

Street-level mobility interactions transfer across cities with minimal local data

Hongrong Yang¹

Markus Schläpfer¹

¹Department of Civil Engineering and Engineering Mechanics
Columbia University, New York, USA

Abstract

The street-level movement of people between places is a fundamental organizing principle of cities, shaping social and economic life, driving transportation demand and facilitating the spread of infectious diseases. However, predicting mobility flows at such fine-grained resolution has remained an open challenge because they are highly heterogeneous across space and time, and existing methods are not designed to learn transferable interactions between individual places. These challenges become particularly severe when large-scale mobility data are scarce, as is the case for most cities worldwide. Here we show that learning transferable interaction representations enables the prediction of street-level mobility flows across structurally different cities using only minimal local mobility data. Leveraging 1.1 billion anonymized GPS records from 4.4 million devices across three contrasting cities in California, we develop a probabilistic framework that learns directed flows between points of interest in dynamic mobility graphs, and adapts these predictions across cities through pretraining, cold-start adaptation and reinforcement-learning refinement. A model pretrained on Los Angeles generalizes to San Francisco and Fresno without any local mobility data, outperforming fully supervised training on one-month target-city data in overall accuracy. With only a single day of local data, the framework delivers stable predictions on high-intensity flows that are competitive with full one-month supervision, while reducing error by up to 35% relative to single-day supervised training from scratch. More broadly, our findings suggest that fine-grained mobility flow models can transfer across cities through local scale recalibration rather than relearning interaction structure. This principle may extend to other complex networked systems with heterogeneous data availability.

USE-Lab Working Paper #2026-07-002

<https://urbansystems.civil.columbia.edu>

Street-level mobility interactions transfer across cities with minimal local data

Hongrong Yang^{1,2} and Markus Schläpfer^{1,2,*}

¹Department of Civil Engineering and Engineering Mechanics, Columbia University,
New York, NY 10027, USA

²Data Science Institute, Columbia University, New York, NY 10027, USA

*Email: m.schlaepfer@columbia.edu

The street-level movement of people between places is a fundamental organizing principle of cities, shaping social and economic life, driving transportation demand and facilitating the spread of infectious diseases. However, predicting mobility flows at such fine-grained resolution has remained an open challenge because they are highly heterogeneous across space and time, and existing methods are not designed to learn transferable interactions between individual places. These challenges become particularly severe when large-scale mobility data are scarce, as is the case for most cities worldwide. Here we show that learning transferable interaction representations enables the prediction of street-level mobility flows across structurally different cities using only minimal local mobility data. Leveraging 1.1 billion anonymized GPS records from 4.4 million devices across three contrasting cities in California, we develop a probabilistic framework that learns directed flows between points of interest in dynamic mobility graphs, and adapts these predictions across cities through pretraining, cold-start adaptation and reinforcement-learning refinement. A model pretrained on Los Angeles generalizes to San Francisco and Fresno without any local mobility data, outperforming fully supervised training on one-month target-city data in overall accuracy. With only a single day of local data, the framework delivers stable predictions on high-intensity flows that are competitive with full one-month supervision, while reducing error by up to 35% relative to single-day supervised training from scratch. More broadly, our findings suggest that fine-grained mobility flow models can transfer across cities through local scale recalibration rather than relearning interaction structure. This principle may extend to other complex networked systems with heterogeneous data availability.

32 Across cities, the constant movement of people between places mediates social, economic
 33 and cultural exchanges, shapes urban form, drives transportation demand, and accelerates the
 34 spread of infectious diseases¹⁻⁴. All these processes are fundamentally anchored in specific
 35 places and follow distinct temporal rhythms: people pass through transit hubs during the
 36 morning commute, shop in grocery stores during the day, gather in restaurants in the evenings,
 37 and visit parks on weekends. It is therefore hard to overstate the importance of understanding
 38 and predicting these flows at the scale of individual places. Indeed, the growing availability
 39 of GPS and mobile-phone records has enabled increasingly high-resolution analyses of human
 40 mobility^{5,6}, revealing how individual places such as restaurants and fitness centres can act as
 41 superspreaders during disease outbreaks⁷, how museums, stadiums and other popular venues
 42 can mitigate social segregation by attracting visitors from different income groups^{8,9}, and how
 43 businesses can economically benefit from shared visitor flows¹⁰.

44 However, despite these recent empirical insights, predicting street-level mobility flows be-
 45 tween individual places remains an open challenge, particularly in cities where fine-grained
 46 mobility data are scarce. Because training high-capacity spatiotemporal models generally
 47 requires large-scale mobility observations, this creates a fundamental barrier for most cities
 48 worldwide, where such data are limited, fragmented or unavailable¹¹. Existing models, rang-
 49 ing from classical spatial-interaction formulations such as gravity¹² and radiation models¹³, to
 50 more recent formulations like the universal visitation law¹⁴ and deep-learning approaches¹⁵⁻¹⁷,
 51 typically model time-averaged (equilibrium) flows between aggregated origin-destination zones
 52 rather than between individual places. As a result, they are not designed to predict the sparse,
 53 directed and time-varying movement patterns between individual points of interest at street-
 54 level resolution. Forecasting such dynamics is fundamentally challenging: between individual
 55 places, mobility flows are highly imbalanced⁵, following heavy-tailed distributions where most
 56 location pairs carry negligible flow while a small fraction forms high-intensity corridors with
 57 distinct temporal patterns.

58 Spatiotemporal graph neural networks¹⁸⁻²⁰, widely applied to traffic forecasting at sensor
 59 locations on road networks, can learn complex spatial and temporal dependencies in mobil-
 60 ity systems, and their architectures can, in principle, be extended to operate at the level of
 61 individual places. However, they are generally formulated on a fixed graph structure defined
 62 by the training city, preventing direct transfer across cities with different spatial organization
 63 and interaction topology. More recently, large language models have been applied to urban
 64 spatiotemporal forecasting^{21,22}, but these foundation-model approaches operate at aggregated
 65 regional resolution and are not designed for directed flow prediction between individual loca-
 66 tions. Existing cross-city transfer methods^{16,17} focus on zero-shot deployment and provide no
 67 principled mechanism to improve predictions when limited target-city training data become
 68 available. More fundamentally, transferring mobility predictions across cities is intrinsically
 69 difficult: cities differ in spatial organization²³, temporal activity patterns and mobility flow

distributions, inducing structural and distributional shifts that cause models to degrade when applied to new urban environments^{24–27}. These conditions expose a central tension: interaction dynamics must be captured at fine spatiotemporal resolution yet remain adaptable across cities under limited supervision. Resolving this tension requires models that can learn transferable interaction rules from one dynamic graph and apply them to structurally different graphs under limited target supervision. To our knowledge, no existing framework provides this capability for directed, sparse and heavy-tailed networks.

Here, we show that transferable interaction representations enable street-level mobility predictions across structurally different cities with distinct flow distributions, even when target-city data are scarce. Leveraging 1.1 billion anonymized GPS records from 4.4 million devices across three cities in California, we develop a transferable probabilistic framework that learns directed flows between individual places, or points of interest (POIs), from one source city and adapts them to new urban environments under limited supervision. The framework combines three components: a spatiotemporal graph encoder paired with an attention-based edge decoder, whose parameters are independent of graph size and topology, enabling prediction across cities with different POI sets and interaction structures; a heavy-tail-aware zero-truncated Poisson objective that captures the sparse, directed and highly imbalanced nature of POI-to-POI flows; and a staged transfer strategy combining source-city pretraining, cold-start (CS) adaptation and reinforcement learning (RL) refinement. Pretraining establishes reusable interaction representations, CS adaptation calibrates higher-level components under limited target supervision, and RL refinement enables targeted optimization through reward design. A model pretrained on Los Angeles and evaluated on San Francisco and Fresno without any local data, already outperforms fully supervised training on one-month target-city data in overall accuracy. With only one day of local data, the approach delivers stable predictions on high-intensity flows that are competitive with full one-month supervision (up to 35% error reduction relative to single-day supervised training from scratch). More broadly, our findings suggest that cities share structural regularities of mobility, so that transferring mobility predictions becomes largely a matter of recalibrating local scales rather than re-learning the underlying interaction structure. This principle may extend to other networked systems where structure varies across instances.

Results

Street-level mobility network dynamics transfer across cities

To enable street-level mobility flow prediction and cross-city model transfer, we construct directed POI-to-POI mobility networks for three cities in California, using a dataset of 1.1 billion anonymized GPS records from 4.4 million devices collected in March 2023 (Methods).

105 The resulting networks comprise tens of thousands of POIs spanning diverse functional cat-
 106 egories (Supplementary Table 1), with hourly snapshots capturing temporal activity cycles.
 107 Figure 1a-c shows exemplary hourly mobility interaction networks in Los Angeles (38,628
 108 POIs) and San Francisco (28,578 POIs), revealing pronounced differences in urban morphol-
 109 ogy and network topology. While Los Angeles has a polycentric spatial organization with
 110 several mobility hubs alongside more evenly distributed residential areas, San Francisco’s net-
 111 work reflects the city’s compact and centralized urban structure. Individual POI-to-POI flows
 112 are highly sparse and imbalanced across all three cities, following heavy-tailed distributions in
 113 which most edges carry minimal flow while a small subset accounts for disproportionately large
 114 interactions (Extended Data Fig. 1). The cities exhibit systematically different tail regimes:
 115 Los Angeles has the heaviest tail and largest observed flows, San Francisco shows a moder-
 116 ately heavy tail, and Fresno displays a substantially weaker tail. Together, these cities span
 117 contrasting mobility regimes, enabling a systematic evaluation of cross-city transfer under
 118 distinct structural and distributional conditions.

119 Despite such strong differences, we ask whether the underlying interaction dynamics can
 120 transfer across cities under limited supervision. Figure 1d illustrates the cross-city training
 121 pipeline: the framework learns transferable interaction representations from hourly snapshots
 122 of the source-city mobility network and adapts them to a structurally distinct target city with
 123 limited local data.

124 Given the discrete and highly imbalanced nature of the directed POI-level flows, we model
 125 them using a zero-truncated Poisson (ZTP) formulation, providing a principled probabilistic
 126 framework for strictly positive flow counts. To prevent low-flow edges from dominating train-
 127 ing, we adopt a heavy-tail-aware training objective that up-weights high-flow interactions
 128 while preserving the probabilistic formulation. Cross-city transfer is enabled by encoding
 129 sequences of dynamic node and edge representations over past time steps, where edge rep-
 130 resentations are constructed from directed source–target node pairs at each time step. An
 131 attention-based edge decoder then infers directed flow intensities from these representations,
 132 with parameters independent of the size and structure of a given graph, allowing the same
 133 learned interaction rules to be applied across structurally different mobility networks. Transfer
 134 proceeds through three stages: pretraining on a source city establishes transferable interac-
 135 tion representations; cold-start adaptation calibrates higher-level components under limited
 136 target-city supervision; and reinforcement learning refinement enables targeted optimization
 137 through reward design, allowing the model to prioritize high-intensity interactions or overall
 138 accuracy depending on application needs. The model architecture is depicted in Extended
 139 Data Fig. 2 and described in detail, together with the training procedure and evaluation
 140 protocol, in the Methods section.

Source-city pretraining combined with local adaptation yields accurate street-level mobility predictions

Figure 2 illustrates the inferred mobility network in San Francisco alongside bidirectional hourly flow predictions for three exemplary POI pairs, comparing 9-day cross-city adaptation from Los Angeles (CS+RL) with fully supervised training on the complete one-month target-city dataset. Despite the pronounced structural differences between the two cities, the adapted model accurately captures both the spatial organization of POI-to-POI flows and their temporal variation across distinct activity regimes — from the dense midday network at 13:00 to the sparser late-night configuration at 22:00 (Fig. 2a).

At the level of individual POI pairs, 9-day target-city adaptation from a pretrained source-city model reproduces directed asymmetries and temporal activity cycles, achieving performance competitive with fully supervised training on the complete one-month target-city dataset (Fig. 2b). The three example pairs illustrate distinct interaction regimes: asymmetric and phase-shifted dynamics between the school and sports center, directionally synchronized arrivals and departures between the bus stop and school, and broadly co-occurring activity between the fuel station and parking lot pair. Together, these examples demonstrate that source-city pretraining combined with limited target-city adaptation recovers directional, temporal and co-occurrence patterns at the level of individual urban facilities.

Beyond point forecasts, the framework produces full predictive distributions over discrete flow counts (Fig. 2b). The probability heatmaps quantify the likelihood of different flow levels at each time step, enabling assessment of both expected intensity and the probability of extreme movements in each direction. Such distributional outputs capture predictive uncertainty and the likelihood of high-intensity flows, which are particularly relevant for interactions associated with localized capacity stress.

