BAViR: Aligning Urban Visual Representations with Revealed Mobility Behavior

Anonymous authors
Paper under double-blind review

Abstract

Urban environment representations should describe how places are used, not only their visible structure. Mobility records reveal this use but need not be available wherever a representation is deployed. We use mobility behavior as training supervision to learn an environmental representation that can be applied where mobility observation are unavailable. We introduce BAViR (Behavior-Aligned Visual Representations), which fuses satellite and street imagery with points of interest and road structure, and a contrastive objective aligns its output with visitation, dwell, activity, and temporal summaries. Since mobility behavior is used for supervision only during training, the learned encoder can represent new locations from environmental observations alone. We evaluate spatial generalization and reuse for targets excluded from encoder training, using two five-fold spatial protocols and three seeds in metropolitan Orlando. On geographically withheld regions, BAViR improves visit-count R^2 from 0.488 to 0.523 over matched masked-stream self-supervision. Moreover, the learned representation also transfers to transaction spend, a target excluded from encoder training, reaching R^2 of 0.355 compared with 0.314 for self-supervision and 0.339 for behavior regression under the same frozen-feature readout. With only 10% of downstream spend labels, BAViR exceeds raw-feature Ridge trained with all labels. Together, these results support the value of behavior supervision for learning reusable functional representations, rather than a universal advantage of contrastive learning.

1 Introduction

A useful representation of an urban place must connect its physical setting with how the place is used by people over time. Satellite imagery reveals buildings, roads, and land cover; street imagery captures the ground-level environment; points of interest describe nearby destinations. Together, these observations characterize the relatively static physical and semantic setting of a place, but they do not directly share either a common semantics or uniform spatial coverage. A deployable encoder must therefore fuse heterogeneous evidence without assuming that every stream is recorded. Moreover, these observations do not directly reveal the pattern of human use, such as whether people visit frequently, how long they stay, what activities occur, or when those activities take place. Mobility records provide this complementary behavioral and temporal evidence, but their coverage is uneven and their use is constrained by access and privacy. A useful representation should therefore capture information about urban use without requiring mobility observations wherever those representations are deployed.

One way to achieve this is to use mobility as training-only behavior supervision without making mobility a required input at deployment. Multimodal urban encoders already combine imagery and semantic context (Wang et al., 2020; Jenkins et al., 2019; Xiao et al., 2024), and recent work uses mobility to enrich visual or geospatial representations (Yong & Zhou, 2024; Wen et al., 2025; 2026). These advances shift the problem from simply fusing heterogeneous data sources to a question about what behavior supervision adds to the learned environment representation. We ask: *can supervision from revealed urban use teach an environment representation that remains useful beyond the behaviors used to supervise it, without requiring mobility observations at deployment?*

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.

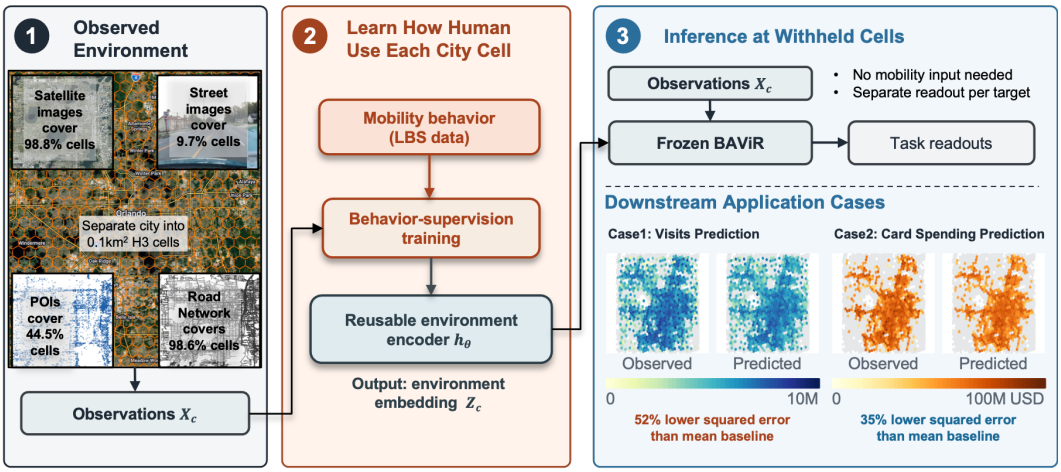

Figure 1: **BAViR learns an environment representation with mobility supervision and reuses it at withheld locations.** Environmental observations form the encoder input, while mobility behavior provides training-only supervision. During inference, the frozen encoder is applied at withheld cells without mobility input, with task-specific readouts for downstream tasks like predictions. Appendix C gives definitions and prediction errors.

We introduce BAViR, which learns to associate environmental evidence with aggregate patterns of visitation, dwell, activity, and time of use. Behavior guides what is learned from imagery, points of interest, and road structure. (Figure 1). In BAViR, we refer to this training paradigm as *behavior supervision*: mobility supervises representation learning during training but is not an input to the environment encoder. This separates the locations where mobility is observed for learning from those where the representation can be applied. The learned encoder is then held fixed across downstream tasks, making information about urban function reusable without retraining the environment model for each outcome.

Evaluating this aim requires more than predicting the behaviors that supplied supervision. We test generalization to spatially withheld locations in metropolitan Orlando and reuse for targets excluded from representation training. Transaction spend provides the clearest evidence: alignment transfers better than matched reconstruction and direct behavior regression, even though behavior regression predicts visits slightly better. This distinction matters because accuracy on the supervisory targets alone does not determine a representation's value for a new task.

Our central claim is that *mobility behavior can supervise an environment representation that retains information useful beyond the mobility summaries used during training, while requiring only environmental observations at inference*. Our contributions are:

1. **Behavior-supervised environment representation learning.** We develop BAViR, which uses human mobility behavior as training supervision to learn multimodal environment representations, while requiring only environmental observations at deployment.
2. **Reusable representations for downstream urban tasks.** We show that the learned representation supports prediction of targets not used to train the encoder, such as recorded card spending and remains useful when downstream training labels are limited.
3. **Controlled evaluation for the learned benefit of behavior supervision.** Two spatial holdout protocols and three seeds distinguish the effect of alignment from fusion architecture, environmental inputs, and direct supervised prediction, including a verified contribution from satellite image content.

2 Related Work

Representing urban environments. Urban2Vec combines street imagery and point-of-interest (POI) information (Wang et al., 2020); RegionEncoder jointly embeds multiple spatial views (Jenk-

ins et al., 2019); ReFound builds on language and vision foundation models (Xiao et al., 2024). Other urban encoders vary the fusion mechanism or spatial and cross-modal pretext objective (Mühlematter et al., 2025; Cao et al., 2026; Li et al., 2024; Liu et al., 2025). This literature establishes the value of combining observations. Our focus is the supervisory signal used to organize the fused representation, holding its inputs and architecture fixed in the principal comparison.

Mobility as supervision and context. Predicting mobility patterns from imagery predates this work (Zhang et al., 2019). More directly related, MuseCL constructs visual contrastive pairs using mobility and POI-distribution similarity (Yong & Zhou, 2024), while MoRA uses a mobility graph as a representation backbone (Wen et al., 2025). MoRAX distills a mobility-aware teacher into a student that requires no mobility input, including for cross-city deployment (Wen et al., 2026). Broader foundation models also incorporate aggregate population dynamics or mobility context (Agarwal et al., 2024; Guo et al., 2026). BAVIR instead aligns a fused environment vector with a per-cell mobility summary. Its evidence comes from controlled objective comparisons, spatial holdout, and downstream reuse within one metropolitan area. We do not claim the first mobility-free deployment setting, and we do not report a head-to-head comparison with MoRAX or MuseCL.

Backbones and incomplete observations. We build on frozen DINOv2 and SigLIP2 image descriptors (Oquab et al., 2023; Tschannen et al., 2025); the learned components are the modality projections and fusion encoder. AlphaEarth (Brown et al., 2025), GeoCLIP (Vivanco Cepeda et al., 2023), and SatCLIP (Klemmer et al., 2025) provide additional feature references. Missing inputs are a known challenge for multimodal Transformers (Ma et al., 2022), and modality dropout is an established response (Neverova et al., 2016). Our learned missing-stream tokens provide a fixed interface for incomplete observations, not a new missing-modality algorithm. Coverage-conditioned evaluation and controlled image replacement distinguish learning from visual content from learning an observation pattern.

3 METHOD

3.1 PROBLEM FORMULATION

We partition the study region into H3 resolution-9 cells (Uber Technologies, 2018). For cell c, the environment is $X_c = \{(x_c^m, a_c^m) : m \in \mathcal{M}\}$, with $\mathcal{M} = \{\text{POI}, \text{road}, \text{satellite}, \text{street}\}$. Here x_c^m is a feature vector and $a_c^m \in \{0, 1\}$ is its recorded stream-presence flag. Let $\mathcal{C}_{\text{mob}}$ denote the training cells with retained mobility behavior observations. Each $c \in \mathcal{C}_{\text{mob}}$ has a 25-dimensional cell-level mobility behavior summary y_c, computed by aggregating visits and stays within that cell over the observation window. We use y_c to supervise an environment encoder $h_\theta(X_c) = z_c \in \mathbb{R}^{256}$; it is never an input to h_θ, including during training. We call z_c the *environment embedding*; the *environment representation* refers to the learned encoder and its resulting feature space.

