HIERARCHICAL URBAN REPRESENTATION PRETRAINING FROM LARGE-SCALE GEOSPATIAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies *how to encode cities with a Transformer for urban representation pretraining on large-scale geospatial data*. Geospatial entities form the urban fabric of blocks, which in turn compose larger regions that collectively constitute cities. This ***entity–block–region–city*** hierarchy provides a natural basis for urban representation learning. Scaling it across thousands of cities requires modeling local functional interactions within blocks and spatial organization within and across regions. To this end, we propose HESTIA, a hierarchical framework for urban representation pretraining. Specifically, an Urban Fabric Tokenizer converts heterogeneous geospatial data into entity-category tokens arranged in a unified tensor reflecting the native city hierarchy. The Urban Interaction Transformer (UiT) applies attention in three nested stages, allowing urban semantics to emerge progressively through the hierarchy: *first from functional interactions among entity categories within blocks, then from geometry-aware interactions among blocks within regions, and finally from city-scale contextualization across regions*. Pretrained on a large-scale U.S. geospatial dataset of 8.18 million blocks and over 263 million entities, HESTIA achieves state-of-the-art performance on the multilevel benchmark, outperforming strong baselines including DeepMind's AlphaEarth, validating its hierarchical pretraining across spatial scales. Code is available **here**.

1 INTRODUCTION

Urban representation pretraining aims to model the functional layouts and spatial structures of cities. Cities are partitioned into regions (e.g., census tracts), which contain blocks that host point entities (POIs), such as schools and restaurants, and areal entities (AOIs), such as parks and water bodies. We refer to these POIs and AOIs collectively as geospatial entities. Entities within blocks form local *urban fabric*, while their containment relations define an ***entity–block–region–city*** spatial hierarchy, providing a natural basis for modeling local functions and multiscale spatial organization.

Existing methods typically use either entities or regions as primary modeling units within the hierarchy. Region-centric methods take regions as their primary units, using either irregular polygons (Li et al., 2024; Lin et al., 2026) or regular cells (Sun et al., 2025; Wen et al., 2026), and aggregate POI, mobility, or check-in data into region-level attributes to model relations among regions. Entity-centric methods instead take geospatial entities as their primary units and model their relations using graph structures (Zhang et al., 2026b) or local contexts (Balsebre et al., 2024). Meanwhile, broad-coverage geospatial foundation models such as AlphaEarth (Brown et al., 2025) and TESSERA (Feng et al., 2026a;b) demonstrate the feasibility of large-scale pretraining, but primarily learn location or environmental representations from coordinates, satellite imagery, and other Earth-observation signals. Across these lines of work, Transformers already appear, but attention stays within a single level of this hierarchy, so cross-level interactions remain hand-designed rather than learned automatically.

Transformers are now the de facto architecture for language and multimodal foundation models, owing in part to modality-matched token organizations: text forms ordered token sequences, whereas images are divided into regular patch grids—an organization that urban data lack. Cities, in contrast, contain varying numbers of regions and blocks with different geometries, as well as different types, counts, and spatial arrangements of geospatial entities within blocks. Therefore, Transformers cannot directly batch such irregular urban data into fixed-shape tensors. Regular grids can regularize inputs, but may overlook urban morphology and fragment spatially coherent environments. Encoding

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.

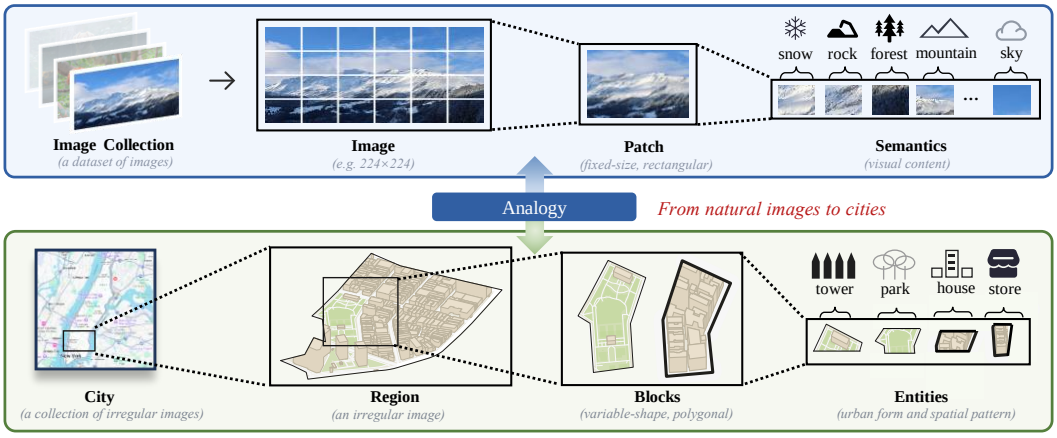


Figure 1: **Image–city analogy.**

every entity is computationally expensive, while coarse regional summaries can obscure fine-grained functions within blocks. Moreover, entity, block, and region interactions have different semantics, so flattening them into a single attention sequence would obscure the hierarchy. These observations raise two challenges: *(Q1) How can heterogeneous and irregular urban data be organized into fixed-shape Transformer inputs while preserving their native hierarchy? (Q2) How can a Transformer model hierarchical interactions along the entity, block, and region axes while retaining local entity composition and supporting large-scale pretraining across thousands of cities?*

Inspired by Vision Transformers (Dosovitskiy et al., 2021), we draw an image analogy (Figure 1): a city is a collection of irregular images, with each region as an image, each block as a patch, and block fabric as semantic content. To implement this view, we propose HESTIA, a hierarchical framework for urban representation pretraining on large-scale geospatial data. Specifically, the Urban Fabric Tokenizer converts heterogeneous geospatial entities into entity-category tokens and arranges them in a fixed-shape tensor that preserves the native city hierarchy. The Urban Interaction Transformer (UiT) then applies attention successively along the entity, block, and region axes within a single model, allowing urban semantics to emerge progressively through the hierarchy: *first from functional interactions among entity categories within blocks, then from geometry-aware interactions among blocks within regions, and finally from city-scale contextualization across regions*. HESTIA is trained to reconstruct masked local functions and regional fabric from the remaining visible observations.

Our contributions are threefold. (i) **Tokenizing Cities.** We introduce an Urban Fabric Tokenizer that converts complex, heterogeneous urban data into a Transformer-compatible format, enabling batched computation while preserving the native city hierarchy. (ii) **Modeling Cities Like Images.** We introduce the Urban Interaction Transformer (UiT), a simple yet effective hierarchical architecture that enables batched modeling of cities' functional layouts and spatial structures, providing a backbone for scalable urban representation pretraining and, in turn, for urban foundation models. (iii) **Large-scale Dataset and Multilevel Benchmark.** We combine Overture Maps with U.S. Census geographic boundaries to construct a dataset of over 263 million entities across 8.18 million blocks and evaluate representations at entity, block, and region levels. On the benchmark, HESTIA achieves state-of-the-art performance (Figure 2), outperforming strong baselines including DeepMind's AlphaEarth and Google Research's S2Vec.

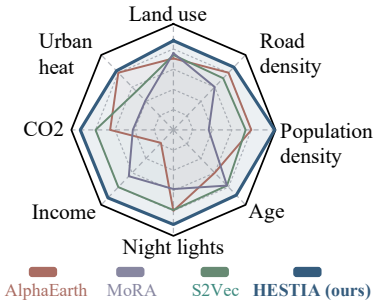


Figure 2: **Performance comparison.**

2 Related Work

One line of work focuses on urban representation methods that use either entities or regions as their primary modeling units. **Region-centric** methods aggregate POI, mobility, check-in, or demographic attributes within semantically bounded polygons (e.g., census tracts), and model inter-region relations with multi-view, graph, or attention architectures (Zhang et al., 2020; Wu et al., 2022; Li et al., 2024;

Lin et al., 2026). Regular grids (e.g., square or hexagonal) instead simplify batching and spatial pretraining (Sun et al., 2025; Wen et al., 2026; Choudhury et al., 2026). Aggregation mixes distinct local functions and obscures interactions among entity categories, while imposed grid boundaries misalign with urban morphology and introduce resolution dependence. **Entity-centric** methods instead model individual objects, learning from POI transitions, spatial co-occurrence, or categorical context (Feng et al., 2017; Yan et al., 2017; Huang et al., 2022), propagating information over entity and spatial graphs (Huang et al., 2023; Zhang et al., 2026b), or enriching entities with text, geometry, and imagery (Balsebre et al., 2024; Jin et al., 2025). They retain local functional detail, but their cost grows with the entity count and none provides a fixed-shape computation spanning the full hierarchy.

Another line of work studies **broad-coverage geospatial foundation models**, which scale self-supervised learning across locations, encoding coordinates or spatial functions, aligning locations with imagery or language, and learning from satellite and other global environmental signals (Mai et al., 2020; Vivanco Cepeda et al., 2023; Klemmer et al., 2025; Brown et al., 2025). They establish the feasibility of large-scale geospatial pretraining, but their inputs describe locations, environments, or mobility patterns rather than the nested composition of structured urban entities.

Overall, existing works largely focus on individual levels of the entity–block–region–city hierarchy, leaving large-scale urban representation pretraining across the full hierarchy unexplored. They lack both *a fine-to-coarse formalization of cities as inputs* and *a Transformer architecture suited to urban data*, which motivates us to design HESTIA. Additional related work can be found in Appendix S1.

3 METHOD

Figure 3 summarizes HESTIA. The Urban Fabric Tokenizer first converts heterogeneous geospatial data into hierarchy-preserving entity-category inputs (§3.1). UiT then applies attention along the entity, block, and region axes to build block, region, and city contexts (§3.2). Finally, two objectives reconstruct local functions and regional fabric from the remaining visible observations (§3.3).

3.1 TOKENIZING CITIES: AN URBAN FABRIC TOKENIZER

Cities expose an **entity–block–region–city** hierarchy, where the entities inside a block form its local *urban fabric*. This hierarchy is irregular in the numbers of regions and blocks, block geometries and layouts, entity-category compositions, and entity geometries. Such irregular data cannot be directly batched into fixed-size tensors, motivating the tokenization and input construction below.

➤ **Urban Fabric Tokenization.** All entity categories form a vocabulary $\mathcal{V}$ (Tables S21–S23). Within each block, the Urban Fabric Tokenizer aggregates same-category entities into a single entity-category token carrying an *aggregate value*. The resulting sequence length depends on the number of categories rather than the number of raw entities. However, aggregate value alone does not determine a category's local significance. First, *spatial support* specifies the physical extent over which the value is aggregated: the same count may be dense in a small block but sparse in a large one. Second, *category semantics* determine how the same value is interpreted across categories: two libraries may constitute a substantial presence, whereas two restaurants may constitute only a modest one. The tokenizer therefore learns how aggregate value and spatial support interact with category identity. For block b, let $\mathcal{V}_b \subseteq \mathcal{V}$ be its complete category set. For each $c \in \mathcal{V}_b$, v_{bc} and s_{bc} denote its aggregate value and spatial support: entity count and block area for POIs, category area share and total AOI coverage for AOIs. Each category $c \in \mathcal{V}$ has a learned embedding $\boldsymbol{x}_c$. Projections ϕ_v and ϕ_s and an attribute-fusion feed-forward network (FFN) f_{tok} construct its entity-category token embedding

$$\boldsymbol{x}_{bc} = \text{LN}(\boldsymbol{x}_c + f_{\text{tok}}([\boldsymbol{x}_c; \phi_v(v_{bc}); \phi_s(s_{bc})])) \,. \tag{1}$$

Here, LN denotes layer normalization and $[\,;\,]$ concatenation. The residual $\boldsymbol{x}_c$ preserves category identity, while f_{tok} conditions on $\boldsymbol{x}_c$ so that the same value and support are interpreted differently across categories. $\{\boldsymbol{x}_{bc} \mid c \in \mathcal{V}_b\}$ initialize block b's entity-category input slots, defined below.

➤ **Packing Tokens into a Hierarchical Tensor.** We then organize the block–region–city hierarchy into a batched input format for UiT. A batch contains C cities indexed by i, and $\mathcal{R}$ denotes all regions in the batch. Following Tobler's first law of geography (Tobler, 1970), the input set $\mathcal{B}_r$ of each region $r \in \mathcal{R}$ contains its internal and neighboring blocks outside the region. During pretraining, $\mathcal{B}_r$ splits into visible and masked blocks, $\mathcal{B}_r = \mathcal{B}_r^{\text{vis}} \cup \mathcal{B}_r^{\text{mask}}$, and $\mathcal{B}^{\text{vis}}$ collects the visible blocks of all regions

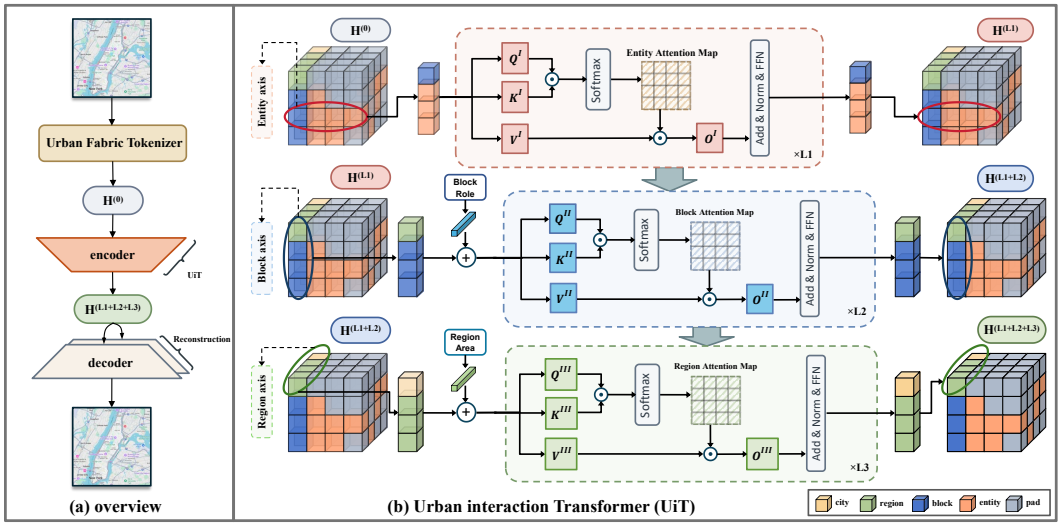


Figure 3: **Overview of HESTIA.** Colors distinguish token types, as listed in the legend at the bottom right.

in the batch, the set on which entity attention operates. Let R, B, and T denote the padded lengths of the region, block, and within-block token axes, including their leading special-token slots, and D the hidden dimension. Each batch forms $\boldsymbol{H}^{(0)} \in \mathbb{R}^{C\times R\times B\times T\times D}$. Its leading slots encode the hierarchy:

$$\boldsymbol{H}^{(0)}_{i,0,0,0,:} \leftarrow \texttt{[city]}, \quad \boldsymbol{H}^{(0)}_{i,r,0,0,:} \leftarrow \texttt{[reg]}\ (r>0), \quad \boldsymbol{H}^{(0)}_{i,r,b,0,:} \leftarrow \texttt{[blk]}\ (r>0,\ b>0).$$

The `[city]`, `[reg]`, and `[blk]` tokens use learnable embeddings. For a visible block b, its $t>0$ slots are initialized by the tokenizer outputs $\{\boldsymbol{x}_{bc} \mid c \in \mathcal{V}_b\}$, with masked categories replaced by `[mask_ent]`. For a masked block, `[mask_blk]` occupies the leading slot. Its remaining slots and all other unused positions contain `[pad]`. $\boldsymbol{H}^{(0)}$ is drawn slot by slot for one batch in Table S4. At inference, no block or entity category is masked. §3.3 details the masking strategy.

3.2 Modeling Cities Like Images: An Urban Interaction Transformer (UiT)

UiT realizes the image analogy of §1: regions are images, blocks are patches, and block fabric is their semantic content. Given $\boldsymbol{H}^{(0)}$, UiT keeps its five-dimensional shape throughout the forward pass while applying entity-, block-, and region-level Transformer layers, with L_1, L_2, and L_3 layers respectively, along the corresponding axes. Each stage updates only its active slice of $\boldsymbol{H}$ and leaves the remaining slots unchanged, so attention acts on $T\times T$, $B\times B$, and $R\times R$ matrices instead of one flattened sequence: the per-city cost drops from $O(R^2B^2T^2)$ to $O(RBT^2+RB^2+R^2)$, which keeps pretraining across thousands of cities tractable. The equations omit layer indices and the standard Transformer sublayers. Below we trace one forward pass and the slots each stage updates:

$$\underset{\text{hierarchical input}}{\boldsymbol{H}^{(0)}} \xrightarrow{\textit{entity axis}} \underset{\text{updated [blk] states}}{\boldsymbol{H}^{(L_1)}} \xrightarrow{\textit{block axis}} \underset{\text{updated [blk] and [reg] states}}{\boldsymbol{H}^{(L_1+L_2)}} \xrightarrow{\textit{region axis}} \underset{\text{updated [city] and [reg] states}}{\boldsymbol{H}^{(L_1+L_2+L_3)}}$$

$$\underbrace{\qquad\qquad}_{\text{(I) Entity attention}} \quad \underbrace{\qquad\qquad}_{\text{(II) Block attention}} \quad \underbrace{\qquad\qquad}_{\text{(III) Region attention}}$$

➤ **(I) Entity Attention.** Because urban functions emerge from interactions among entity categories, we first apply entity attention along the entity axis of $\boldsymbol{H}_{i,r,b,:,:}$ for each fixed (i,r,b) with $b\in\mathcal{B}^{\text{vis}}$. Masked block rows bypass this level, while the leading `[blk]` participates in the same attention and summarizes the block's urban fabric. Entity-category order has no inherent semantics, so the entity-attention stage uses no positional encoding. The attention is computed as

$$\boldsymbol{Q}^{\text{I}}_{i,r,b} = \boldsymbol{H}_{i,r,b,:,:}\boldsymbol{W}^{\text{I}}_{Q}, \quad \boldsymbol{K}^{\text{I}}_{i,r,b} = \boldsymbol{H}_{i,r,b,:,:}\boldsymbol{W}^{\text{I}}_{K}, \quad \boldsymbol{V}^{\text{I}}_{i,r,b} = \boldsymbol{H}_{i,r,b,:,:}\boldsymbol{W}^{\text{I}}_{V},$$
$$\boldsymbol{O}^{\text{I}}_{i,r,b} = \text{softmax}_T\left(\frac{\boldsymbol{Q}^{\text{I}}_{i,r,b}(\boldsymbol{K}^{\text{I}}_{i,r,b})^\top}{\sqrt{d_h}} + \boldsymbol{M}^{\text{I}}_{i,r,b}\right)\boldsymbol{V}^{\text{I}}_{i,r,b}. \tag{2}$$

$\boldsymbol{M}^{\mathrm{I}}_{i,r,b}$ masks padded tokens, and d_h is the head dimension, with the same meaning in the later stages. Each visible block therefore forms a $T \times T$ attention matrix. After L_1 layers, the `[blk]` state $\boldsymbol{H}^{(L_1)}_{i,r,b,0,:}$ summarizes functional interactions among entity categories and enters the next stage.