Staged transfer outperforms full supervision under limited target-city data

To systematically evaluate cross-city transfer, we vary the amount of target-city supervision for two target cities, San Francisco and Fresno (Fig. 3, Supplementary Note 5). Performance is evaluated separately on overall flows and on the highest-intensity subset (Top-1%), reflecting the distinct challenges of predicting common low-flow edges and rare but important high-flow corridors.

Zero-shot transfer from Los Angeles to both San Francisco and Fresno outperforms fully supervised training on the complete one-month target-city dataset across overall flow metrics, demonstrating that pretrained interaction representations generalize effectively even without local adaptation. However, overall metrics are dominated by low-intensity edges, and zero-shot transfer degrades in the Top-1% flow regime, reflecting the greater sensitivity of high-intensity corridors to cross-city structural and distributional shifts. Indeed, the pretrained

177 model achieves stronger zero-shot performance in Fresno than in San Francisco, consistent
178 with a smaller distributional shift relative to Los Angeles.

179 Purely supervised training from scratch does not overcome these limitations under sparse
180 supervision. With one day of San Francisco data, supervised training improves ZTP NLL
181 but remains worse than zero-shot transfer on magnitude-based errors (MSE and MAE). With
182 three days, tail error decreases slightly but overall performance deteriorates further, becoming
183 the worst among all methods. The same instability is even more pronounced in Fresno, where
184 extending training from 1 to 3 days substantially increases overall error metrics, consistent
185 with overfitting in sparse regimes where limited high-flow observations amplify sensitivity to
186 noise. These results suggest that scarce supervision biases learning toward a small subset of
187 high-intensity edges while limiting generalization across the broader interaction structure.

188 In San Francisco, cold-start adaptation improves consistently as target-city supervision
189 increases. From 1, 3 and 9 days of target-city data, both overall and Top-1% errors decrease
190 monotonically, demonstrating balanced gains across flow regimes. Remarkably, with only one
191 day of target data, cold-start adaptation reduces Top-1% error by up to 35% relative to one-
192 day supervised training from scratch. With 9 days of target data, it not only surpasses 9-day
193 supervised training across all metrics but also achieves lower overall error than fully supervised
194 training on the complete one-month dataset, while reaching comparable Top-1% performance.
195 In Fresno, cold-start adaptation plays a complementary role: rather than producing monotonic
196 gains, it mitigates the instability observed under naive supervision. Performance remains
197 stable across 1, 3 and 9 days, with only minor non-monotonic variation, indicating that
198 pretrained source-city representations regularize optimization under sparse and weakly heavy-
199 tailed conditions.

200 RL refinement further enables targeted optimization across flow regimes (Fig. 3b). Re-
201 weighting the reward systematically shifts performance between extreme-tail and overall flows,
202 allowing the model to prioritize either high-intensity interactions or more uniform accuracy
203 depending on application needs. The same mechanism adapts to the underlying flow distribu-
204 tion: in San Francisco, improvements concentrate in the heavy tail, whereas in Fresno – where
205 high-flow edges are rarer – identical reward weighting shifts optimization toward reductions
206 in overall error (Supplementary Tables 3–4).

207 Overall, performance follows a clear progression. Zero-shot transfer captures shared in-
208 teraction structure but underestimates extreme flows. Supervised training from scratch be-
209 comes unstable under scarce data, degrading on both overall and tail performance. Cold-start
210 adaptation stabilizes optimization and progressively narrows the gap. RL refinement selec-
211 tively reshapes the objective to the local flow regime, improving tail accuracy where flows are
212 strongly heavy-tailed (San Francisco) and overall accuracy where they are less so (Fresno).

213 Source-city performance against extended graph baselines

214 To our knowledge, no existing method predicts directed POI-level mobility flows, let alone
215 transfers them across cities. Existing spatiotemporal graph forecasting architectures were
216 typically designed to predict node-level traffic at sensor locations on fixed road networks, not
217 directed flows between individual locations. To enable a source-city comparison, we extend
218 three representative graph baselines – Graph WaveNet²⁰, STGCN¹⁸ and DCRNN¹⁹ – with
219 an edge-level prediction layer and a heavy-tail-aware loss (see Methods), allowing them to
220 operate on POI interaction graphs. On overall flows, the proposed model achieves the lowest
221 MSE while remaining competitive on ZTP NLL and MAE (Table 1). In the Top-1% regime,
222 it achieves the best performance on both MSE and ZTP NLL while ranking second on MAE,
223 whereas several baselines deteriorate substantially. The proposed architecture thus maintains
224 strong accuracy across both common low-flow edges and structurally important high-flow
225 corridors, with residual errors concentrated on long-range cross-region edges (Extended Data
226 Fig. 3a). Crucially, the extended baselines remain tied to a fixed adjacency structure defined
227 by the training city and therefore cannot generalize to a structurally different city without
228 retraining. The proposed framework employs a graph-size-independent edge decoder that
229 enables transfer across cities with different POI sets and network topologies. The framework
230 therefore not only matches or exceeds the source-city performance of the extended baselines,
231 but also enables accurate transfer under limited target-city supervision, a capability no exist-
232 ing approach provides.

233 Pretraining learns transferable mobility regularities, while adaptation cor- 234 rects city-specific scale biases

235 To identify what is transferred across cities, we compare mobility regularities learned by
236 the pretrained model of Los Angeles with those recovered after cold-start adaptation in San
237 Francisco (Fig. 4). A consistent picture emerges: pretraining learns structural regularities
238 of urban mobility that transfer directly, while adaptation primarily corrects for city-specific
239 scale biases rather than relearning interaction patterns from scratch.

240 From a representation perspective, node embeddings remain largely stable under transfer
241 (Fig. 4a). As more target-city data are introduced, the embedding space shifts gradually away
242 from the pretrained representation, but the magnitude of this shift remains small, indicating
243 that the geometric structure of representations is largely preserved throughout adaptation.

244 Several mobility regularities transfer directly without requiring recalibration. Population-
245 driven outflow scaling shows near-perfect agreement across zero-shot and adapted models
246 (Spearman’s $\rho \approx 0.93$), confirming that the relationship between the dynamic local popula-
247 tion (number of people present) and outflow intensity is a robust structural regularity that
248 generalizes across cities (Fig. 4b). Weather sensitivity follows a similar pattern: pretrained

and adapted models alike track ground-truth mobility responses across temperature, precipitation and wind speed (Fig. 4c), indicating that environmental modulation of mobility transfers without explicit recalibration. In both cases, very limited supervision introduces minor fluctuations that diminish as additional target-city data are introduced. Distance decay presents a different pattern (Fig. 4d). The decay form — the systematic decrease in flow intensity with increasing distance — transfers directly across cities. However, the absolute flow magnitudes differ systematically between Los Angeles and San Francisco. Increasing target-city data corrects this scale mismatch, shifting the distance-decay curve while preserving its underlying form. This distinguishes scale bias correction from structural relearning: the model does not need to relearn how distance affects mobility, only to recalibrate the magnitude of flows in the new city.

At the level of individual POI-type interactions, transferability is partial (Fig. 4e). Although approximately 73% of POI types overlap between the two cities, the pairings of source and target types differ substantially, limiting direct transfer at the level of exact POI-type pairs. Quantitatively, about half of POI-type pairs show near-identical systematic biases before and after adaptation (residual shifts below 0.02 in log-flow space), indicating that the pretrained model already captures the correct interaction pattern for these pairs. For the remaining pairs, adaptation adjusts the model’s predictions, reflecting city-specific differences in how certain facility types interact.

Comparing the exemplary POI pairs in San Francisco (Fig. 2b) with corresponding pairs in Los Angeles (Extended Data Fig. 3b) illustrates both sides of this finding. The school–sports center and fuel–parking lot pairs show directional patterns broadly consistent across cities, though at lower magnitude in San Francisco, reflecting overall scale differences. In contrast, the bus stop–school pair shows notably different temporal dynamics between the two cities, demonstrating that identical POI categories do not guarantee identical interaction patterns. Functional categories therefore provide a useful prior, but accurate cross-city prediction requires the contextual signals captured by the graph encoder together with local adaptation data to recalibrate city-specific interaction patterns.

Together, these findings reframe cross-city mobility forecasting as the recalibration of shared interaction rules under new structural and distributional constraints, rather than relearning city-specific connectivity from scratch. Transferable regularities — population scaling, weather sensitivity, distance decay form — provide a stable foundation that cold-start adaptation builds upon, correcting for city-specific scale biases while preserving the learned interaction structure.

283 Discussion

284 In this work, we have shown that street-level mobility flow dynamics transfer across struc-
285 turally different cities with distinct flow distributions, even under severe target-city data
286 scarcity. A model pretrained on Los Angeles and evaluated on San Francisco or Fresno with-
287 out any local data already outperforms fully supervised training on a complete one-month
288 target-city dataset in overall accuracy. Limited local data becomes important primarily on
289 high-intensity flows: with only one day of San Francisco data, the framework is competitive
290 with full one-month supervision (up to 35% error reduction relative to supervised training
291 from scratch). In Fresno, where the tail regime is markedly weaker, the same adaptation pro-
292 cedure stabilizes optimization that would otherwise degrade under sparse naive supervision.

293 The cross-city transfer demonstrated here rests on a series of mobility regularities that
294 hold across structurally different urban environments. For instance, pretraining captures
295 population-driven outflow scaling, weather sensitivity, and the form of distance decay —
296 patterns the adapted model reproduces in San Francisco without explicit recalibration. Cold-
297 start adaptation primarily corrects systematic biases in flow magnitudes, while preserving
298 the underlying interaction structure. The result is that limited target-city data are primarily
299 needed to recalibrate how strongly places interact, rather than to relearn which places interact.

300 Fine-grained mobility data at the resolution required for street-level analysis are typi-
301 cally difficult to obtain in most cities, for instance due to privacy regulations. A pretrained
302 model from a single data-rich source city, calibrated using limited target-city observations —
303 for example from traditional travel surveys or volunteered mobility traces — may therefore
304 provide a practical route toward reconstructing fine-grained mobility interactions in cities
305 where comprehensive GPS records are unavailable. In such data-scarce environments, the
306 reconstructions can inform longer-term planning decisions: siting social infrastructure such
307 as parks and community centres, or designing transit networks informed by actual interaction
308 patterns. Independent of data availability, the framework also generates short-term forecasts
309 at hourly resolution, supporting operational use cases such as outbreak response (predicting
310 where mobility may amplify superspreading during an emerging outbreak), demand-responsive
311 transit operations, and crowd-flow management.

312 Our work has several limitations that motivate future research. First, all three cities
313 studied are located in California. Although they differ substantially in spatial organization,
314 they share relatively similar cultural, economic and infrastructural contexts. An important
315 next step is to evaluate transferability across more heterogeneous urban contexts worldwide,
316 where mobility dynamics, infrastructure availability, and data quality may differ substan-
317 tially. Another promising direction is multi-source pretraining, in which diverse cities are
318 leveraged jointly or sequentially to learn more generalizable interaction representations. Such
319 approaches may also enable the integration of smaller but high-resolution mobility datasets

collected across multiple cities¹¹, reducing dependence on a small number of comprehensive proprietary datasets. Second, the present analysis focuses on a one-month observation window. Extending the framework to longer temporal horizons would help determine whether the transferable interaction patterns identified here remain stable across seasons and in response to major infrastructure changes, such as the introduction of new transit lines. Third, GPS traces may carry biases, such as uneven coverage of different population groups²⁸, that may propagate into predictions. The reinforcement learning stage of the framework provides a natural mechanism to mitigate such effects through reward design, for example by incorporating fairness objectives that promote equitable predictive performance across socioeconomic groups²⁹.

In conclusion, our work demonstrates that fine-grained, street-level mobility flows can be transferred across cities, paving the way for effective, privacy-respecting mobility simulations with broad applications. Beyond human mobility networks, many other real-world systems exhibit similar dynamic, directed, and heavy-tailed interactions, in which a small subset of high-intensity connections dominates system behavior. Examples span biological systems (gene regulatory, brain, and metabolic networks^{30–32}), technological systems (software dependency networks³³), economic systems (global supply chains³⁴), and social systems (communication networks³⁵). The framework introduced here therefore points toward a broader class of transferable forecasting approaches for complex networked systems under heterogeneous data availability.

Methods

Spatiotemporal graph formulation

To model the evolving dynamics of urban mobility, we represent the city as a sequence of dynamic graphs that capture both the spatial distribution of population and movement patterns over time.