The 15-dimensional POI stream is constructed from SafeGraph Places (SafeGraphPlaces) with semantic annotations from Liu et al. (2024). It contains 14 activity-category fractions and $\log(1 + n_{\text{POI}})$; the road-geometry stream contains four log-counts of total network elements, footways, roads, and intersections, derived from OpenStreetMap (OpenStreetMap contributors, 2024). The visual streams are a 768-dimensional DINOv2 satellite descriptor and a 768-dimensional SigLIP2 street descriptor with four capture-time fractions appended; the source images are from NAIP and Mapillary, respectively (of Agriculture, 2024; Mapillary, 2026; Oquab et al., 2023; Tschannen et al., 2025). Image descriptors are pooled per cell and their backbones remain frozen. The mobility behavior summary y_c contains 14 activity fractions, four time-of-day fractions, log mean and 90th-percentile dwell, and five log-counts (visits, POI visits, home visits, unmatched visits, and devices). It is derived from Veraset visit/stay records (Veraset, 2022), whose stays are annotated by the semantic-trajectory pipeline of Liu et al. (2024). That pipeline infers the activity associated with each stay from nearby POI semantics. To minimize privacy risk before model training, records are aggregated by cell and the encoder receives neither device identifiers nor individual trajectories.

The source taxonomy indexes 14 activity types. The processed Orlando extract contains POI-linked codes 2–15 and no records for code 1; consequently, the model uses 14 observed activity fractions rather than inserting an unsupported fifteenth fraction. Home stays are identified from stay type and retained separately in the home-visit count.

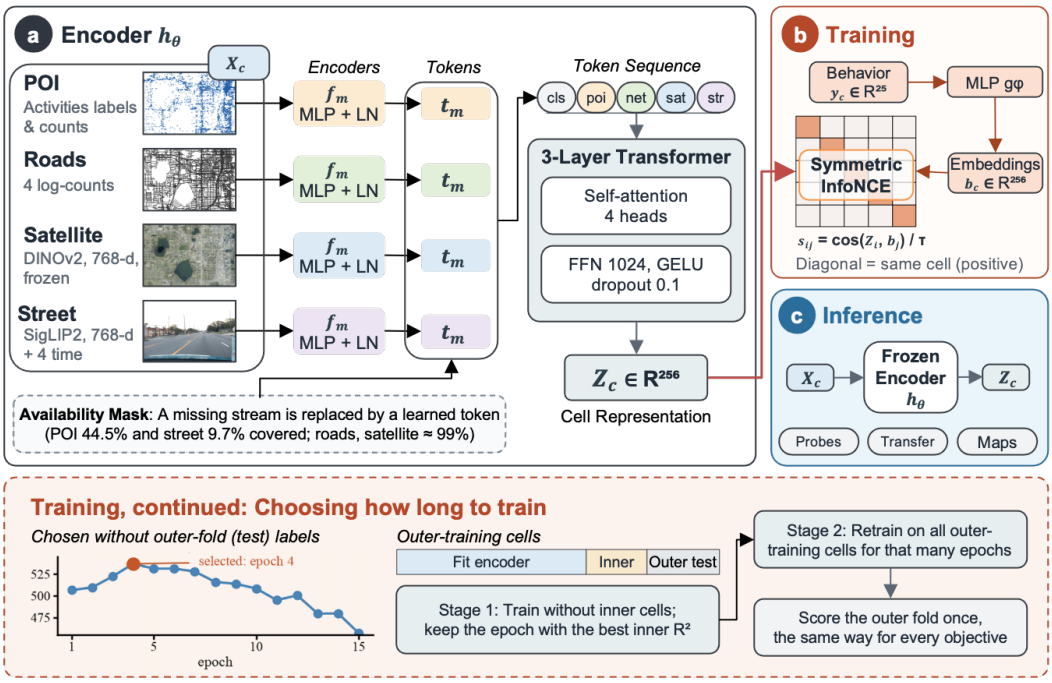

Figure 2: **BAViR architecture and training procedure.** (a) Four modality-specific projectors map POI, road-network, satellite, and street observations to 256-dimensional tokens. Missing streams are replaced by learned tokens, and a Transformer fuses the available observations into the environment representation z_c. (b) For a mobility behavior-observed training cell, the observed summary y_c is encoded as $b_c = g_\phi(y_c)$. Symmetric InfoNCE aligns environment and behavior embeddings from the same cell while contrasting other cells in the batch. The image backbones remain frozen. (c) After training, the behavior branch is removed and the frozen environment encoder produces z_c from environmental observations alone. Task-specific readouts use this representation for downstream prediction, transfer, and mapping. Checkpoint selection uses only an inner split of the outer-training region. The selected training length is then used to retrain the encoder on all outer-training cells before the outer fold is evaluated once.

Activity shares aggregate the supplied top-category probabilities over POI-linked visits, whereas counts and dwell also include home and unmatched stays. They summarize inferred activity composition, not directly observed trip purpose. Cells without a POI-linked activity distribution use zeros for those entries; the other behavior features remain available.

3.2 Multimodal Environment Encoder

Because the four streams differ in dimension and semantics, we first map each available observation to a common token space before fusion, as shown in Figure 2(a). Each stream has its own MLP f_m: a linear layer of width 512, ReLU, a 256-output linear layer, and LayerNorm. Its token is

$$t_c^m = \begin{cases} f_m(x_c^m), & a_c^m = 1, \text{ with record,} \\ u_m, & a_c^m = 0, \text{ no record.} \end{cases} \tag{1}$$

The sequence $[t_{\text{CELL}}, t_c^{\text{POI}}, t_c^{\text{road}}, t_c^{\text{satellite}}, t_c^{\text{street}}]$ enters a Transformer (Vaswani et al., 2017), whose first output is z_c. Here u_m is a learned stream-specific token and t_{CELL} is a learned summary token. The Transformer has three layers, four self-attention heads, feed-forward width 1024, GELU, and dropout 0.1. The other token outputs are not used by the downstream readout. Separate projectors and missing tokens identify the streams; self-attention exchanges information among their tokens.

For images, a_c^m indicates a stored descriptor. For POIs and road geometry, it indicates a positive recorded count: zero-count and absent records both use u_m. Consequently, the implementation

cannot distinguish a surveyed cell with no recorded element from a cell for which that stream was not observed. The fixed token interface permits learning and inference when one or more streams are unavailable, but does not establish robustness when deployment exhibits a stream-availability pattern absent from training.

3.3 Behavior supervision through contrastive alignment

To supervise the environment embedding with information about how places are used, we contrastively align it with the observed cell-level mobility behavior summary during training (Figure 2(b)). A mobility behavior-summary encoder g_ϕ, with the same hidden width and output normalization as the stream projectors, maps the observed summary y_c to the mobility behavior embedding $b_c \in \mathbb{R}^{256}$. For a batch of B mobility behavior-observed cells, define $\hat{z}_i = z_i/\|z_i\|_2$, $\hat{b}_i = b_i/\|b_i\|_2$, and $s_{ij} = \hat{z}_i^\top \hat{b}_j/\tau$, with $\tau = 0.07$. We jointly train the environment and mobility behavior-summary branches with symmetric InfoNCE (van den Oord et al., 2018):

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{e^{s_{ii}}}{\sum_{j=1}^{B} e^{s_{ij}}} + \log\frac{e^{s_{ii}}}{\sum_{j=1}^{B} e^{s_{ji}}}\right]. \tag{2}$$

Each environment embedding must identify its own cell's mobility behavior embedding among the batch, and vice versa. The diagonal gives positive pairs; all other cells are negatives. This is a matching objective, not direct regression of the 25 behavior entries, and it does not encode a causal model of why a person visits a place.

After training, g_ϕ is discarded and h_θ is frozen. At deployment, Figure 2(c), environmental observations X_c are the sole input to h_θ and produce the reusable representation z_c. For a downstream target q with labels v_c^q, we fit a readout r_q on labeled training cells and evaluate $\hat{v}_c^q = r_q(z_c)$ on held-out cells. Section 4 introduces the supervised behavior targets and the separate transfer targets, including spending and assessed value. Changing the task changes r_q, not h_θ; neither fitting nor applying a readout requires the behavior branch. Figure 2 separates these training and deployment paths.