➤ **(II) Block Attention.** Entity attention captures each block's composition in isolation, but a block's function also depends on its spatial relations to surrounding blocks. For each city–region pair (i, r), the block-attention stage applies attention along the block axis of $\boldsymbol{H}_{i,r,:,0,:}$. This slice contains `[reg]` at $b = 0$ and the Level-I `[blk]` or unchanged `[mask_blk]` states at the remaining valid positions. Each block row's leading slot receives a learned role embedding that distinguishes internal from neighboring blocks. 2D RoPE (Su et al., 2024; Heo et al., 2024) injects each block's offset from the region center into Query and Key, with `[reg]` at the origin. The attention is computed as

$$\begin{aligned} \boldsymbol{Q}^{\mathrm{II}}_{i,r} = \boldsymbol{H}_{i,r,:,0,:}\boldsymbol{W}^{\mathrm{II}}_{Q}, \quad \boldsymbol{K}^{\mathrm{II}}_{i,r} = \boldsymbol{H}_{i,r,:,0,:}\boldsymbol{W}^{\mathrm{II}}_{K}, \quad \boldsymbol{V}^{\mathrm{II}}_{i,r} = \boldsymbol{H}_{i,r,:,0,:}\boldsymbol{W}^{\mathrm{II}}_{V}, \\ \boldsymbol{O}^{\mathrm{II}}_{i,r} = \mathrm{softmax}_B\left(\frac{\mathrm{RoPE}_{\mathrm{2D}}(\boldsymbol{Q}^{\mathrm{II}}_{i,r})\,\mathrm{RoPE}_{\mathrm{2D}}(\boldsymbol{K}^{\mathrm{II}}_{i,r})^{\top}}{\sqrt{d_h}} + \boldsymbol{M}^{\mathrm{II}}_{i,r}\right)\boldsymbol{V}^{\mathrm{II}}_{i,r}. \end{aligned} \tag{3}$$

$\boldsymbol{M}^{\mathrm{II}}_{i,r}$ masks padded block slots, so each region forms a $B \times B$ attention matrix. After L_2 layers, $\boldsymbol{h}_b = \boldsymbol{H}^{(L_1+L_2)}_{i,r,b,0,:}$ denotes the contextualized representation of each $b \in \mathcal{B}^{\mathrm{vis}}$. The updated `[reg]` at $\boldsymbol{H}^{(L_1+L_2)}_{i,r,0,0,:}$ summarizes the spatial organization of the input blocks.

➤ **(III) Region Attention.** Block attention captures within-region organization, but a region's fabric also reflects its city-scale role, such as being central or peripheral. For each city i, the region-attention stage applies attention along the region axis of $\boldsymbol{H}_{i,:,0,0,:}$. Because regions vary in physical extent, each `[reg]` state is first augmented with a region-area embedding. We then apply 2D RoPE using region-centroid offsets from the city center, with `[city]` at the origin. The shared `[city]` slot acts as a city-level hub that every region can attend to, shaping the region states rather than serving as a readout. The attention is computed as

$$\begin{aligned} \boldsymbol{Q}^{\mathrm{III}}_{i} = \boldsymbol{H}_{i,:,0,0,:}\boldsymbol{W}^{\mathrm{III}}_{Q}, \quad \boldsymbol{K}^{\mathrm{III}}_{i} = \boldsymbol{H}_{i,:,0,0,:}\boldsymbol{W}^{\mathrm{III}}_{K}, \quad \boldsymbol{V}^{\mathrm{III}}_{i} = \boldsymbol{H}_{i,:,0,0,:}\boldsymbol{W}^{\mathrm{III}}_{V}, \\ \boldsymbol{O}^{\mathrm{III}}_{i} = \mathrm{softmax}_R\left(\frac{\mathrm{RoPE}_{\mathrm{2D}}(\boldsymbol{Q}^{\mathrm{III}}_{i})\,\mathrm{RoPE}_{\mathrm{2D}}(\boldsymbol{K}^{\mathrm{III}}_{i})^{\top}}{\sqrt{d_h}} + \boldsymbol{M}^{\mathrm{III}}_{i}\right)\boldsymbol{V}^{\mathrm{III}}_{i}. \end{aligned} \tag{4}$$

$\boldsymbol{M}^{\mathrm{III}}_{i}$ masks padded region slots, so each city forms an $R \times R$ attention matrix. After L_3 layers, the updated `[reg]` at $\boldsymbol{H}^{(L_1+L_2+L_3)}_{i,r,0,0,:}$ gives the contextualized region representation $\boldsymbol{u}_r$.

3.3 Reconstructing Cities: Hierarchy-Aligned Objectives

We sample which blocks and entity categories to mask within each region's input set $\mathcal{B}_r$. Two objectives then reconstruct local functions and regional fabric from the remaining observations.

➤ **Sampling & Masking.** We perform breadth-first search on the nearest-neighbor graph of $\mathcal{B}_r$, starting from a randomly selected internal block and stopping when the selected set covers a ρ_{blk} fraction of $\mathcal{B}_r$. The selected blocks become $\mathcal{B}^{\mathrm{mask}}_r$ and carry `[mask_blk]`. For each visible block b, we replace a ρ_{ent} fraction of its present categories with `[mask_ent]` to form positive queries $\mathcal{V}^{+}_b \subseteq \mathcal{V}_b$. We then sample an equal number of negative queries $\mathcal{V}^{-}_b \subseteq \mathcal{V} \setminus \mathcal{V}_b$ from absent categories. These positive and negative block–category pairs form the query set $\mathcal{Q} = \{(b, c) \mid b \in \mathcal{B}^{\mathrm{vis}},\ c \in \mathcal{V}^{+}_b \cup \mathcal{V}^{-}_b\}$.

➤ **Local Function Reconstruction.** Category presence and absence jointly characterize a block's local function. For each $(b, c) \in \mathcal{Q}$, the local-function decoder, comprising an FFN maps the concatenation $[\boldsymbol{h}_b; \boldsymbol{x}_c]$ of the block representation and queried category embedding to a logit whose sigmoid yields $\widehat{p}_{bc}$. The target is $y_{bc} = \mathbf{1}[c \in \mathcal{V}_b]$, where $\mathcal{V}_b$ is the category set before masking. Binary cross-entropy on these predictions trains the tokenizer, UiT, and decoder end to end to infer category occurrence from the visible local categories and the surrounding spatial context.

➤ **Regional Fabric Reconstruction.** Regional fabric is jointly determined by all internal blocks, whereas $\boldsymbol{u}_r$ is derived from incomplete observations. For each vocabulary category c, the target v_{rc} is the pre-masking aggregate value of c over the internal blocks of region r, and is zero when c is absent. The regional-fabric decoder, comprising an FFN followed by a linear regression head, maps

$\boldsymbol{u}_r$ to $\widehat{v}_{rc}$ for every vocabulary category $c \in \mathcal{V}$. Mean-squared error trains the same modules end to end, encouraging $\boldsymbol{u}_r$ to encode the missing regional fabric from the remaining context.

After the two decoders, the total pretraining objective sums the two losses:

$$\mathcal{L} = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{(b,c)\in\mathcal{Q}} [y_{bc}\log\widehat{p}_{bc} + (1-y_{bc})\log(1-\widehat{p}_{bc})]}_{\text{Local Function Reconstruction}} + \underbrace{\frac{1}{|\mathcal{R}|\,|\mathcal{V}|}\sum_{r\in\mathcal{R}}\sum_{c\in\mathcal{V}}(\widehat{v}_{rc}-v_{rc})^2}_{\text{Regional Fabric Reconstruction}}. \tag{5}$$

At inference no masking is applied. Region- and block-level tasks use the UiT states $\boldsymbol{u}_r$ and $\boldsymbol{h}_b$ with both decoders discarded, while entity-level category ranking reuses the frozen local-function decoder.

4 Experiments

4.1 Experimental Settings

Pretraining Data and Downstream Tasks. We spatially assign 263 million Overture Maps entities across the United States to Census TIGER/Line blocks and tracts, giving a pretraining corpus of 8.18 million blocks. The entities span six Overture themes: places, buildings, infrastructure, land cover, land, and water. Census tracts instantiate regions, and the 3,143 county equivalents that cover the 50 states and Washington, DC instantiate cities, since municipal boundaries neither cover the country exhaustively nor nest in the Census hierarchy. The benchmark evaluates frozen representations at three levels: 12 region-level tasks spanning morphology, demographics, economy, and environment; land-use composition and population at the block level; and contextual category recovery over places and buildings at the entity level. The last targets a concrete gap: many places carry no category and building function is rarely recorded, so inferring one from its surroundings completes map inventories without new survey effort. Full details in Appendices S3.1–S3.2.

Baselines and Evaluation Protocol. We compare against 21 baselines spanning region-centric, entity-centric, and broad-coverage geospatial representations, and each evaluation uses the subset applicable to its spatial granularity. For region- and block-level tasks, we freeze every representation and fit a one-hidden-layer MLP probe, checked against a linear ridge probe, reporting mean and standard deviation over seven fixed geographically disjoint splits. At the entity level, we encode the other entity categories in the anchor's block into a contextualized representation, and the frozen local-function decoder ranks candidate categories under Hit@1 and MRR. Full details in Appendices S3.3–S3.4.

Implementation Details. HESTIA contains 14 Transformer layers with a hidden dimension of 256 and eight attention heads. The category and block mask ratios are 0.50 and 0.20, respectively. We pretrain for 60 epochs on seven NVIDIA A100 GPUs using bf16 and AdamW with a peak learning rate of 2×10^{-4} decayed by a cosine schedule after a 5% linear warmup, and weight decay of 0.05, processing up to 16 regions per GPU for an effective batch of up to 112, which gives 971 optimizer steps per epoch over the 80,181 pretraining regions. Full details in Appendix S2.

4.2 Performance Comparison

Parameter Efficiency. HESTIA reaches its accuracy with 12.9M parameters (Figure 4), a compact budget next to AlphaEarth (480M), while S2Vec stays competitive at 4.8M. Performance therefore does not track capacity across this range, suggesting that hierarchy-aligned modeling contributes more than scale. Tables S14 and S15 sweep our own capacity and its cost.

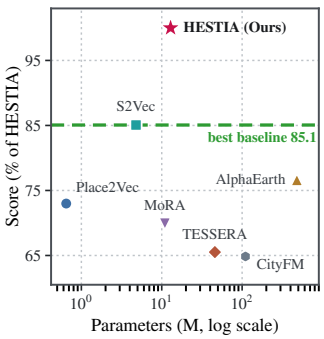


Figure 4: **Parameter efficiency.**

Fine-to-Coarse Evaluation. HESTIA ranks first at all three evaluated levels of the hierarchy, taken here from the finest upward. At the entity level, the frozen decoder recovers a category removed before encoding, reaching 40.35 Hit@1 on Places against 36.39 for Place2Vec and 50.85 on Buildings against 46.26 for CityFM, gains of 11% and 10%. Entity-category tokens therefore reflect what surrounds a category rather than how common it is: the gains concentrate on categories with a characteristic neighbourhood, such as elementary schools (83.8 against

Table 1: **Panel A** is the main benchmark and pretrains on all cities. **Panel B** tests generalization to unseen cities. Tasks span morphology (♠), demographics (♡), economy (♢), and environment (♣). The last column orders methods by their mean task rank. Each task marks its **best** and **second-best** result.

	♠ Morphology		♡ Demographics		♢ Economy		♣ Environment		
Model	Land use (AUPRC ↑)	Road density (R^2 ↑)	Pop. density (R^2 ↑)	Age (KL ↓)	Night lights (R^2 ↑)	Income (R^2 ↑)	CO_2 (R^2 ↑)	Urban heat (R^2 ↑)	**Rank**
Panel A: All-city pretraining with downstream evaluation									
Region-Centric Urban Representations									
MVURE	0.311 ± 0.02	-0.040 ± 0.12	-0.102 ± 0.09	0.103 ± 0.01	0.193 ± 0.08	-0.085 ± 0.06	-0.017 ± 0.06	-0.008 ± 0.03	15
HREP	0.293 ± 0.01	-0.137 ± 0.13	-0.178 ± 0.09	0.104 ± 0.01	0.043 ± 0.11	-0.137 ± 0.09	-0.117 ± 0.08	-0.053 ± 0.01	17
HAFusion	0.294 ± 0.01	-0.158 ± 0.17	-0.174 ± 0.10	0.100 ± 0.01	-0.032 ± 0.08	-0.053 ± 0.08	-0.108 ± 0.14	0.005 ± 0.01	16
ReCP	0.408 ± 0.01	0.282 ± 0.43	-0.029 ± 0.11	0.088 ± 0.01	0.489 ± 0.04	0.162 ± 0.06	0.011 ± 0.08	0.153 ± 0.05	10
ComSRE	0.416 ± 0.02	0.281 ± 0.08	-0.995 ± 0.81	0.084 ± 0.01	0.369 ± 0.05	0.308 ± 0.08	0.038 ± 0.03	0.178 ± 0.04	11
MoRA	0.499 ± 0.02	0.642 ± 0.03	0.261 ± 0.12	0.074 ± 0.01	0.546 ± 0.04	0.428 ± 0.06	0.136 ± 0.04	0.286 ± 0.05	5
MVJC	0.314 ± 0.02	0.025 ± 0.15	-0.148 ± 0.11	0.101 ± 0.01	0.185 ± 0.09	0.001 ± 0.07	0.001 ± 0.02	0.027 ± 0.02	14
READ	0.385 ± 0.02	0.381 ± 0.09	-0.155 ± 0.56	0.096 ± 0.01	0.413 ± 0.15	0.054 ± 0.19	0.121 ± 0.09	0.066 ± 0.09	12
Entity-Centric Urban Representations									
Place2Vec	**0.526 ± 0.02**	0.648 ± 0.05	0.161 ± 0.13	**0.072 ± 0.01**	0.589 ± 0.02	0.439 ± 0.06	0.161 ± 0.05	0.303 ± 0.05	4
CityFM	0.425 ± 0.02	0.570 ± 0.06	0.370 ± 0.26	0.078 ± 0.01	0.584 ± 0.07	0.332 ± 0.05	0.177 ± 0.10	0.243 ± 0.07	7
GURPP	0.423 ± 0.02	0.639 ± 0.03	0.307 ± 0.04	0.086 ± 0.01	0.509 ± 0.05	0.332 ± 0.06	0.123 ± 0.06	0.225 ± 0.06	9
UrbanMFM	0.340 ± 0.02	0.143 ± 0.09	-0.681 ± 0.63	0.094 ± 0.01	0.357 ± 0.07	0.017 ± 0.06	0.054 ± 0.07	0.129 ± 0.04	13
Broad-Coverage Geospatial Representations									
SatCLIP	0.302 ± 0.02	-0.543 ± 0.58	-1.346 ± 2.55	0.103 ± 0.01	-0.366 ± 0.47	-0.040 ± 0.18	-0.159 ± 0.25	-0.078 ± 0.11	18
SIREN-M	0.307 ± 0.03	-0.309 ± 0.22	-1.847 ± 3.92	0.108 ± 0.01	-0.488 ± 0.75	-0.168 ± 0.23	-0.371 ± 0.40	0.053 ± 0.18	19
UrbanCLIP	0.422 ± 0.02	0.801 ± 0.04	0.375 ± 0.18	0.095 ± 0.01	0.650 ± 0.07	0.067 ± 0.15	0.123 ± 0.06	0.385 ± 0.07	8
AlphaEarth	0.465 ± 0.02	**0.856 ± 0.03**	**0.576 ± 0.11**	0.096 ± 0.02	**0.747 ± 0.09**	0.120 ± 0.26	0.212 ± 0.18	**0.561 ± 0.09**	3
S2Vec	0.487 ± 0.02	0.773 ± 0.03	0.534 ± 0.08	0.074 ± 0.01	0.740 ± 0.07	**0.529 ± 0.05**	**0.261 ± 0.12**	0.378 ± 0.07	2
TESSERA	0.424 ± 0.02	0.778 ± 0.04	0.513 ± 0.12	0.095 ± 0.01	0.665 ± 0.06	0.115 ± 0.17	0.149 ± 0.11	0.402 ± 0.12	6
HESTIA (Ours)	**0.554 ± 0.02**	**0.896 ± 0.03**	**0.743 ± 0.07**	**0.066 ± 0.01**	**0.836 ± 0.02**	**0.602 ± 0.06**	**0.299 ± 0.09**	**0.563 ± 0.06**	**1**
Panel B: Seen-city pretraining with unseen-city evaluation (Unseen-City Generalization)									
Place2Vec	**0.533 ± 0.03**	0.682 ± 0.06	0.170 ± 0.25	**0.068 ± 0.01**	0.579 ± 0.09	0.422 ± 0.04	-0.543 ± 0.83	0.304 ± 0.27	3
CityFM	0.443 ± 0.03	0.609 ± 0.06	-0.115 ± 0.88	0.073 ± 0.01	0.639 ± 0.09	0.314 ± 0.13	-0.265 ± 0.63	0.273 ± 0.22	4
S2Vec	0.494 ± 0.03	**0.793 ± 0.05**	**0.575 ± 0.11**	0.072 ± 0.01	**0.780 ± 0.08**	**0.513 ± 0.08**	**-0.080 ± 0.43**	**0.400 ± 0.31**	2
HESTIA (Ours)	**0.554 ± 0.03**	**0.934 ± 0.02**	**0.796 ± 0.05**	**0.063 ± 0.01**	**0.843 ± 0.07**	**0.583 ± 0.09**	**0.004 ± 0.28**	**0.547 ± 0.16**	**1**

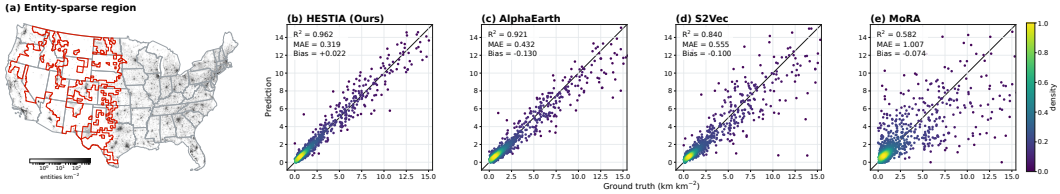


Figure 5: **(a)** Entity density per tract, with the red outline bounding the 460 sparse counties. **(b)–(e)** Predicted against observed road density there, with $y = x$ in black and color showing point density.

75.1) and parks (59.3 against 27.4). At the block level, the granularity HESTIA encodes directly, land-use AUPRC rises from the 0.26 of AlphaEarth to 0.41 and block population R^2 from 0.41 to 0.62, improvements of 58% and 51% (Table 3), the widest relative gains we report.

Region Level and Unseen Cities. At the region level the margin tracks how much a label depends on functional composition, that is, on what a place contains rather than on how it looks from above. Under all-city pretraining (Panel A of Table 1), population density rises from 0.576 to 0.743 R^2 over the strongest baseline, a gain of 29%, followed by CO_2 intensity (15%), household income (14%) and nighttime lights (12%); the lead narrows to 5% over Place2Vec on land use, where overhead radiometry already suffices. The two runners-up, S2Vec and AlphaEarth, also cover the whole country, but through regular cells and satellite pixels, so the gap is not an artifact of coverage. Under the strict unseen-city protocol (Panel B), where each held-out city must be encoded from its own entities, the gaps widen rather than shrink: population density improves from 0.575 to 0.796 over S2Vec (38%), urban heat by 37% and road density by 18%. A new city needs no per-region table

Table 2: Entity-level contextual category recovery.

Model	Places		Buildings	
	Hit@1 ↑	MRR ↑	Hit@1 ↑	MRR ↑
Place2Vec	36.39	57.57	41.84	60.90
SpaBERT	25.73	48.99	43.25	63.62
HGI	14.59	38.90	33.62	53.91
CityFM	32.04	54.35	46.26	63.70
GURPP	23.62	47.29	38.77	58.33
UrbanMFM	26.59	49.41	42.24	60.62
HESTIA (Ours)	**40.35**	**60.92**	**50.85**	**68.07**

Table 3: Block-level evaluation.