Data. We use anonymized GPS mobility traces from March 2023 collected by CITYDATA.ai across three cities: Los Angeles, San Francisco and Fresno. The dataset comprises 1.10 billion raw location records, reduced to 103.07 million after cleaning and preprocessing, covering 4.43 million unique anonymized devices over 27 days. We additionally remove distance-based outliers during graph construction. These traces are aggregated into time-indexed population-flow graphs at a fixed temporal resolution, with each snapshot summarizing flows over the corresponding interval. The real-world points of interest are extracted from OpenStreetMap using the Python package OSMnx³⁶. We obtain meteorological data from the ERA5 reanalysis dataset provided by the Copernicus Climate Data Store, including air temperature at 2 meters

above ground, total precipitation, and 10-meter eastward and northward wind components. Wind speed is computed as the Euclidean norm of the two wind vectors.

Node representation. Each graph snapshot comprises a fixed set of nodes representing real-world POIs distributed across the city. For each node, we construct a multi-dimensional feature vector capturing both static and dynamic attributes. These include a categorical encoding of POI type, geographic coordinates, the number of people assigned to the node at each time interval, and three meteorological variables: 2-meter air temperature, total precipitation, and wind speed. With the exception of the node-level population counts, which are calculated from mobility data, all other node attributes are obtained from publicly available or external data sources. As the population counts are defined at the node level, they can be replaced by alternative proxies when mobility data are unavailable.

User locations are mapped to the nearest POI using geodesic great-circle distance on the WGS-84 ellipsoid. To avoid spurious assignments in POI-sparse areas, we apply city-specific snapping thresholds determined by urban density (800 m for Los Angeles and San Francisco, 1000 m for Fresno). Locations farther than the threshold are excluded from node and edge construction, ensuring that each individual is associated with a POI only when spatial proximity is plausible. An assignment to a POI represents spatial proximity rather than an explicit visit, since individuals may also pass through the surrounding area.

All real-value variables are normalized to the range $[0, 1]$ using min-max scaling based on training data statistics. To reduce the influence of extreme values, precipitation is first capped at 50 mm/h and then log-transformed before normalization. In addition to aligning heterogeneous feature scales and enabling efficient model training, normalization also supports cross-city comparability across differing spatial layouts and population densities³⁷. In particular, normalizing geographic coordinates preserves the relative spatial configuration of POIs, allowing the model to reason about urban structure in a city-agnostic manner.

The inclusion of weather-related features is essential for capturing external environmental influences on human movement³⁸. Temperature and precipitation, in particular, strongly affect the temporal variability of mobility flows³⁹, wind speed serves as a proxy for transportation constraints and has been shown to influence travel behavior, particularly during non-work periods⁴⁰. Integrating these exogenous signals allows the model to better anticipate short-term fluctuations in population flow under varying conditions.

Edge construction. We model urban mobility as a directed, edge-valued process over POIs. In principle, the model can operate on the complete POI interaction graph. However, explicitly enumerating all POI pairs would be computationally prohibitive. We therefore introduce a time-indexed candidate edge set $\tilde{\mathcal{E}}_t$ to specify admissible interaction channels, constructed from historically observed transfers or domain constraints such as spatial proximity or transport

390 network connectivity.

391 For any candidate pair $(i, j) \in \tilde{\mathcal{E}}_t$, the prediction target is the directed flow intensity w_{ij}^t ,
 392 defined as the number of users transitioning from POI i to POI j within the subsequent
 393 prediction interval. Transitions are inferred from temporally consecutive GPS records of the
 394 same user after map-matching to POIs, with each valid consecutive POI pair contributing one
 395 count to the corresponding directed edge.

396 To ensure spatial plausibility, transitions whose origin–destination distance exceeds a city-
 397 specific threshold are discarded, and GPS records that cannot be matched to any POI within
 398 a predefined snapping radius are ignored.

399 Evaluation is conducted under an oracle active-edge setting, where candidate edges are
 400 restricted to those active in the prediction window. This protocol isolates conditional flow-
 401 intensity forecasting from edge-activation prediction; modeling edge activation is left for future
 402 work.

403 **Prediction objective.** To construct training samples, we slide a fixed-length window of T
 404 over the sequence of dynamic graphs. Each sample comprises T consecutive snapshots of node
 405 attributes $\{\mathbf{X}_\tau\}_{\tau=t-T+1}^t$ as the input context, where t denotes the reference time at which the
 406 prediction is made. The model takes dynamic node attributes as input signals and predicts
 407 future edge intensities from the evolving node and graph context, rather than by directly
 408 extrapolating historical edge-flow trajectories. At inference time, the candidate edge set can
 409 be specified to define the potential POI interactions of interest.

410 The supervised targets are the edge flows aggregated over the subsequent interval $[t, t+\Delta)$,
 411 where Δ is the temporal resolution. Formally, given the past T graph snapshots, the model
 412 with parameters Θ learns a mapping

$$f_\Theta : \{\mathbf{X}_\tau\}_{\tau=t-T+1}^t \mapsto \widehat{\mathbf{W}}_{t \rightarrow t+\Delta},$$

413 where $\widehat{\mathbf{W}}_{t \rightarrow t+\Delta} = [\hat{w}_{ij}^{t \rightarrow t+\Delta}]_{(v_i, v_j) \in \tilde{\mathcal{E}}_t}$ denotes the predicted flow intensities between all POI
 414 pairs over the next interval $[t, t + \Delta)$, where $\Delta = 1$ h denotes the temporal resolution. Here,
 415 $\hat{w}_{ij}$ and w_{ij} represent the predicted and observed numbers of users moving from POI i to POI
 416 j , respectively.

417 Temporal graph encoder

418 The encoder processes node-level dynamics through four sequential components:

419 **(1) POI embedding and feature fusion** For a mini-batch of B temporal windows, the
 420 raw node features are denoted as $\mathbf{X} \in \mathbb{R}^{B \times N \times F \times T}$, where N is the number of POIs, F the fea-
 421 ture dimension, and T the window length. Among these features, the POI type index, which

422 categorically describes urban functions, is mapped to a dense embedding $\mathbf{E}_{\text{poi}} \in \mathbb{R}^{B \times N \times M}$,
 423 where M denotes the embedding dimension. This embedding is then broadcast along the
 424 temporal dimension. The remaining numerical attributes $\mathbf{X}_r$ are concatenated with this em-
 425 bedding and projected through a linear layer to obtain the fused representation:

$$\mathbf{X}_{\text{proj}} = \text{Linear}([\mathbf{X}_r; \mathbf{E}_{\text{poi}}]), \quad (1)$$

426 where $\mathbf{X}_{\text{proj}} \in \mathbb{R}^{B \times N \times D_t \times T}$ and D_t denotes the input dimension of the temporal encoder.
 427 This representation integrates both semantic and numerical features of POIs, forming the
 428 input to the subsequent temporal encoding stage.

429 **(2) Temporal encoding.** To encode the temporal position of each node feature, fixed
 430 sinusoidal positional encodings $\mathbf{PE}(t)$ are added:

$$\mathbf{Z} = \mathbf{X}_{\text{proj}} + \mathbf{PE}(t). \quad (2)$$

431 **(3) Temporal convolution.** To capture short-range temporal dependencies, we apply a
 432 stack of one-dimensional convolutions along the time axis with Root-mean-square (RMS)
 433 normalization for stable feature scaling. Each layer enables a node to integrate signals from
 434 its recent history by aggregating across neighboring time steps:

$$\mathbf{H}^{(l)} = \text{Dropout}\left(\text{ReLU}\left(\text{RMSNorm}\left(\mathbf{H}^{(l-1)} * \mathbf{W}_e^{(l)}\right)\right)\right), \quad l = 1, \dots, L_e, \quad (3)$$

435 where $\mathbf{H}^{(0)} = \mathbf{Z}$ and $\mathbf{W}_e^{(l)}$ denotes the learnable convolution weights. RMS normalization
 436 stabilizes feature scales, while ReLU introduces nonlinearity and dropout regularizes the rep-
 437 resentation. The module produces temporally enriched node features $\mathbf{H}^{(L_e)}$.

438 **(4) Spatial graph convolution.** To capture spatial dependencies, we apply graph convo-
 439 lutional layers at each time step independently, following the temporal convolution module.
 440 This design explicitly decouples temporal and spatial modeling: temporal patterns are first
 441 captured by 1D convolutions, and spatial relations are then extracted within each snapshot
 442 of the graph. At time step t and for batch b , node embeddings are updated as:

$$\mathbf{H}_{b,t}^{(l)} = \text{ReLU}(\hat{\mathbf{A}}_{b,t} \mathbf{H}_{b,t}^{(l-1)} \mathbf{W}_g^{(l)}) + \mathbf{R}^{(l)} \mathbf{H}_{b,t}^{(l-1)}, \quad l = 1, \dots, L_g, \quad (4)$$

443 where $\mathbf{H}_{b,t}^{(0)} = \mathbf{H}^{(L_e)}[b, :, :, t]$, $\hat{\mathbf{A}}_{b,t}$ denotes the symmetrically normalized adjacency matrix,
 444 $\mathbf{W}_g^{(l)}$ are learnable graph convolution weights, and $\mathbf{R}^{(l)}$ denotes the residual projection⁴¹.
 445 The resulting $\mathbf{H}_{b,t}^{(L_g)}$ encodes spatial structure conditioned on recent temporal context.

446 **(5) Edge representation construction.** To enable edge-level flow prediction, we con-
 447 struct directed edge embeddings from the final node representations. For each edge $(i, j) \in \tilde{\mathcal{E}}_{b,t}$,
 448 we retrieve the corresponding source and target node embeddings $\mathbf{h}_i^{b,t}$ and $\mathbf{h}_j^{b,t}$ from the out-
 449 put of the final graph convolutional layer. A composite edge representation is formed by
 450 concatenating the source embedding, the target embedding and their element-wise absolute
 451 difference:

$$\mathbf{u}_{ij}^{b,t} = [\mathbf{h}_i^{b,t} \parallel \mathbf{h}_j^{b,t} \parallel |\mathbf{h}_i^{b,t} - \mathbf{h}_j^{b,t}|]. \quad (5)$$

452 This representation vector is passed through a shared linear projection to obtain the edge
 453 embedding $\mathbf{e}_{ij}^{b,t} \in \mathbb{R}^{D_g}$, where D_g is the edge embedding dimension. The resulting edge
 454 embeddings are organized into an edge-token sequence $\mathbf{E}$, and a binary validity mask $\mathbf{A}$
 455 indicates valid edge entries across time steps and prevents attention to padded positions.
 456 Together, $\mathbf{E}$ and $\mathbf{A}$ are passed to the attention-based edge decoder.

457 Because prediction is formulated at the edge level, the output space is decoupled from fixed
 458 node indexation. The decoder operates on directed edge embeddings rather than outputs tied
 459 to a fixed set of nodes, so its parameters are independent of the number or ordering of nodes
 460 in a particular graph. This formulation allows the same decoding architecture to be applied
 461 across graphs with varying sizes and topologies.

462 Attention-based edge decoder

463 **Decoder processing.** The decoder operates on edge embeddings to model dependencies
 464 across candidate interactions, inferring edge-level flow intensities conditioned on global spa-
 465 tiotemporal dynamics rather than directly extrapolating individual edge trajectories. This
 466 conditioning is achieved through cross-attention over the encoder outputs, which provide the
 467 global context for interaction inference. For readability, we omit the prediction-time super-
 468 script from decoder representations and intensity parameters when no ambiguity arises.

469 A stack of L_d pre-normalized Transformer decoder layers⁴² is used to implement this mech-
 470 anism. Each layer comprises attention and feed-forward modules with residual connections.

471 Given the edge embedding sequence $\mathbf{E}$, encoder memory $\mathbf{M}$ and validity mask $\mathbf{A}$, the
 472 decoder updates representations as

$$\mathbf{Z}^{(l)} = \text{TransformerLayer}(\mathbf{Z}^{(l-1)}, \mathbf{M}; \mathbf{A}), \quad l = 1, \dots, L_d, \quad (6)$$

473 with $\mathbf{Z}^{(0)} = \mathbf{E}$.

474 The final decoded representation $\mathbf{Z}^{(L_d)} = [\mathbf{z}_{ij}^{(L_d)}]_{(v_i, v_j) \in \tilde{\mathcal{E}}_t}$ is linearly projected to produce
 475 nonnegative edge-wise intensities $\boldsymbol{\lambda} = [\lambda_{ij}]_{(v_i, v_j) \in \tilde{\mathcal{E}}_t}$, which parameterize the subsequent ZTP
 476 distribution.