4 Experimental Setup

Study region and prediction targets. We evaluate BAV1R in the Orlando metropolitan area, comprising 42,187 H3 cells, of which 7,495 have cell-level mobility summaries aggregated from 69.0 million retained 2019 visit records (Veraset, 2022). Environmental inputs come from the semantically annotated SafeGraph Places catalog (SafeGraphPlaces; Liu et al., 2024), OpenStreetMap road geometry (OpenStreetMap contributors, 2024), NAIP imagery (of Agriculture, 2024), and Mapillary street imagery (Mapillary, 2026). Appendix F provides detailed sources and coverage. We evaluate three components of the cell-level mobility summary used for behavior supervision: the recorded visit count per cell over the collection window, mean dwell in minutes, and the 14-category activity-share vector. The held-out aggregates, not simulated behavior, are the reference labels. Activity categories are inferred from POI semantics, not independently reported travel intentions.

To evaluate whether the learned representation supports downstream tasks beyond the behavior used for encoder training, we consider two transfer targets: total recorded transaction spend per cell (SafeGraph, 2022) and median assessed parcel value (ATTOM Data, 2025). Neither enters encoder training or the visit-based checkpoint-selection criterion defined below. These measure transfer to a new target type, whereas refitting a readout on a later mobility snapshot measures temporal reuse of a frozen encoder. Additional physical and risk targets test the scope of the representation (Appendix B). Counts, monetary values, and dwell are fitted after a $\log(1+x)$ transform; activity shares are fitted directly. Visits and spend denote recorded panel aggregates, not population traffic rates or area-normalized densities.

Spatial holdout. Two five-fold protocols test cells excluded from encoder training while keeping resolution-6 H3 parent groups intact. Interleaved holdout randomly assigns groups to folds; geo-block holdout withholds five contiguous regions. Seeds 0–2 vary partitions and initialization, giving 30 runs per configuration.

Encoder model selection. We reserve 20% of outer-training parent groups for inner validation and select the epoch whose frozen representation gives the best inner-split visit prediction. A fresh encoder is then trained on all outer-training cells for that duration. Every objective uses the same cells, architecture, maximum duration, and rule; objectives without mobility in their loss still use visit labels for selection. Readout regularization is likewise selected without outer-fold labels.

Implementation. Frozen image backbones are reused; all learned BAViR components are initialized anew. Appendix A gives optimization and baseline details.

Objective controls isolate behavior supervision. Holding architecture and inputs fixed, masked-stream SSL tests reconstruction without mobility; environment InfoNCE tests contrastive learning without mobility; and behavior regression (MSE) directly predicts the same mobility summary used to supervise BAVIR. An untrained fusion encoder measures information present before representation learning. All use identical frozen-feature readouts.

Task-specific controls test the need for representation learning. Raw-feature Ridge, a supervised MLP, and boosted trees test direct prediction from the same environmental inputs; a training-mean predictor anchors error. These controls ask whether a reusable representation is needed at all.

Downstream label budgets. Frozen-encoder readouts for spend and assessed value use common nested subsets of 1–100% of the outer-training target labels. For spend, 10% denotes one tenth of eligible training cells with recorded transaction totals; the test population is unchanged. Budgets cover readout fitting and selection, not the mobility data used for behavior supervision; details appear in Appendix A.

Metrics and uncertainty. For a scalar target, $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, so 0.55 means 45% of the held-out-mean predictor's squared error remains, not 55% accuracy. We also report original-unit MAE and activity-share MAE. Results are means and standard deviations over seeds; spatial bootstrap intervals resample parent groups (Appendix B). Seeds and folds remain within one city and are not independent city-level replications.

5 RESULTS

5.1 BEHAVIOR SUPERVISION CHANGES WHAT THE ENVIRONMENT ENCODER RETAINS

Table 1 first tests whether behavior supervision makes the environment representation more informative about urban use. Relative to matched masked-stream SSL, BAVIR improves visit-count R^2 from 0.514 to 0.550 under interleaved holdout and from 0.488 to 0.523 under geo-block holdout. Dwell MAE falls by 4.3 and 3.8 minutes, and activity-share MAE by 0.68 and 0.72 percentage points. The visit improvement occurs in each seed; a paired spatial bootstrap also supports a positive difference (Table 11). Environment InfoNCE is weaker than masked SSL on these targets, indicating that contrastive matching of environmental views alone does not reproduce the benefit of behavior supervision. Figure 3 makes the visit-count probe concrete with two selected held-out cells and then reports the comparison over the full held-out region.

Two controls qualify this result. The untrained fusion encoder already outperforms raw-feature Ridge on visits, so not all of BAVIR's improvement over raw features is learned through behavior supervision. Behavior supervision adds 0.058 and 0.060 R^2 beyond that reference. More importantly, behavior regression (MSE) reaches visit R^2 of 0.559 and 0.526, slightly above BAVIR. Boosted trees yield lower activity-share error than any frozen-encoder readout. These results support the value of behavior supervision but do not establish a uniquely superior objective for reconstructing its own targets.

The remaining error also matters. For geo-block visit prediction, BAVIR's original-unit MAE is 6,870 recorded visits per cell over the collection window, compared with 7,158 for masked SSL and 6,734 for behavior MSE (Table 2). Log-scale predictability is therefore not evidence of accurate operational traffic counts. These probes measure information in a representation, not readiness for event forecasting or traffic control.

Table 1: **Behavior alignment against matched controls.** Held-out cells, mean over three seeds (subscript: seed SD). Encoder stopping epochs and direct-model hyperparameters use inner spatial validation. The untrained row retains pretrained image descriptors but leaves fusion at initialization. Dwell MAE is in minutes and activity-share MAE in percentage points; lower is better for both. Spend and assessed value are excluded from encoder supervision. Best in each column is bold; differences are computed before rounding.

Method	Visit R^2	Dwell	Activity	Spend R^2	Value R^2
Interleaved holdout					
DIRECT PREDICTION					
Training mean	$-0.006_{0.004}$	$85.7_{0.0}$	$10.70_{0.00}$	$-0.004_{0.003}$	–
Raw-feature Ridge	$0.448_{0.004}$	$71.1_{0.5}$	$7.44_{0.01}$	$0.206_{0.009}$	$0.134_{0.016}$
Supervised MLP	$0.316_{0.013}$	$78.3_{0.3}$	$8.33_{0.03}$	$0.096_{0.014}$	$-0.065_{0.012}$
Boosted trees	$0.536_{0.005}$	$\mathbf{62.9}_{0.3}$	$\mathbf{5.74}_{0.01}$	$0.329_{0.002}$	$0.199_{0.005}$
UNTRAINED FUSION					
Untrained encoder	$0.492_{0.003}$	$68.5_{0.1}$	$7.47_{0.01}$	$0.292_{0.005}$	$0.184_{0.007}$
TRAINED ENCODERS					
Masked-stream SSL	$0.514_{0.004}$	$67.4_{1.0}$	$6.92_{0.03}$	$0.296_{0.005}$	$0.163_{0.010}$
Environment InfoNCE	$0.467_{0.002}$	$68.7_{0.4}$	$7.39_{0.03}$	$0.328_{0.008}$	$0.192_{0.007}$
Behavior MSE	$\mathbf{0.559}_{0.004}$	$63.4_{1.0}$	$6.51_{0.02}$	$0.338_{0.009}$	$0.204_{0.004}$
BAVIR (ours)	$0.550_{0.005}$	$63.1_{0.2}$	$6.23_{0.04}$	$\mathbf{0.355}_{0.006}$	$\mathbf{0.209}_{0.005}$
Δ BAVIR $-$ *SSL*	$+0.037$	-4.3	-0.68	$+0.059$	$+0.047$
Geo-block holdout					
DIRECT PREDICTION					
Training mean	$-0.097_{0.001}$	$85.8_{0.0}$	$10.71_{0.00}$	$-0.012_{0.000}$	–
Raw-feature Ridge	$0.414_{0.003}$	$72.7_{0.0}$	$7.50_{0.00}$	$0.197_{0.001}$	$-0.021_{0.006}$
Supervised MLP	$0.286_{0.010}$	$79.7_{0.7}$	$8.43_{0.03}$	$0.073_{0.006}$	$-0.355_{0.014}$
Boosted trees	$0.510_{0.001}$	$65.6_{0.1}$	$\mathbf{5.77}_{0.00}$	$0.346_{0.011}$	$0.077_{0.006}$
UNTRAINED FUSION					
Untrained encoder	$0.463_{0.001}$	$70.3_{0.0}$	$7.54_{0.02}$	$0.294_{0.007}$	$0.061_{0.015}$
TRAINED ENCODERS					
Masked-stream SSL	$0.488_{0.005}$	$68.5_{0.2}$	$6.98_{0.03}$	$0.314_{0.007}$	$0.039_{0.011}$
Environment InfoNCE	$0.427_{0.007}$	$70.9_{0.1}$	$7.47_{0.10}$	$0.328_{0.004}$	$0.071_{0.015}$
Behavior MSE	$\mathbf{0.526}_{0.005}$	$65.9_{0.6}$	$6.59_{0.03}$	$0.339_{0.003}$	$\mathbf{0.095}_{0.007}$
BAVIR (ours)	$0.523_{0.003}$	$\mathbf{64.8}_{0.4}$	$6.26_{0.03}$	$\mathbf{0.355}_{0.005}$	$0.094_{0.006}$
Δ BAVIR $-$ *SSL*	$+0.035$	-3.8	-0.72	$+0.040$	$+0.055$

5.2 THE FROZEN REPRESENTATION TRANSFERS BEYOND ITS SUPERVISION

Transaction spend tests the central reuse claim because the encoder has never seen spend labels. With all downstream labels, BAVIR reaches spend R^2 of 0.355 under both protocols, above masked SSL (0.296 and 0.314), behavior MSE (0.338 and 0.339), and the tested direct supervised predictors. The gain over behavior MSE is also positive in the paired bootstrap (Table 11). The ordering differs from visit prediction: optimizing the encoder to regress its cell-level mobility summary most accurately is not necessarily best for transferring to this new target.

