–340, 2023.

Gengchen Mai, Xiaobai Yao, Yiqun Xie, Jinmeng Rao, Hao Li, Qing Zhu, Ziyuan Li, and Lao Ni. Srl: Towards a general-purpose framework for spatial representation learning. In *32nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2024.

OpenStreetMap contributors. Planet dump from openstreetmap. https://planet.osm.org, 2017. Accessed: 2026-03-26.

OpenStreetMap contributors, Overture Maps Foundation. Overture maps open data platform. https://overturemaps.org, 2026. Accessed: 2026-03-26.

David A Randell, Zhan Cui, Anthony G Cohn, et al. A spatial logic based on regions and connection. *KR*, 92(165-176):40–40, 1992.

Blake Regalia, Krzysztof Janowicz, and Grant McKenzie. Computing and querying strict, approximate, and metrically refined topological relations in linked geographic data. *Transactions in GIS*, 23(3):601–619, 2019.

Maria Despoina Siampou, Jialiang Li, John Krumm, Cyrus Shahabi, and Hua Lu. Poly2vec: Polymorphic fourier-based encoding of geospatial objects for geoAI applications. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kWyov6XrXs.

Nicolas Tempelmeier, Simon Gottschalk, and Elena Demidova. Geovectors: A linked open corpus of openstreetmap embeddings on world scale. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pp. 4604–4612, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482004. URL https://doi.org/10.1145/3459637.3482004.

Waldo Tobler. On the first law of geography: A reply. *Annals of the association of American geographers*, 94(2):304–310, 2004.

Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Huaiyu Wan, Yan Lin, Shengnan Guo, and Youfang Lin. Pre-training time-aware location embeddings from spatial-temporal trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 34 (11):5510–5523, 2022. doi: 10.1109/TKDE.2021.3057875.

Meng-xiang Wang, Wang-Chien Lee, Tao-yang Fu, and Ge Yu. Learning embeddings of intersections on road networks. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, pp. 309–318, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369091. doi: 10.1145/3347146.3359075. URL https://doi.org/10.1145/3347146.3359075.

Zhangyu Wang, Gengchen Mai, Krzysztof Janowicz, and Ni Lao. Mc-gta: Metric-constrained model-based clustering using goodness-of-fit tests with autocorrelations. In *41st International Conference on Machine Learning, ICML 2024*, 2024.

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*, pp. 1–10, 2017.

Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2014.

Yifan Yang, Jingyuan Wang, Xie Yu, and Yibang Tang. Hygmap: representing all types of map entities via heterogeneous hypergraph. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/1049. URL https://doi.org/10.24963/ijcai.2025/1049.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28877–28888. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f1c1592588411002af340cbaedd6fc33-Paper.pdf.

Liang Zhang and Cheng Long. Road network representation learning: A dual graph-based approach. *ACM Transactions on Knowledge Discovery from Data*, 17(9):1–25, 2023.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Haicang Zhou, Weiming Huang, Yile Chen, Tiantian He, Gao Cong, and Yew-Soon Ong. Road network representation learning with the third law of geography. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 11789–11813. Curran Associates, Inc., 2024. doi: 10.52202/079017-0376. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/15cc8e4a46565dab0c1a1220884bd503-Paper-Conference.pdf.

A METHOD DETAILS

A.1 TOPOLOGICAL RELATION DEFINITIONS

We adopt a symmetric topology schema $\mathcal{T} = \{\text{DISJOINT}, \text{CONTACT}, \text{CONTAINMENT}\}$, constructed from standard intersection and containment predicates derived from DE-9IM (Egenhofer, 1990). The relations are mutually exclusive under the precedence CONTAINMENT $>$ CONTACT $>$ DISJOINT:

- DISJOINT: $g_i \cap g_j = \varnothing$.
- CONTACT: the geometry intersect, but neither is within the other. This relation includes both boundary contact and interior intersection.
- CONTAINMENT: one geometry is within the other. We map WITHIN and CONTAINS to the same relation.

The labels are symmetric: (v_i, v_j) and (v_j, v_i) receive the same relation. This design represents shared spatial configuration rather than directional geometric roles. For example, a shop within a mall and the mall enclosing the shop describe the same spatial configuration from complementary perspectives. This design aligns with topology-modulated distance bias and geometry-aware spatial relation modeling.