Flow prediction. POI-to-POI flows are positive count observations within fixed temporal windows and display pronounced skewness. Poisson-family models are widely used in mobility⁴³ and transportation research⁴⁴ to represent such count data. We therefore model interaction intensities using a zero-truncated Poisson likelihood, providing a principled probabilistic formulation for positive discrete counts rather than treating flows as continuous regression targets⁴⁵. Note that, at the aggregate level, mixing heterogeneous Poisson processes across edges and time yields heavy-tailed marginal behavior, even though each conditional distribution is light-tailed. To account for this imbalance, we introduce a heavy-tail-aware training objective that up-weights high-intensity interactions.

The decoder outputs a nonnegative intensity λ_{ij} for each directed edge (v_i, v_j) :

$$\lambda_{ij} = \text{Softplus}(\mathbf{W}_{\text{out}} \mathbf{z}_{ij}^{(L_d)} + \mathbf{b}_{\text{out}}) + \varepsilon, \quad (7)$$

where $\mathbf{z}_{ij}^{(L_d)}$ is the final edge embedding, $\mathbf{W}_{\text{out}}$ and $\mathbf{b}_{\text{out}}$ are learnable parameters, and ε denotes a small positive constant introduced to maintain numerical stability. The output λ_{ij} serves as the intensity parameter of the ZTP distribution governing the expected flow from POI i to POI j .

Given observed counts $w_{ij} \in \{1, 2, \dots\}$, the unweighted negative log-likelihood is

$$\ell(w_{ij}, \lambda_{ij}) = - \sum_{(v_i, v_j) \in \tilde{\mathcal{E}}_t} \left[w_{ij} \log(\lambda_{ij} + \varepsilon) - \lambda_{ij} - \log(1 - e^{-\lambda_{ij}}) - \log(w_{ij}!) \right]. \quad (8)$$

For evaluation or prediction, we compute the conditional mean of the ZTP distribution and round it to the nearest integer to obtain discrete flow intensity:

$$\mu_{ij} = \frac{\lambda_{ij}}{1 - e^{-\lambda_{ij}}}, \quad \hat{w}_{ij} = \text{round}(\mu_{ij}). \quad (9)$$

This formulation preserves the discrete, strictly positive support of POI-to-POI flows, mitigates bias under heavy-tailed count distributions, and provides explicit predictive distributions over positive discrete flows.

The detailed model architecture, including the full specification of encoder and decoder components and the layer-wise cold-start adaptation protocol, is provided in Extended Data Fig. 2b.

Three-stage transfer learning

To enable robust cross-city generalization and adaptive transfer, we design a three-stage transfer learning procedure: (1) supervised pretraining on a source city, (2) cold-start adaptation on a data-scarce target city, and (3) RL refinement under multi-objective feedback.

504 Stage 1: Pretraining

505 We pretrain the model on large-scale POI–POI population flows from Los Angeles to learn
 506 transferable spatiotemporal representations. Given the heavy-tailed distribution of positive
 507 edge flows (Extended Data Fig. 1), we adopt a weighted zero-truncated Poisson (wZTP) NLL
 508 to rebalance training contributions across flow regimes.

509 Specifically, we assign each edge a deterministic weight $\beta(w_{ij})$ that downweights low-flow
 510 interactions and applies controlled emphasis to larger flows. In the low- and medium-flow
 511 regime, the weight takes the piecewise-constant form $\beta(w_{ij}) = \zeta_k$ for $w_{ij} \in [\tau_{k-1}, \tau_k)$, with
 512 thresholds $0 = \tau_0 < \tau_1 < \dots < \tau_K < \tau_M$.

513 For higher flows, i.e., $w_{ij} \in [\tau_K, \tau_M]$, the weighting transitions to a logarithmic tail-
 514 weighting function $\phi(w_{ij})$, which increases smoothly with flow intensity and is capped to
 515 ensure numerical stability:

$$\beta(w_{ij}) = \begin{cases} \zeta_k, & w_{ij} \in [\tau_{k-1}, \tau_k), \quad k = 1, \dots, K, \\ \phi(w_{ij}), & \tau_K \leq w_{ij} \leq \tau_M, \\ 1, & w_{ij} > \tau_M. \end{cases} \quad (10)$$

516 This weighting scheme moderately amplifies medium- and high-flow mobility edges while
 517 preventing domination by extreme outliers, and downscales ubiquitous low-flow edges to im-
 518 prove training balance. The thresholds and weights are determined from empirical flow statis-
 519 tics in the available training data. The corresponding weighted objective becomes

$$\ell_w(w_{ij}, \lambda_{ij}) = \beta(w_{ij}) \ell(w_{ij}, \lambda_{ij}). \quad (11)$$

520 Finally, the overall pretraining objective averages the per-edge wZTP NLL over all directed
 521 edges in the training batch set $\mathcal{E}_{\text{batch}}$:

$$\mathcal{L}_{\text{train}} = \frac{1}{|\mathcal{E}_{\text{batch}}|} \sum_{(i,j) \in \mathcal{E}_{\text{batch}}} \ell_w(w_{ij}, \lambda_{ij}). \quad (12)$$

522 We jointly optimize the graph encoder and attention-based edge decoder in an end-to-end
 523 fashion using AdamW with decoupled weight decay⁴⁶. A moderate learning rate and weight
 524 regularization are employed to ensure stable convergence during pretraining.

525 Stage 2: Cold-start adaptation

526 In the second stage, the pretrained model is adapted to each target city through cold-start
 527 adaptation, using the same wZTP NLL formulation as in Eq. (11), with stage-specific pa-
 528 rameter settings. This setting simulates realistic deployment conditions in which only limited

labeled flow data is available in the target domain. Rather than reinitializing the model or retraining all parameters, we update a subset of the architecture while freezing the remaining layers. Specifically, lower layers of the temporal convolution, graph convolution, and edge decoder are frozen to preserve pretrained representations, while parameter updates are restricted to higher layers and the final output projection. This selective fine-tuning allows the model to specialize to local mobility dynamics while preserving shared spatiotemporal priors and reducing overfitting under limited supervision.

Stage 3: RL refinement

We formulate the refinement stage as a single-step RL task, in which the agent observes a compact summary of the pretrained model’s predictive statistics and graph dynamics, proposes continuous adjustments to the decoder’s output-layer parameters, and receives an immediate mixed reward that explicitly weights global error against tail errors on high-impact edges. This closed-loop single-step feedback enables direct policy optimization without multi-step rollouts, allowing reward-driven refinement to serve as an objective-shaping mechanism that adapts to varying heavy-tail regimes.

State. The state is represented as a 1,024-dimensional vector that summarizes the short-term dynamics of the urban mobility graph. Specifically, it is constructed by computing the mean, standard deviation, maximum, and minimum of the 256-dimensional edge embedding differences between consecutive time steps, with each statistic contributing 256 dimensions. During RL refinement, the data loader stride is set to a fraction of the sliding window size to balance temporal continuity and sample diversity while reducing the risk of state overfitting. The resulting state vector is standardized to zero mean and unit variance to stabilize policy optimization.

Action. At each environment step, the agent outputs a four-dimensional continuous action $\mathbf{a} = [a_{\text{low}}, a_{\text{mid}}, a_{\text{high}}, a_{\text{bias}}]$ to perform structured updates on the decoder’s edge-output layer. Specifically, the weight vector of the final linear head is partitioned into low-, mid-, and high-variance index groups according to the principal component structure derived from source-city training data. These groups facilitate low-dimensional, structured modulation of the output parameters, thereby reducing the action space while preserving the pretrained representation structure. The first three action components apply group-wise multiplicative scaling to the corresponding subsets of weights, i.e., $w_g \leftarrow w_g(1 + a_g)$ for $g \in \{\text{low}, \text{mid}, \text{high}\}$, while the last component applies an additive bias shift, $b \leftarrow b + a_{\text{bias}}$. After each action, both weights and bias are clamped within fixed bounds to avoid disruptive parameter drift, thereby preserving the pretrained structure while enabling stable and sample-efficient updates during RL refinement.

Reward. We use a weighted combination of MSE and MAE as the RL reward. The weighted MSE term provides sensitivity to deviations on structurally important medium- and high-flow edges, while the weighted MAE term offers robustness to outliers and complements the sensitivity of MSE^{47,48}. To maintain consistency with supervised pretraining, edge-wise weights $\beta_{\text{RL}}(w_{ij})$ follow the same piecewise tail-aware weighting design as pretraining, downweighting ubiquitous low-flow edges and applying controlled emphasis to larger flows.

Given predicted and observed flows $\hat{w}_{ij}$ and w_{ij} , we compute

$$e_{\text{wmse}} = \frac{\sum_{(i,j)} \beta_{\text{RL}}(w_{ij}) (\hat{w}_{ij} - w_{ij})^2}{\sum_{(i,j)} \beta_{\text{RL}}(w_{ij})}, \quad (13)$$

$$e_{\text{wmae}} = \frac{\sum_{(i,j)} \beta_{\text{RL}}(w_{ij}) |\hat{w}_{ij} - w_{ij}|}{\sum_{(i,j)} \beta_{\text{RL}}(w_{ij})}. \quad (14)$$

The scalar reward returned to the agent is

$$r_t = -(\delta_{\text{mse}} e_{\text{wmse}} + \delta_{\text{mae}} e_{\text{wmae}}), \quad (15)$$

where δ_{mse} and δ_{mae} are non-negative coefficients controlling the relative contribution of the weighted MSE and MAE terms. Each episode corresponds to a single prediction window, providing immediate reward feedback for stable policy optimization.

Learning algorithm. We adopt Proximal Policy Optimization (PPO)⁴⁹ to optimize the policy in a stable and sample-efficient manner. PPO updates a stochastic policy using a clipped surrogate objective that constrains policy changes between iterations, improving training stability in continuous action spaces. The actor and value networks are jointly trained using standard PPO procedures, with entropy regularization to encourage exploration.

Evaluation

The Los Angeles dataset is used for pretraining and the San Francisco and Fresno datasets are reserved for transfer evaluation. For pretraining, the time series is divided into training, validation and test sets in approximate proportions of 70%, 15% and 15%, respectively. For cross-city adaptation, models are initialized from Los Angeles pretraining and evaluated under three target-city data regimes — 1, 3 and 9 days — reflecting realistic deployment conditions ranging from minimal to moderate local supervision (Supplementary Fig. 1). For the 9-day scenario, target-city data are split into 7 days for cold-start training and 2 days for validation; RL refinement is then performed on the full 9-day dataset. For the 3- and 1-day scenarios, cold-start training and RL refinement are conducted on the full dataset; given the limited data regime, the cold-start model is trained for a fixed five epochs without a separate validation

split. In all cases, the test set corresponds to the final 15% of the target-city time series, ensuring that test data always follows the training and adaptation period in time. Performance is reported separately on overall flows and on the highest-intensity subset (Top-1%), reflecting the distinct challenges of predicting common low-flow edges and rare but important high-flow corridors.

Training and validation NLL trajectories of the pretraining are provided in Supplementary Fig. 2. Reward dynamics during RL refinement in San Francisco are shown in Supplementary Fig. 3. An ablation study assessing the contribution of each model component is provided in Supplementary Note 4.

Baseline models

We compare the proposed model against three representative spatiotemporal graph baselines: Graph WaveNet²⁰, STGCN¹⁸ and DCRNN¹⁹. These models were originally designed for node-level or aggregate traffic forecasting on fixed road networks and do not natively support directed edge-level prediction. We therefore extend each architecture with an explicit edge-level prediction layer for pairwise flow estimation, and reformulate the training objective using the same weighted zero-truncated Poisson loss as the proposed model to account for the heavy-tailed flow distribution. All baselines are trained under full supervision on the Los Angeles dataset, using the same data representation, evaluation protocol and prediction setting – forecasting the next one-hour flow snapshot from the past twelve hourly intervals – with hyperparameters individually tuned for fair comparison. Because these baselines rely on a fixed adjacency matrix tied to the training city’s mobility network, they cannot be applied to a structurally different city without retraining.

References

- [1] Batty, M. *The New Science of Cities* (MIT Press, 2013).
- [2] Bettencourt, L. M. A. *Introduction to Urban Science: Evidence and Theory of Cities as Complex Systems* (MIT Press, 2021).
- [3] Fujita, M., Krugman, P. R. & Venables, A. *The Spatial Economy: Cities, Regions, and International Trade* (MIT Press, 2001).
- [4] Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, 1991).
- [5] Barbosa, H. *et al.* Human mobility: Models and applications. *Physics Reports* **734**, 1–74 (2018).