Model	Land use		Population	
	AUPRC ↑	F1 ↑	R^2 ↑	MAE ↓
HGI	0.18 ± 0.01	0.11 ± 0.01	0.21 ± 0.05	57.7 ± 11.6
UrbanMFM	0.14 ± 0.00	0.10 ± 0.01	0.08 ± 0.02	37.6 ± 4.9
S2Vec	0.19 ± 0.01	0.12 ± 0.01	0.24 ± 0.05	34.4 ± 4.3
SatCLIP	0.08 ± 0.00	0.05 ± 0.01	-0.01 ± 0.12	37.3 ± 4.1
AlphaEarth	$\mathbf{0.26 \pm 0.01}$	$\mathbf{0.18 \pm 0.01}$	$\mathbf{0.41 \pm 0.04}$	$\mathbf{32.4 \pm 4.3}$
TESSERA	0.21 ± 0.01	0.14 ± 0.01	0.34 ± 0.04	33.5 ± 4.2
HESTIA	$\mathbf{0.41 \pm 0.02}$	$\mathbf{0.34 \pm 0.01}$	$\mathbf{0.62 \pm 0.02}$	$\mathbf{23.8 \pm 2.9}$

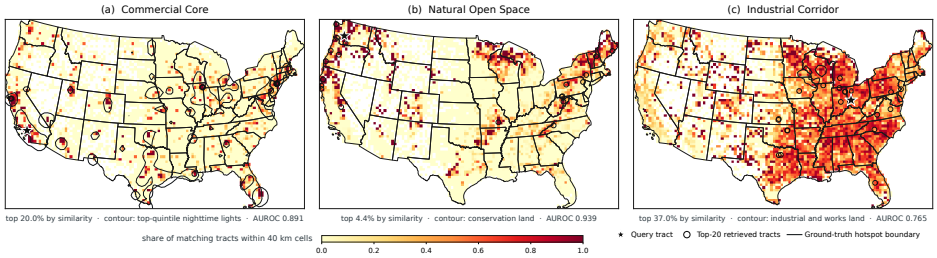


Figure 6: **Nationwide retrieval against independent references.** Color is the share of tracts per 40 km cell that HESTIA ranks closest to the query, thresholded at each reference's own national prevalence. Black contours bound the reference hotspots. The three references—top-quintile nighttime lights, conservation land, and industrial land—enter neither pretraining nor query selection. Stars mark queries, circles the top-20 retrievals.

re-estimated, since its representation comes from the entities observed inside it rather than from a stored per-region vector that a held-out region has no way to fill.

Entity-Sparse Regions. Because the hierarchy is built from observed entities, we test it where they are scarcest: the 460 contiguous counties in the lowest quintile of entity density (Figure 5). Over their 1,374 tracts, HESTIA reaches $R^2 = 0.962$ with $+0.022$ bias, ahead of AlphaEarth (0.921), S2Vec (0.840) and MoRA (0.582). These fits pool all sparse tracts rather than averaging the seven splits, so only the ordering is comparable with Table 1. Appendices S4.1–S4.9 give the per-level breakdowns, the ridge-probe readout, the twelve-task unseen-city table and the other sparse-region tasks.

4.3 FURTHER ANALYSIS

Objectives, Not Hyperparameters. The lead reported above comes from the two pretraining objectives rather than from tuning. Across all twelve region-level tasks (Table S18 of Appendix S5), removing the regional objective loses on all eight main tasks, with the largest drop, 6.8 points, on household income; removing the category objective loses on seven of the eight. Removing the joint token modulation loses on five of the eight, and removing the region stage, so that a city's regions never attend to one another, costs 13% of urban heat and 3% of nighttime lights (Table S19). The position encoding is decisive one level down: without the role it gives context blocks, block population R^2 falls from 0.36 to -0.48. Hyperparameters move it far less: sweeping the category and block mask ratios shifts the eleven-task mean, which excludes the population count variant, by at most 1.9% and 1.4%, and every setting preserves the ranking above (Figure 7, Table S20). The two ratios differ in shape: the category ratio plateaus, while the block ratio has an interior optimum, since hiding larger connected groups trades local targets for diffuse ones.

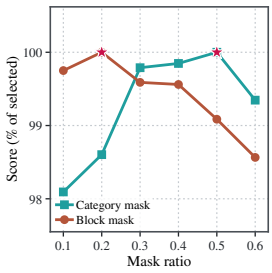


Figure 7: **Sensitivity study.**

Similarity Structure across the Hierarchy. When HESTIA calls two units similar, they are also similar in the entities they actually contain. Figure 8 encodes each category as a block holding only that category at its median count and area, so agreement cannot be driven by how many entities a block holds, and compares it against pointwise mutual information in 123,512 real blocks nationwide. The learned and measured matrices share the same block-diagonal structure, with Spearman rank

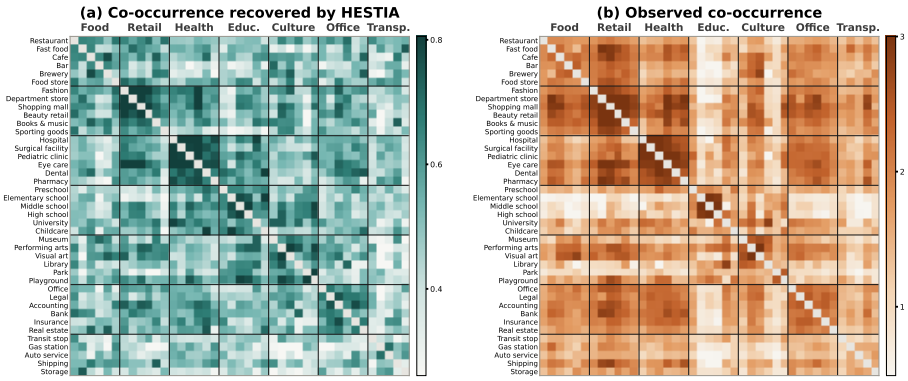


Figure 8: **Learned category similarity against observed co-occurrence.** Both panels cover the same 41 representative categories in the same order, grouped by function, drawn from the 185 categories the diagnostic covers. **(a)** Cosine similarity between single-category block states. **(b)** Pointwise mutual information measured in real blocks. Colors are scaled within each panel.

agreement 0.63 across the 185 categories frequent enough to estimate, and an AUC of 0.81 for separating within-group from cross-group pairs. Categories reach the model only as identifiers, so what is testable is how they sit relative to one another, and that arrangement recovers a functional grouping neither the tokenizer nor the objectives ever receive. Appendix S6 runs the same test at the block and region levels. The three together show urban semantics emerging where the architecture places them, as an entity category acquires a functional organisation it was never given, a block state still resembles the categories inside it, and a region state no longer does.

Geographic Coverage Sets the Ceiling. What the model can learn is bounded by how much of the country it sees. Varying how many cities enter pretraining (Figure 9) separates two regimes: across a fivefold increase in coverage under a fixed architecture and schedule, morphology improves by 2%, whereas economy improves by 11%, demographics by 9% and environment by 8%, and every domain is still improving when the last cities enter. The labels that depend most on coverage are therefore exactly those a single-metro corpus cannot supply (Table S12).

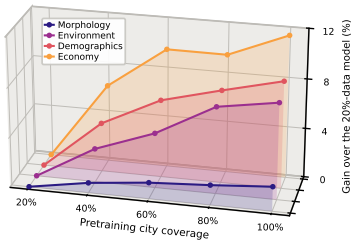


Figure 9: **Pretraining data scaling.**

Nationwide Retrieval. Region embeddings retrieve functionally similar tracts nationwide rather than spatially nearby ones. We rank every tract outside a prototype query's county by cosine similarity after removing the national mean embedding and compare the result against an independent reference used in neither pretraining nor query selection (Figure 6). Retrieval separates the reference class far above chance in every case, with AUROC 0.891 for commercial cores, 0.939 for natural open space, and 0.765 for industrial corridors. The effect is functional rather than geographic: similarity correlates with category composition at partial ρ of 0.61, 0.42, and 0.61 after controlling for distance, whereas correlation with proximity alone is only 0.13, 0.08, and 0.28. The similarity and retrieval tests both describe a representation organized by what places contain and how those contents co-occur. Appendix S6 maps the full similarity fields, and Appendix S4 the robustness to missing inputs, the stratified error analysis and the measured cost of pretraining the reported model.

5 Conclusion

This paper asks how to encode cities with a Transformer, and answers it by treating the entity–block–region–city hierarchy as the input structure rather than as post-processing. HESTIA tokenizes geospatial entities inside a fixed-shape tensor that preserves that hierarchy, attends successively along the entity, block, and region axes, and learns by reconstructing masked local functions and regional fabric. Pretrained on 8.18 million U.S. blocks, it ranks first at all three levels, and its advantage is largest on labels determined by what a place contains and on cities held out of pretraining, where a representation must be computed from a city's own entities. Beyond these results, UiT provides a scalable architectural backbone for pretraining urban foundation models at national scale, which we hope supports the wider development of urban AI.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

We used generative AI for two purposes. First, for data processing: places lacking an Overture `basic_category` were assigned to the frozen vocabulary by a constrained classifier, under the acceptance rule of Appendix S3.1. Second, for copy-editing and polishing the presentation. The authors verified all AI-assisted output and take responsibility for the final content of this work.

REFERENCES

Mohit Agarwal, Mimi Sun, Chaitanya Kamath, Arbaaz Muslim, Prithul Sarker, Joydeep Paul, Hector Yee, Marcin Sieniek, Kim Jablonski, Swapnil Vispute, Atul Kumar, Yael Mayer, David Fork, Sheila de Guia, Jamie McPike, Adam Boulanger, Tomer Shekel, David Schottlander, Yao Xiao, Manjit Chakravarthy Manukonda, Yun Liu, Neslihan Bulut, Sami Abu-El-Haija, Bryan Perozzi, Monica Bharel, Von Nguyen, Luke Barrington, Niv Efron, Yossi Matias, Greg Corrado, Krish Eswaran, Shruthi Prabhakara, Shravya Shetty, and Gautam Prasad. General geospatial inference with a population dynamics foundation model. *arXiv preprint arXiv:2411.07207*, 2024.

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*, pp. 87–97, 2024. doi: 10.1145/3627673.3679662.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Meng Chen, Hongwei Jia, Zechen Li, Weiming Huang, Kai Zhao, Yongshun Gong, Haoran Xu, and Hongjun Dai. Region embedding with adaptive correlation discovery for predicting urban socioeconomic indicators. *IEEE Transactions on Knowledge and Data Engineering*, 38(2):1280–1291, 2026. doi: 10.1109/TKDE.2025.3631025.

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*, 12(3):1–21, 2026. doi: 10.1145/3787217.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. POI2Vec: Geographical latent representation for predicting future visitors. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, volume 31, pp. 102–108, 2017. doi: 10.1609/aaai.v31i1.10500.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, Toby Jackson, James Ball, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34818–34831, 2026a.

Zhengpeng Feng, Sadiq Jaffer, Ira Shokar, Jovana Knezevic, James Ball, Pedro Sousa, Mark Elvers, Madeline Lisaius, Clement Atzberger, Robin Young, et al. Tessera v2: Scaling pixel-wise earth foundation models. *arXiv preprint arXiv:2607.03949*, 2026b.

Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. Urbanvlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28061–28069, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305, 2024. doi: 10.1007/978-3-031-72684-2_17.

Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, et al. Olmoearth: Stable latent image modeling for multimodal earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34806–34817, 2026.

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. doi: 10.1080/13658816.2022.2040510.

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. doi: 10.1016/j.isprsjprs.2022.11.021.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7383–7394. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00693.

Jiahui Jin, Yifan Song, Dong Kan, Haojia Zhu, Xiangguo Sun, Zhicheng Li, Xigang Sun, and Jinghui Zhang. Urban region pre-training and prompting: A graph-based approach. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 1071–1082, 2025. doi: 10.1145/3711896.3737176.

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. doi: 10.1609/aaai.v39i4.32457.

Zechen Li, Weiming Huang, Kai Zhao, Min Yang, Yongshun Gong, and Meng Chen. Urban region embedding via multi-view contrastive prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8724–8732, 2024. doi: 10.1609/aaai.v38i8.28718.

Zechen Li, Hongwei Jia, Kai Zhao, Weiming Huang, and Meng Chen. Multi-view urban region embedding via commonality-specificity disentanglement. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 748–758, 2026. doi: 10.1145/3770854.3780234.

Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. SpaBERT: A pretrained language model from geographic data for geo-entity representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2757–2769, 2022. doi: 10.18653/v1/2022.findings-emnlp.200.

Yingde Lin, Yuanbo Xu, Lu Jiang, and Pengyang Wang. Comprehensive urban region representation learning via multi-view joint learning and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15261–15268, 2026. doi: 10.1609/aaai.v40i18.38551.

Junyuan Liu, Xinglei Wang, Zichao Zeng, Jiazhuang Feng, Quan Qin, Ilya Ilyankou, Guangsheng Dong, and Tao Cheng. Cityrep: A unified benchmark for urban representations across cities, tasks, and modalities. *arXiv preprint arXiv:2605.26036*, 2026.

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.

Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23498–23515. PMLR, 2023a.

Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2Vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023b. doi: 10.1016/j.isprsjprs.2023.06.016.

Julia Moska, Oleksii Furman, Kacper Kozaczko, Szymon Leszkiewicz, Jakub Polczyk, Piotr Gramacki, and Piotr Szymański. OBSR: Open benchmark for spatial representations. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 670–681, 2025. doi: 10.1145/3748636.3762783.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1):4540, 2020. doi: 10.1038/s41467-020-18321-y.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4065–4076, 2023. doi: 10.1109/ICCV51070.2023.00378.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *International Conference on Learning Representations*, volume 2024, pp. 1746–1759, 2024.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Fengze Sun, Jianzhong Qi, Yanchuan Chang, Xiaoliang Fan, Shanika Karunasekera, and Egemen Tanin. Urban region representation learning with attentive fusion. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 4409–4421, 2024. doi: 10.1109/ICDE60146.2024.00336.

Fengze Sun, Yanchuan Chang, Egemen Tanin, Shanika Karunasekera, and Jianzhong Qi. FlexiReg: Flexible urban region representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2702–2713, 2025. doi: 10.1145/3711896.3736965.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2023. doi: 10.1109/TGRS.2022.3194732.

Waldo R Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234, 1970. doi: 10.2307/143141.

Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232, 2019. doi: 10.1111/2041-210X.13107.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8690–8701, 2023. doi: 10.52202/075280-0379.

Xinglei Wang, Tao Cheng, Stephen Law, Zichao Zeng, Lu Yin, and Junyuan Liu. Multi-modal contrastive learning of urban space representations from POI data. *Computers, Environment and Urban Systems*, 120:102299, 2025. doi: 10.1016/j.compenvurbsys.2025.102299.

Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. Urban2Vec: Incorporating street view imagery and POIs for multi-modal urban neighborhood embedding. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pp. 1013–1020, 2020. doi: 10.1609/aaai.v34i01.5450.

Ya Wen, Jixuan Cai, Qiyao Ma, Linyan Li, Xinhuan Chen, Chris Webster, and Yulun Zhou. Mora: Mobility as the backbone for geospatial representation learning at scale. In *International Conference on Learning Representations*, volume 2026, pp. 37647–37664, 2026.

Nemin Wu, Qian Cao, Zhangyu Wang, Zeping Liu, Yanlin Qi, Jielu Zhang, Joshua Ni, Xiaobai Yao, Hongxu Ma, Lan Mu, Stefano Ermon, Tanuja Ganu, Akshay Nambi, Ni Lao, and Gengchen Mai. TorchSpatial: A location encoding framework and benchmark for spatial representation learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 81437–81460, 2024. doi: 10.52202/079017-2588.

Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. Multi-graph fusion networks for urban region embedding. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2312–2318, 2022. doi: 10.24963/ijcai.2022/321.

Congxi Xiao, Jingbo Zhou, Yixiong Xiao, Jizhou Huang, and Hui Xiong. ReFound: Crafting a foundation model for urban region understanding upon language and visual foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3527–3538, 2024. doi: 10.1145/3637528.3671992.

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. doi: 10.1145/3139958.3140054.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*, pp. 4006–4017, 2024. doi: 10.1145/3589334.3645378.

Xixian Yong and Xiao Zhou. MuseCL: Predicting urban socioeconomic indicators via multi-semantic contrastive learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 7536–7544, 2024. doi: 10.24963/ijcai.2024/834. URL https://www.ijcai.org/proceedings/2024/834.

Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 4431–4437, 2020. doi: 10.24963/ijcai.2020/611.

Yimei Zhang, Guojiang Shen, Kaili Ning, Tongwei Ren, Xuebo Qiu, Mengmeng Wang, and Xiangjie Kong. Improving region representation learning from urban imagery with noisy long-caption supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16397–16405, 2026a. doi: 10.1609/aaai.v40i19.38678.

Zhaoqi Zhang, Miao Xie, Pasquale Balsebre, Weiming Huang, Siqiang Luo, and Gao Cong. Urbanmfm: Spatial graph-based multiscale foundation models for learning generalized urban representation. *IEEE Transactions on Knowledge and Data Engineering*, 38(3):2064–2078, 2026b. doi: 10.1109/TKDE.2026.3656202.

Silin Zhou, Dan He, Lisi Chen, Shuo Shang, and Peng Han. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4981–4989, 2023. doi: 10.1609/aaai.v37i4.25625.

CONTENTS OF THE APPENDIX

S1 Extended Related Work **16**
- S1.1 Region-Centric Urban Representation Learning . . . 16
- S1.2 Entity-Centric and Hierarchical Urban Modeling . . . 16
- S1.3 Broad-Coverage Geospatial Foundation Models . . . 16
- S1.4 Evaluation Benchmarks and Spatial Validation . . . 17
- S1.5 Pretraining Relations and the Position of HESTIA . . . 17

S2 Algorithmic Flow of HESTIA **17**

S3 Experimental Settings **18**
- S3.1 Dataset Construction and Pretraining Corpus . . . 18
- S3.2 Downstream Benchmark Inventory . . . 20
- S3.3 Spatial Splits, Probes, and Reporting . . . 20
- S3.4 Baselines under a Common Protocol . . . 23

S4 Extended Results, Scaling, and Robustness **23**
- S4.1 Extended Region-Level Results . . . 23
- S4.2 Extended Block-Level Results . . . 26
- S4.3 Extended Entity-Level Results . . . 26
- S4.4 City-Count Scaling . . . 27
- S4.5 Pretraining Convergence . . . 27
- S4.6 Pretraining Efficiency . . . 28
- S4.7 Model Capacity and Pretraining Cost . . . 29
- S4.8 Robustness to Missing Observations . . . 29
- S4.9 Entity-Sparse Regions . . . 30
- S4.10 Stratified Error Analysis . . . 31

S5 Design and Hyperparameter Controls **31**
- S5.1 Core Ablation . . . 32
- S5.2 Additional Design Controls . . . 34
- S5.3 Mask-Ratio Sensitivity . . . 34

S6 Representation Diagnostics **35**
- S6.1 Nationwide Similarity Fields . . . 35
- S6.2 Similarity Structure at the Tract and Block Levels . . . 36

The appendix is organized as follows: §S1–§S3 document what HESTIA is and how it is evaluated, §S4–§S5 establish that the reported result is robust to geography, scale and design choices, and §S6 inspects what the learned representation contains.

S1 Extended Related Work

This section reviews urban representation methods and broad-coverage geospatial models, organized by modeling unit, spatial abstraction, pretraining objective, and evaluation protocol.