A.2 DISTANCE NORMALIZATION AND RADIAL BASIS FEATURES

The proposed topology-modulated distance bias encodes the minimum distance between geometry with $K = 16$ Gaussian radial basis functions. We expand distance into overlapping Gaussian radial basis features so that a learned weighted sum can express smooth, nonlinear distance effects on attention, rather than restricting the bias to a linear function of distance. In NARA, distances are computed after the fixed affine normalization to $[-1, 1]$ applied per city. Let $\bar{d}_{ij}$ denote this normalized distance and let $D_{\max}$ be the largest normalized distance retained by the dense relation representation ($D_{\max} = 0.105$ for NYC and $D_{\max} = 0.144$ for Singapore). The centers and shared width are

$$\mu_k = \frac{kD_{\max}}{K-1}, \qquad \sigma = \frac{D_{\max}}{K-1}, \qquad k = 0, \ldots, K-1, \tag{14}$$

and the k-th feature is

$$\phi_k(\bar{d}_{ij}) = \exp\left[-\frac{1}{2}\left(\frac{\text{clip}(\bar{d}_{ij}, 0, D_{\max}) - \mu_k}{\sigma}\right)^2\right]. \tag{15}$$

Thus $\boldsymbol{\phi}(\bar{d}_{ij}) = [\phi_0(\bar{d}_{ij}), \ldots, \phi_{K-1}(\bar{d}_{ij})]$.

A.3 ANCHOR AND SIBLING-GROUP CONSTRUCTION

We construct three reference-relation spaces:

1. polyline anchors and polygon members within a 30-meter buffer (e.g., buildings near roads)
2. polyline anchors and point members within a 30-meter buffer (e.g., POIs near roads)
3. polygon anchors and point members within a 5-meter buffer to capture containment (e.g., POIs inside buildings)

Points and polygons serve as sibling members, whereas polylines serve only as anchors. Common vector datasets provide no general relation that groups multiple polylines under a shared reference entity without task-specific semantics (e.g., road segments sharing a transit route); we therefore do not treat polylines as members.

These thresholds are evaluated with `ST_DWithin` on complete geometry in EPSG:3857. A geoentity may belong to the sibling group, but anchor entities are not added to a context solely because they define a group. For the non-sibling baseline used by $\mathcal{L}_{\text{RSR}}$, we retain same-type point–point and polygon–polygon base-context pairs whose complete-geometry distance is at most 50 projected meters, excluding every pair that shares any sibling group.

A.4 RELATIONAL SEMIVARIOGRAM DETAILS

The relation-conditioned semivariogram regularizer compares sibling and non-sibling dispersion at similar distances. We formalize the connection between cosine distance and squared Euclidean

distance under ℓ_2 normalization, and show that enforcing lower semivariance translates to a margin-based similarity gap.

Proposition 1 (Equivalence under normalization)**.** *If* $\|\mathbf{h}_i^{\text{sem}}\|_2 = \|\mathbf{h}_j^{\text{sem}}\|_2 = 1$*, then*

$$\frac{1}{2}\left\|\mathbf{h}_i^{\text{sem}} - \mathbf{h}_j^{\text{sem}}\right\|_2^2 = 1 - \text{sim}\big(\mathbf{h}_i^{\text{sem}}, \mathbf{h}_j^{\text{sem}}\big)\,. \tag{16}$$

Proof. Expanding the squared distance gives

$$\left\|\mathbf{h}_i^{\text{sem}} - \mathbf{h}_j^{\text{sem}}\right\|_2^2 = \|\mathbf{h}_i^{\text{sem}}\|_2^2 + \left\|\mathbf{h}_j^{\text{sem}}\right\|_2^2 - 2(\mathbf{h}_i^{\text{sem}})^\top \mathbf{h}_j^{\text{sem}} \tag{17}$$

$$= 2\left[1 - \text{sim}\big(\mathbf{h}_i^{\text{sem}}, \mathbf{h}_j^{\text{sem}}\big)\right], \tag{18}$$

since $\|\mathbf{h}_i^{\text{sem}}\|_2 = \|\mathbf{h}_j^{\text{sem}}\|_2 = 1$. Dividing by two proves the result.