- [6] Alessandretti, L., Aslak, U. & Lehmann, S. The scales of human mobility. *Nature* **587**, 402–407 (2020).
- [7] Chang, S. *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2021).
- [8] Moro, E., Calacci, D., Dong, X. & Pentland, A. Mobility patterns are associated with experienced income segregation in large us cities. *Nature Communications* **12**, 4633 (2021).
- [9] Nilforoshan, H. *et al.* Human mobility networks reveal increased segregation in large cities. *Nature* **624**, 586–592 (2023).
- [10] Yabe, T., García Bulle Bueno, B., Frank, M. R., Pentland, A. & Moro, E. Behaviour-based dependency networks between places shape urban economic resilience. *Nature Human Behaviour* **9**, 496–506 (2025).
- [11] Yabe, T. *et al.* Enhancing human mobility research with open and standardized datasets. *Nature Computational Science* **4**, 469–472 (2024).
- [12] Zipf, G. K. The P1P2/D hypothesis: on the intercity movement of persons. *American Sociological Review* **11**, 677–686 (1946).
- [13] Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).
- [14] Schlöpfer, M. *et al.* The universal visitation law of human mobility. *Nature* **593**, 522–527 (2021).
- [15] Simini, F., Barlacchi, G., Luca, M. & Pappalardo, L. A deep gravity model for mobility flows generation. *Nature Communications* **12**, 6576 (2021).
- [16] Rong, C., Feng, J. & Ding, J. Goddag: Generating origin-destination flow for new cities via domain adversarial training. *IEEE Transactions on Knowledge and Data Engineering* **35**, 10048–10057 (2023).
- [17] Yang, J. *et al.* Transferable human mobility network reconstruction with neuroGravity. Preprint at <https://arxiv.org/abs/2604.23678> (2026).
- [18] Yu, B., Yin, H. & Zhu, Z. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3634–3640 (2018).

- [19] Li, Y., Yu, R., Shahabi, C. & Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations* (2018).
- [20] Wu, Z., Pan, S., Long, G., Jiang, J. & Zhang, C. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 1907–1913 (2019).
- [21] Li, Z. *et al.* Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5351–5362 (2024).
- [22] Liu, C. *et al.* Spatial-temporal large language model for traffic prediction. In *2024 25th IEEE International Conference on Mobile Data Management (MDM)*, 31–40 (IEEE, 2024).
- [23] Xue, J. *et al.* Quantifying the spatial homogeneity of urban road networks via graph neural networks. *Nature Machine Intelligence* **4**, 246–257 (2022).
- [24] Shao, J., Li, S., Zhang, K., Wang, A. & Li, M. Cross-city traffic prediction based on deep domain adaptive transfer learning. *Transportation Research Part C: Emerging Technologies* **176**, 105152 (2025).
- [25] Yang, Y., Zhan, J., Liu, Y. & Wang, Q. Cross-city transfer learning: Applications and challenges for smart cities and sustainable transportation. *Communications in Transportation Research* **5**, 100206 (2025).
- [26] Wang, Y., Zheng, T., Liang, Y., Liu, S. & Song, M. Cola: Cross-city mobility transformer for human trajectory simulation. In *Proceedings of the ACM Web Conference 2024*, 3509–3520 (2024).
- [27] Jin, Y., Chen, K. & Yang, Q. Selective cross-city transfer learning for traffic prediction via source city region re-weighting. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 731–741 (2022).
- [28] Chen, C. *et al.* From biases to opportunities: leveraging location-based-service (LBS) data for next-generation transportation planning. *Transportation Research Part C: Emerging Technologies* **182**, 105416 (2026).
- [29] Liu, Z., Huang, L., Fan, C. & Mostafavi, A. Generating equitable urban human flows with a fairness-aware deep learning model. *Cities* **167**, 106296 (2025).
- [30] Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nature Methods* **9**, 796–804 (2012).

- [31] Bassett, D. S. & Sporns, O. Network neuroscience. *Nature Neuroscience* **20**, 353–364 (2017).
- [32] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
- [33] Decan, A., Mens, T. & Grosjean, P. An empirical comparison of dependency network evolution in seven software packaging ecosystems. *Empirical Software Engineering* **24**, 381–416 (2019).
- [34] Pichler, A. *et al.* Building an alliance to map global supply networks. *Science* **382**, 270–272 (2023).
- [35] Onnela, J.-P. *et al.* Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* **104**, 7332–7336 (2007).
- [36] Boeing, G. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* **65**, 126–139 (2017).
- [37] Jiang, W. & Luo, J. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications* **207**, 117921 (2022).
- [38] Pappalardo, L., Manley, E., Sekara, V. & Alessandretti, L. Future directions in human mobility science. *Nature Computational Science* **3**, 588–600 (2023).
- [39] Zhang, F., Li, Z., Li, N. & Fang, D. Assessment of urban human mobility perturbation under extreme weather events: A case study in Nanjing, China. *Sustainable Cities and Society* **50**, 101671 (2019).
- [40] Brum-Bastos, V. S., Long, J. A. & Demšar, U. Weather effects on human mobility: a study using multi-channel sequence analysis. *Computers, Environment and Urban Systems* **71**, 131–152 (2018).
- [41] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
- [42] Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems* (Curran Associates, 2017).
- [43] Xie, Z., Huang, B., Cai, J. & Lee, H. F. Commuting flow patterns across a triad of chinese megacities: Evidence from poisson gravity modeling using mobile device data. *Journal of Transport Geography* **128**, 104380 (2025).

- 717 [44] Ulvi, H., Yerlikaya, M. A. & Yildiz, K. An innovative mathematical model for inte-
 718 grated traffic flow optimization in organized industrial zones. *Scientific Reports* **15**,
 719 29427 (2025).
- 720 [45] Zhou, Q. *et al.* Explainable origin-destination crowd flow interpolation via variational
 721 multi-modal recurrent graph auto-encoder. In *Proceedings of the AAAI Conference on*
 722 *Artificial Intelligence*, vol. 38, 9422–9430 (2024).
- 723 [46] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International*
 724 *Conference on Learning Representations* (2019).
- 725 [47] Willmott, C. J. & Matsuura, K. Advantages of the mean absolute error (MAE) over
 726 the root mean square error (RMSE) in assessing average model performance. *Climate*
 727 *Research* **30**, 79–82 (2005).
- 728 [48] Hodson, T. O. Root mean square error (RMSE) or mean absolute error (MAE): When
 729 to use them or not. *Geoscientific Model Development Discussions* **2022**, 1–10 (2022).
- 730 [49] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy
 731 optimization algorithms. Preprint at <https://arxiv.org/abs/1707.06347> (2017).
- 732 [50] Healy, J. & McInnes, L. Uniform manifold approximation and projection. *Nature Reviews*
 733 *Methods Primers* **4**, 82 (2024).

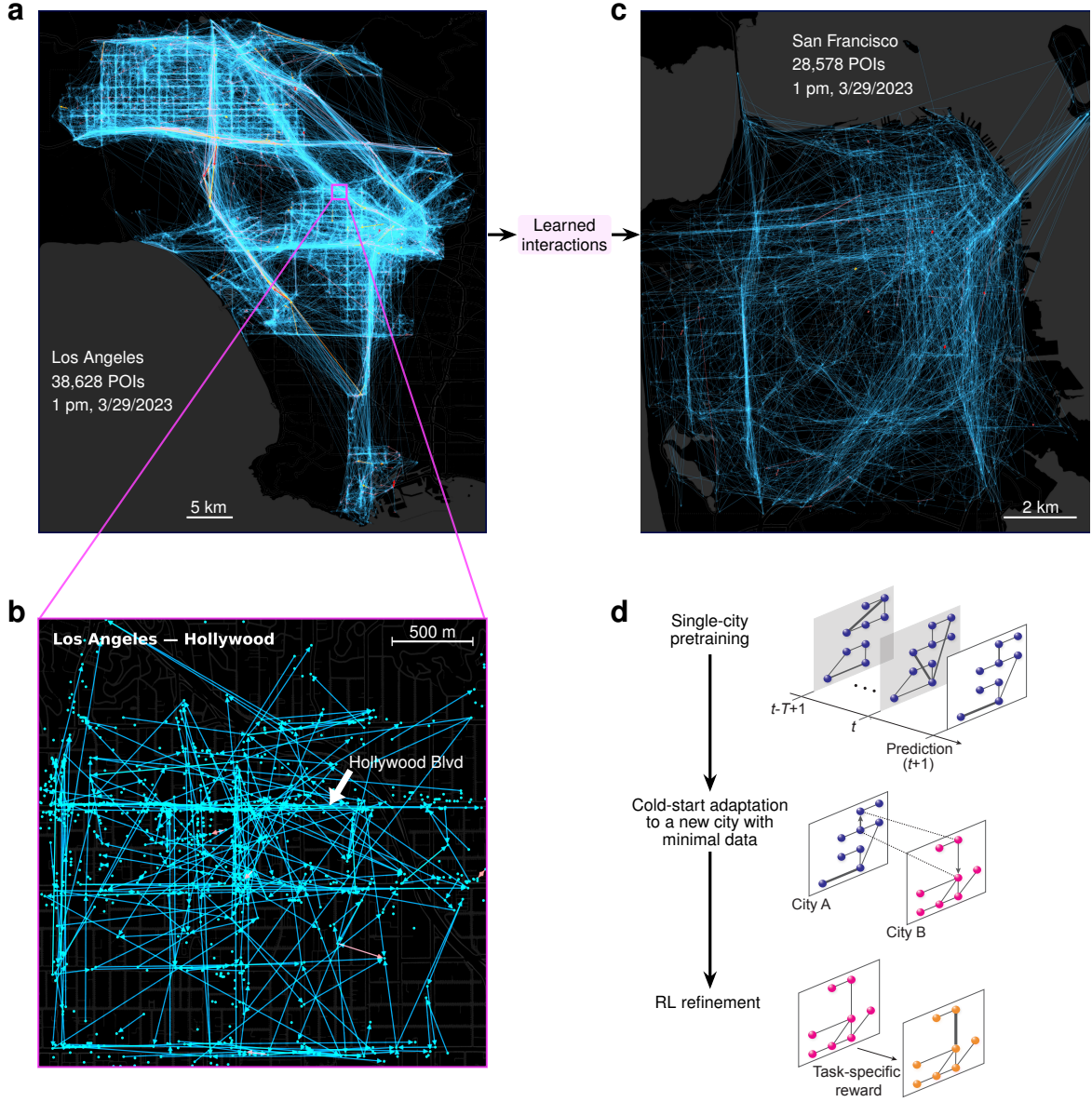


Fig. 1 | Street-level mobility network dynamics transfer across structurally different cities. **a**, Directed POI-to-POI mobility network in Los Angeles (38,628 POIs) at 13:00 on 29 March 2023. Edge color and opacity encode flow intensity; arrows indicate direction. **b**, Zoom into the Hollywood neighborhood illustrating fine-grained POI-level mobility interactions. **c**, POI-to-POI mobility network in San Francisco (28,578 POIs) at the same timestamp, showing substantial differences in spatial organization and topology relative to Los Angeles. **d**, Cross-city training pipeline: the model learns transferable interaction representations from hourly source-city snapshots and adapts them to a structurally distinct target city under limited supervision.