S1.1 Region-Centric Urban Representation Learning

Region-centric methods use predefined spatial units as their primary modeling objects. Early multi-view approaches aggregate points of interest, mobility flows, check-ins, or demographic attributes within census tracts or other administrative regions, and model inter-region relations with multi-graph, heterogeneous-relation, or attention architectures (Zhang et al., 2020; Wu et al., 2022; Zhou et al., 2023; Sun et al., 2024). This makes it efficient to represent many regions, but aggregation can mix local functions and discard relations among entity categories within blocks. Recent work strengthens regional learning with cross-view and inter-region objectives. ReCP aligns complementary POI and mobility views (Li et al., 2024); MVJC learns structured correlations to reduce false negatives (Lin et al., 2026); and ComSRE separates shared from view-specific information, while READ models adaptive regional relations (Li et al., 2026; Chen et al., 2026). These methods improve regional dependencies but still operate on pre-aggregated regions. Multimodal variants enrich regional semantics. Urban2Vec combines urban attributes with imagery (Wang et al., 2020); UrbanCLIP aligns satellite imagery with language (Yan et al., 2024); and MuseCL, ReFound, and CaLLiPer use multimodal or foundation-model supervision (Yong & Zhou, 2024; Xiao et al., 2024; Wang et al., 2025). These variants still operate primarily on aggregated regions and do not resolve the granularity trade-off with entity-level modeling. UrbanVLP further explores multi-granularity vision–language pretraining for urban socioeconomic prediction, while Zhang et al. study noisy long-caption supervision for urban imagery (Hao et al., 2025; Zhang et al., 2026a). Regular cells provide uniform spatial inputs. Region-centric methods may use square or hexagonal tessellations, while FlexiReg adaptively aggregates multimodal cell representations into flexibly defined regions (Sun et al., 2025). S2Vec treats built-environment vectors assigned to Google S2 cells as patches and applies masked autoencoding within parent cells (He et al., 2022; Choudhury et al., 2026). These approaches simplify batching and broad-area pretraining, but imposed boundaries may misalign with urban morphology, fragment coherent environments, and introduce resolution dependence. Pre-aggregation within cells can also obscure the native boundaries and containment relations of regions, blocks, and entities.

S1.2 Entity-Centric and Hierarchical Urban Modeling

Entity-centric methods retain finer-grained geographic objects as their basic units. POI2Vec and Place2Vec learn from POI sequences, spatial contexts, or co-occurrence patterns (Feng et al., 2017; Yan et al., 2017), while SPPE incorporates spatial and categorical context (Huang et al., 2022). More recent models use multimodal contexts or graphs: CityFM aligns map entities with textual, geometric, and visual information (Balsebre et al., 2024). GURPP instead performs graph-based urban-region pretraining and prompting with spatial context (Jin et al., 2025). These approaches preserve local information but incur higher computational and memory costs as the entity count grows. A related line of work explicitly models spatial hierarchy. HGI propagates information along a POI–region–city hierarchy (Huang et al., 2023), while UrbanMFM connects micro-level and global contexts through spatial graphs (Zhang et al., 2026b). These works demonstrate the value of multi-scale urban relations, but hierarchies are represented as graph structures rather than a fixed-shape Transformer input and computation schedule.

S1.3 Broad-Coverage Geospatial Foundation Models

Broad-coverage geospatial models demonstrate that self-supervised pretraining can scale across large geographic areas. Location encoders learn from coordinates or spatial functions (Mai et al., 2020; 2023b; Rußwurm et al., 2024; Wu et al., 2024); coordinate–image and language-alignment methods connect locations with imagery or text (Mai et al., 2023a; Vivanco Cepeda et al., 2023; Klemmer et al., 2025); and Earth-observation models learn from satellite or global environmental signals (Sun et al., 2023; Reed et al., 2023; Brown et al., 2025; Jakubik et al., 2025; Herzog et al., 2026; Feng et al., 2026a;b). Population-dynamics and mobility models add temporal or graph-based signals (Agarwal et al., 2024; Wen et al., 2026). These models provide broad geographic coverage, but their inputs

primarily describe locations, environments, or pre-aggregated mobility patterns. They do not directly model how entity categories form blocks, blocks form regions, and regions provide city-scale context.

S1.4 Evaluation Benchmarks and Spatial Validation

Unified benchmarks such as CityRep evaluate urban representations across multiple cities, tasks, modalities, and spatially structured splits (Liu et al., 2026). OBSR provides a broader modality-agnostic benchmark spanning multiple geospatial datasets and continents (Moska et al., 2025). Beyond benchmark design, spatial-validation studies show that random cross-validation can underestimate prediction error when nearby samples are spatially dependent (Roberts et al., 2017; Valavi et al., 2019; Ploton et al., 2020). These findings motivate our geographically disjoint frozen-probe protocol and the separate unseen-city evaluation.

S1.5 Pretraining Relations and the Position of HESTIA

Urban pretraining methods also differ in their supervisory relations. Contrastive methods may construct relations among distinct regions: GURPP uses spatial context for positive and negative examples (Jin et al., 2025); MuseCL derives multimodal regional relations from POI and mobility observations (Yong & Zhou, 2024); and MVJC learns structure-aware relations to mitigate false negatives (Lin et al., 2026). Other methods align multiple views of the same unit, including ReCP and ComSRE for regional views, UrbanCLIP for regional imagery and text, and CityFM for entity-level modalities (Li et al., 2024; 2026; Yan et al., 2024; Balsebre et al., 2024). These objectives provide useful supervision, but cross-view consistency or predefined similarity does not necessarily capture functional relations across spatial levels. HESTIA is predictive rather than contrastive: it reconstructs masked local functions and regional fabric from the visible observations, so no positive or negative region pairs have to be specified. The representation space is therefore never committed to a predefined notion of inter-region similarity, while the hierarchy-preserving input keeps fine-grained entity composition available at every level.

Table S1: **Notation.** Symbols used in the main text, grouped by the component that introduces them.

Urban fabric tokenizer (§3.1)			
$\mathcal{V}$	entity-category vocabulary	$\boldsymbol{x}_c$	learned category embedding
$\mathcal{V}_b$	categories present in block b	ϕ_v, ϕ_s	value and support projections
v_{bc}	aggregate value of c in b	f_{tok}	attribute-fusion MLP
s_{bc}	spatial support of c in b	$\boldsymbol{x}_{bc}$	entity-category token
Hierarchical input (§3.1)			
C	cities per batch	$\mathcal{B}_r^{\text{vis}}, \mathcal{B}_r^{\text{mask}}$	visible and masked blocks of r
$\mathcal{R}$	regions in the batch	$\mathcal{B}^{\text{vis}}$	all visible blocks in the batch
$\mathcal{B}_r$	input blocks of region r	R, B, T	padded region, block, token axes
$\boldsymbol{H}^{(0)}$	input tensor	D	hidden dimension
Urban Interaction Transformer (§3.2)			
L_1, L_2, L_3	layers per attention stage	$\boldsymbol{h}_b$	contextualized block state
d_h	attention head dimension	$\boldsymbol{u}_r$	region representation
$\boldsymbol{M}^{\text{I}}, \boldsymbol{M}^{\text{II}}, \boldsymbol{M}^{\text{III}}$	padding masks per stage		
Pretraining objectives (§3.3)			
$\rho_{\text{blk}}, \rho_{\text{ent}}$	block and category mask ratios	$\mathcal{V}_b^+, \mathcal{V}_b^-$	positive and negative queries
v_{rc}	regional aggregate target	$\widehat{v}_{rc}$	its prediction

S2 Algorithmic Flow of HESTIA

Table S1 collects the symbols used throughout the paper. Algorithm 1 summarizes end-to-end HESTIA pretraining over E epochs and mini-batches of C cities. Each batch is converted into one hierarchy-preserving tensor and passed through all three UiT stages before the reconstruction losses jointly update the tokenizer, UiT, and decoders. The tensor keeps its five-dimensional shape throughout the forward pass. Only the active attention axis changes from entity to block to region.

Algorithm 1: End-to-end HESTIA pretraining.

Input: Cities $\mathcal{D}$ with nested regions, blocks, and POI/AOI entities; entity-category vocabulary $\mathcal{V}$
Hyperparameters: Epochs E; cities per batch C; mask ratios $\rho_{\text{blk}}, \rho_{\text{ent}}$; depths L_1, L_2, L_3
Output: Urban Fabric Tokenizer, UiT, and reconstruction decoders

for $e \leftarrow 1$ **to** E **do**
  foreach *mini-batch* $\mathcal{S} \subset \mathcal{D}$ *of* C *cities, with region set* $\mathcal{R}$ **do**
    `/* Sampling & masking (Sec. 3.3) */`
    foreach *region* $r \in \mathcal{R}$ **do**
      $\mathcal{B}_r \leftarrow$ internal blocks of r together with neighboring blocks outside r
      Traverse the nearest-neighbor graph of $\mathcal{B}_r$ breadth-first from a random internal block until the visited set covers a ρ_{blk} fraction of $\mathcal{B}_r$
      $\mathcal{B}_r^{\text{mask}} \leftarrow$ visited blocks; $\mathcal{B}_r^{\text{vis}} \leftarrow \mathcal{B}_r \setminus \mathcal{B}_r^{\text{mask}}$
      foreach *visible block* $b \in \mathcal{B}_r^{\text{vis}}$ **do**
        $\mathcal{V}_b^+ \leftarrow$ a ρ_{ent} fraction of $\mathcal{V}_b$; $\mathcal{V}_b^- \leftarrow$ equally many categories drawn from $\mathcal{V} \setminus \mathcal{V}_b$
        Aggregate same-category entities and encode $\boldsymbol{x}_{bc}$ for every $c \in \mathcal{V}_b \setminus \mathcal{V}_b^+$ `// Eq. 1`
    $\mathcal{Q} \leftarrow \{(b, c) \mid b \in \mathcal{B}^{\text{vis}},\ c \in \mathcal{V}_b^+ \cup \mathcal{V}_b^-\}$ with targets $y_{bc} = \mathbf{1}[c \in \mathcal{V}_b]$ taken before masking
    Assemble $\boldsymbol{H}^{(0)}$ of shape $C \times R \times B \times T \times D$ from `[city]`, `[reg]`, `[blk]`/`[mask_blk]`, the tokens $\boldsymbol{x}_{bc}$, `[mask_ent]`, and `[pad]`
    `/* UiT forward pass: the tensor keeps its shape and only the attention axis changes */`
    for $\ell \leftarrow 1$ **to** L_1 **do**
      Entity attention along the token axis of $\boldsymbol{H}_{i,r,b,:,:}$ for every (i, r, b) with $b \in \mathcal{B}^{\text{vis}}$ `// Eq. 2`
    Add an internal/neighboring role embedding to each block-leading slot $\boldsymbol{H}_{i,r,b,0,:}$
    for $\ell \leftarrow L_1 + 1$ **to** $L_1 + L_2$ **do**
      Block attention with 2D RoPE along the block axis of $\boldsymbol{H}_{i,r,:,0,:}$ for every city–region pair (i, r) `// Eq. 3`
    $\boldsymbol{h}_b \leftarrow \boldsymbol{H}_{i,r,b,0,:}$ for every $b \in \mathcal{B}^{\text{vis}}$
    Add a region-area embedding to each `[reg]` slot $\boldsymbol{H}_{i,r,0,0,:}$
    for $\ell \leftarrow L_1 + L_2 + 1$ **to** $L_1 + L_2 + L_3$ **do**
      Region attention with 2D RoPE along the region axis of $\boldsymbol{H}_{i,:,0,0,:}$ for every city i `// Eq. 4`
    $\boldsymbol{u}_r \leftarrow \boldsymbol{H}_{i,r,0,0,:}$ for every $r \in \mathcal{R}$
    `/* Hierarchy-aligned reconstruction */`
    Local-function decoder: $\widehat{p}_{bc} \leftarrow \sigma(\text{FFN}([\boldsymbol{h}_b; \boldsymbol{x}_c]))$ for every $(b, c) \in \mathcal{Q}$
    Regional-fabric decoder: a linear head on $\text{FFN}(\boldsymbol{u}_r)$ yields $\widehat{v}_{rc}$ for every $r \in \mathcal{R}$ and $c \in \mathcal{V}$, regressed against corpus-standardized targets with POI counts entering as $\log(1 + \cdot)$ and AOI categories as area shares
    Evaluate $\mathcal{L}$ and backpropagate through the decoders, UiT, and tokenizer `// Eq. 5`
return the tokenizer, UiT, and the local-function decoder used by entity-level ranking; the regional-fabric decoder is discarded and no masking is applied at inference

S3 EXPERIMENTAL SETTINGS

This section states everything needed to reproduce the reported numbers: how the corpus and the entity vocabulary are built, what each downstream task is, how the spatial splits and the frozen probe are defined, and how every baseline is read out under the same protocol.

S3.1 DATASET CONSTRUCTION AND PRETRAINING CORPUS

The frozen corpus aligns Overture Maps with 2020 Census TIGER/Line boundaries across the 50 states and Washington, DC. Table S2 summarizes the approximately 263 million U.S. Overture entities and the spatial units they populate. Geographic scope, geometry validation, category mapping, and deduplication are applied before spatial assignment.

The entity-category vocabulary $\mathcal{V}$ of Section 3.1 is built once and then frozen. An entity becomes a POI category when it is a local countable facility, organization, or service, and an AOI category when it is a continuous area that may span several blocks. Nothing appears in both, so the same

Table S2: **Frozen pretraining corpus.** The corpus aligns the U.S. Overture Maps snapshot with 2020 Census TIGER/Line boundaries.

Source or component	Snapshot / coverage	Quantity	Role or definition
Overture Maps entities	Release 2026-07-22.0; United States	$\approx$ 263M	Entity records in the frozen U.S. snapshot
Census blocks	2020 Census TIGER/Line	8,180,866	Finest spatial units in the pretraining hierarchy
Overture-observed tracts	50 states and Washington, DC	84,414	Tracts used for U.S. pretraining and evaluation
Canonical tract rows	Complete Census tract index	85,528	Matrix rows, including locations without Overture observations
County equivalents	50 states and Washington, DC	3,143	Highest-level city contexts used by UiT
Entity-category vocabulary	Frozen urban vocabulary	178 POI + 17 AOI	195 shared entity-category inputs

signal never enters as a count and as an area. Large water bodies, airport grounds, parking areas, power plants, beaches, and land cover are therefore areas, while retail, education, health, transport nodes, and small infrastructure are points. On the POI side we start from the 264 non-empty Overture `basic_category` values observed nationwide and reduce them to 177 named categories under four rules. We merge synonyms and any distinction that no downstream target resolves. We split a category whose members have visibly different spatial distributions. We keep the categories that separate urban from rural, and commercial from industrial, civic and institutional use. We drop the ones an area category already represents. The 17 AOI categories are assembled from the Overture `water`, `infrastructure`, `land`, and `land_cover` themes. A reserved slot for entities whose category cannot be resolved completes the 178 point inputs. Selection uses only the raw taxonomy and national entity counts. No downstream label is consulted. Tables S21 and S23 list the full vocabulary.

POIs that lack an Overture `basic_category` are passed to a constrained classifier built on Qwen3-VL-32B-Thinking, served locally with vLLM in bfloat16. It may return only a category from the frozen vocabulary or an unknown label. We accept a prediction when its mean output-token log-probability exceeds -0.0054, a threshold calibrated on 1,000 manually labeled examples at 95.09% precision, and every rejected case stays unknown rather than being guessed. AOI categories are not imputed in this way because polygon errors would be amplified by area aggregation. POIs are aggregated into within-block category counts. Same-category AOI polygons are unioned before their coverage and spatial support are computed, preventing overlapping geometries from being counted repeatedly. The tokenizer inputs of Eq. 1 follow this construction: for a POI category, v_{bc} is its entity count within the block and s_{bc} is the block area. For an AOI category, v_{bc} is its share of the block's AOI area and s_{bc} is the block's total AOI coverage. HESTIA therefore processes one token per present entity category rather than one token per raw object. The frozen block and tract matrices contain 195 columns ordered as 178 POI categories followed by 17 AOI categories. Block-to-tract membership follows the deterministic Census GEOID hierarchy. No fuzzy spatial join is used. In our experiments, Census tracts instantiate regions and county equivalents instantiate the highest-level city contexts.

What a city is in this corpus. The corpus covers the entire United States, and the top level of the hierarchy has to cover it without gaps. That level has to satisfy three properties. It must partition the country exhaustively, so that no tract is left without a city context. It must nest exactly with tracts and blocks, so that membership follows from the identifier rather than from a spatial join. And it must be defined uniformly nationwide, so that the same unit means the same thing in every state. Municipal boundaries satisfy none of the three. Incorporated places cover only part of the land area, leaving rural and unincorporated tracts outside any place. Places also cross Census lines, so a single place can span several counties and split individual tracts, and incorporation law itself differs by state, which makes a place in one state incomparable to a place in another. Metropolitan statistical areas are exhaustive only over metropolitan and micropolitan territory and leave the remaining tracts unassigned. County equivalents satisfy all three: the 3,143 of them tile the 50 states and Washington, DC, and the Census GEOID nests blocks within tracts within counties by construction.

We therefore instantiate a city by its county equivalent, and every count of cities, every spatial split and every batch in this paper uses that unit. A phrase such as pretraining across thousands of cities refers to the 3,143 county equivalents that tile the country, not to a separate municipal inventory. This choice fixes a context window, not a label. The top level enters the model only through region attention (Eq. 4), where it decides which tracts may attend to one another. It never appears in a downstream target, a probe feature or a split key. Renaming the level would therefore leave every reported number unchanged. In dense metropolitan areas the county contains the municipality

Table S3: **One block as HESTIA reads it.** Census block `360610111002000` in tract `36061011100`, Manhattan, covering 0.0082 km^2. Its ten Overture point entities and two area themes collapse into ten entity-category tokens, one per present category, which occupy the $t > 0$ slots of this block's row in $\boldsymbol{H}^{(0)}$. A point category carries its entity count and the block area. An area category carries its share of the block and the block's total area coverage. Both quantities enter Eq. 1 through ϕ_v and ϕ_s.

Entity category c	Kind	Aggregate value v_{bc}	Spatial support s_{bc} (km^2)
Restaurant or casual food venue	POI	3	0.0082
Cafe or coffee shop	POI	1	0.0082
Non-alcoholic beverage venue	POI	1	0.0082
Home maintenance, repair and improvement service	POI	1	0.0082
Real estate service	POI	1	0.0082
Park or garden	POI	1	0.0082
Multi-unit housing	POI	1	0.0082
Public transit stop	POI	1	0.0082
Parking area	AOI	0.274	0.0105
Built-up cover	AOI	1.000	0.0105

together with its immediate surroundings, and in sparse areas it is the smallest exhaustive container available above the tract.

A worked example. Table S3 shows one block exactly as the tokenizer receives it. Its ten point entities fall into eight categories and two area themes cover it, so the block occupies ten entity-category slots. The number of raw objects inside a block therefore does not enter the sequence length, which is what keeps the entity axis short enough to batch. The parent tract contains 22 such blocks, whose `[blk]` slots form one region row of $\boldsymbol{H}^{(0)}$ together with the neighboring blocks outside the tract, and the tracts of New York County form one city row. Table S4 then shows where those tokens sit in $\boldsymbol{H}^{(0)}$, next to the special and padded slots that surround them.

Only Overture-derived POI/AOI observations and Census geometry enter pretraining. Overture `land_use` polygons and road segments construct downstream land-use and road-density labels, respectively, but never enter HESTIA inputs or objectives. ACS, Decennial Census population, environmental products, downstream labels, and split metadata are likewise excluded. The Overture `land_cover` input theme remains distinct from the `land_use` label source.

S3.2 Downstream Benchmark Inventory

Following CityRep's domain taxonomy (Liu et al., 2026), the region-level benchmark spans morphology, demographics, economy, and environment. Land-use composition is multilabel classification, age distribution predicts an eight-bin probability vector, and the remaining tasks are regression. Table S5 reports each task's source, eligible coverage, target, metric, and spatial split. The main paper presents eight representative tasks and reserves four for the appendix.