Figure 4 asks whether this information remains useful with fewer downstream labels. With 10% of spend labels, BAVIR reaches 0.302 and 0.310, exceeding raw-feature Ridge trained with all labels (0.206 and 0.197) and the untrained-fusion readout with all labels (0.292 and 0.294). Behavior MSE is close at the same 10% budget (0.289 and 0.301). This comparison demonstrates label-efficient reuse within the tested readout family; it does not imply that a 10%-label readout beats every fully supervised model. In particular, full-label boosted trees reach 0.329 and 0.346. Table 8 provides the full curves, including failed small-budget settings.

Transfer is not equally strong for every target. For assessed value, BAVIR improves on masked SSL but remains close to behavior MSE, and geo-block R^2 is only 0.094. At 10% of assessed-value labels, BAVIR's geo-block score is 0.032, so the spend result should not be generalized to all socioeconomic outcomes. Reusing the 2019-trained encoder with a readout refitted on 2026 visit labels gives R^2 0.554 under interleaved holdout, compared with 0.504 for raw features. This is

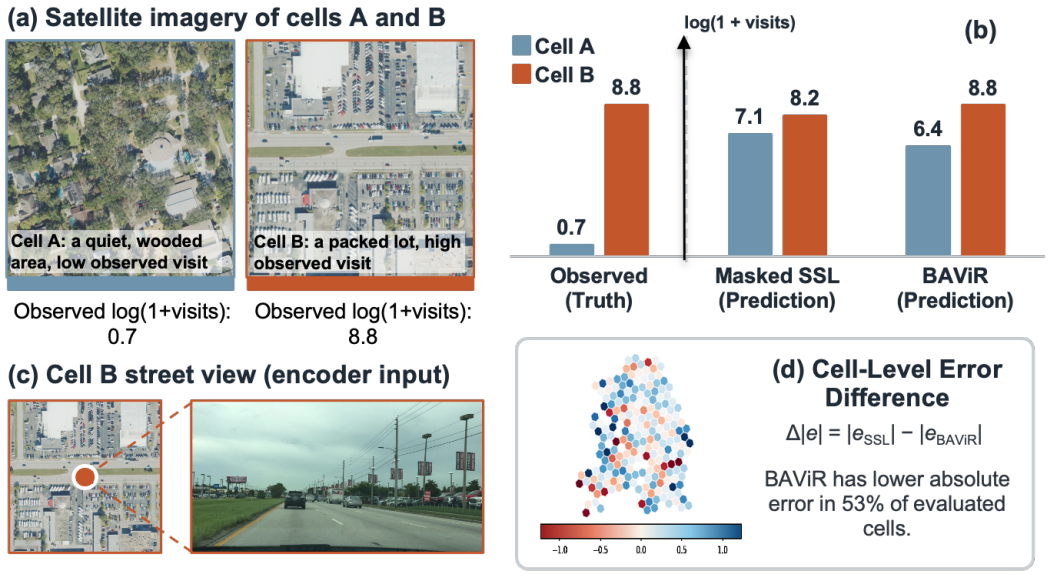

Figure 3: **Illustrative held-out predictions and cell-level error comparison.** Panels (a) and (c) show the environmental observations for two example cells, while panel (b) compares the observed and predicted visit targets, expressed as $\log(1 + n_{\text{visits}})$, for masked SSL and BAVIR. The two cells are illustrative examples and are not used to support the aggregate claim. Panel (d) reports the cell-level difference in absolute prediction error, $|y - \hat{y}_{\text{SSL}}| - |y - \hat{y}_{\text{BAViR}}|$, across the held-out region. Blue indicates lower error for BAVIR, red indicates lower error for masked SSL, and BAVIR has lower absolute error in 53% of evaluated cells. Both models use the same inputs, folds, architecture, and readout; only the representation-learning objective differs. Cell A is a failure case for both methods: although BAVIR has lower absolute error than masked-stream SSL, both predictions remain far from the observed target. Appendix D.1 provides the exact log-scale values, raw-count interpretation, map aggregation, color range, and sample sizes.

reuse across collection windows with new training labels for the readout, not zero-shot temporal forecasting; Table 9 reports both spatial protocols and the other controls.

5.3 SATELLITE IMAGE CONTENT CONTRIBUTES TO THE GAIN

Adding satellite imagery to POI and road features raises interleaved visit R^2 from 0.492 to 0.534 and spend R^2 from 0.300 to 0.349 (Table 3). A matched replacement of satellite descriptors with noise reduces the full model's visit R^2 by 0.045 and 0.051 across the two protocols (Table 4). This supports a contribution from visual content rather than fusion of semantic features alone.

Street imagery requires a narrower interpretation. Replacing its image descriptor with noise costs only 0.005 and 0.006 visit R^2, while retaining its presence flag and capture-time features. Much of the street stream's benefit thus survives without its visual content. Because imagery was collected preferentially at visited cells, observation patterns can convey behavioral information. The no-street comparison still favors BAVIR over masked SSL on visits and spend under both protocols, so this pattern does not explain the entire alignment advantage. Appendix E examines natural coverage strata and training subsets; these diagnostics do not establish robustness to arbitrary missing inputs.

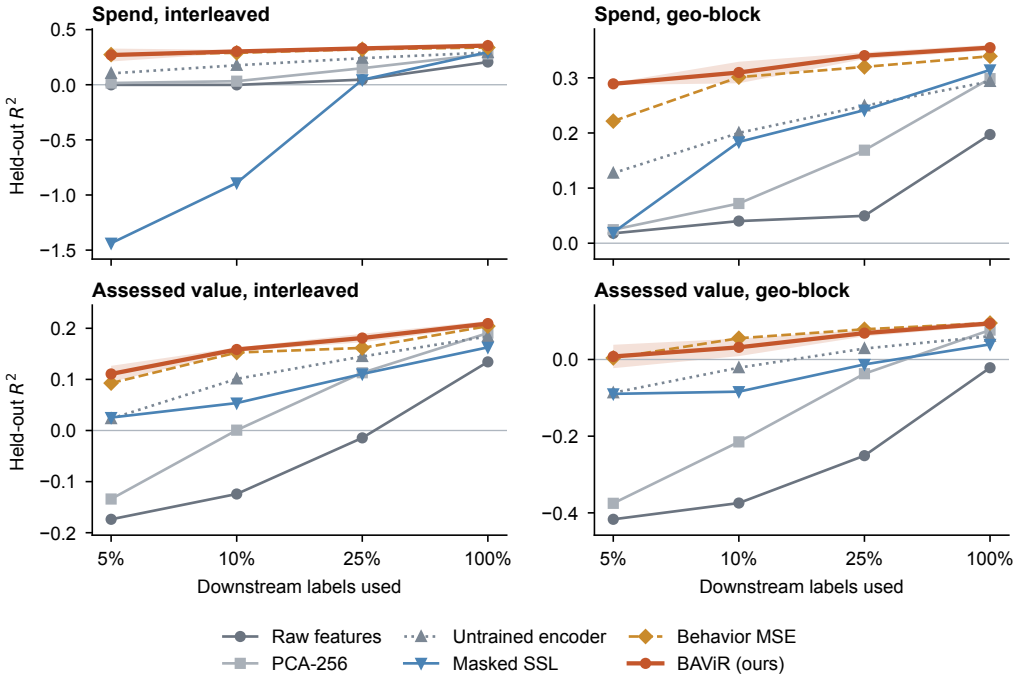

Figure 4: **Reuse for targets excluded from encoder learning.** Spend (top) and assessed-value (bottom) R^2 versus the downstream label budget, using a frozen representation and a separately fitted Ridge readout. Shading is seed standard deviation for BAVIR. The mobility data used for behavior supervision are not included in these percentages; full-label direct supervised comparisons appear in Table 1.

6 Discussion and Conclusion

BAVIR uses mobility as training-only behavior supervision for an environment representation rather than requiring mobility at every location where it is applied. The clearest test of this reuse is a target excluded from encoder training: the frozen representation supports recorded card-spend prediction with fewer downstream labels. Spending is an evaluation target, not the final purpose of the representation; it tests whether the behavior-supervised representation retains functional information beyond the mobility summaries that supplied supervision. Matched reconstruction, regression, and untrained-fusion controls bound what can be attributed to alignment, while direct supervised models show when fitting the downstream task alone is competitive or better.