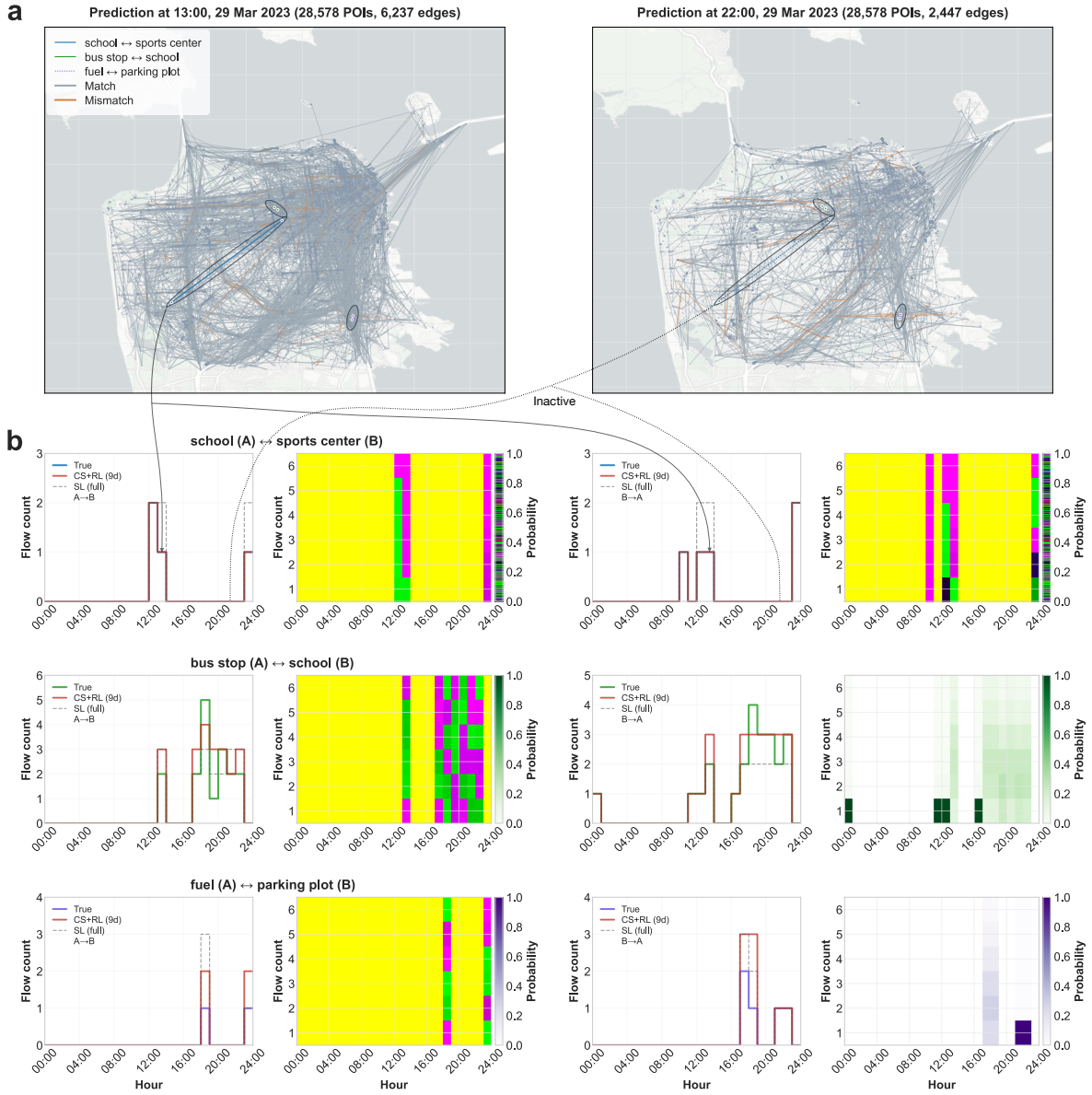


Fig. 2 | Cross-city adaptation accurately resolves directed mobility interactions at street level. **a**, POI-to-POI flows in San Francisco at 13:00 and 22:00 on 29 March 2023 after 9-day cross-city adaptation from Los Angeles. Despite San Francisco’s more compact and centralized mobility structure relative to Los Angeles, the adapted model accurately captures the spatial organization of POI-to-POI flows and their temporal variation. Edges are labeled as matches if the absolute prediction error in both directions is at most 1; otherwise they are labeled as mismatches. **b**, Bidirectional hourly flow counts for three exemplary POI pairs, comparing 9-day cross-city adaptation (CS+RL) against fully supervised training on the complete one-month target-city dataset (SL full). For each pair, flow counts in both directions are shown together with the ZTP probability heatmap over discrete flow intensities. Nine-day cross-city adaptation achieves performance competitive with full target-city supervision.

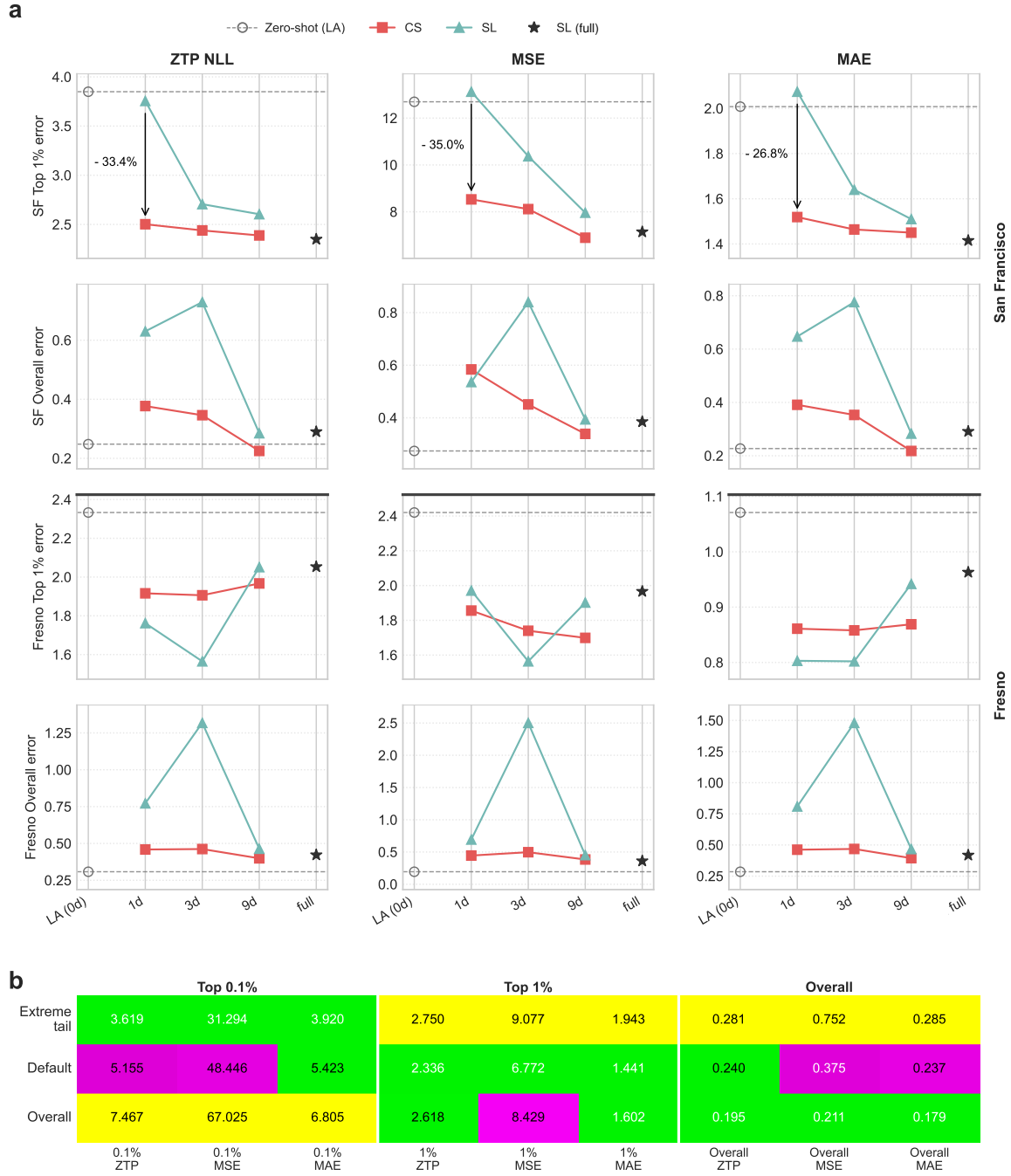


Fig. 3 | Staged transfer with limited local data approaches or exceeds supervised learning with full local supervision. **a**, Cross-city transfer performance in San Francisco (top) and Fresno (bottom). Models pretrained on Los Angeles are adapted to each city using 1, 3 and 9 days of local data. For each city, the upper row reports errors on the Top-1% flows and the lower row reports overall errors; columns show ZTP NLL, MSE and MAE. The dashed line denotes zero-shot transfer and the star indicates full one-month supervision. **b**, Reward reweighting in San Francisco under 9-day adaptation steers the trade-off between extreme-tail and overall performance. Darker colors indicate lower error within each metric. CS, cold-start adaptation; MSE, mean squared error; MAE, mean absolute error; ZTP NLL, zero-truncated Poisson negative log-likelihood.

Table 1 | Source-city performance compared against extended spatiotemporal graph baselines in Los Angeles

Method	Top-1% ZTP NLL	Top-1% MSE	Top-1% MAE	Overall ZTP NLL	Overall MSE	Overall MAE
Proposed model	2.105	3.439	<u>1.002</u>	<u>0.271</u>	0.208	<u>0.253</u>
Extended GWN	2.140	5.852	1.068	0.339	<u>0.278</u>	0.329
Extended STGCN	2.343	<u>4.965</u>	1.089	0.216	0.358	0.191
Extended DCRNN	<u>2.119</u>	6.032	0.998	0.293	0.330	0.282

Note: Bold indicates the best result; underline indicates the second best. Extended baselines are adapted to operate on POI interaction graphs with an edge-level prediction layer and heavy-tail-aware loss formulation (see text for details).