The finer-level benchmarks include only tasks supported at their native resolution. Tract labels and coarse environmental rasters are never copied to blocks. As summarized in Table S6, block land use uses exact polygon–block intersections, while block population uses a label-independent deterministic GEOID sample. Their union contains 3,740,243 blocks, and the common finite-row mask used by all methods retains 3,731,700. Blocks inherit the spatial fold of their parent tract.

S3.3 Spatial Splits, Probes, and Reporting

Following recommendations for structured spatial data (Roberts et al., 2017; Valavi et al., 2019; Ploton et al., 2020), we construct seven state-disjoint folds using only Census divisions and tract counts, without reading downstream targets. In rotation k, fold k is test, fold $(k+1) \bmod 7$ is validation, and the other five folds are training. Urban heat follows the same rotation rule but groups complete CBSAs rather than states, preventing a metropolitan area from crossing partitions. Table S7 reports the frozen fold sizes and the separate unseen-city partition.

Table S4: **What one batch looks like inside $H^{(0)}$.** Each row is one (i, r, b) triple of the $C \times R \times B \times T \times D$ tensor and each column one position on the entity axis. Colour marks the token type. Reading the tensor by axis recovers the three stages: Level I attends *along a row* for visible blocks, Level II attends *down the $t = 0$ column* within one (i, r), and Level III attends down that column at $b = 0$ within one city. Region and block slots a city or region does not fill stay `[pad]` and are masked out of every attention. Block-leading slots also carry a role embedding separating internal from neighbouring blocks, drawn here as two shades of blue. In the reported run a process holds $C = 16$ counties, a region carries a median of 144 internal and 937 neighbouring blocks, a block holds a median of 1.5 point and 2.3 area categories, and 18.5% of the tensor is `[pad]`.

Slot (i, r, b)	What the slot holds	$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$		$t = T-1$
City i: New York County, one of the C counties in the batch								
$(i, 0, 0)$	city hub	`[city]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 1, 0)$	region token, tract `36061011100`	`[reg]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 1, 1)$	internal block of Table S3	`[blk]`	`x(restaurant)`	`x(cafe)`	`x(beverage)`	`x(park)`	$\cdots$	`x(transit)`
$(i, 1, 2)$	internal block, one category hidden	`[blk]`	`x(pharmacy)`	`[mask_ent]`	`x(bank)`	`[pad]`	$\cdots$	`[pad]`
$(i, 1, 3)$	internal block, hidden as a whole	`[mask_blk]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 1, 4)$	neighbouring block outside the tract	`[blk]`	`x(hotel)`	`x(parking)`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 1, 5)$	neighbouring block, one category hidden	`[blk]`	`x(office)`	`[mask_ent]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 1, 6)$	neighbouring block, hidden as a whole	`[mask_blk]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 1, 7)$	block whose categories fill the entity axis	`[blk]`	`x(office)`	`x(bank)`	`x(clinic)`	`x(school)`	$\cdots$	`x(built-up)`
$(i, 1, B-1)$	block slot this region does not use	`[pad]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 2, 0)$	region token, next tract	`[reg]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 2, 1)$	internal block of that tract	`[blk]`	`x(school)`	`x(built-up)`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, 3, 0)$	region token, third tract	`[reg]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i, R-1, 0)$	region slot this city does not use	`[pad]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
City $i+1$: the next county in the same batch								
$(i+1, 0, 0)$	city hub	`[city]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i+1, 1, 0)$	region token	`[reg]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i+1, 1, 1)$	internal block	`[blk]`	`x(warehouse)`	`x(cropland)`	`[pad]`	`[pad]`	$\cdots$	`[pad]`
$(i+1, 1, 2)$	internal block, hidden as a whole	`[mask_blk]`	`[pad]`	`[pad]`	`[pad]`	`[pad]`	$\cdots$	`[pad]`

Table S5: **Complete region-level benchmark.** † marks tasks reported in the appendix rather than the eight-task main table. State denotes the shared state-disjoint rotations. CBSA denotes the separate metropolitan-area-disjoint rotations used for urban heat.

Group	Task	Source / year	Tracts	Target	Primary metric	Split
Morphology	Land-use composition	Overture `land_use`, 2026-07-22.0	57,581	20-label meaningful presence	Macro-AUPRC ↑	State
Morphology	Road density	Overture road segments, 2026-07-22.0	84,095	Drivable-road km / land km^2	R^2 ↑	State
Demographics	Population density	ACS 2024 5-year, B01003	83,202	Persons / land km^2	R^2 ↑	State
Demographics	Age distribution	ACS 2024 5-year, B01001	82,594	Eight-bin population distribution	KL ↓	State
Economy	Nighttime lights	NASA Black Marble VNP46A4, 2024	84,069	Area-weighted mean radiance	R^2 ↑	State
Economy	Median household income	ACS 2024 5-year, B19013	81,994	Inflation-adjusted 2024 USD	R^2 ↑	State
Environment	CO_2 emission intensity	EDGAR 2025 product, 2024 emissions	84,069	tonnes CO_2 / land km^2 / year	R^2 ↑	State
Environment	Urban heat anomaly	Landsat 8/9 Collection 2, 2024 JJA	69,653	Tract LST minus CBSA median	R^2 ↑	CBSA
Appendix	$PM_{2.5}$ concentration†	ACAG V6GL02.04, 2023	84,068	Mean surface $PM_{2.5}$ (μg/m^3)	R^2 ↑	State
Appendix	Poverty rate†	ACS 2024 5-year, B17001	82,313	Population below poverty / supported population	R^2 ↑	State
Appendix	Unemployment rate†	ACS 2024 5-year, B23025	82,418	Unemployed / civilian labor force	R^2 ↑	State
Appendix	Bachelor-or-higher rate†	ACS 2024 5-year, B15003	82,464	Share of age-25+ population	R^2 ↑	State

Table S6: **Block- and entity-level benchmarks.** "Examples" denotes labeled blocks for block-level evaluation and matched single-entity queries, over places or buildings, for entity-level recovery. Entity evaluation is frozen native scoring without a downstream classifier.

Level	Task	Source	Examples	Target / protocol	Metrics	Spatial setting
Block	Land-use composition	Overture `land_use` polygons	3,124,491	20-label presence; 50 m^2 minimum area	Macro-AUPRC, Macro-F1	State rotations
Block	Population	2020 Decennial Census `POP20`	1,000,000	Deterministic GEOID sample; $\log(1+y)$ target	R^2, original-scale MAE	State rotations
Entity	Places	Overture-derived queries	1,814,537	Contextual category recovery; target evidence removed	Hit@1, MRR	Nationwide, shared queries
Entity	Buildings	Overture-derived queries	550,183	Contextual category recovery; target evidence removed	Hit@1, MRR	Nationwide, shared queries

Table S7: **Frozen spatial partitions.** The seven state folds partition all 51 state units and 84,414 Overture-observed tracts. In rotation k, fold k is test, fold $(k+1) \bmod 7$ is validation, and the remaining folds are training. The unseen-city protocol fixes a separate 2,011/503/629 county partition.

(a) State-disjoint folds

Fold	0	1	2	3	4	5	6
States	8	6	7	7	8	7	8
Tracts	12,274	12,108	12,026	11,964	12,011	12,083	11,948

(b) Strict unseen-city partition

Role	Counties	Pretraining
Probe training	2,011	Yes
Probe validation	503	Yes
Unseen test	629	No
Total	3,143	—

The main benchmark is transductive and spatially disjoint: probe training and testing are geographically disjoint, while all unlabeled U.S. tracts may participate in representation pretraining. The unseen-city experiment instead excludes its 629 test counties from both encoder pretraining and probe fitting. Its ten repeated probe fits keep the county partition fixed and vary only the optimization seed.

For region- and block-level evaluation, the common probe is a one-hidden-layer MLP with width 1,024, GELU activation, dropout 0.1, and a linear output head. Multilabel land use uses independent sigmoid outputs. We optimize with AdamW at learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, at most 100 epochs, and validation patience 10. Rotation k uses seed $42 + k$. Features are standardized and skewed non-negative targets are modeled in $\log(1+\cdot)$ and mapped back for reporting. Both transformations are fitted on the training fold only, and validation loss alone selects the probe checkpoint. Multilabel land use uses a fixed decision threshold of 0.5, which is never tuned on any split. We report the arithmetic mean and population standard deviation across the seven test rotations. The ridge probe is a control only and never replaces the common MLP for an individual baseline.

Entity-level evaluation is frozen contextual category recovery rather than a trained downstream classifier. Before encoding, we remove the target category's identity, numeric value, and spatial support, then rank the fixed candidate vocabulary with each method's native score. The context is the rest of the anchor's own block: the remaining entity categories keep their values and spatial supports, the unresolved-category slot is zeroed so that unlabeled entities cannot serve as evidence, and no neighboring block is supplied. Each query is encoded as its own single-block city, so the block and region stages cannot borrow information across benchmark rows. The queries and the candidate vocabulary come from the frozen national corpus and are identical for every method. No method is fine-tuned for the task: each is read out by its own native scoring rule on inputs from which the target category has been deleted.

S3.4 Baselines under a Common Protocol

The comparison covers 21 baselines, grouped by the unit each method represents. Each evaluation uses the subset applicable to its granularity: 18 of them enter the region-level comparison, while SpaBERT, HGI and SPPE appear only at the block and entity levels.

- **Region-centric representations** encode a bounded polygon from its aggregated attributes: MVURE (Zhang et al., 2020), HREP (Zhou et al., 2023), HAFusion (Sun et al., 2024), ReCP (Li et al., 2024), ComSRE (Li et al., 2026), MoRA (Wen et al., 2026), MVJC (Lin et al., 2026), and READ (Chen et al., 2026).
- **Entity-centric representations** encode individual geospatial objects and their context: Place2Vec (Yan et al., 2017), SPPE (Huang et al., 2022), SpaBERT (Li et al., 2022), HGI (Huang et al., 2023), CityFM (Balsebre et al., 2024), GURPP (Jin et al., 2025), and UrbanMFM (Zhang et al., 2026b).
- **Broad-coverage geospatial representations** encode locations from imagery, coordinates, or language: SatCLIP (Klemmer et al., 2025), SIREN-M (Rußwurm et al., 2024), UrbanCLIP (Yan et al., 2024), AlphaEarth (Brown et al., 2025), S2Vec (Choudhury et al., 2026), and TESSERA (Feng et al., 2026a).

Baselines that release pretrained weights are used as released. The remaining baselines are pretrained under their published recipes on the same national corpus HESTIA uses, which grants them the same geographic coverage rather than the single-city corpora they were originally reported on. They are then frozen, so every method reaches the probe through the identical spatial protocol, probe architecture, and metric. No baseline representation is retrained on downstream labels. For AlphaEarth, we average available pixel embeddings within each target unit. For S2Vec, we use the corrected SRAI-aligned encoder and pool its cell embeddings. UrbanMFM encodes canonical Census blocks with its frozen national encoder, followed by an unweighted within-tract mean for region-level evaluation. GURPP uses the POI and road inputs its public release supports. That road input overlaps the source of the road-density label, which works in its favour on that task. Place2Vec pools its frozen category table with the original POI-count weighting. UrbanCLIP uses its released 768-dimensional tract representation without appending tract area or handcrafted features. Other coordinate and image baselines follow their released inference procedures without task-specific feature engineering. Each baseline's input modality, checkpoint, pooling rule, and source manifest are stored with its fold metrics.

All methods share the same eligible IDs, labels, spatial rotations, probe hyperparameters, and training-only transformations within each task. Downstream labels, GEOIDs, and split metadata never enter probe features. Test partitions are excluded from normalization, threshold selection, early stopping, and model selection. Entity-level scoring removes all target-category evidence before encoding, and unseen-city test regions never enter the corresponding encoder's unsupervised pretraining. HESTIA checkpoints are selected solely by the unsupervised validation objective, without downstream validation or test statistics.

S4 Extended Results, Scaling, and Robustness

The main text reports one number per task per level. This section gives the full comparison at each of the three levels, then asks how far the result survives changes it was not tuned for: a different probe, an unseen city, a smaller corpus, a smaller encoder, missing inputs, and the sparsest fabric in the country.

S4.1 Extended Region-Level Results

All-city comparison. Panel A of Table 1 compares HESTIA with eighteen baselines under one frozen-probe protocol, and HESTIA is first on all eight tasks. The size of the margin, rather than its sign, is what separates the task families. Where a label is determined by what a place contains, the gain over the best baseline is large: population density improves by 0.167 R^2 (29% relative to AlphaEarth), nighttime lights by 0.089, household income by 0.073 over S2Vec, and CO_2 intensity by 0.038 (15%). Where a label is largely an overhead radiometric quantity, the gain shrinks to the width

Table S8: **The four reserved region-level tasks.** All entries use the same spatially disjoint frozen-probe protocol as the main table. Values are R^2 with standard deviations over the seven spatial folds. The best value per row is dark red and the runner-up is underlined blue.

Task	HESTIA	S2Vec	MoRA	CityFM	ComSRE	TESSERA	AlphaEarth
$PM_{2.5}$	**0.436 ± 0.15**	0.387 ± 0.16	0.265 ± 0.13	0.213 ± 0.10	0.083 ± 0.08	**0.417 ± 0.38**	0.315 ± 0.63
Poverty	**0.473 ± 0.05**	**0.394 ± 0.05**	0.329 ± 0.04	0.310 ± 0.05	0.273 ± 0.07	0.110 ± 0.09	0.047 ± 0.23
Unemployment	**0.163 ± 0.06**	**0.116 ± 0.07**	0.105 ± 0.05	0.099 ± 0.05	0.057 ± 0.07	0.017 ± 0.04	0.049 ± 0.05
Bachelor+	**0.699 ± 0.04**	**0.619 ± 0.04**	0.549 ± 0.04	0.492 ± 0.03	0.371 ± 0.04	0.206 ± 0.08	0.256 ± 0.13
Mean	**0.443**	**0.379**	0.312	0.279	0.196	0.188	0.167

of the error bars: urban heat is a tie with AlphaEarth (0.563 versus 0.561) and land use improves 0.028 AUPRC over Place2Vec.

The ordering of the baselines is equally informative. The two runner-ups, S2Vec and AlphaEarth, are broad-coverage models that see the whole country through regular cells and satellite pixels. They dominate the radiometric tasks but fall behind precisely on the composition-driven ones, which is the gap that an entity-composed hierarchy is meant to close. Region-centric baselines built on aggregated tract attributes occupy the bottom of the table on the regression tasks, with several producing negative R^2 on population density, because a single pooled vector per tract cannot recover a quantity that depends on how fabric is distributed inside it. Place2Vec is the strongest entity-centric reference and is second on land use and age, confirming that raw category composition already carries much of the morphological signal, but it has no mechanism for spatial context and trails by 0.58 R^2 on population density.

The four reserved tasks. Table S8 reports the region-level tasks held out of the main table, under the same spatially disjoint frozen-probe protocol. The columns list the six strongest baselines on these tasks, including every one that is the best baseline on at least one of them. HESTIA attains the best value on all four and the best mean (0.443 against 0.379 for S2Vec). The margins follow the same pattern as the main table: the two socioeconomic rates that depend on fine-grained composition gain most, with poverty at 0.473 against 0.394 and the education rate at 0.699 against 0.619, whereas $PM_{2.5}$, a smooth atmospheric field, is the narrowest case at 0.436 against 0.417 for TESSERA. Unemployment is the hardest target for every method, since a tract-level rate derived from survey estimates is only partly determined by the built environment any representation can observe; HESTIA still attains the best value on it. The satellite baselines behave inconsistently here: AlphaEarth reaches 0.315 on $PM_{2.5}$ but collapses to 0.047 on poverty with a 0.23 spread, so their advantage is confined to targets that are visible from above.

Probe choice. Because every method is read out by one frozen-representation probe, the ranking could in principle be a property of that probe rather than of the representations. Table S9 repeats the comparison with a ridge probe on the same representations and folds. Nearly every column loses a little under the linear readout, HESTIA included: road density falls from 0.896 to 0.873 and household income from 0.602 to 0.582, which is expected when a linear readout replaces a nonlinear one. The ordering is unchanged: HESTIA is first on all eight tasks under both probes. The two closest columns are urban heat against AlphaEarth (0.563 against 0.561 under the MLP and 0.540 against 0.538 under ridge) and CO_2 intensity against S2Vec (0.286 against 0.283 under ridge), the two targets least determined by composition. The margins therefore come from what the representations encode, not from the capacity of the probe that reads them.

Unseen-city generalization. The strict protocol removes 629 counties from encoder pretraining and from probe fitting, so a held-out city is encoded from its own entities by a model that has never seen it. The neighboring blocks that Algorithm 1 adds to a region are filtered to the same partition, so a retained tract never reads blocks from a held-out county. Panel B of Table 1 restricts the comparison to the baselines that can satisfy it, namely those we pretrain ourselves on the retained cities, together with Place2Vec, whose frozen category table carries no city-specific parameters. Two groups cannot be held to this protocol. Methods that learn a per-region embedding table cannot encode a city they never saw. Released global encoders have already been pretrained over the held-out cities, so admitting them would credit them with cities that HESTIA is denied. Panel A already reports that comparison on equal terms. The panel reports means over the nine Census divisions of the held-out cities. HESTIA leads all eight tasks, by 0.221 R^2 on population density (38% over S2Vec), 0.147 on

Table S9: **Probe control.** The same frozen representations and spatial folds are read out twice, by the common one-hidden-layer MLP and by a ridge probe. The MLP column reproduces Table 1 on these tasks, so the two columns isolate the readout. Age distribution is a KL divergence (lower is better). The best value per column within each block is dark red. Population is the count variant of that target. Every other column follows the protocol of Table 1, which the MLP block reproduces.

Variant	Road	Land use	Lights	Income	CO_2	Heat	Pop.	Age↓
Common MLP probe								
HESTIA	**0.896**	**0.554**	**0.836**	**0.602**	**0.299**	**0.563**	**0.380**	**0.066**
S2Vec	0.773	0.487	0.740	0.529	0.261	0.378	-0.048	0.074
AlphaEarth	0.856	0.465	0.747	0.120	0.212	0.561	-0.336	0.096
MoRA	0.642	0.499	0.546	0.428	0.136	0.286	-0.142	0.074
Ridge probe on the same folds								
HESTIA	**0.873**	**0.554**	**0.806**	**0.582**	**0.286**	**0.540**	**0.254**	**0.069**
S2Vec	0.764	0.471	0.730	0.490	0.283	0.360	-0.035	0.077
AlphaEarth	0.784	0.442	0.718	0.044	0.240	0.538	-0.202	0.092
MoRA	0.532	0.489	0.414	0.343	0.118	0.273	-0.027	0.078

Table S10: **Unseen-city generalization.** HESTIA, S2Vec and CityFM are pretrained only on 2,514 seen counties. Place2Vec contributes its frozen category table. All are evaluated on 629 counties that never appear in pretraining. Probes are trained on 2,011 seen counties, selected on 503 held-out seen counties, and tested on the unseen counties. We report tract-micro means over ten probe seeds. Age distribution is a KL divergence (lower is better). All other tasks are higher-is-better. The best value per row is bold dark red.