Let $\|\mathbf{h}_{i,\ell}^{\text{sem}} - \mathbf{h}_{j,\ell}^{\text{sem}}\|^2 = 2(1 - c_{ij})$. For a pair set Π in context ℓ, define

$$\hat{\gamma}(\Pi) = 1 - \bar{c}(\Pi), \qquad \bar{c}(\Pi) = \frac{1}{|\Pi|}\sum_{(v_i,v_j)\in\Pi} c_{ij}, \tag{19}$$

where $\bar{c}(\Pi)$ is the average cosine similarity over Π. By Proposition 1, this is the empirical semivariance of the normalized contextual semantic embeddings.

For a sibling group $\mathcal{S}_{a,r}$ and bin b, let $\Pi_{\text{rel}}^{(a,r)}(b)$ denote its sibling pairs in bin b and $\Pi_{\text{glob}}^{(\eta_{a,r})}(b)$ the corresponding type-matched non-sibling pairs in the same context, so that

$$\gamma_{\text{rel}}^{(a,r)}(b) = \hat{\gamma}\big(\Pi_{\text{rel}}^{(a,r)}(b)\big), \qquad \gamma_{\text{glob}}^{(\eta_{a,r})}(b) = \hat{\gamma}\big(\Pi_{\text{glob}}^{(\eta_{a,r})}(b)\big). \tag{20}$$

Proposition 2 (Dispersion ordering implies a similarity gap)**.** *Fix a sibling group* $\mathcal{S}_{a,r}$ *and a valid bin* b*. The corresponding hinge term in Eq.* (12) *is zero if and only if*

$$\gamma_{\text{rel}}^{(a,r)}(b) \le \gamma_{\text{glob}}^{(\eta_{a,r})}(b) - \delta, \tag{21}$$

or, equivalently,

$$\bar{c}\big(\Pi_{\text{rel}}^{(a,r)}(b)\big) \ge \bar{c}\big(\Pi_{\text{glob}}^{(\eta_{a,r})}(b)\big) + \delta. \tag{22}$$

Proof. The hinge term $\max\big(0,\, \gamma_{\text{rel}}^{(a,r)}(b) - \gamma_{\text{glob}}^{(\eta_{a,r})}(b) + \delta\big)$ is zero if and only if $\gamma_{\text{rel}}^{(a,r)}(b) - \gamma_{\text{glob}}^{(\eta_{a,r})}(b) + \delta \le 0$, which is Eq. (21). Substituting $\hat{\gamma}(\Pi) = 1 - \bar{c}(\Pi)$ from Eq. (19) into both sides gives

$$1 - \bar{c}\big(\Pi_{\text{rel}}^{(a,r)}(b)\big) \le 1 - \bar{c}\big(\Pi_{\text{glob}}^{(\eta_{a,r})}(b)\big) - \delta, \tag{23}$$

which rearranges to Eq. (22).

Interpretation. Proposition 2 shows that minimizing $\mathcal{L}_{\text{RSR}}$ enforces a margin-based similarity gap between sibling and non-sibling pairs at each distance bin: sibling pairs must be at least δ more similar (in cosine similarity) than type-matched non-sibling pairs at the same distance, ensuring that relational structure enhances semantic coherence beyond metric proximity alone.

NARA uses five equal-width bins over $[0, 50]$ projected meters of geometry-to-geometry distance, matching the range of the non-sibling pairs (Appdx. A.3). A bin is valid for $\mathcal{S}_{a,r}$ if both pair sets contain more than five pairs; only valid bins contribute to Eq. (12). For valid bins, the weights are

$$\pi_b^{(a,r)} = \frac{\big|\Pi_{\text{rel}}^{(a,r)}(b)\big|}{\sum_{b'}\big|\Pi_{\text{rel}}^{(a,r)}(b')\big|}, \tag{24}$$

where b' ranges over the valid bins of $\mathcal{S}_{a,r}$, so that the weights of each group sum to one.