The scope is functional rather than universal. Behavior MSE is stronger on visit counts, and several physical or risk targets favor raw features (Appendix B). Missing-stream tokens make incomplete records processable, not invariant to changes in observation coverage. Spatial holdout remains within one metropolitan area whose data were inspected during development; neither it nor temporal readout refitting demonstrates zero-shot transfer to another city. Within these bounds, the results support behavior supervision as a useful strategy for learning reusable urban environment features from training-only mobility data, with transfer to new target types as the relevant test of their value.

AI use statement

We used generative AI solely to correct grammar and spelling, improve sentence structures, and review the writing of the manuscript. Data processing used original source records, not AI-generated observations, and reported numerical results come strictly from executed experiments and saved outputs, not language-model estimates. The authors retain full responsibility for all content, analyses, and claims in the final manuscript.

Ethics statement

The study uses licensed mobility, card-spend, and property-assessment records aggregated to hexagonal cells of roughly 0.1 km^2. No device identifier, individual trajectory, or personal attribute enters the representation model. The model describes places, not individual intentions. Mobility panels, imagery, and commercial records cover the city unevenly; low-coverage regions require particular caution. These representations should not be used for individual surveillance or operational resource allocation without independent validation and an assessment of coverage and bias. Raw licensed records are not redistributed.

Reproducibility statement

Section 4 and Appendix A specify the spatial splits, training-duration selection, baseline configurations, and aggregation rules. Appendices B–F give additional results, coverage diagnostics, and data sources. Tables are generated from saved experiment outputs. Reproducing the full evaluation requires access to the licensed source records; aggregate results alone do not reproduce data preparation.

Acknowledgments

The Veraset Visits, SafeGraph Spend Patterns, ATTOM Tax Assessor, NatureQuant Urban Heat Index, and ClimateCheck U.S. Climate Risk Data used in this paper were accessed under the terms of the University of California, Berkeley's institutional subscription to Dewey Data's platform.

References

Mohit Agarwal, Mimi Sun, Chaitanya Kamath, Arbaaz Muslim, Prithul Sarker, Joydeep Paul, Hector Yee, Marcin Sieniek, Kim Jablonski, 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.

ATTOM Data. Tax assessor. Dewey Data, 2025. URL https://doi.org/10.82551/395J-S146. Dataset.

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.

Jinzhou Cao, Jiashi Chen, Xiangxu Wang, Weiming Huang, Dongsheng Chen, Tianhong Zhao, Wei Tu, and Qingquan Li. UrbanMMCL: Urban region representations via multi-modal and multi-graph self-supervised contrastive learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232:75–93, 2026. doi: 10.1016/j.isprsjprs.2025.11.012.

ClimateCheck. U.S Climate Risk Data. Dewey Data, 2022. URL https://doi.org/10.82551/8TST-ZT67. Dataset.

Baoshen Guo, Donghang Li, Zhiqing Hong, Kailai Sun, Heye Huang, Alok Prakash, and Shenhao Wang. Enhancing the socioeconomic understanding of foundation models with urban mobility. *arXiv preprint arXiv:2606.01745*, 2026.

Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. Unsupervised representation learning of spatial data via multimodal embedding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1993–2002, 2019. doi: 10.1145/3357384.3358001.

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*, 2025.

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.

Junyuan Liu, Quan Qin, Guangsheng Dong, Xinglei Wang, Jiazhuang Feng, Zichao Zeng, and Tao Cheng. Beyond AlphaEarth: Toward human-centered spatial representation via POI-guided contrastive learning. *arXiv preprint arXiv:2510.09894*, 2025.

Yifan Liu, Chenchen Kuai, Xishun Liao, Haoxuan Ma, Brian Yueshuai He, and Jiaqi Ma. Semantic trajectory data mining with llm-informed poi classification. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 207–213. IEEE, 2024.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Mapillary. Mapillary: Street-level imagery platform. `https://www.mapillary.com/`, 2026. Accessed: 2026-08-10.

Dominik J. Mühlematter, Lin Che, Ye Hong, Martin Raubal, and Nina Wiedemann. UrbanFusion: Stochastic multimodal fusion for contrastive learning of robust spatial representations. *arXiv preprint arXiv:2510.13774*, 2025. Accepted to ICML 2026.

NatureQuant. Urban heat index. Dewey Data, 2025. URL `https://doi.org/10.82551/ZJWK-2G94`. Dataset.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. ModDrop: Adaptive multimodal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (8):1692–1706, 2016.

U.S. Department of Agriculture. National agriculture imagery program (naip). `https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/`, 2024. https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/.

OpenStreetMap contributors. Openstreetmap. `https://www.openstreetmap.org`, 2024. `https://www.openstreetmap.org`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, and Alaaeldin El-Nouby. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

SafeGraph. Spend patterns. Dewey Data, 2022. URL `https://doi.org/10.82551/NSF5-R186`. Dataset.

SafeGraphPlaces. Safegraph places. `https://www.safegraph.com/data-examples/places`, 2024. https://www.safegraph.com/data-examples/places.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Uber Technologies. H3: A hexagonal hierarchical geospatial indexing system. `https://h3geo.org`, 2018.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Veraset. Visits. Dewey Data, 2022. URL https://doi.org/10.82551/TP7E-VA02. Dataset.

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*, 2023.

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

Ya Wen, Jixuan Cai, Qiyao Ma, Linyan Li, Xinhua Chen, Chris Webster, and Yulun Zhou. MoRA: Mobility as the backbone for geospatial representation learning at scale. *arXiv preprint arXiv:2506.01297*, 2025. Accepted to ICLR 2026.

Ya Wen, Jixuan Cai, Yulun Zhou, and Alec Kirkley. MoRAX: Mobility-based representation augmentation for geospatial foundation models. *arXiv preprint arXiv:2608.17848*, 2026.

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.

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 (IJCAI)*, pp. 7536–7544, 2024. doi: 10.24963/ijcai.2024/834.

Fan Zhang, Lun Wu, Di Zhu, and Yu Liu. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS Journal of Photogrammetry and Remote Sensing*, 153:48–58, 2019. doi: 10.1016/j.isprsjprs.2019.04.017.

A Evaluation Protocol

A.1 Splits, seeds, and metrics

Cells are grouped by resolution-6 H3 parent so that cells sharing a parent never straddle a split. Adjacent cells belonging to different parents can still lie on opposite sides. Interleaved holdout randomly assigns parent groups to five folds; geo-block holdout clusters parent centroids into five regions. Within each seed, R^2 averages five fold scores, MAE pools cells with count weights, and RMSE pools squared errors. Across three seeds we report the mean and standard deviation. Original-unit predictions use $\exp(\max(0, \hat{y})) - 1$. Visit counts refer to the recorded collection window, not all travelers. Activity-share predictions are clipped to be nonnegative and renormalized before computing share errors.

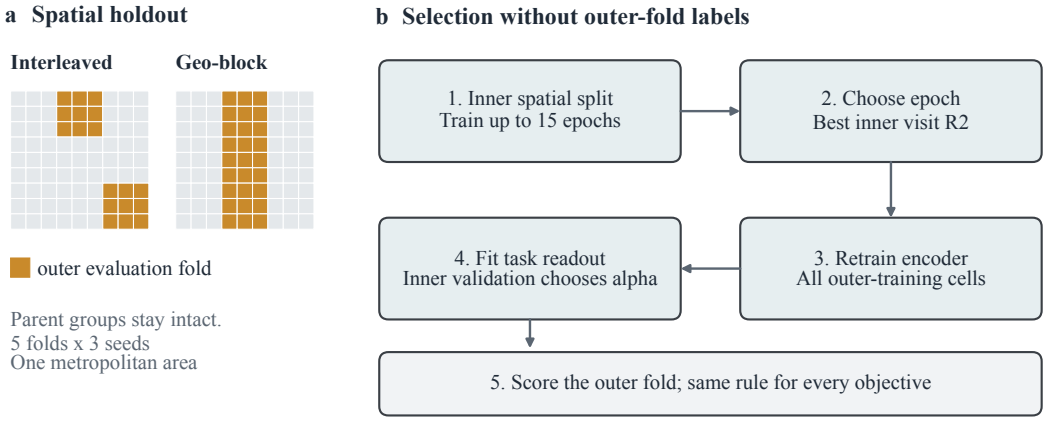

Figure 5: Spatial evaluation and model selection. The grids illustrate the two split types, not the actual folds. Encoder training length and readout regularization are chosen within the outer-training region.

A.2 Training and readout hyperparameters

The two-stage selection procedure in Section 4 uses AdamW with learning rate 3×10^{-4}, batch size 256, weight decay 10^{-4}, and gradient clipping 1.0. The inner visit readout used to select the encoder epoch is standardized Ridge with $\alpha = 10$. The selected training length averages 4.4–6.1 epochs across objectives. Downstream Ridge readouts choose $\alpha \in \{0.1, 1, 10, 100\}$ on the inner partition and are refitted on all eligible outer-training labels, or on the specified downstream label subset. Masked SSL and environment InfoNCE use behavior only for selection, not in their encoder-training loss. Their full training procedures are therefore not label-free.