Task	**HESTIA**	S2Vec	CityFM	Place2Vec
Road density	**0.933 ± 0.00**	0.800 ± 0.01	0.642 ± 0.01	0.697 ± 0.01
Land-use composition	**0.557 ± 0.00**	0.494 ± 0.00	0.446 ± 0.00	0.536 ± 0.00
Population density	**0.867 ± 0.01**	0.595 ± 0.04	0.530 ± 0.17	0.541 ± 0.03
Age distribution (KL ↓)	**0.064 ± 0.00**	0.071 ± 0.00	0.071 ± 0.00	0.069 ± 0.00
Nighttime lights	**0.869 ± 0.00**	0.804 ± 0.01	0.694 ± 0.01	0.635 ± 0.01
Median income	**0.648 ± 0.01**	0.579 ± 0.00	0.414 ± 0.01	0.488 ± 0.00
CO_2 intensity	**0.242 ± 0.05**	0.184 ± 0.07	0.171 ± 0.03	-0.057 ± 0.04
Urban heat anomaly	**0.554 ± 0.01**	0.436 ± 0.01	0.298 ± 0.01	0.341 ± 0.01
$PM_{2.5}$ concentration	**0.472 ± 0.01**	0.439 ± 0.01	0.244 ± 0.01	0.262 ± 0.01
Poverty rate	**0.482 ± 0.01**	0.431 ± 0.01	0.369 ± 0.01	0.376 ± 0.01
Unemployment rate	**0.178 ± 0.01**	0.153 ± 0.01	0.139 ± 0.00	0.149 ± 0.01
Bachelor-or-higher rate	**0.692 ± 0.00**	0.602 ± 0.00	0.498 ± 0.01	0.556 ± 0.01

urban heat and 0.141 on road density, and by as little as 0.021 AUPRC on land use, where the frozen Place2Vec category table remains a strong reference.

Two further properties matter more than the ranking. First, the geographic spread of HESTIA is the narrowest on five of the eight tasks, most clearly on population density (± 0.05 against ± 0.11) and road density (± 0.02 against ± 0.05). It is comparable on land use and age and wider only on household income, where Place2Vec varies less. A model whose quality is even across divisions is the one that can be deployed to an arbitrary new city. Second, grouping by division exposes failures that seed averaging hides: CityFM reaches -0.115 ± 0.878 on population density and Place2Vec -0.543 ± 0.826 on CO_2, which are not uniform degradations but complete breakdowns on entire divisions. CO_2 is the one task we decline to claim: at 0.004 ± 0.278 HESTIA is nominally first yet explains essentially none of the variance, and its ordering against S2Vec lies well inside the spread.

Table S10 extends the same experiment to all twelve tasks under the complementary aggregation, taking the mean over test tracts and the spread over ten probe seeds. Averaging over tracts is dominated by the dense divisions that hold most of them, whereas averaging over divisions weights a sparse division as heavily as a metropolitan one. The two therefore diverge wherever quality is uneven across the country, most visibly on CO_2 intensity, which reads 0.242 here against 0.004 in Panel B. We report both rather than the more favorable one. HESTIA is best on all twelve tasks under both aggregations, and the four additional tasks follow the established pattern, with the education rate at 0.692 against 0.602 and $PM_{2.5}$ at 0.472 against 0.439. The two panels of the main table also show that the unseen-city margins are the wider ones on composition-driven targets. That is expected: once a baseline can no longer rely on a region table fitted during pretraining, encoding a city from its own entities becomes decisive.

Table S11: **Representative entity-category recovery.** Concrete categories are selected to cover common commercial, civic, educational, medical, mobility, cultural, and green functions. Each cell reports Hit@1 / MRR (percent) under the frozen native contextual scoring protocol after removing all target-category evidence. Queries and Share are computed over the combined category-recovery benchmark. Each category is shown only once. Bold marks the best displayed value for the category.

Entity category	Queries	Share	HESTIA	CityFM	Place2Vec	UrbanMFM	HGI	SPPE
Restaurant	120,669	5.1%	76.2 / 86.4	24.1 / 50.1	**93.1 / 96.2**	32.5 / 55.5	4.2 / 25.4	6.2 / 23.4
Gas station	22,896	1.0%	31.1 / 53.9	**52.2 / 68.6**	9.0 / 33.9	51.0 / 68.2	22.2 / 45.8	20.9 / 45.6
Park or garden	28,705	1.2%	**59.3 / 75.4**	27.4 / 50.1	14.8 / 39.2	1.7 / 28.2	16.6 / 34.7	16.0 / 32.9
Bank or credit union	17,702	0.7%	30.6 / 57.0	47.8 / 67.7	24.3 / 51.8	34.7 / 55.3	4.2 / 29.0	**62.5 / 78.2**
Place of worship	74,268	3.1%	**39.7 / 60.8**	23.1 / 45.0	26.7 / 51.5	34.6 / 53.4	10.7 / 27.2	8.2 / 25.2
Elementary school	2,528	0.1%	**83.8 / 91.5**	41.5 / 62.8	63.8 / 79.7	45.3 / 66.0	75.1 / 86.1	46.4 / 67.6
Hospital	16,742	0.7%	57.7 / 72.4	58.0 / 74.2	**70.2 / 83.5**	43.1 / 64.2	53.4 / 71.7	50.8 / 70.0
Fire station	58,661	2.5%	**72.7 / 85.0**	47.0 / 67.2	48.2 / 69.8	34.8 / 59.9	33.2 / 57.3	47.4 / 67.7
Library	62,491	2.6%	**70.0 / 83.2**	46.5 / 66.8	45.3 / 66.9	33.9 / 58.9	31.6 / 55.9	46.1 / 66.8
Museum or science attraction	11,956	0.5%	67.7 / 82.2	**69.9 / 83.0**	67.6 / 82.5	64.8 / 78.7	36.3 / 61.3	23.3 / 47.5
Police station	55,865	2.4%	**75.3 / 86.9**	47.9 / 68.0	50.4 / 71.8	35.6 / 60.7	34.3 / 58.3	48.5 / 68.6
Warehouse / large-format retail	94,153	4.0%	89.1 / 94.4	89.8 / 94.5	93.3 / **96.6**	83.6 / 91.0	**93.8** / 96.6	81.2 / 90.1

S4.2 Extended Block-Level Results

A block is the unit HESTIA encodes directly, so the block level is where the representation is read with the least aggregation between the model and the label. Table 3 evaluates two labels that a single pooled vector cannot express: the land-use composition of a block, scored as multilabel classification over its land-use classes, and its resident population. Both labels are constructed at block resolution from independent sources and are never copied down from a tract. None of the six baselines represents a block natively. The broad-coverage models cover the block by pooling the cells or pixels that fall inside it, and the region- and entity-centric models must subdivide a coarser vector or aggregate individual objects. Both routes discard the same thing: which categories co-occur inside the block, and in what proportion. That is exactly what the two labels depend on.

The margins are the largest in the paper. Land-use AUPRC rises from the 0.26 of AlphaEarth, the strongest baseline at this level, to 0.41, and F1 nearly doubles from 0.18 to 0.34. Block population R^2 rises from 0.41 to 0.62 and the mean absolute error falls from 32.4 to 23.8 residents, a 27% reduction. The baseline ordering is also informative: the two satellite models, which reach 0.41 and 0.34 R^2 on block population, beat the urban-representation models at this level because a block is small enough for overhead imagery to resolve, whereas SatCLIP, which encodes only coordinates, is at -0.01 and confirms that location alone carries no block-level signal. Standard deviations across the held-out regions are at most 0.02 for HESTIA on the AUPRC, F1 and R^2 columns, so the gap is not produced by one favourable fold.

S4.3 Extended Entity-Level Results

The entity level tests whether an entity-category token keeps a meaning of its own or merely contributes to a block vector. Before encoding we delete a target category from a block, together with its count and its spatial support, and ask the frozen model to name it from the categories that remain. Nothing is retrained: HESTIA scores candidates with the pretrained local decoder and every baseline with its own native scoring rule, over the same fixed candidate vocabulary and the same queries. The setting is deliberately unforgiving, since the block is encoded in isolation and no neighbouring block is supplied, so the only evidence is the functional company the missing entity keeps.

Table 2 reports the two query sets. HESTIA recovers the deleted category with 40.35 Hit@1 on places and 50.85 on buildings, 11% and 10% ahead of the strongest entity-centric references, Place2Vec and CityFM. Its MRR on places is 60.92 against a Hit@1 of 40.35, so when the exact category is missed it is still ranked near the top rather than far away.

Table S11 breaks the same protocol down over twelve concrete urban functions, chosen to span commercial, civic, educational, medical, mobility, cultural and green uses, with each category appearing once. The gains are concentrated rather than uniform. HESTIA leads on the categories whose presence is implied by a characteristic functional neighbourhood: elementary schools (83.8 Hit@1 against 75.1 for the best baseline), police stations (75.3 against 50.4), fire stations (72.7 against

Table S12: **City-count scaling on all twelve region-level tasks.** Metrics are multiplied by 100. Age distribution is a KL divergence, where lower is better. Each row pretrains on the stated fraction of U.S. cities under a fixed architecture, schedule and mask setting, and every fraction is evaluated on the identical tract IDs, labels and seven spatial rotations. Levels are comparable across the sweep rather than with Table 1, and Figure 9 plots the same trend on the scale of the reported model. The population column is the count variant of that target.

	Morphology		Demographics		Economy		Environment		Additional appendix tasks			
Cities	Road	Land use	Pop.	Age↓	Lights	Income	CO_2	Heat	$PM_{2.5}$	Poverty	Unemp.	Bach.+
20%	91.9	56.1	44.5	6.69	86.4	59.3	49.5	52.2	61.6	46.4	19.1	69.1
40%	92.0	56.9	48.5	6.46	87.4	63.1	51.3	53.3	62.4	48.5	19.7	71.3
60%	92.0	57.6	49.5	6.33	88.0	**65.5**	51.5	53.2	64.9	49.5	20.1	72.6
80%	**92.1**	**58.1**	**50.0**	**6.23**	**88.0**	65.4	**51.5**	**54.1**	**65.4**	**49.8**	**20.6**	**72.9**

48.2), libraries (70.0 against 46.5), parks (59.3 against 27.4) and places of worship (39.7 against 34.6). These are civic and green functions that rarely appear twice in a block and therefore cannot be guessed from frequency, but that sit inside recognisable compositions. The remaining six are settled by simpler cues. Restaurants are recovered at 93.1 by Place2Vec, whose frozen co-occurrence table is close to optimal for the single most common category in the vocabulary, and gas stations at 52.2 by CityFM. Banks (62.5 for SPPE) and hospitals (70.2 for Place2Vec) are recovered better by scores that key on one strong companion category rather than on the composition of the whole block. Warehouses are near the ceiling for every method (81.2 to 93.8), so no method is distinguished by them. The breakdown therefore locates the mechanism behind the aggregate result: the hierarchy contributes precisely where a category has to be inferred from its functional context, which is the regime the entity level is meant to test, while a frequency prior or a single strong companion is already sufficient elsewhere.

S4.4 City-Count Scaling

The corpus spans the whole United States, which is unusual for urban representation learning and expensive to assemble. This experiment asks whether that breadth is necessary or whether a few metropolitan areas would have produced the same representation, and, if it is necessary, for which targets.

Figure 9 varies how many cities enter pretraining while holding the architecture, the schedule, the mask ratios and the entire downstream protocol fixed, so the only difference between points is how much of the country the encoder has seen. Cities are sampled without replacement, so every smaller corpus is a subset of every larger one and no point benefits from a different sampling scheme. Each point is then evaluated with the same frozen probe on the same spatial rotations. The curves report each fraction's gain over the 20% corpus, so they describe a trend rather than absolute levels. Table S12 lists the task-level values for the sampled fractions on this suite's own scale, so they are read across the sweep, while the full-corpus point is the model reported throughout the paper.

Coverage is not uniformly useful, and the split follows the same line as every other experiment in this paper. Morphological targets saturate after a few hundred cities, whereas the socioeconomic and environmental domains keep improving up to full national coverage. Every task improves from 20% to 80% of the cities and ten of the twelve improve monotonically, but the size of the gain separates the domains: road density moves by 0.2 points and land use by 2.0, whereas household income gains 6.1, population 5.5, and $PM_{2.5}$ and the education rate 3.8 each. Street layout repeats itself quickly across American cities, so a model sees most of the available morphological variety early. The composition that predicts income or air quality varies between regions and keeps supplying new signal. The practical reading is that a single-metro corpus is sufficient for morphology and insufficient for exactly the targets that motivate urban representation learning.

S4.5 Pretraining Convergence

Two things about a pretrained checkpoint should be checkable by a reader: that it was selected without touching downstream information, and that the objective it optimizes actually converges rather than being stopped at a convenient point. The loss curves also expose something the benchmark cannot, namely how the two reconstruction terms behave relative to each other during training.

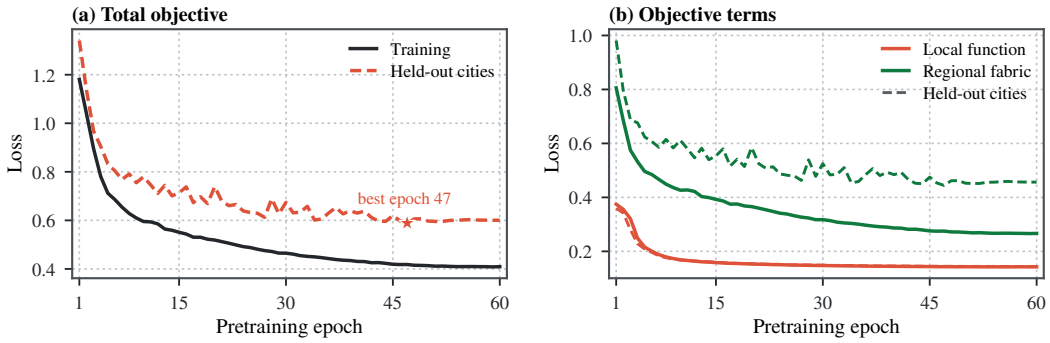


Figure S1: **Pretraining loss of the reported run.** **(a)** Total objective on training cities and on the held-out validation cities, with the selected checkpoint starred. **(b)** The two objective terms, held-out cities dashed.

Table S13: **Pretraining cost of the reported model.** All values are measured from the training log of the released checkpoint. Throughput and padding are per-step medians with their observed range. Peak memory is the maximum over all steps and processes. The padding fraction is the share of the fixed-shape tensor occupied by `[pad]` slots, which is what the hierarchy-preserving layout costs relative to a ragged representation.

Quantity	Value
Setup	
Hardware	7× NVIDIA A100 (80GB), bf16
Pretraining / validation regions	80,181 / 4,233
Schedule	60 epochs, 971 steps per epoch
Cost	
Wall-clock time	7.5 h
Total compute	53 A100-hours
Throughput	180 regions/s (99–235)
Peak memory per GPU	26.9 GiB
Input utilization	
Padding fraction of the tensor	0.185 (0.03–0.38)
Blocks per region	144 internal + 937 surrounding
Longest block axis per batch	1529
Entity categories per block	1.5 POI + 2.3 AOI

Figure S1 traces the reported run: 60 epochs of 971 optimizer steps, of which the first 2,913 are linear warmup, after which the learning rate follows a cosine decay to zero over the remaining steps. Validation is the same unsupervised objective evaluated on 4,233 tracts from 162 counties that never enter training, and it alone selects the checkpoint. The total loss falls from 1.181 to 0.409 on training cities and from 1.341 to 0.589 on held-out ones, where it reaches its minimum at epoch 47 and then stays flat, so the gap between the two curves reflects geographic shift rather than progressive overfitting. The two terms converge at different rates. Local-function reconstruction falls from 0.375 to 0.169 within ten epochs and only to 0.143 over the remaining fifty, and its held-out value tracks the training one to within 0.002. Regional-fabric reconstruction instead keeps improving to the end of the schedule, from 0.806 to 0.267, and carries the entire train–validation gap. Regional structure is therefore the slower and harder half of the objective. That asymmetry runs through the paper: it is the same term whose removal is the most damaging control in Table S18, and the same one whose targets keep improving with corpus size in Figure 9. A model that only had to reconstruct local composition would have finished training in ten epochs.

S4.6 Pretraining Efficiency

A hierarchy over 8.18 million blocks is only useful if it can be trained at that scale, and the tokenization of Section 3.1 was designed around that constraint rather than chosen for elegance. This section reports what the design actually costs and where the cost goes, so that the claim of national-scale pretraining is verifiable rather than asserted.

Pretraining the reported model over the entire United States takes 7.5 hours on seven A100s, or 53 A100-hours, and peaks at 26.9 GiB per process (Table S13). Two properties of the input keep the

cost at this scale. The entity axis is short, because a block contributes one token per present category rather than one per object, and a block holds a median of 1.5 POI and 2.3 AOI categories. Cost is therefore dominated by the block axis: a region carries a median of 144 internal blocks and 937 surrounding ones, and the longest block sequence in a batch has a median of 1,529.

Batching irregular cities into a fixed-shape tensor is paid for in padding, and the measured overhead is 18.5% of the tensor at the median, between 3% and 38% across steps. Context-aware length bucketing holds it in that range by grouping regions of similar block count into the same batch, so a batch is shaped by regions of comparable size rather than by the largest region in the corpus. Throughput averages 180 regions per second, between 99 and 235 across steps as bucket composition changes. The hierarchy-preserving layout of Section 3.1 therefore costs roughly one fifth of the tensor, which is the price of making attention batchable along all three axes. The comparison worth drawing is not against a cheaper layout but against the alternative of not aggregating at all: with one token per raw object rather than per present category, the entity axis would grow by more than an order of magnitude and the same corpus would not fit the same hardware.

S4.7 Model Capacity and Pretraining Cost

The most economical explanation of any gain is capacity, and it has to be ruled out explicitly. Figure 4 of the main text makes the cross-method version of this point. This section makes the within-method version, where everything except size is held fixed and no confound from a different corpus or objective remains.

Table S14 sweeps encoder capacity over a 17× parameter range under an identical recipe, and Table S15 records what each point costs. Downstream accuracy does not improve monotonically with size: the 12.1M encoder attains the best mean on seven of the eleven tasks shown, while the 86.0M one needs 5.2× its pretraining time. The reported model keeps the width and depth of the 12.1M point and adds the region stage, which brings it to the 12.9M reported in the main text. Every count in this paper excludes the reconstruction decoders, which are discarded at inference. This is not a tuning artifact. Every point shares the corpus, objectives, schedule and effective batch, and the widest encoder also attains the highest validation loss of the series (0.6127 against 0.5952), so the extra capacity is not converted into a better fit of the pretraining objective either. Accuracy on this benchmark is therefore set by the input abstraction rather than by encoder size, which is why Hestia reports the compact operating point: it spends no parameters that the hierarchy does not use. Scaling the encoder further would require scaling the supervision alongside it, by extending the corpus beyond one country and adding modalities along the entity axis. The conclusion identifies that as the next step.

The same conclusion holds across methods rather than only within our series. Hestia reaches its accuracy with 12.9M parameters, against roughly 480M for AlphaEarth and 110M for CityFM, both of which trail the strongest baseline on the region-level benchmark, while S2Vec remains competitive at 4.8M (Figure 4). Accuracy on these tasks is therefore uncorrelated with encoder size over two orders of magnitude, and what separates the methods is the input abstraction: models that read a city as entity composition inside a preserved hierarchy do more with fewer parameters than models that read it as pixels or as regular cells.

S4.8 Robustness to Missing Observations

A representation meant for national deployment will be applied to places whose map coverage is thinner than average, so the practical question is how much of it survives when part of the observed fabric is absent from the input. This experiment answers that question by degrading the input on purpose, well past the coverage any maintained corpus would present.