B EXPERIMENTAL SETUP

B.1 NARA IMPLEMENTATION DETAILS

We encode semantic attributes with a frozen `bert-base-uncased` model using mean pooling, producing a normalized 768-dimensional embedding. Complete point, polyline, and polygon geometry are converted to 210-dimensional complex Fourier features and passed to Poly2Vec (Siampou et al., 2025), whose hidden and output dimensions are 256 and 128. The semantic and geometric representations are projected to $d_{\text{model}} = 128$ and combined by the adaptive gate in Eq. (5); the final gate bias is initialized to -2.0.

The contextual encoder has three Transformer layers, four attention heads, a feed-forward dimension of 256, and dropout 0.1. Queries and keys are shared and are computed from the fused embeddings. The semantic and fused embeddings use separate value and output projections, residual connections, layer normalization, and feed-forward networks. The topology-modulated distance bias uses the 16 radial basis features defined in Appdx. A.2; its topology intercepts and distance weights are initialized to zero. We set the masking ratio to 40%.

We pretrain NARA for 100 epochs with AdamW (Loshchilov & Hutter, 2019), learning rate 2×10^{-4}, and weight decay 0.01. Training uses seed 42 and an optimizer batch of 256 contexts. The finalized loss is

$$\mathcal{L} = \mathcal{L}_{\text{MGSM}} + 0.5\mathcal{L}_{\text{topo}} + 200\mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{ACC}} + 25\mathcal{L}_{\text{RSR}}. \tag{25}$$

We set $\tau_{\text{MGSM}} = 0.15$, $\tau_{\text{ACC}} = 0.4$, and use the ACC weight $\exp(-d_{ij}/20)$, where d_{ij} is the complete-geometry distance in projected meters. For RSR, we use margin $\delta = 0.1$, five 10-meter intervals over $[0, 50]$ meters. Their different magnitudes compensate for differences in the numerical scales of the individual objectives and should not be interpreted as relative semantic importance. We share all pretraining hyperparameters across NYC and Singapore. Pretraining NARA requires approximately 7 hours for NYC and 4 hours for Singapore on a single NVIDIA RTX A5000 GPU.

B.2 PRETRAINING DATASET STATISTICS

Polylines are noded at intersections before pretraining, and each noded entity remains a separate model token and reference entity. The finalized 250-meter construction contains 87,135 spatial contexts for NYC and 44,311 for Singapore.

Table 5: Pretraining geoentities and context statistics after preprocessing and one-hop reference-relation expansion. Context means are computed over nonempty contexts.

	NYC	Singapore
Point geoentities	43,549	23,621
Polygon geoentities	49,486	40,594
Polyline geoentities	186,667	181,592
Total # of contexts	87,135	44,311
Mean entities per context	25.213	38.133
Mean mask-eligible entities	3.210	5.547
Maximum context size	573	498

B.3 DOWNSTREAM TASK PROTOCOLS

No downstream labels are used during NARA pretraining. Target road geoentities for traffic-speed estimation may appear in the unlabeled OSM pretraining data, but their Uber Movement traffic speeds are never used during pretraining. Untagged target buildings are excluded from pretraining following the CityFM protocol (Balsebre et al., 2024), and target Foursquare POIs are absent from the OSM pretraining data. Building function classification and next POI prediction therefore additionally evaluate geoentities unseen during pretraining.

Task A: Traffic Speed Estimation. We use the New York City traffic-speed benchmark introduced by CityFM (Balsebre et al., 2024). The targets are Uber Movement speed measurements, in miles per hour (mph), mapped to OSM parent-road segments. Roads with fewer than 10 observations are excluded, leaving 29,755 road segments (mean speed 19.50 mph, standard deviation 8.69 mph, and range 1.46–57.37 mph). Following the CityFM protocol (Balsebre et al., 2024), we use one partition of 17,853 training, 5,951 validation, and 5,951 test roads, corresponding to a 60/20/20 split. For each

seed, model selection uses only the validation set, and the selected checkpoint is evaluated once on the test set. Evaluation metrics include root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and mean absolute percentage error (MAPE), reported as mean and standard deviation over 10 seeds (42–51).