Spend and assessed-value readouts use common nested subsets of the available outer-training labels. A minimum of 32 labeled cells means that the smallest nominal fractions can exceed their stated percentage; settings without enough inner-validation labels are not scored. The held-out test population remains fixed across budgets.

The direct MLP has one ReLU hidden layer of width 64 or 128, 100 AdamW epochs, learning rate 10^{-3}, and weight decay 10^{-4}. Features and targets are standardized on the training partition. Histogram gradient boosting uses 100 iterations and 15 or 31 leaves, without random-validation early stopping; multivariate targets use one regressor per category. Width and leaf count are selected by inner mean squared error, separately for each target. These are the tested supervised configurations, not an exhaustive search over task-specific models.

The image-noise controls replace 768 descriptor dimensions with independent Gaussian draws matched to the per-dimension mean and variance of covered outer-training cells. Presence flags and the four street capture-time fractions are retained. Noise matches marginal moments, not covariance.

A.3 Development history

An earlier implementation selected checkpoints using outer-fold visit scores. Those selected scores are excluded from the paper. A fixed-length rerun removed this selection leak; the diagnostic in Figure 6 then motivated the inner-validation procedure used in the main comparison. Because the same metropolitan data were inspected during development, the final spatial cross-validation is not a fresh confirmatory sample, even though outer-fold labels are excluded from the final selection procedure.

B Additional Results

Unless stated otherwise, every table below averages three seeds under the aggregation convention of Appendix A, and every trained encoder uses its stopping rule.

B.1 Errors in original units

Table 2: **Prediction errors in interpretable units.** Visits are panel-recorded visits per collection window, not traffic per hour. MAE/RMSE are pooled across held-out cells within seed; MedAE averages fold medians. Entries average three seeds.

Method	Visit MAE	Visit RMSE	Visit MedAE	Dwell MAE (min)
interleaved				
Training mean	8,687.7	38,919.0	1,319.5	85.7
Raw-feature Ridge	7,882.8	39,552.5	959.9	71.1
Supervised MLP	9,087.5	47,348.4	1,155.6	78.3
Boosted trees	6,856.5	37,311.6	829.5	62.9
Untrained encoder	7,185.6	37,728.3	892.7	68.5
Masked-stream SSL	7,242.0	42,626.3	853.3	67.4
Environment InfoNCE	7,260.6	38,072.1	881.7	68.7
Behavior MSE	6,680.4	36,932.6	782.8	63.4
BAVIR (ours)	6,813.1	37,283.0	792.1	63.1
geo-block				
Training mean	8,742.2	38,954.3	1,356.0	85.8
Raw-feature Ridge	8,031.0	39,032.5	908.3	72.7
Supervised MLP	8,706.5	40,103.6	1,029.3	79.7
Boosted trees	6,964.1	37,463.7	739.3	65.6
Untrained encoder	7,234.8	37,778.5	821.5	70.3
Masked-stream SSL	7,158.0	38,075.7	771.0	68.5
Environment InfoNCE	7,351.8	38,585.6	817.4	70.9
Behavior MSE	6,733.9	36,891.6	720.9	65.9
BAVIR (ours)	6,870.1	37,337.2	721.0	64.8

B.2 Which streams matter

Table 3 adds streams one at a time and Table 4 replaces the frozen image features with matched noise while preserving availability flags and street capture-time features.

Table 3: **Adding streams one at a time.** Each row trains BAVIR on the listed streams with the same stopping rule. Mean over three seeds.

	Interleaved			Geo-block		
Streams	Visit R^2	Dwell	Spend R^2	Visit R^2	Dwell	Spend R^2
POI	0.473	65.5	0.283	0.425	67.9	0.274
+ road geometry	0.492	64.8	0.300	0.449	67.3	0.305
+ satellite	0.534	62.8	0.349	0.509	64.8	0.342
+ street (full)	0.550	63.1	0.355	0.523	64.8	0.355

Table 4: Image-content controls. Noise replaces the 768 image-descriptor dimensions with Gaussian draws matched to per-feature moments in outer-training cells; presence flags and street capture-time features are retained. Same stopping rule throughout.

	Interleaved		Geo-block	
Input	Visit R^2	Spend R^2	Visit R^2	Spend R^2
Availability flags only	0.126	0.103	0.055	0.110
Street features → noise	0.545	0.352	0.517	0.347
Satellite features → noise	0.505	0.307	0.472	0.312
Masked SSL, no street	0.510	0.292	0.476	0.306
BAVIR, no street	0.534	0.349	0.509	0.342
BAVIR, all streams	0.550	0.355	0.523	0.355

B.3 Published encoders and coordinate features

Table 5 passes frozen features from published models, alone and combined with POI and network features, to the same folds and readout, with no encoder training. Table 6 instead trains the tower on top of AlphaEarth satellite features and includes the untrained-tower floor. Table 7 uses coordinate-based location encoders.

Table 5: Off-the-shelf and single-stream features under the same folds and Ridge readout, with no encoder training. Compare with BAVIR in Table 1.

	Interleaved			Geo-block		
Features	Visit	Spend	Value	Visit	Spend	Value
POI + network (raw)	0.457	0.264	0.063	0.418	0.259	−0.029
DINOv2 satellite (raw)	0.318	0.181	0.160	0.260	0.165	0.028
SigLIP2 street (raw)	0.139	0.062	−0.005	0.059	0.061	−0.097
AlphaEarth	0.307	0.161	0.152	0.256	0.159	−0.024
AlphaEarth + POI + network	0.513	0.300	0.162	0.483	0.294	−0.012

Table 6: Replacing the frozen satellite features with AlphaEarth (64 dimensions instead of DINOv2's 768). The satellite projector's input width changes accordingly; fusion, other streams, folds, and selection rule are unchanged. Mean over three seeds. Dwell is MAE in minutes.

	Interleaved			Geo-block		
Encoder	Visit R^2	Dwell	Spend R^2	Visit R^2	Dwell	Spend R^2
DINOv2 satellite stream (main model)						
Untrained tower	0.492	68.5	0.292	0.463	70.3	0.294
Masked SSL	0.514	67.4	0.296	0.488	68.5	0.314
Behavior MSE	0.559	63.4	0.338	0.526	65.9	0.339
BAVIR	0.550	63.1	0.355	0.523	64.8	0.355
AlphaEarth satellite stream						
Untrained tower	0.519	67.0	0.286	0.485	68.8	0.280
Masked SSL	0.522	66.5	0.298	0.485	68.3	0.293
Behavior MSE	0.553	63.1	0.328	0.513	66.1	0.321
BAVIR	0.549	63.6	0.330	0.514	65.6	0.324

Table 7: Coordinate-based location encoders under the same folds and Ridge readout. Their only input is the cell centroid. All entries are R^2, averaged over three seeds.

	Interleaved			Geo-block		
Features	Visit	Spend	Value	Visit	Spend	Value
Coordinates (quadratic)	0.039	−0.008	0.021	−0.096	−0.021	−0.087
GeoCLIP	0.020	−0.121	−0.062	−0.668	−0.744	−1.318
SatCLIP	0.056	−0.009	0.023	−0.090	−0.036	−0.311

B.4 LABEL EFFICIENCY IN FULL

Table 8: Label efficiency in full. Held-out R^2 of a Ridge readout fit on a fraction of the downstream labels, mean over three seeds; negative values are kept. A dash means at least one fold has too few inner-validation labels.

Method	1%	5%	10%	25%	100%
Interleaved holdout, spend					
Raw features	–	−0.001	−0.001	0.046	0.206
PCA-256	–	0.016	0.032	0.150	0.280
Untrained encoder	–	0.103	0.176	0.241	0.292
Masked SSL	–	−1.437	−0.890	0.045	0.296
Environment InfoNCE	–	0.152	0.228	0.286	0.328
Behavior MSE	–	0.273	0.289	0.322	0.338
BAVIR	–	0.270	0.302	0.330	0.355
Interleaved holdout, assessed value					
Raw features	−0.152	−0.174	−0.124	−0.014	0.134
PCA-256	−0.095	−0.134	0.001	0.113	0.191
Untrained encoder	−0.054	0.023	0.101	0.145	0.184
Masked SSL	−0.050	0.025	0.054	0.111	0.163
Environment InfoNCE	−0.137	0.091	0.131	0.165	0.192
Behavior MSE	0.009	0.093	0.153	0.162	0.204
BAVIR	−0.078	0.111	0.159	0.181	0.209
Geo-block holdout, spend					
Raw features	–	0.018	0.040	0.050	0.197
PCA-256	–	0.025	0.072	0.169	0.299
Untrained encoder	–	0.127	0.200	0.249	0.294
Masked SSL	–	0.020	0.184	0.241	0.314
Environment InfoNCE	–	0.209	0.254	0.299	0.328
Behavior MSE	–	0.222	0.301	0.320	0.339
BAVIR	–	0.289	0.310	0.340	0.355
Geo-block holdout, assessed value					
Raw features	−0.433	−0.417	−0.374	−0.251	−0.021
PCA-256	−0.404	−0.375	−0.215	−0.037	0.077
Untrained encoder	−0.425	−0.087	−0.021	0.029	0.061
Masked SSL	−2.764	−0.090	−0.084	−0.013	0.039
Environment InfoNCE	−0.216	−0.082	0.001	0.048	0.071
Behavior MSE	−0.099	0.004	0.055	0.079	0.095
BAVIR	−0.334	0.008	0.032	0.069	0.094

B.5 REUSE ACROSS COLLECTION WINDOWS

Table 9: Reusing the 2019-trained encoder for the 2026 mobility snapshot. Readouts are refitted on 2026 outer-training labels; encoders are not retrained. Direct predictors also use 2026 training labels. Values are visit-count R^2 (mean over three seeds, subscript: seed SD). This is not zero-shot forecasting.