Table S16 deletes a random fraction of observed category tokens at inference and reuses the clean probe, isolating representation robustness from probe refitting. The erasure is applied to the input only. Block and region structure is unchanged, so the model still knows where the missing fabric would have been. Erasure of the magnitude a practitioner actually meets is close to free: at 10% the representation loses 0.008 R^2 on road density and 0.006 AUPRC on land use, and the decline stays gradual rather than cliff-like throughout. Even deleting half of every block's observed categories, far beyond any realistic gap in a maintained map, costs road density only 0.038 R^2 and land use

Table S14: **Parameter-count scaling.** Four encoders share the identical pretraining recipe, corpus and effective batch, and differ only in width, depth and head count. Columns are labelled by encoder parameter count, excluding the reconstruction decoders, which is the convention used throughout the paper. All are evaluated with the same frozen probe over the seven spatial rotations. Downstream accuracy does not increase monotonically with capacity: the 12.1M encoder attains the best mean on seven of the eleven tasks shown, and the reported design adopts its width and depth and adds the region stage, so levels are comparable across the sweep rather than with Table 1. Population is omitted because this suite scores the count variant of that target.

	Model parameters			
Task	5.0M	**12.1M**	37.8M	86.0M
Road density	**0.907 ± 0.02**	0.904 ± 0.02	0.905 ± 0.02	0.903 ± 0.02
Land-use composition	**0.553 ± 0.02**	0.552 ± 0.02	0.550 ± 0.02	0.544 ± 0.02
Age distribution (KL ↓)	0.068 ± 0.01	**0.066 ± 0.01**	0.067 ± 0.01	0.069 ± 0.01
Nighttime lights	0.802 ± 0.04	**0.809 ± 0.03**	0.793 ± 0.04	0.798 ± 0.04
Median income	0.586 ± 0.06	**0.606 ± 0.06**	0.586 ± 0.07	0.562 ± 0.07
CO_2 intensity	0.293 ± 0.09	**0.298 ± 0.10**	0.281 ± 0.11	0.285 ± 0.10
Urban heat anomaly	0.494 ± 0.08	0.491 ± 0.07	**0.508 ± 0.07**	0.496 ± 0.07
$PM_{2.5}$ concentration	0.383 ± 0.13	0.397 ± 0.14	0.386 ± 0.14	**0.406 ± 0.13**
Poverty rate	0.466 ± 0.05	**0.471 ± 0.05**	0.470 ± 0.05	0.454 ± 0.05
Unemployment rate	0.161 ± 0.07	**0.166 ± 0.06**	0.159 ± 0.07	0.159 ± 0.07
Bachelor-or-higher rate	0.692 ± 0.05	**0.697 ± 0.05**	0.697 ± 0.05	0.674 ± 0.05

Table S15: **Pretraining cost across the capacity series.** All four encoders share the corpus, objectives, sixty-epoch schedule, effective batch and hardware, and differ only in width and depth. Rows are labelled by hidden width and *Layers* counts the entity and block stages. Encoder parameters exclude the reconstruction decoders, which are discarded at inference and are the count compared against baselines. Throughput is the steady-state median of the training log and peak memory is the single-process maximum. Gradient accumulation keeps the effective batch fixed, so memory is not monotone in size. All four rows are two-level encoders of this series. The cost of the reported model, which adds the region stage, is reported separately in Table S13. Cost grows almost nine-fold across the series while downstream accuracy does not (Table S14), which is why the reported design keeps the 256-wide encoder. Values are comparable within the series. Table S13 profiles the reported run itself.

Width d	Layers	Encoder (M)	Total (M)	Time (h)	Tracts/s	Peak memory (GiB)	Embeddings (MiB)	Val. loss↓
192	2 / 6	5.04	5.15	4.04	373.2	18.79	61.8	0.6018
256	4 / 8	12.09	12.27	6.82	222.3	47.86	82.4	**0.5952**
384	6 / 12	37.78	38.15	16.73	86.8	42.20	123.7	0.6093
512	8 / 16	86.00	86.63	35.35	39.5	37.30	165.0	0.6127

0.026 AUPRC, so the composition-driven targets hold up under an erasure rate no operational corpus would exhibit. The diffuse environmental fields are the ones that depend on erasure rate, with $PM_{2.5}$ at 0.256 and urban heat at 0.348 once half the fabric is gone, and the split is consistent with what each target needs. Road density and land use are determined by the fabric of the block itself, which retains enough evidence after random deletion, whereas an atmospheric or thermal field is inferred from the surrounding composition, so removing categories across the neighborhood removes the very context the region stage aggregates. Socioeconomic targets fall in between, with household income at 0.526 and the education rate at 0.622 under the same 50% erasure. Each column degrades the same representation's own input, so a row is read across from its clean column rather than against Table 1, where every method keeps its full input.

S4.9 Entity-Sparse Regions

HESTIA builds every representation from observed entities, so the obvious objection is that it should degrade wherever entities are scarce, which is most of the country by area. This is the stress test that objection implies, and it is the one setting where a satellite model has the clearest structural advantage, since imagery is available at uniform quality everywhere.

We rank counties by observed entity density, take the lowest quintile, and keep the largest spatially contiguous component, which yields 460 counties across the interior West and the 1,374 tracts they contain. The region is selected from the input alone. No downstream label enters its definition, and the tracts are test tracts under the ordinary spatial protocol, so nothing about this region was optimized for. The fits pool all of these tracts instead of averaging the seven rotations, so their levels are on their own scale and only the ordering is comparable with Table 1.

Table S16: **Robustness to missing entity-category observations.** At inference we delete a random fraction of the observed category tokens of every block and reuse the clean probe, so the degradation reflects the representation rather than probe refitting. The erased categories are removed from the input only. Block and region structure is unchanged. The suite runs its own corruption harness, so the clean column is the reference for the columns beside it rather than a restatement of Table 1.

Task	0% (clean)	10%	30%	50%
Road density	0.897	0.889	0.882	0.859
Land-use composition	0.557	0.551	0.545	0.531
Age distribution (KL ↓)	0.065	0.067	0.069	0.073
Nighttime lights	0.834	0.810	0.801	0.777
Median income	0.614	0.587	0.567	0.526
CO_2 intensity	0.313	0.284	0.284	0.276
Urban heat anomaly	0.574	0.495	0.462	0.348
$PM_{2.5}$ concentration	0.453	0.364	0.332	0.256
Poverty rate	0.477	0.454	0.439	0.408
Unemployment rate	0.166	0.159	0.152	0.133
Bachelor-or-higher rate	0.700	0.680	0.659	0.622

The advantage does not shrink here; it widens. Figure 5 of the main text shows road density, and Figure S2 repeats the comparison on the two other tasks measured over the same tracts against the three strongest baselines. On nighttime lights HESTIA reaches $R^2 = 0.740$ while AlphaEarth and S2Vec reach 0.370 and 0.059, and on population density it reaches 0.837 against 0.451 and 0.753. Its bias is also the smallest on both tasks, $+0.94$ against $+2.37$ and $+3.24$ on lights and -1.53 against -3.98 and -23.76 on population density, so the advantage holds for calibration as well as for fit. The scatter plots show why: the baselines do not fail randomly in these tracts but flatten towards the conditional mean, predicting a narrow band of values for a wide range of truth, which is what a representation does when its input carries little information about the target. Sparse fabric is still fabric, and a handful of categories in the right arrangement remains more informative about a rural tract than a uniformly available pixel.

S4.10 STRATIFIED ERROR ANALYSIS

A mean over the country could in principle hide an uneven representation, and for an entity-based hierarchy the specific question a reader will ask is whether quality tracks how much input a place happens to have. This analysis answers it directly. Table S17 groups the test tracts by three properties of the input alone, none of which uses a label: how densely entities are observed, how mixed their categories are, and how irregular the tract geometry is.

Absolute scores move across strata, but not in the direction a coverage story would predict. The lowest R^2 on road density, nighttime lights, household income and population density appears in the *densest* third of tracts, not the sparsest. Much of this movement is a property of the strata rather than of the model. R^2 is measured against the target variance inside each group, and dense cores compress the range of exactly those targets, while widening it for land use and CO_2, where the same rows score highest. The comparison that survives this is the gap to the baselines, since every method faces the same within-stratum variance. Measured against AlphaEarth, S2Vec and MoRA on the same tracts, that gap never collapses: its mean across the three groups of each partition stays within 0.260–0.276 on household income and 0.186–0.289 on population density. It is widest where a tract is hardest to compose, namely the top decile by block count, where HESTIA leads by 0.340 on household income, 0.367 on population density and 0.196 on CO_2 intensity. The advantage therefore holds in every stratum, and it is largest exactly where composing a tract is most demanding.

S5 DESIGN AND HYPERPARAMETER CONTROLS

A benchmark result shows that a system works; it does not show which of its parts produce the result, and a hierarchical model offers many places to hide an effect. Every control in this section therefore retrains the encoder from scratch with exactly one decision changed, then evaluates the frozen result under the probe, spatial rotations and metrics of the main benchmark, so a row differs from the reported design only in the stated component. No variant is tuned in its own favour: the corpus, mask ratios, schedule and optimizer follow Section 4.1, and the probe, splits and metrics

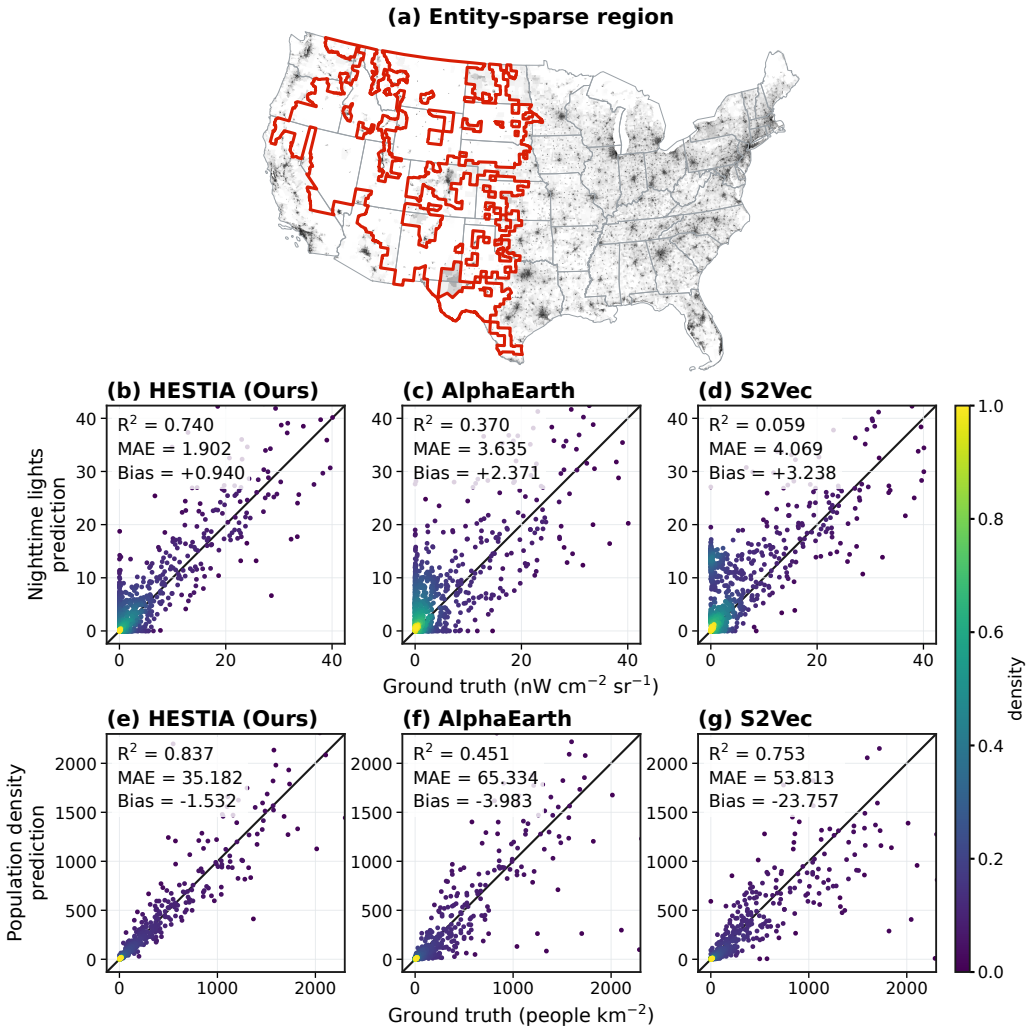


Figure S2: **Out-of-sample prediction in the entity-sparse region.** **(a)** Entity density per tract, with the red outline bounding the 460 sparse counties. **(b)–(d)** Nighttime lights and **(e)–(g)** population density, predicted against observed over the 1,374 tracts inside that outline, with $y = x$ in black and color encoding point density. Rows share a scale within a task. Panels and color follow Figure 5.

follow Appendix S3, so each row isolates one decision instead of one decision plus a retuned recipe. The mask-ratio sweep of Section S5.3 follows the same principle. Its twelve points share one matched schedule, half the length of the reported run, so every setting receives an identical budget and the points are read against one another for the trend they establish. We separate two kinds of decisions: the *objectives* determine what the model must reconstruct and therefore what it has to represent, while the *representational components* determine how the urban fabric is encoded once the objectives are fixed. Reading the two groups apart is what lets us say which choices carry the result and which are interchangeable.

S5.1 Core Ablation

The method makes two kinds of commitment. It claims that reconstructing urban semantics at two levels is the right pretraining signal, and it claims that the tokenizer and the position encoding are the right way to present urban fabric to a Transformer. These are independent claims and they can fail independently, so the core ablation removes one component at a time and reads the two groups apart. Table S18 reports the four controls of Section 4.3 on all twelve region-level tasks, including the four held out of the main comparison, because a design claim should not be settled on the subset of tasks that happens to appear in the main table. Removing the category objective deletes the local

Table S17: **Stratified error analysis.** Test tracts are grouped by three properties of the input alone, with no label used to form the groups and no change to the model. Entity coverage is the density of observed entities and functional mixture is the entropy of a tract's category composition. Each is cut at its terciles, so the three rows hold equally many tracts. The geometry rows are not a partition: they contrast the least compact tenth and the top decile by block count against an ordinary band in which both properties lie between their quartiles, leaving intermediate tracts out. Age distribution is a KL divergence (lower is better). Scores are comparable within a column across rows of the same partition, and the best value per column is dark red.

Stratum	Tracts	Road	Land use	Lights	Income	CO_2	Heat	Pop. dens.	Age↓
Entity coverage									
Lowest third	27,823	0.904	0.502	0.819	**0.669**	0.316	**0.589**	0.784	0.064
Middle third	28,100	0.917	0.542	0.857	0.634	0.339	0.569	0.811	0.057
Highest third	28,083	0.863	0.562	0.806	0.534	0.436	0.550	0.657	0.074
Functional mixture									
Lowest third	27,693	0.901	0.441	0.823	0.593	**0.438**	0.578	0.690	0.068
Middle third	27,822	0.896	0.478	0.828	0.621	0.372	0.569	0.689	0.065
Highest third	27,834	0.896	0.565	0.852	0.641	0.347	0.582	0.689	0.062
Tract geometry									
Ordinary	22,299	0.904	0.535	0.833	0.644	0.241	0.583	0.792	0.063
Least compact	8,290	0.915	0.574	0.868	0.641	0.341	0.587	0.801	0.063
Most blocks	6,788	**0.956**	**0.604**	**0.885**	0.523	0.168	0.548	**0.898**	**0.041**

Table S18: **Core ablation on all twelve region-level tasks.** Metrics are multiplied by 100. Age distribution is a KL divergence, where lower is better. Every other task is higher-is-better. The best value in each column is shown in bold dark red. Each row removes one component: the category objective is local function reconstruction, the regional objective is regional fabric reconstruction, the joint token modulation couples a category's value and spatial support, and the last row replaces 2D rotary position encoding with an additive one.

	Morphology		Demographics		Economy		Environment		Additional appendix tasks			
Variant	Road	Land use	Pop.	Age↓	Lights	Income	CO_2	Heat	$PM_{2.5}$	Poverty	Unemp.	Bach.+
w/o category objective	88.6	**55.9**	71.4	6.71	80.6	57.1	27.9	53.5	39.3	44.3	14.4	67.9
w/o regional objective	88.7	52.9	72.4	7.07	79.5	53.4	27.8	55.3	41.3	43.6	15.2	67.6
w/o joint token modulation	89.5	54.9	74.4	6.75	82.9	57.5	**30.5**	56.5	**44.8**	46.8	15.3	69.1
w/o 2D rotary encoding	89.1	55.1	**78.3**	6.64	82.2	58.7	29.8	**57.9**	41.2	46.9	15.8	68.6
HESTIA	**89.6**	55.4	74.3	**6.62**	**83.6**	**60.2**	29.9	56.3	43.6	**47.3**	**16.3**	**69.9**

reconstruction of which entity categories a block contains; removing the regional objective deletes the reconstruction of a region's aggregate fabric; a third variant replaces the joint token modulation with separable value and support branches; and a fourth replaces 2D rotary encoding with an additive position term.

The two objectives behave very differently from the two representational components. Removing the regional objective loses on all twelve tasks, by 6.8 points of household income, 4.1 of nighttime lights and 3.7 of poverty, which makes it the single most consequential decision here. Removing the category objective loses on eleven of twelve, most heavily on $PM_{2.5}$ (-4.3), household income (-3.1), and nighttime lights and poverty (-3.0 each). Its one gain is land-use composition ($+0.5$), the task most directly readable from raw composition. The two are complementary rather than redundant: the category objective matters most where a target depends on what sits inside a block, the regional objective where it depends on how blocks combine.

The two representational components behave differently, and informatively so. Exchanging either of them moves region-level scores by a fraction of what the objectives move, and without a consistent direction across tasks: replacing the joint modulation stays within 1.2 points either way apart from household income, and replacing 2D rotary encoding within 2.4 points apart from population density. Region-level aggregation is therefore not the scale at which these two choices express themselves, which is expected, since both govern how an individual block is presented to the encoder rather than how blocks combine into a region. We accordingly settle them one level down, where the spatial-encoding controls below are decisive.

Table S19: **Additional design controls.** Each row changes one decision and keeps everything else fixed. The first block removes the region stage, so the regions of a city never attend to one another; the second changes what surrounding blocks a block may read; the third changes how position and role enter the block encoder. Age distribution is a KL divergence (lower is better). The reported-design row reproduces the HESTIA row of Table 1 on these tasks, so each control is read directly against the headline model. Population is reported separately in the text because these suites score the count variant of that target.

Variant	Road	Land use	Lights	Income	CO_2	Heat	Age↓
Region stage							
HESTIA (reported design)	0.896	0.554	0.836	0.602	0.299	0.563	0.066
w/o region stage	**0.904**	0.552	0.809	0.606	0.298	0.491	0.066
Context blocks							
w/o context blocks	0.891	0.550	0.780	0.489	0.258	0.541	0.070
Masked context blocks	0.897	0.552	0.829	0.580	0.299	**0.571**	0.067
Spatial encoding							
Geodesic local coordinates	0.895	0.553	0.832	0.594	0.304	0.569	0.066
w/o role embedding	0.898	0.547	**0.838**	0.597	0.320	0.559	0.066
w/o region-area descriptor	0.890	0.554	0.835	0.598	0.297	0.569	0.066
Absolute geographic position	0.902	**0.555**	0.830	**0.626**	**0.324**	0.556	**0.065**
w/o position	0.897	0.552	0.832	0.587	0.295	0.568	0.066

S5.2 ADDITIONAL DESIGN CONTROLS

Table S19 groups the remaining one-at-a-time controls into three blocks. Every row is read against the reported design in the first row, so levels are comparable within the table rather than with Table 1.

Region stage. The core ablation removes objectives and encodings, but not the third attention stage itself, which is the newest structural component and the one a reader is most entitled to question. The control keeps entity and block attention and pools the resulting block states into a region vector, so a city's regions never attend to one another while everything else is held fixed. Removing it costs 7.2 points of urban heat and 2.7 of nighttime lights, the two targets whose signal extends beyond a single tract, and returns at most 0.8 elsewhere, on road density and household income, which are settled inside it. That exchange is what the third level is for: it buys cross-tract signal at a negligible price on the targets that do not need it.