Task B: Building Function Classification. We use the Singapore building-function benchmark introduced by CityFM (Balsebre et al., 2024), consisting of 64,384 OSM building polygons labeled according to an eight-class land-use taxonomy. Following the CityFM protocol (Balsebre et al., 2024), we use the 10 folds, each containing 38,628 training, 12,876 validation, and 12,880 test buildings, corresponding approximately to a 60/20/20 split. Within each fold, the checkpoint with the highest validation Macro-F1 is selected and evaluated once on the held-out test set. Evaluation metrics include Accuracy, Macro-F1, and Weighted-F1, reported as mean and standard deviation over 10 folds. In the released CityFM evaluation code, Weighted-F1 is computed as `f1_score(y_pred, y_true, average='weighted')`, passing arguments in the order `(y_pred, y_true)`. Because `sklearn`'s weighted averaging draws class support from the first argument, this weights per-class F1 scores by predicted class frequencies rather than ground-truth class frequencies. We recompute Weighted-F1 identically for every evaluated method, CityFM included, as `f1_score(y_true, y_pred, labels=range(8), average='weighted', zero_division=0)`. This choice affects only Weighted-F1: Macro-F1 is unchanged, since each per-class F1 and its unweighted average are invariant to exchanging precision and recall, and Accuracy is order-independent. Published CityFM Weighted-F1 values therefore reflect a different weighting and are not directly comparable to those reported here.

Task C: Next POI Prediction. We use the Foursquare check-in datasets from Yang et al. (2014) following the task formulation of POI-Enhancer (Cheng et al., 2025). Foursquare-NY contains 15,171 users, 24,118 POIs, and 641,005 check-ins, and Foursquare-SG contains 10,909 users, 20,154 POIs, and 696,306 check-ins. Our protocol uses fixed 80/10/10 train/validation/test trajectory partitions with split seed 42, producing 91,352/11,419/11,420 trajectories for New York and 110,810/13,851/13,852 trajectories for Singapore. For each configuration, we run downstream training with seeds 42, 43, and 44 for at most 30 epochs. Validation Hit@1 is measured after every epoch, and we select, independently per seed, the earliest epoch that achieves the maximum validation Hit@1. The test set is evaluated once using that checkpoint. Hit@1 and Hit@5 are taken from the same selected epoch and reported as mean and standard deviation over the three seeds (42–44).

B.4 Baseline Details

Task A: Traffic Speed Prediction Baselines.

- *Topology (Road Network):*
  - **Node2Vec** (Grover & Leskovec, 2016): learns node representations from biased random walks over the road-network graph by encouraging neighboring nodes in the walk context to have similar embeddings.
  - **GCWC** (Hu et al., 2019): a graph convolutional method that captures edge correlations and local spatial dependencies in road networks.
  - **RFN** (Jepsen et al., 2022): a graph neural network that models relational dependencies between connected road segments.
- *Geometry:*
  - **Poly2Vec** (Siampou et al., 2025): a geometry encoder that represents coordinate sequences in the spectral domain using Fourier transforms.
  - **Geo2Vec** (Chu & Shahabi, 2026): a geometry encoder based on Signed Distance Fields (SDFs) that captures fine-grained spatial boundary information.
- *Semantics + Metric:*
  - **IRN2Vec** (Wang et al., 2019): learns road-intersection embeddings from random-walk sequences generated over adjacent intersections.
  - **GeoVectors** (Tempelmeier et al., 2021): learns embeddings of OSM entities by jointly modeling geographic coordinates and semantic tags.
- *Semantics + Geometry:*
  - **CityFM** (Balsebre et al., 2024): combines hand-crafted polyline features, positional encodings, and learned semantic representations of surrounding point and polygon geoentities.

Task B: Building Function Classification Baselines.

- *Non-Spatial Semantics:*
  - **BERT (fine-tuned)** (Devlin et al., 2019): a pretrained language model used as the textual encoder backbone.
- *Geometry:*
  - **Poly2Vec** (Siampou et al., 2025), **Geo2Vec** (Chu & Shahabi, 2026): geometry encoders that map spatial coordinates into continuous latent representations.
- *Semantics + Metric:*
  - **GeoVectors** (Tempelmeier et al., 2021): encodes buildings using geographic coordinates and semantic information from neighboring entities.
  - **SpaBERT** (Li et al., 2022): encodes surrounding entity names together with relative distance information from the target building.
- *Semantics + Geometry:*
  - **CityFM** (Balsebre et al., 2024): combines polygon geometry features, positional encodings, polygon area, and learned semantic representations of nearby geographic entities.