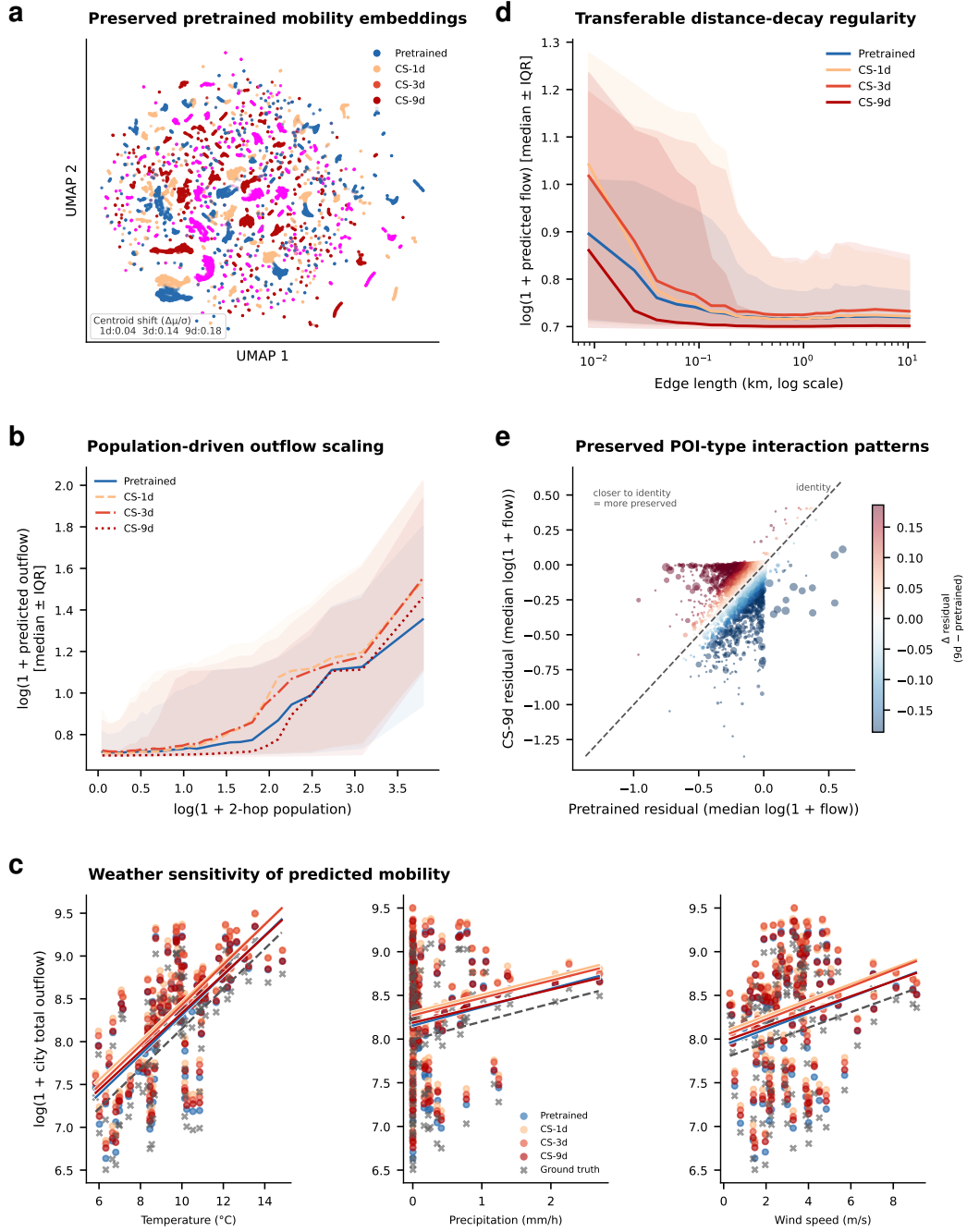
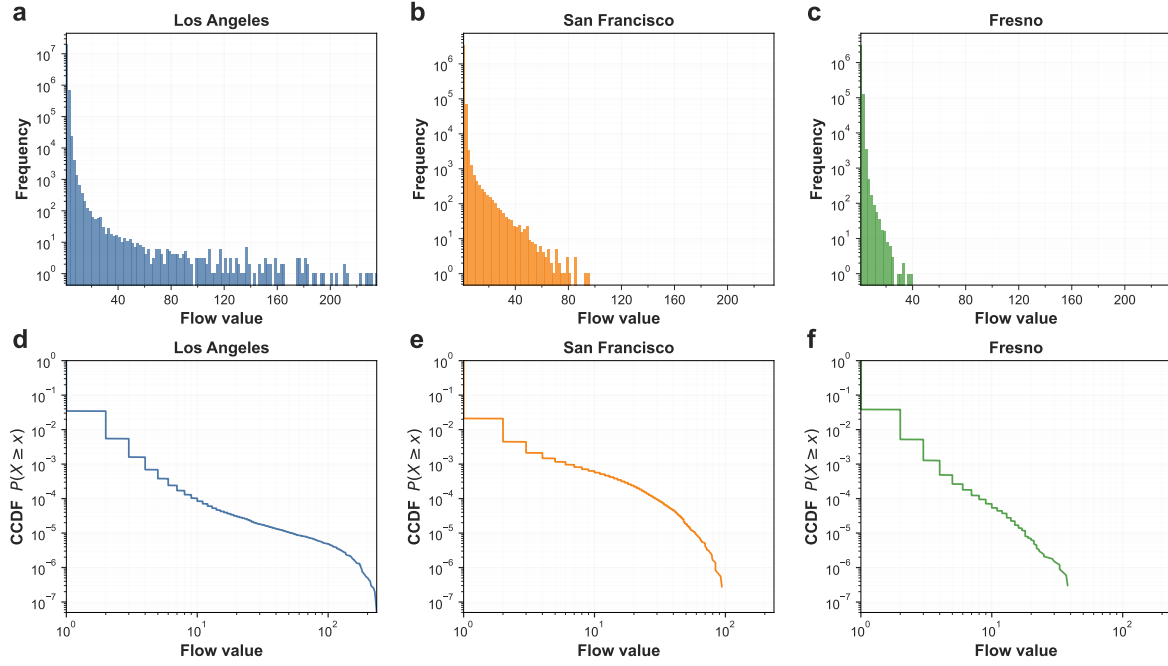
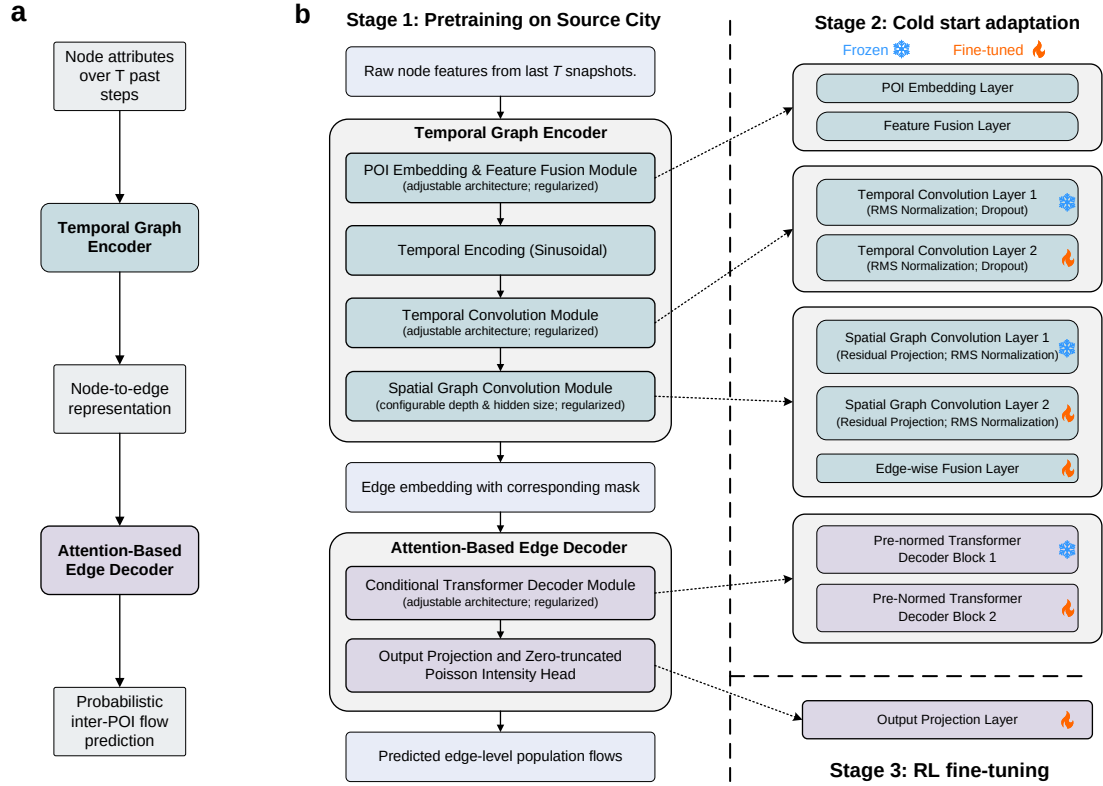


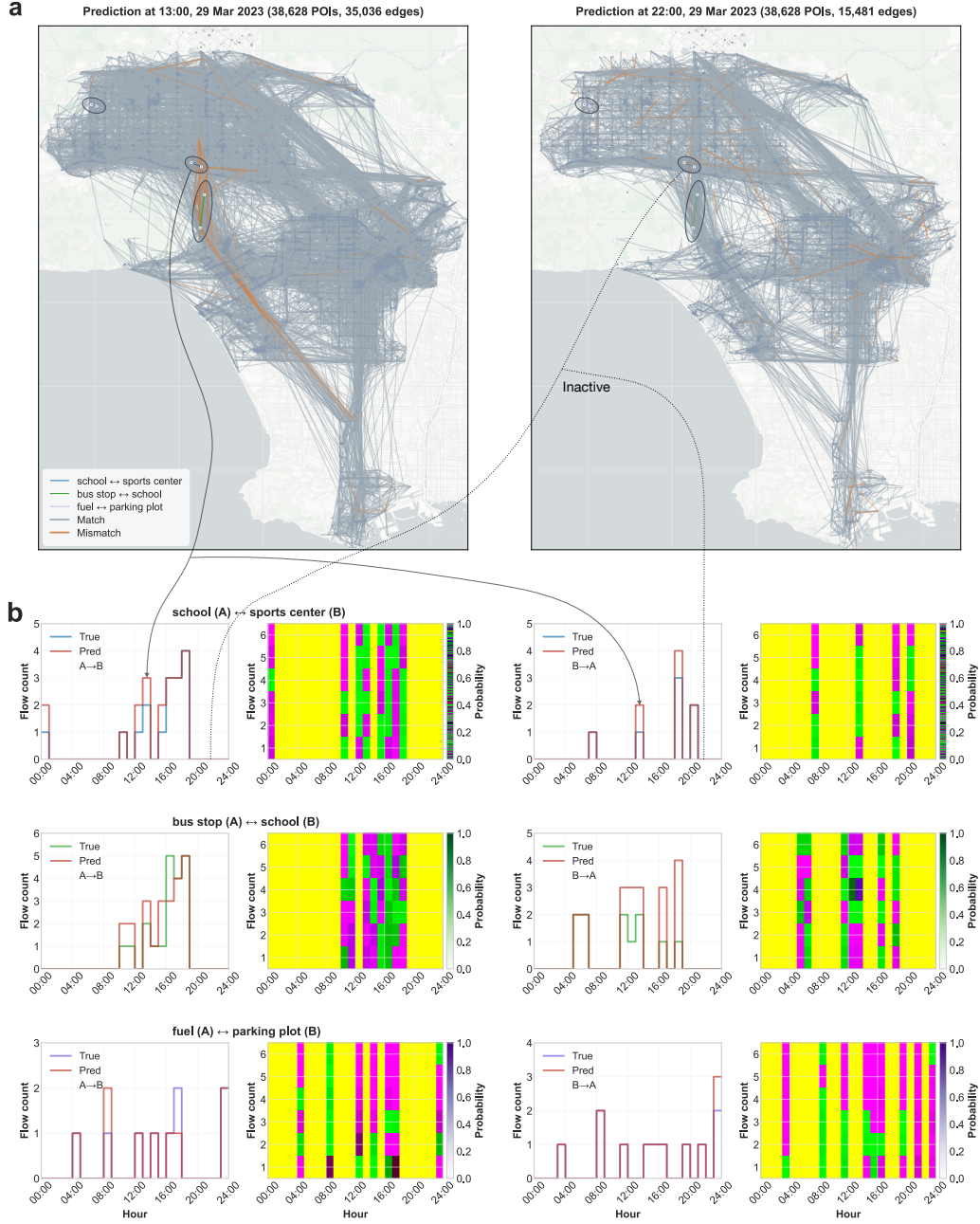
Fig. 4 | Pretraining learns structural regularities that transfer directly, while adaptation corrects city-specific scale biases. **a**, Two-dimensional UMAP⁵⁰ projection of San Francisco’s POI embeddings under zero-shot transfer and cold-start adaptation with 1, 3 and 9 days of local data; embeddings remain close to the pretrained distribution (normalized centroid shift of $\Delta\mu/\omega = 0.04 - 0.18$). **b**, Outflow scaling with dynamic two-hop node population (Spearman’s $\rho \approx 0.93$). **c**, Weather sensitivity of predicted city-level mobility; pretrained and adapted models follow ground-truth trends. CS: cold-start adaptation; 1d, 3d, 9d: number of adaptation days. **d**, Distance decay of predicted flows (median and IQR); the decay form transfers directly, while adaptation corrects scale mismatches in absolute magnitude. **e**, POI-type interaction residuals comparing zero-shot transfer and CS-9d; about half of POI-type pairs show residual shifts below 0.02 in log-flow space.



Extended Data Fig. 1 | Heavy-tailed distribution of POI-to-POI population flows in three cities. **a-c**, Histogram of edge-level population flows (LA, SF, and FRE) with a logarithmic y -axis, showing a highly skewed distribution with a large mass at small counts. **d-f**, Complementary cumulative distribution function (CCDF) of LA, SF, FRE, $P(X \geq x)$, on log-log axes, highlighting the heavy-tailed behavior where a small fraction of edges carry disproportionately large flows. Together, the panels indicate strong over-dispersion (variance exceeding the mean) and pronounced inequality across edges, consistent with real-world urban mobility patterns.



Extended Data Fig. 2 | Model architecture for transferable POI-to-POI mobility flow forecasting. a, Overview of the proposed framework. b, Detailed model architecture.



Extended Data Fig. 3 | Spatial and temporal validation of POI-to-POI flow predictions in Los Angeles. **a**, Flows at 13:00 and 22:00 on 29 March 2023, predicted by the pretrained model. The model reproduces the city’s modular spatial organization (hubs surrounded by residential districts). Prediction errors concentrate on long-range cross-region edges, which depend less on local neighborhood structure and are inherently more difficult to forecast than densely connected intra-zone flows. Edges are shown as match if the absolute prediction error in both directions is at most 1, and as mismatch otherwise. **b**, Bidirectional flow prediction for three representative POI pairs. For each POI pair, hourly population transfers (rounded conditional mean) in both directions ($A \rightarrow B$ and $B \rightarrow A$) are shown, together with the corresponding zero-truncated Poisson (ZTP) probability heatmap over discrete flow intensities. The model resolves distinct temporal patterns across functionally different POI pairs.