Method	Interleaved	Geo-block
Training mean	$-0.007_{0.006}$	$-0.122_{0.003}$
Raw-feature Ridge	$0.504_{0.003}$	$0.457_{0.000}$
Supervised MLP	$0.341_{0.005}$	$0.275_{0.014}$
Boosted trees	$0.555_{0.002}$	$0.522_{0.001}$
Untrained encoder	$0.527_{0.001}$	$0.496_{0.002}$
Masked-stream SSL	$0.525_{0.003}$	$0.493_{0.002}$
Environment InfoNCE	$0.535_{0.004}$	$0.500_{0.006}$
Behavior MSE	$0.549_{0.002}$	$0.519_{0.003}$
BAVIR (ours)	$0.554_{0.005}$	$0.525_{0.001}$

B.6 PHYSICAL, RISK, AND CRIME TARGETS

The advantage on spend does not extend uniformly to physical or risk targets. Raw-feature Ridge is stronger than BAVIR on interleaved transit distance, heat index, fire rating, and flood rating. BAVIR improves transit-stop density and some crime scores, but the crime R^2 values remain low and use limited geographic coverage. Under geo-block holdout, transit-stop density, flood, and fire have negative R^2 for all listed methods. ZIP-level climate ratings should not be interpreted as independent cell-level measurements.

Table 10: **Physical, risk, and crime targets.** Held-out R^2 of a Ridge readout on the frozen encoders, mean over three seeds. Crime data cover the city police area, so geo-block scores use the two folds with enough labeled cells. Best per row in bold.

Target	Raw	Untrained	SSL	MSE	BAVIR
Interleaved					
Transit stop density	0.212	0.234	0.243	0.277	**0.315**
Transit distance (log)	**0.546**	0.491	0.458	0.484	0.500
Heat index	**0.497**	0.453	0.420	0.438	0.447
Fire rating (ZIP)	**0.427**	0.383	0.337	0.366	0.388
Flood rating (ZIP)	**0.217**	0.127	0.085	0.101	0.118
Violent crime density	−0.128	**0.059**	0.056	0.052	0.049
Total crime density	−0.139	**0.027**	0.018	0.006	0.010
Geo-block					
Transit stop density	−0.641	−0.459	−0.383	−0.368	**−0.320**
Transit distance (log)	**0.058**	−0.026	−0.066	−0.023	−0.006
Heat index	**0.409**	0.363	0.340	0.357	0.362
Fire rating (ZIP)	**−0.207**	−0.275	−0.353	−0.314	−0.234
Flood rating (ZIP)	**−0.756**	−0.950	−1.072	−1.028	−0.970
Violent crime density	−0.354	0.003	0.068	**0.110**	0.081
Total crime density	−0.355	0.042	0.088	**0.108**	0.089

B.7 TRAINING LENGTH

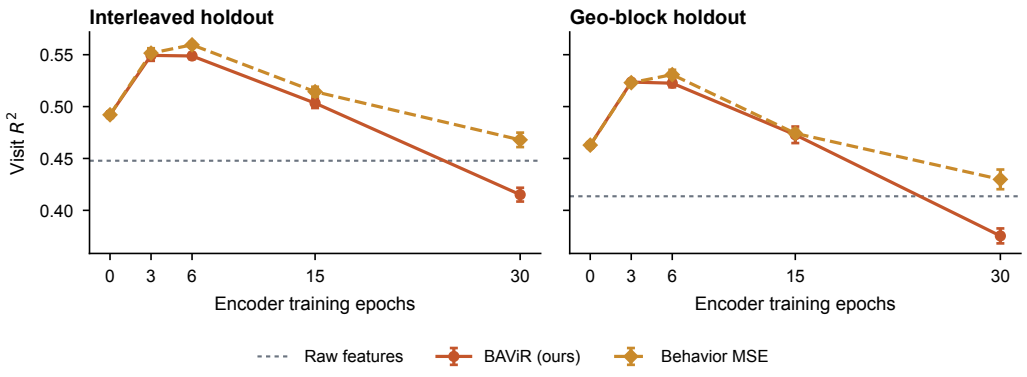

Figure 6: Alignment overfits quickly. Visit R^2 against encoder training epochs on the outer folds (mean and SD over seeds); epoch 0 is the untrained tower. These development curves motivated the inner-selection protocol, but do not select checkpoints in the final runs.

B.8 SPATIAL BOOTSTRAP

Table 11: Spatial bootstrap for paired differences in R^2. Resolution-6 parent groups are resampled with replacement (2,000 draws), using the same groups for all three seeds. Scores pool held-out cells across folds within each seed, then average seed differences; they are not differences of fold-mean R^2 in Table 1. Brackets are 95% percentile intervals.

	Difference	Interleaved	Geo-block
Visits	BAVIR − masked SSL	$+0.037$ [$+0.029$, $+0.045$]	$+0.037$ [$+0.028$, $+0.046$]
Visits	BAVIR − behavior MSE	-0.008 [-0.013, -0.004]	-0.003 [-0.008, $+0.002$]
Visits	BAVIR − untrained encoder	$+0.059$ [$+0.049$, $+0.068$]	$+0.062$ [$+0.053$, $+0.071$]
Spend	BAVIR − masked SSL	$+0.060$ [$+0.047$, $+0.074$]	$+0.048$ [$+0.036$, $+0.060$]
Spend	BAVIR − behavior MSE	$+0.017$ [$+0.011$, $+0.024$]	$+0.015$ [$+0.007$, $+0.023$]

C PREDICTION EVIDENCE IN THE OVERVIEW

Figure 1 visualizes two evaluated outputs, not a generic *busyness* score. Visits are device-panel visit/stay counts per cell over the observed 2019 collection window. Card spending is the sum of recorded transaction amounts at POIs in a cell, in USD, from SafeGraph Spend Patterns (SafeGraph, 2022). Neither is a complete population total, and neither map depicts live conditions. Spend labels fit a separate predictor after the environment encoder has been frozen; they never supervise or select that encoder.

The metric labels in Figure 1 are rounded three-seed means under geo-block holdout. Visit and spending R^2 are computed on their $\log(1 + x)$ targets and compare the model's squared error with that of the held-out target mean. The dwell and activity errors are reported separately in Table 12: dwell MAE returns predictions to minutes, whereas activity MAE averages errors in category shares and is therefore in percentage points. These quantities are not directly comparable across units.

The maps include all saved outer-test predictions from the five geo-block folds for seed 0: 7,495 visit cells and 4,561 spend cells. Each cell appears once, predicted by the model that withheld its region. No cells are selected according to prediction accuracy. The spend source has 4,562 cells before restricting it to the model grid; the one outside the grid is not evaluated. For display only, we invert the natural-log transform and sum the evaluated resolution-9 cells within each resolution-7 H3 parent. Observed and predicted maps use the same contributing cells, geographic extent, and $\log_{10}(1 + x)$ color scale within each target. Gray areas contain no evaluated cells; a displayed parent's sum does not impute missing observations. This aggregation makes spatial patterns legible but can conceal cell-level errors, so Table 12 reports metrics at the original resolution, across all three seeds. Comparative results, including stronger alternatives on some targets, remain in Table 1.

Table 12: **Evidence behind Figure 1.** Scores use the original resolution-9 cells, not the coarser display maps. Mean over three seeds and five geo-block folds; subscript is seed SD. N counts the unique evaluated cells shown for seed 0. Log MAE is on $\ln(1 + x)$; original-unit MAE pools cells within seed. Activity is the share MAE after projecting predictions onto valid mixes.

Target	In encoder training	N	R^2	Log MAE	Original-unit MAE
Recorded visits	yes	7,495	$0.523_{0.003}$	$0.893_{0.001}$	6,870 visits
Mean dwell	yes	7,495	$0.366_{0.004}$	$0.466_{0.003}$	64.8 min
Activity mix	yes	7,231	–	–	$6.26_{0.03}$ pts
Recorded card spending	no	4,561	$0.355_{0.005}$	$1.417_{0.002}$	54,805 USD

The displayed seed-0 predictions sum to 32.7M visits versus 69.0M recorded, and 99.8M USD versus 290.9M recorded, so these log-trained readouts capture spatial variation but do not recover aggregate totals accurately. The predicted main activity matches the recorded one in 72% of displayed areas, compared with 38% for always predicting the most common class.