Task C: Next POI Prediction Baselines.

- **TALE** (Wan et al., 2022): learns POI embeddings from visit co-occurrence patterns while modeling temporal periodicity through hierarchical time discretization.
- **CTLE** (Lin et al., 2021): a self-supervised Transformer that learns contextual POI representations by reconstructing masked location and temporal attributes from visit sequences.
- **POI-Enhancer** (Cheng et al., 2025): enhances POI representations using auxiliary textual knowledge generated by large language models (e.g., Llama-2-7B (Touvron et al., 2023)). It incorporates information such as addresses, nearby POIs, and check-in statistics, and aligns text-enhanced representations with pretrained POI embeddings (e.g., TALE or CTLE).

C Additional Results

C.1 Polyline analyses: performance by road type

Tab. 6 shows per-type MAE on the NYC test set. NARA reduces MAE relative to CityFM in 13 of 16 road types, including all four most frequent classes, namely residential (5.4%), primary (6.4%), secondary (9.9%), and tertiary (8.8%) roads. These four classes contain 83% of test roads and account for 67% of the total reduction in absolute error. Gains are also large on high-speed classes such as trunk roads (15.9%), motorway links (13.8%), and motorways (11.9%); motorway links alone contribute 16% of the total reduction. The three types where NARA underperforms (construction, living street, and service) together contain only 20 test roads (0.3%). Without CityFM's engineered road features, NARA † still outperforms CityFM on all four most frequent classes, and the engineered features mainly help on high-speed classes such as motorway links, trunk roads, and primary links.

Table 6: Per-type MAE for traffic speed prediction on the NYC test set, using representative downstream runs. NARA † excludes CityFM's engineered road features. ΔMAE = MAE(CityFM) $-$ MAE(NARA); positive values denote lower error than CityFM.

Road type	Count	CityFM	NARA†	NARA	ΔMAE
Pedestrian	1	2.605	0.852	0.030	+2.575
Tertiary Link	12	3.946	1.876	2.606	+1.340
Trunk Link	30	4.613	4.162	3.496	+1.117
Secondary Link	37	3.897	3.234	3.123	+0.775
Motorway Link	377	4.942	5.007	4.262	+0.680
Trunk	71	3.959	3.987	3.331	+0.628
Motorway	279	3.768	3.642	3.320	+0.449
Secondary	1,100	3.182	2.922	2.868	+0.315
Primary Link	59	3.907	4.068	3.610	+0.297
Tertiary	602	3.202	2.941	2.921	+0.281
Unclassified	89	3.694	3.484	3.474	+0.220
Primary	1,114	3.117	3.025	2.918	+0.199
Residential	2,160	2.767	2.651	2.618	+0.150
Construction	4	2.216	3.360	2.689	-0.473
Living Street	4	3.343	2.893	3.907	-0.564
Service	12	3.629	3.477	4.711	-1.082

C.2 Polygon Analyses: performance by land use label

Tab. 7 evaluates performance across land-use categories. NARA outperforms CityFM in 5 of 8 classes, with particularly large gains for Civic & Community Institutions (+27.49 F1), Commercial & Residential (+20.61 F1), and Educational buildings (+16.74 F1), suggesting that relational spatial context is especially valuable for less frequent or functionally ambiguous building types. Gains are not uniform: NARA underperforms on Transport (-18.39) and Sports & Recreation (-8.05), low-support classes whose footprints limit reliable contextualization.

Table 7: Per-class F1 score for building function classification. Counts denote the number of test instances per class.