Supplementary Information:
**Street-level mobility interactions transfer across
cities with minimal local data**

Hongrong Yang^{1,2} and Markus Schläpfer^{1,2,*}

¹Department of Civil Engineering and Engineering Mechanics, Columbia University,
New York, NY, USA

²Data Science Institute, Columbia University, New York, NY, USA

*Email: m.schlaepfer@columbia.edu

Contents

Supplementary Note 1: Point of interest (POI) categories	3
Supplementary Note 2: Cross-city evaluation setup	4
Supplementary Note 3: Training dynamics	5
Supplementary Note 4: Ablation and component analysis	7
Supplementary Note 5: Full numerical results for the cross-city performance	8

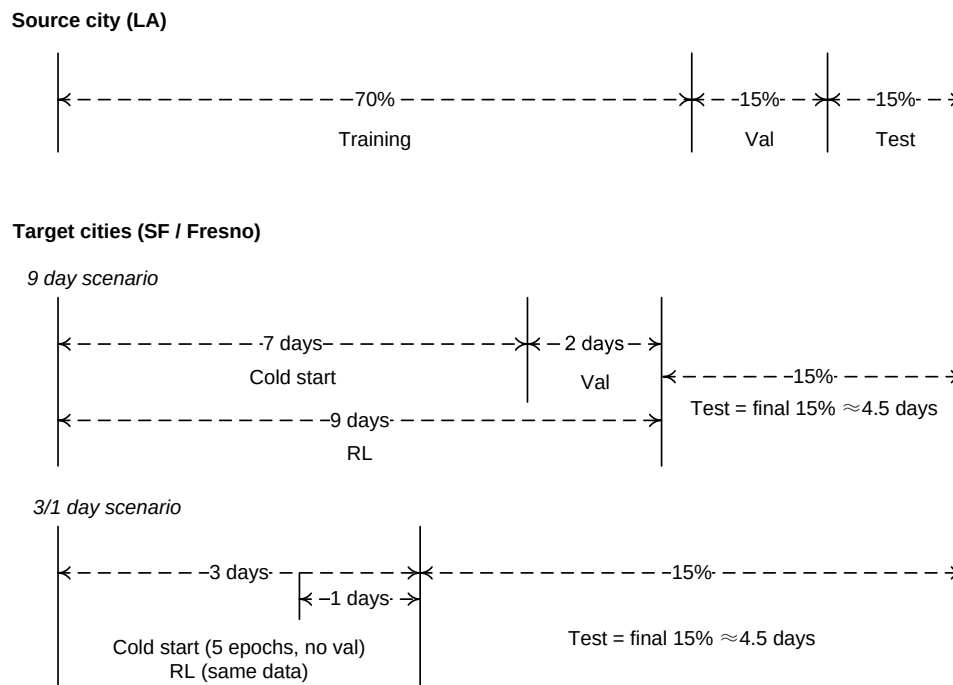
Supplementary Note 1: Point of interest (POI) categories

Supplementary Table 1 | Top-visited point of interest (POI) categories in Los Angeles based on aggregated incoming mobility flows.

Category	Example POIs	Share (%)
Transportation	Bus stops, parking, parking entrances	43.98
Education	Schools	5.85
Community	Places of worship	5.29
Food and dining	Restaurants, fast food, cafés	9.81
Recreation	Parks	3.29
Services	Fuel stations, convenience stores	4.00

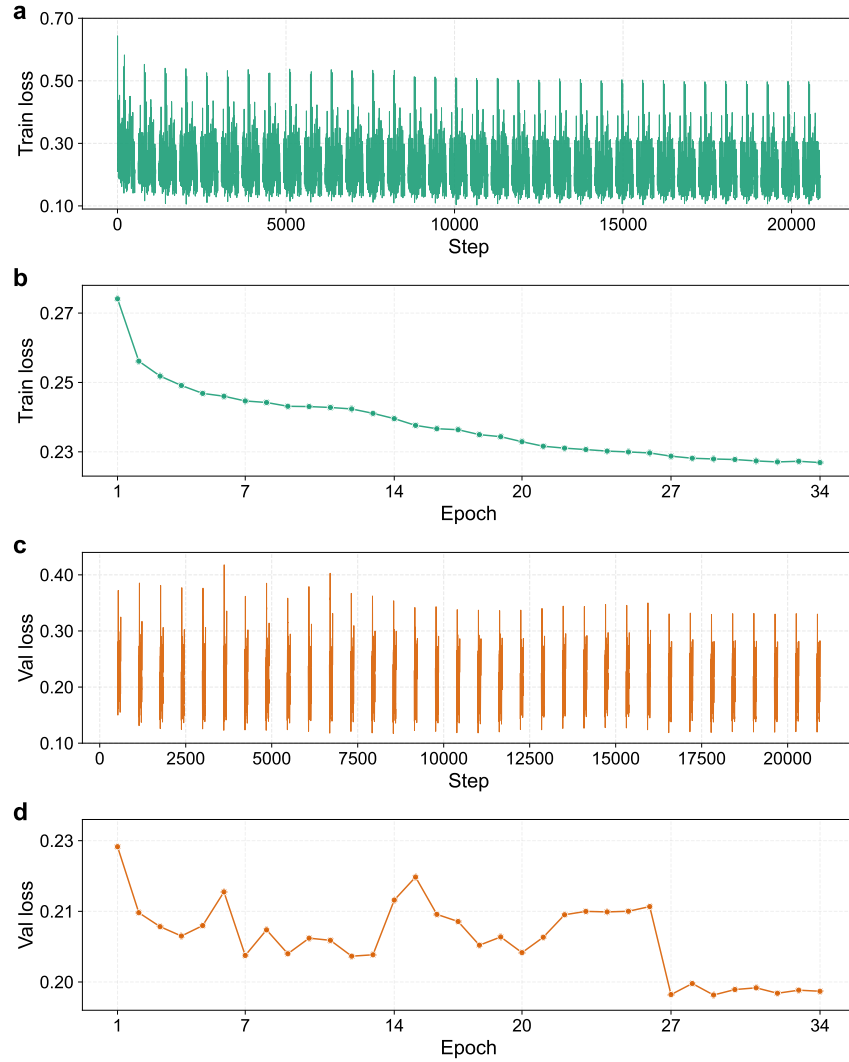
Supplementary Note 2: Cross-city evaluation setup

The cross-city evaluation setup is presented in Supplementary Fig. 1.

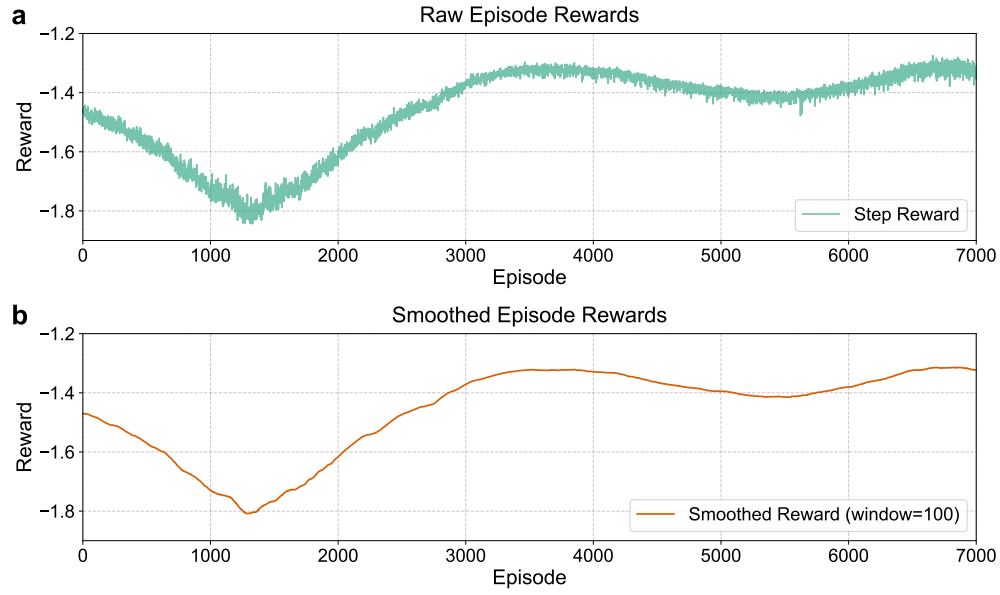


Supplementary Figure 1: **Overview of the cross-city evaluation setup under different data regimes.** Val denotes the validation dataset.

Supplementary Note 3: Training dynamics



Supplementary Figure 2: **Training and validation NLL dynamics on the Los Angeles dataset.** **a,b**, Step- and epoch-wise training wZTP NLL. **c,d**, Step- and epoch-wise validation wZTP NLL. Training wZTP NLLs (**a,b**) decrease smoothly, whereas validation wZTP NLLs (**c,d**) fluctuate during early optimization but converge overall. The checkpoint used for evaluation is selected according to the minimum validation ZTP NLL.



Supplementary Figure 3: **Reward dynamics during RL refinement in San Francisco.** **a**, Step-wise rewards received by the RL agent during the refinement stage. **b**, Smoothed reward trajectory computed using a moving average with a window size of 100 steps, revealing the underlying reward trend. Results based on the 9-day cross-city evaluation scenario (Supplementary Fig. 1).

Supplementary Note 4: Ablation and component analysis

We conduct an ablation study on the Los Angeles dataset to assess the contribution of each component, see Supplementary Table 2. Removing the POI embedding leads to the most pronounced degradation, particularly in the Top-1% regime, confirming that functional node representations are central to modeling high-intensity interactions. Eliminating meteorological features results in moderate increases in both tail and overall errors, indicating that environmental context provides complementary information for edge-level flow estimation.

Removing temporal encoding degrades performance across both tail and overall metrics, demonstrating that capturing periodic temporal patterns supports predictive stability. In contrast, excluding temporal convolution yields a trade-off: improvements in certain Top-1% metrics come at the cost of increased overall error. This pattern suggests that temporal convolution primarily contributes to stabilizing global prediction, preventing over-optimization toward extreme flows at the expense of distributional balance.

Removing the weighting scheme reveals a clear trade-off: overall errors decrease due to the dominance of single-count edges, whereas Top-1% performance degrades substantially. This indicates that, without weighting, learning is biased toward low-flow edges and fails to capture rare but important high-flow interactions.

The ablation highlights that accurate tail modeling and global stability rely on different components, confirming that accurate tail modeling and global stability depend on complementary components that must be jointly optimized.

Supplementary Table 2 | Ablation study on Los Angeles

Metric	Full	No PE	No MF	No TE	No TC	No W
Top-1% ZTP NLL	<u>2.1049</u>	2.9672	2.2494	2.1820	2.0835	4.9970
Top-1% MSE	3.4393	7.5760	3.6550	4.3533	4.3680	6.7275
Top-1% MAE	1.0022	1.3743	<u>1.0012</u>	1.0322	0.9978	1.6189
Overall ZTP NLL	0.2707	0.2875	0.2747	0.2911	0.3115	0.1439
Overall MSE	<u>0.2082</u>	0.1873	0.2402	0.2233	0.2421	0.0885
Overall MAE	0.2531	0.2614	0.2554	0.2740	0.2982	0.0625

PE = POI embedding; MF = meteorological features; TE = temporal encoding; TC = temporal convolution; No W = training without weighting. It is shown for reference only and is not included in best-value highlighting. Metrics are scaled by 10^2 . Bold values indicate the best performance; underline indicates the second best.

Supplementary Note 5: Full numerical results for the cross-city performance

Supplementary Tables 3 and 4 provide the complete quantitative results for cross-city transfer performance in San Francisco and Fresno respectively, complementing the performance curves shown in Fig. 3. Results are reported for cold-start alignment (CS), reinforcement learning refinement (CS+RL) and supervised training from scratch (SL) under 1, 3 and 9 days of target-city data, alongside zero-shot transfer from Los Angeles and full supervision baselines.

Supplementary Table 3 | Cross-city performance comparison on the San Francisco test set.

Metric	LA (zero)	1 day			3 days			9 days			SL (full)
		CS	CS+RL	SL	CS	CS+RL	SL	CS	CS+RL	SL	
Top 1%											
ZTP NLL	3.849	2.502	2.503	3.755	2.439	2.423	2.706	2.388	2.336	2.604	2.349
MSE	12.696	8.530	8.544	13.126	8.115	8.094	10.370	6.893	6.772	7.960	7.141
MAE	2.008	1.519	1.519	2.074	1.464	1.458	1.640	1.450	1.441	1.510	1.415
Overall											
ZTP NLL	0.248	0.377	0.376	0.630	0.346	0.353	0.729	0.225	0.240	0.285	0.290
MSE	0.274	0.584	0.579	0.536	0.451	0.460	0.840	0.339	0.375	0.394	0.385
MAE	0.227	0.391	0.390	0.647	0.353	0.361	0.776	0.218	0.237	0.283	0.291

Note: LA = Los Angeles, SF = San Francisco. Bold values indicate the best performance within each setting.

Supplementary Table 4 | Cross-city performance comparison on the Fresno test set.

Metric	LA (zero)	1 day			3 days			9 days			FRE (full)
		CS	CS+RL	SL	CS	CS+RL	SL	CS	CS+RL	SL	
Top 1%											
ZTP NLL	2.333	1.916	1.899	1.762	1.906	1.871	1.565	1.967	1.947	2.051	2.053
MSE	2.420	1.856	1.893	1.971	1.740	1.829	1.564	1.699	1.714	1.902	1.966
MAE	1.071	0.861	0.863	0.803	0.858	0.857	0.802	0.869	0.866	0.942	0.963
Overall											
ZTP NLL	0.308	0.459	0.456	0.772	0.462	0.451	1.318	0.399	0.399	0.464	0.422
MSE	0.194	0.445	0.407	0.693	0.497	0.397	2.502	0.384	0.368	0.452	0.361
MAE	0.285	0.462	0.457	0.809	0.468	0.453	1.480	0.394	0.394	0.467	0.418

Note: FRE = Fresno. Bold values indicate the best performance within each setting.