Context blocks. Exposing a block to its neighbours is what makes the second level spatial rather than merely a second pooling step, and it is also the design choice with the largest cost in sequence length, so it should be shown to pay for itself. The two context rows separate the two things it could supply, the presence and arrangement of surrounding blocks and the contents of those blocks, by removing them entirely and by keeping them while masking what they contain. Removing the surrounding blocks is by far the most damaging change in the table, costing 11.3 points of household income, 5.6 of nighttime lights and 4.1 of CO_2. Keeping those blocks but masking their contents recovers most of it. What a block needs from its surroundings is therefore their layout at least as much as their contents.

Spatial encoding. Geometry can enter a block encoder in several ways, and the reported design uses relative 2D rotary encoding with a learned role embedding that distinguishes internal blocks from context. These rows vary one such component at a time, replacing relative positions with absolute geographic coordinates or with geodesic local coordinates, removing the role embedding, removing the tract-area descriptor, and removing position altogether. Region-level performance is robust to all of them: no variant shifts a region-level task by more than 2.5 points in either direction, so no region-level conclusion in this paper rests on this choice. The choice is settled one level down, where it is decisive: removing the role embedding that separates internal blocks from context collapses block population prediction from $R^2 = 0.36$ to -0.48. We therefore adopt the relative encoding for the block-level behaviour it alone supports.

S5.3 MASK-RATIO SENSITIVITY

Masked modelling has one free parameter per objective, and a reader is right to ask whether a strong result reflects the architecture or a fortunate choice of them. Each ratio also has a concrete meaning here. ρ_{ent} sets how much of a block's own category evidence is hidden, and therefore how hard local

Table S20: **Mask-ratio sensitivity on all twelve region-level tasks.** Metrics are multiplied by 100. Age distribution is a KL divergence, where lower is better. Each sweep point is pretrained under an identical matched schedule, shortened to half the length of the reported run, and evaluated with ten probe seeds on the seven spatial rotations of the main benchmark. Levels are therefore comparable across the sweep rather than with Table 1. The best value in each column is shown in bold dark red and the selected setting is shaded.

	Morphology		Demographics		Economy		Environment		Additional appendix tasks			
Setting	Road	Land use	Pop.	Age↓	Lights	Income	CO_2	Heat	$PM_{2.5}$	Poverty	Unemp.	Bach.+
Panel A: entity-category mask ratio ρ_{ent}, with $\rho_{\text{blk}} = 0.20$ fixed												
0.10	91.7	57.4	48.2	6.28	86.7	64.2	49.5	52.5	60.7	48.9	19.9	71.4
0.20	91.9	57.7	48.6	6.24	87.0	64.6	50.2	52.7	62.2	48.9	20.0	71.7
0.30	91.9	**58.1**	48.3	6.21	87.6	65.2	51.3	52.4	63.6	49.6	20.5	72.6
0.40	92.0	57.9	48.9	**6.20**	**87.9**	**66.1**	52.1	52.8	64.9	**50.0**	20.6	73.1
0.50	92.1	57.7	**50.7**	6.21	87.8	66.0	52.0	**53.3**	65.2	49.7	**20.7**	**73.4**
0.60	**92.1**	57.5	50.7	6.28	87.9	65.3	**52.8**	52.1	**65.5**	49.6	20.6	73.4
Panel B: connected block mask ratio ρ_{blk}, with $\rho_{\text{ent}} = 0.50$ fixed												
0.10	92.0	**57.7**	**51.3**	**6.20**	87.6	65.6	51.0	52.9	63.7	49.8	20.6	73.2
0.20	**92.1**	57.7	50.7	6.21	87.8	**66.0**	52.0	53.3	65.2	49.7	20.7	73.4
0.30	92.0	57.5	49.5	6.26	88.1	65.6	52.1	**53.5**	65.0	**50.1**	20.7	73.2
0.40	92.0	57.3	50.7	6.27	88.1	65.7	52.2	52.2	65.1	49.9	20.7	73.5
0.50	91.9	57.1	49.8	6.32	88.1	65.1	52.7	52.8	66.1	49.8	20.6	**73.6**
0.60	91.8	56.7	48.5	6.36	**88.1**	65.1	**53.5**	52.1	**66.7**	49.7	**20.7**	73.2

reconstruction is. ρ_{blk} hides whole connected groups of blocks, and therefore decides how far across a region the model must reach to fill a hole. They pull in different directions, so the sweep is also a statement about what each level is for.

Table S20 sweeps each ratio over six values from 0.10 to 0.60 with the other held at the reported setting, retraining every one of the twelve points from scratch under a matched schedule and evaluating all of them with ten probe seeds on the seven spatial rotations. Figure 7 of the main text summarises the same sweep as an eleven-task mean, taken over the tasks whose targets are identical across the sweep and the main benchmark. The population column of Table S20 is the count variant and stays out of that mean.

We read that mean as a percentage of the selected setting, so 100 is the reported model. The entity-category sweep rises from 98.1 at $\rho_{\text{ent}} = 0.10$ to its peak at 0.50 and stays within 0.7 points of that peak across 0.30–0.60. The block sweep instead peaks at $\rho_{\text{blk}} = 0.20$ and declines steadily to 98.6 at 0.60. Neither response is knife-edged: the selected settings stay within 0.8 and 1.6 points of the per-task best everywhere. Where they do disagree the pattern is systematic rather than noisy. Raising the block ratio from 0.10 to 0.60 gains 2.5 points of CO_2 and 3.0 of $PM_{2.5}$, the two diffuse environmental targets, while costing 2.8 points of population and 1.0 of land use, which depend on nearby fabric. That trade is the sweep's most informative outcome: hiding larger connected groups forces the model to reconstruct from farther away and yields a representation tuned to fields that vary slowly over space, at the price of the detail that local targets need. We keep $\rho_{\text{blk}} = 0.20$ because it sits at the balance point rather than at either extreme, and $\rho_{\text{ent}} = 0.50$ because the entity sweep has plateaued by then and a harder local task costs nothing.

S6 Representation Diagnostics

This section inspects what the learned representation encodes, using no new training and no label beyond those of the benchmark.

S6.1 Nationwide Similarity Fields

A representation trained on geolocated data can succeed for the wrong reason, by encoding where a place is rather than what it is. If that were happening here, tracts near a query would look similar to it and distant ones would not, and the downstream gains would be a smoothed map rather than an urban representation. We pick three prototype queries using input categories only, with no downstream label in the selection, and compute the centred cosine similarity between the query and every other tract in the country, shown in Figure S3. Tract similarities are aggregated to 25 km equal-area cells

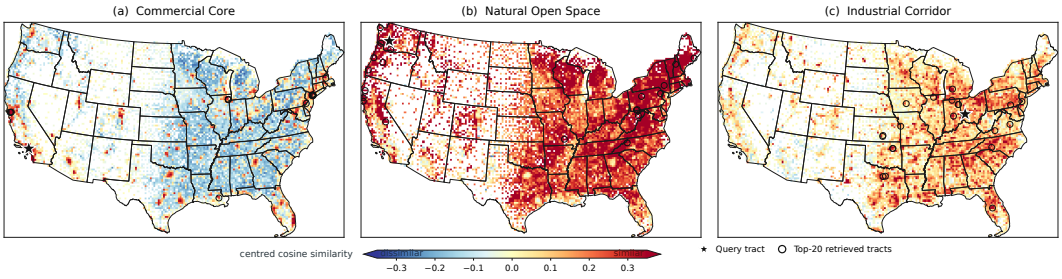


Figure S3: **Nationwide similarity fields for three prototype queries.** Centred cosine similarity between each tract and the query, aggregated to 25 km cells and shown on a symmetric scale about the national mean. Blue is below average and red above. Queries are selected from tracts passing each theme's entity and signature minima, using only input categories and no downstream labels. Query categories are commercial (restaurants, apparel, beauty), natural (parks, rural), and industrial (home services, automotive, manufacturing).

because the largest decile of tracts covers 82.4% of the land area, so filling raw polygons would let the most rural tracts dominate the map. Each cell reports the 90th percentile of the similarities falling inside it, which gives every tract an equal vote regardless of how large its polygon is.

The fields follow urban function rather than proximity. Every query brightens cells across the whole country instead of a neighbourhood around itself: the commercial query lights up metropolitan cores nationwide, the natural query the mountain and forest belts, and the industrial query the manufacturing corridors, none of which is a contiguous region around the query. Section 4.3 quantifies the same effect, with similarity correlating with category composition at partial ρ of 0.61, 0.42 and 0.61 for the commercial, natural and industrial queries after controlling for distance, against 0.13, 0.08 and 0.28 for proximity alone. Figure 6 then checks the retrieved sets against references that enter neither pretraining nor query selection.

S6.2 Similarity Structure at the Tract and Block Levels

Downstream accuracy shows that a representation is useful but not that it encodes the structure we designed it around. This diagnostic asks a question with an observable answer: when the model calls two units similar, are they similar in the entities they actually contain? It is run at all three levels, so the answer also shows how the hierarchy transforms information as it moves upward. The test has to be relational. A category or a unit reaches the model only as an identifier, so no single embedding means anything on its own and only the arrangement of the embeddings can be checked.

At each level we compare the similarity the model assigns with a similarity measured directly from the data, and report their Spearman rank agreement so that the two need not share a scale. Two agreements appear here and should not be confused. Each figure annotates one anchor's own similarity profile, as a guide for the reader following that map. The numbers quoted below are computed over all unit pairs at that level, and the claims rest on those. The main text tests the entity level in Figure 8. This section completes the hierarchy at the two coarser levels, reading each unit through the state the model actually produces there. Blocks use their spatially contextualized Level-II states, compared against the cosine similarity of their raw category compositions over $3{,}298$ Manhattan blocks, and the learned and observed maps brighten the same corridors (Figure S4). Tracts use the exported region embeddings, shown for two contrasting anchors over 310 Manhattan tracts, where the three tracts the model ranks highest sit at observed ranks 1, 2, 5 and 2, 1, 24 (Figure S5).

Rank agreement falls from 0.63 at the entity level to 0.61 at the block level and 0.44 at the region level. The decline is what a working hierarchy should produce: a block state should still look like the categories inside it, whereas a region state that merely reproduced its blocks' composition would add nothing over pooling them. The level at which agreement drops is also the level at which the region stage earns its gains in Table S19, on the targets whose signal extends beyond a single tract.



Figure S4: **Learned block similarity against observed composition.** Two contrasting query blocks over the 3,298 Manhattan blocks around them, rotated so the long axis is horizontal. **(a)** A restaurant and design-services anchor; **(b)** a legal and professional-services anchor. In each pair the upper map is the HESTIA block state and the lower map the raw category composition, and the star marks the query. The line under each pair reports the agreement of that anchor's own similarity profile. The 0.61 quoted in the text is the agreement over all block pairs. Colors are scaled within each panel.

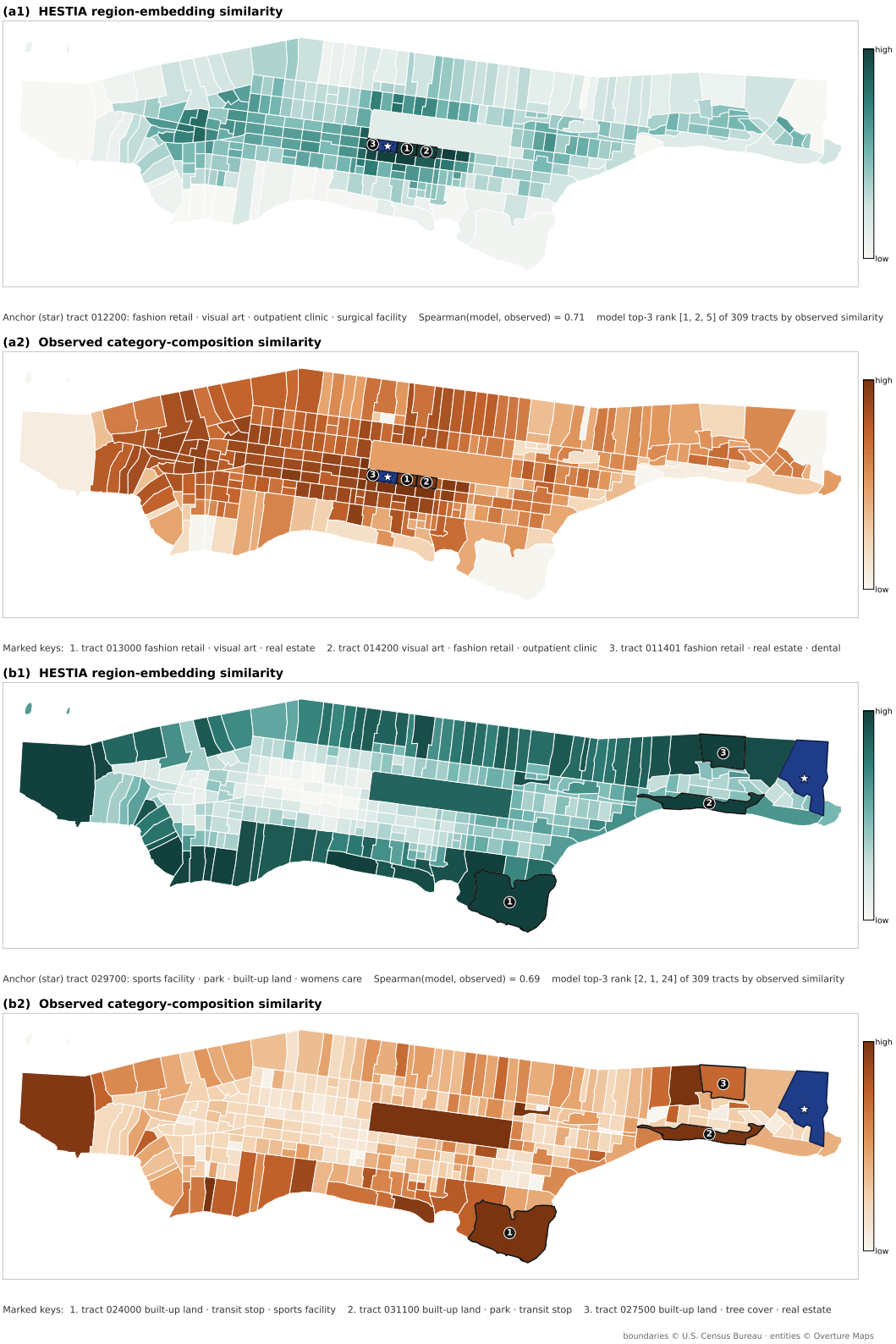


Figure S5: **Learned tract similarity against observed composition.** Two contrasting query tracts over the 310 Manhattan tracts, rotated so the long axis is horizontal. **(a)** A retail and creative-services anchor; **(b)** a recreation and open-space anchor. In each pair the upper map is the exported region embedding and the lower map the raw category composition, the star marks the query and numbered outlines mark the three tracts the model ranks highest. The line under each pair reports the agreement of that anchor's own similarity profile. The 0.44 quoted in the text is the agreement over all tract pairs. Colors are scaled within each panel.

Table S21: **Frozen entity-category vocabulary (1 of 3).** The 178 point and 17 area categories of Section 3.1 index the 195 columns of every block and tract matrix. A point category receives an entity count and an area category an area share, and no entity type appears in both. Entries read down each column in alphabetical order. The grey heading gives the initial-letter span of the column below.

Point categories (POI)

A–B	C–D	D–F
ATM	Cafe or coffee shop	Digital and telecommunication service
Accounting advisory and investment service	Campground	Discount store
Advertising marketing and public relations service	Cannabis or vape store	Dog park
Agricultural service	Casino	EV charging station
Amusement attraction	Cemetery	Electronics store
Animal and pet store	Childcare and family support service	Elementary school
Animal or pet service	Civic or political organization	Emergency or urgent care facility
Arts crafts and hobby store	College university	Employment and staffing service
Auto dealer	Communication tower	Environmental or ecological service
Automotive service	Community or civic center	Event or festival venue
B2B energy and utility service	Complementary and alternative medicine	Event support or ticketing service
B2B transportation and storage service	Corporate or business office	Farm
Bank or credit union	Courthouse	Farmers market
Bar	Creative media production service	Fashion and apparel store
Behavioral or mental health clinic	Cultural or civic public space	Fast food restaurant
Books music and video store	Dam or weir	Ferry terminal
Brewery or distillery	Dental clinic	Fire station
Broadcast and publishing service	Department store	Flowers and gifts store
Building or construction service	Design service	Food and beverage store
Bus station	Diagnostics imaging or lab service	Food bank

Table S22: **Frozen entity-category vocabulary (2 of 3).** Point categories continued.

Point categories (POI), continued

F–I	J–O	O–P
Food delivery service	Jail or prison	Other financial service
Food truck stand	Labor union	Other rental service
Funeral cremation and mortuary service	Laundry service	Other specialty store
Gaming venue	Legal service	Package locker
Gas station	Lending credit and debt service	Park or garden
General lodging or accommodation	Library	Pediatric clinic
Golf or country club	Manufacturing or industrial facility	Performing arts or theatre venue
Government or diplomatic office	Marina	Personal care and beauty store
Gun ammunition or fireworks store	Medical and accessibility supply store	Personal or beauty service
Gym or fitness studio	Middle school	Pharmacy and drug store
Hardware home and garden store	Military site	Physical medicine and rehabilitation
Helipad	Mobile home park or special housing	Place of worship
High school	Movie theater	Playground
Historic or memorial site	Multi-unit housing	Police station
Home health and medical support service	Museum or science attraction	Power substation
Home maintenance repair and improvement service	Music venue	Preschool
Hospital	Musical instrument and pro audio store	Primary care or general clinic
Hotel	Nightlife venue	Printing service
Industrial or specialized technical service	Non-alcoholic beverage venue	Private or religious school
Insurance service	Office supply store	Professional and business support service

Table S23: **Frozen entity-category vocabulary (3 of 3).** The remaining point categories and the 17 area categories. One point slot is reserved for entities whose category cannot be resolved. It is zeroed in the entity-level protocol so that unlabeled entities never serve as evidence.

Point categories (POI), continued

P–S	S	T–Z
Public or specialty market	Skate park	Taxi ride share or shuttle service
Public or unspecified school	Small format convenience retail	Toys and games store
Public transit stop	Social fraternal or veterans organization	Travel service
RV park	Social or community service	Unresolved place (reserved)
Real estate service	Specialized medical or surgical facility	Utility provider
Recreational equipment rental	Specialty or outpatient clinic	Vacation rental or guest house
Recycling point or facility	Specialty school	Vehicle parts store
Religious organization	Specialty vehicle sales and service	Vehicle rental service
Remittance currency and check cashing service	Sport or recreation organization	Vision or eye care clinic
Reproductive perinatal and women's care	Sporting goods store	Visual art and creative space
Research institute or scientific service	Sports and recreation facility	Warehouse club or superstore
Resort	Spring or geothermal feature	Waste or septic service
Restaurant or casual food venue	Stadium arena	Water tower
Rural event or attraction	Storage facility	Waterfall
Rural or small scale lodging	Storage tank	Wellness spa or spiritual service
Second-hand store	Subway station	Winery
Senior living facility	Supplementary education service	Youth organization
Shipping or delivery service	Supplier wholesaler or distributor	Zoo or aquarium
Shopping mall	Surface rail station	
Silo	Swimming pool	

Area categories (AOI)

Aviation facility area	Grassland cover	Shrubland cover
Bare or sparse vegetation cover	Moss and lichen cover	Snow and ice cover
Beach area	Natural flowing water area	Standing inland water area
Built-up cover	Parking area	Tree or forest cover
Cropland cover	Power plant area	Wetland or mangrove cover
Engineered channel area	Reservoir or basin area	