Category	Count	CityFM	NARA	ΔF1
Residential	8,645	96.03	96.77	+0.74
Industrial	2,087	94.65	92.40	-2.25
Commercial	1,038	87.07	90.23	+3.16
Commercial & Residential	329	46.29	66.90	+20.61
Educational	286	73.00	89.74	+16.74
Civic & Community Inst.	241	24.92	52.41	+27.49
Sports & Recreation	151	69.76	61.71	-8.05
Transport	103	61.53	43.14	-18.39

C.3 Point analyses

C.3.1 Performance by check-in POI type

We evaluate next POI prediction across the 15 most frequent venue types in the Foursquare-New York test set (Tab. 8). NARA improves Hit@1 in 13 categories and ties in the remaining two, while improving Hit@5 in all 15. The largest Hit@1 gains occur for Deli / Bodega (+3.01 points), Home (private) (+1.80), and transit- or neighborhood-linked venues such as Subway (+1.62), Park (+1.61), and Neighborhood (+1.47), consistent with NARA benefiting from the surrounding OSM spatial context.

Table 8: Per-category Hit@1 and Hit@5 for next POI prediction (NY-CTLE). Counts denote test prediction events.

Foursquare POI Type	Count	POI-Enhancer		NARA	
		Hit@1	Hit@5	Hit@1	Hit@5
Subway	927	16.757	34.808	18.375	38.799
Bar	891	1.534	3.330	1.534	4.003
Home (private)	816	17.198	32.598	18.995	38.766
Office	689	20.077	32.898	20.174	35.897
Coffee Shop	649	4.212	8.783	4.417	10.324
Park	559	6.142	12.403	7.752	13.715
Neighborhood	477	6.499	16.492	7.966	20.266
Train Station	442	13.047	25.792	14.178	30.769
Hotel	422	6.635	11.848	7.583	15.324
American Restaurant	400	2.417	6.833	2.750	8.333
Grocery Store	378	4.145	12.169	4.586	15.168
Deli / Bodega	355	13.333	31.455	16.338	34.178
Building	336	12.302	21.627	13.492	27.480
Italian Restaurant	309	2.050	3.991	2.050	5.394
Plaza	296	1.014	6.306	1.914	8.108

C.3.2 Robustness Across Prediction Conditions

Performance on next POI prediction varies widely across conditions for every model. For example, baseline Hit@5 for NY-CTLE ranges from 6.9% to 30.1% across the slices below, so the overall gain does not show whether a method helps in all conditions. We therefore split test events by one factor at a time (Tab. 9). The five factors are the number of geoentities in the target's context (neighborhood density), the fraction of these geoentities that share an anchor relation with the target (sibling share), how often the target POI is visited (popularity), the number of prior check-ins (1, 2–3, or 4+), and whether the target POI is new or revisited. Slice boundaries are fixed per city before evaluation, and

every slice has at least 1.6k events. Results are averaged over three downstream seeds, and 95% confidence intervals are computed by resampling test trajectories.

Across the 14 slices and the four combinations of city (NYC/SG) and trajectory model (CTLE (Lin et al., 2021)/TALE (Wan et al., 2022)) settings, NARA improves over POI-Enhancer in 54 of 56 comparisons on Hit@1 and 55 of 56 on Hit@5, and the gain is statistically significant in 49 and 51 of them, respectively. Gains grow with neighborhood density in Singapore, where the Hit@1 change rises from a 0.1% decrease to a 9.7% gain for CTLE and from 10.1% to 27.4% for TALE. Gains are similar across sibling-share levels, from 10.7% to 16.8% in Hit@1 for NY-CTLE, so NARA does not need targets with many siblings. Gains also hold for targets with only one prior check-in in every setting and for rare targets in three of four settings, and popular targets gain up to 24.5% (SG-TALE). All three comparisons without improvement are under SG-CTLE, the setting with the smallest overall gain (3.3%). These are sparse neighborhoods on Hit@1 and rare targets on Hit@1 and Hit@5, and none of these decreases is statistically significant.

Table 9: Relative improvement (%) in Hit@1 and Hit@5 of NARA over POI-Enhancer (Cheng et al., 2025) on next POI prediction, averaged over three seeds. Each row is a group of test events. For neighborhood density, sibling share, and target popularity, events are divided into three equal-sized groups (low, mid, high). History length is the number of prior check-ins. Group boundaries are fixed per city before evaluation. Bold values are statistically significant (95% bootstrap CI excludes zero). †Sibling-share groups are formed within each density group, so that sibling share is not confounded with density.