D ADDITIONAL MODEL VARIANTS

We also tested changes to the objective under the same folds and selection rule. A dense latent-prediction loss, in which every stream position regresses the stop-gradient pre-mask token, raised interleaved visit R^2 by about 0.004 and lowered dwell error by 1.3 minutes, with no gain on spend or activity mix (Table 13). Table 13 also reports modality dropout and a three-encoder ensemble. These variants do not consistently improve all reported targets.

Table 13: Objective and training variants under the same folds and stopping rule. Mean over three seeds.

	Interleaved			Geo-block		
Variant	Visit R^2	Dwell	Spend R^2	Visit R^2	Dwell	Spend R^2
BAVIR (reference)	0.550	63.1	0.355	0.523	64.8	0.355
+ dense latent prediction	0.554	61.8	0.355	0.524	64.2	0.353
+ modality dropout, $p = 0.2$	0.549	63.6	0.352	0.527	65.9	0.348
+ modality dropout, $p = 0.4$	0.548	63.2	0.354	0.522	66.0	0.354
3-encoder ensemble (concatenated)	0.555	62.0	0.354	0.528	63.9	0.354

D.1 INTERPRETATION OF THE SELECTED CELLS AND ERROR MAP IN FIGURE 3

Figure 3 uses the same $\log(1 + n_{\text{visits}})$ target as the visit-count readout. For Cell A, the observed value is 0.693, corresponding to one recorded visit. Masked-stream SSL predicts 7.130, corresponding to approximately 1,248 visits after inverting the transform, while BAVIR predicts 6.399, corresponding to approximately 600 visits. Both predictions are therefore poor. BAVIR is only relatively better because its absolute log-scale error is smaller; Cell A should not be interpreted as an accurate prediction.

For Cell B, the observed value is 8.814, corresponding to 6,724 recorded visits. Masked-stream SSL predicts 8.201, corresponding to approximately 3,645 visits, while BAVIR predicts 8.804, corresponding to approximately 6,661 visits. Thus, unlike Cell A, Cell B is predicted closely by BAVIR. These two cells make the target and model outputs concrete; they are not used to establish the aggregate result.

For each resolution-7 H3 parent containing at least three evaluated resolution-9 cells, the map displays

$$\Delta E = \text{MAE}_{\text{SSL}} - \text{MAE}_{\text{BAViR}} . \tag{3}$$

Positive blue values indicate lower error for BAVIR, negative red values indicate lower error for masked-stream SSL, and zero indicates equal mean absolute error. The mapped error is measured on the $\log(1 + n_{\text{visits}})$ scale.

The map contains 207 resolution-7 parent regions. Its symmetric displayed range, $[-0.36, 0.36]$, is determined by the 95th percentile of the absolute mapped values; more extreme values are clipped only when assigning colors. The reported 53.4% is computed before spatial aggregation and color clipping: BAVIR has lower absolute error than masked-stream SSL for 1,611 of the 3,016 underlying held-out resolution-9 cells available for both models. It is not the percentage of the 207 displayed parent regions.

E OBSERVATION COVERAGE

Coverage and model performance describe different populations. Street descriptors exist for 4,076 cells, or 9.7% of the full grid, but these cells make up 54.4% of the 7,495 behavior-observed cells. Positive recorded POI counts occur in 97.5% of that behavior population, compared with 44.5% of the grid. Thus, evaluation on behavior-observed cells does not establish performance on every unvisited or sparsely documented location.

Table 14 conditions visit prediction on recorded stream flags. It pools sums of squares over held-out cells across the five folds within a seed before averaging seeds, unlike the fold-mean R^2 in Table 1. The resulting all-cell entries can therefore differ. The strata overlap and are naturally

Table 15: Training-subset diagnostic. Complete cases have four positive stream flags; the random half ignores these flags. Subset sizes vary by fold and are not exactly matched. Readouts use all outer-training target labels. Mean over three seeds.

	Interleaved		Geo-block	
Encoder trained on	Visit R^2	Spend R^2	Visit R^2	Spend R^2
Untrained tower	0.492	0.292	0.463	0.294
Four positive stream flags	0.537	0.348	0.513	0.352
Random half of labeled cells	0.544	0.346	0.517	0.345
All labeled cells (BAVIR)	0.550	0.355	0.523	0.355

occurring populations, not interventions that remove a stream from identical cells. Behavior MSE outperforms BAVIR in every listed interleaved stratum; the difference is largest among the 129 cells with neither a positive POI record nor a street descriptor, where BAVIR has negative R^2.

Table 14: Visit R^2 within natural coverage strata, pooling held-out cells across folds before averaging three seeds. This differs from the fold-mean R^2 in Table 1. Strata overlap; absent means no positive stream flag. SSL is masked-stream self-supervision and MSE is behavior regression.

Held-out cells	n	Raw	Untrained	SSL	MSE	BAVIR
Interleaved						
All held-out cells	7,495	0.450	0.493	0.514	0.560	0.552
Street imagery present	4,076	0.416	0.482	0.503	0.551	0.546
Street imagery absent	3,419	0.375	0.391	0.419	0.471	0.456
Positive recorded POI count	7,311	0.447	0.492	0.511	0.554	0.549
No positive POI record	184	−0.080	−0.064	0.051	0.227	0.111
Neither POI record nor street image	129	−0.143	−0.197	−0.114	0.099	−0.041
Geo-block						
All held-out cells	7,495	0.422	0.478	0.503	0.543	0.540
Street imagery present	4,076	0.387	0.462	0.487	0.535	0.535
Street imagery absent	3,419	0.343	0.383	0.413	0.449	0.441
Positive recorded POI count	7,311	0.420	0.477	0.499	0.539	0.539
No positive POI record	184	−0.145	−0.081	0.060	0.142	0.041
Neither POI record nor street image	129	−0.293	−0.216	−0.096	−0.046	−0.147

Table 15 compares training on cells with four positive stream flags, a random half of the labeled cells, and all labeled cells. Using all cells gives the highest visit scores among these subsets. Complete-case and random-half sizes are not exactly matched, however, and readouts retain all outer-training target labels. The comparison shows that discarding incomplete cells is unnecessary in this setting; it does not isolate a causal benefit of learned missing tokens or establish robustness under a shifted missingness pattern.

F Data Sources and Coverage

Table 16 reports sources and recorded-feature counts across the 42,187-cell grid. For POIs and roads, zero counts and absent records both produce inactive streams, so these counts are not a measure of exhaustive catalog coverage. Satellite and street descriptors summarize available NAIP and Mapillary imagery; Mapillary denotes the imagery platform, not the Vistas segmentation benchmark.

The source contains 69,649,190 visit/stay records (Veraset, 2022). Cleaning removes 641,015 records with non-positive dwell, out-of-region coordinates, or dates outside January 1–August 1, 2019. The remaining 69,008,175 records are exactly the sum of cell visit counts used by the encoder. The collection is not continuous over that interval: March 2–May 30 is absent. Counts therefore summarize the observed window, not a continuous seven-month mobility panel.

Activity categories use the supplied POI-derived labels and their probabilities, sharing a taxonomy with the POI input features. The activity probe is consequently not an independent measure of in-

dividual intent. Among behavior-observed cells, 264 lack a POI-linked activity distribution; their activity entries are filled with zero in training and evaluation. The later mobility evaluation uses a separate 2026 collection window and independently fitted readouts, without retraining the environment encoder. Spend is the sum of recorded card transactions per cell (SafeGraph, 2022). Assessed value is the median recorded parcel value in the available county data (ATTOM Data, 2025). These observed products are imperfect measurements of urban activity and value, not complete population censuses. Dataset citation years identify the products, not necessarily the year of the observations.

Heat and climate-risk targets come from NatureQuant (NatureQuant, 2025) and ClimateCheck (ClimateCheck, 2022). Flood and fire ratings are assigned at ZIP-code scale, so they do not provide independent cell-level labels. Street images were collected preferentially at visited cells, making street availability correlated with the behavior targets. These limitations motivate the coverage and image-content controls rather than a claim of uniform metropolitan coverage.

Table 16: Sources and record counts across the 42,187-cell grid. POI and road counts indicate positive recorded features, not certified catalog coverage. These percentages differ from those of the behavior-observed evaluation population.

Source	Origin	Cells	Share	Role
Behavior (visits, dwell, activity)	2019 device visit records	7,495	17.8%	encoder supervision
POI semantics	point-of-interest catalog	18,774	44.5%	input
Road geometry	OpenStreetMap	41,606	98.6%	input
Satellite	NAIP imagery, DINOv2 features	41,678	98.8%	input
Street view	Mapillary images, SigLIP2 features	4,076	9.7%	input
Recorded card spending	card-transaction aggregates	4,561	10.8%	transfer target
Assessed value	county tax assessor	13,615	32.3%	transfer target