			NY-CTLE		NY-TALE		SG-CTLE		SG-TALE	
Factor	Slice	n (NY/SG)	H@1	H@5	H@1	H@5	H@1	H@5	H@1	H@5
All events		21.9k/31.9k	**+14.1**	**+17.4**	**+5.8**	**+7.5**	**+3.3**	**+3.3**	**+13.6**	**+15.2**
Density	Sparse	7.5k/10.8k	**+13.8**	**+17.5**	**+5.3**	**+9.1**	−0.1	+0.8	**+10.1**	**+12.8**
	Medium	7.1k/10.5k	**+12.2**	**+17.8**	**+9.3**	**+6.7**	**+5.3**	**+3.9**	**+11.2**	**+13.0**
	Dense	7.3k/10.6k	**+17.5**	**+16.6**	+2.5	**+4.9**	**+9.7**	**+7.0**	**+27.4**	**+22.6**
Sibling share†	Low	7.4k/10.7k	**+16.8**	**+16.4**	**+5.5**	**+7.9**	**+3.0**	**+3.0**	**+15.2**	**+14.8**
	Mid	7.2k/10.7k	**+10.7**	**+17.4**	**+7.4**	**+7.3**	**+4.0**	**+5.5**	**+10.9**	**+14.8**
	High	7.3k/10.5k	**+14.6**	**+18.7**	**+4.5**	**+7.2**	**+3.1**	+1.5	**+14.6**	**+15.9**
Prefix length	1	11.4k/13.9k	**+10.6**	**+15.3**	**+3.8**	**+5.7**	**+3.9**	+1.0	**+12.5**	**+13.5**
	2–3	6.4k/10.2k	**+21.0**	**+24.3**	**+9.8**	**+6.9**	**+3.4**	**+4.9**	**+15.4**	**+17.4**
	4+	4.1k/7.8k	**+12.9**	**+14.2**	**+5.0**	**+10.8**	**+2.9**	**+4.0**	**+13.2**	**+14.9**
Popularity	Rare	7.5k/11.9k	**+16.0**	**+22.1**	+0.6	**+12.5**	−1.6	−0.1	**+11.3**	**+15.3**
	Mid	7.8k/9.4k	**+12.2**	**+19.5**	**+6.4**	**+7.2**	+1.4	+1.2	**+6.4**	**+12.4**
	Popular	6.6k/10.6k	**+14.9**	**+14.1**	**+7.5**	**+5.9**	**+9.9**	**+7.3**	**+24.5**	**+17.6**
Novelty	New POI	20.3k/28.6k	**+14.2**	**+17.7**	**+6.4**	**+8.1**	**+3.7**	**+3.5**	**+13.8**	**+15.5**
	Revisit	1.6k/3.3k	**+13.6**	**+15.7**	+2.6	**+4.3**	+2.0	**+2.1**	**+13.1**	**+14.0**

C.4 Additional Ablations

We compare variants of the attention bias that differ in how they combine distance and topology, keeping all other components of NARA fixed (Tab. 10). Topology is more informative than distance, as topology alone raises Macro-F1 from 70.92 to 72.73, whereas distance alone brings no gain. Adding distance and topology independently, with distance weights shared across topologies, raises Macro-F1 to 73.68. Modulating the distance effect by topology, as in NARA, performs best on all three metrics, consistent with the principle in §3 that the same distance carries different contextual meaning under different topological relations.

Table 10: Variants of the attention bias on building function classification (Singapore), differing in how they combine distance and topology. All other components of NARA are fixed. **Bold** = best.

Pairwise attention bias	Macro-F1 ↑	Weighted-F1 ↑	Accuracy ↑
No pairwise bias	70.916 ± 1.373	91.449 ± 0.296	91.970 ± 0.261
Distance only	70.803 ± 2.416	91.750 ± 0.569	92.308 ± 0.455
Topology only	72.730 ± 1.408	92.379 ± 0.340	92.754 ± 0.296
Independent distance + topology	73.676 ± 1.370	92.591 ± 0.355	92.982 ± 0.314
Topology-modulated distance (**NARA**)	**74.298 ± 1.153**	**92.764 ± 0.306**	**93.134 ± 0.296**