SCOT: Soft Correspondence via Optimal Transport for Multi-Source Cross-City Transfer

Yuyao Wang[1] **Min Yang**[2] **Meng Chen**[2] **Weiming Huang**[3] **Yilong Yin**[2] **Yongshun Gong**[*2]

[1]Department of Mathematics and Statistics, Boston University, Boston, MA, USA
[2]School of Software, Shandong University, Jinan, China
[3]School of Geography, University of Leeds, Leeds, UK

Abstract

Cross-city transfer uses labeled data from well-instrumented cities to improve prediction in label-scarce ones. The task is difficult when cities have incompatible region partitions and no ground-truth correspondence. Existing methods either rely on heuristic anchors, which can create hubness, or align global distributions, which can over-mix heterogeneous regions. We propose **SCOT** (**S**oft **C**orrespondence via **O**ptimal **T**ransport), an OT-based framework for learning explicit, mass-controlled soft correspondences between unequal region sets. SCOT combines an OT-weighted contrastive loss, a one-sided cycle regularizer, and a shared prototype hub for multi-source transfer. Across real-world cities and tasks, SCOT consistently outperforms strong baselines, reducing MAE and MAPE by 5–50%. Additional analyses show that the learned correspondences are selective, region-specific, and robust to heterogeneous settings. Code is available at `https://github.com/olivia3395/SCOT-Multi-Source-Cross-City-Transfer`.

1 Introduction

Many urban computing tasks build city-scale predictors from heterogeneous data (human mobility, POIs, remote sensing) and rely on high-quality region representations for downstream outcomes such as regional GDP, population, and carbon estimation [An et al., 2025, Yang et al., 2025a, Li et al., 2024a]. Since reliable labels exist only for a few well-instrumented cities, cross-city transfer aims to learn region embeddings that generalize from labeled source cities to a label-scarce target [Li et al., 2025, Jin et al., 2022, Yang et al., 2023, Lu et al., 2022, Wang et al., 2019, Jin et al., 2023, Fang et al., 2022, Yao et al., 2019, Li et al., 2022]. This is harder than standard domain adaptation: city pairs rarely share a natural region correspondence and often have unequal region counts (Fig. 1a), regions are graph nodes rather than i.i.d. samples, and only part of the semantics transfers (commuting corridors generalize while tourist districts are often city-specific), so alignment must be local and selective [Saito et al., 2018, Chen et al., 2024, 2025, Zhang et al., 2023, Chen et al.].

Where does the difficulty actually live? Modern GNN encoders already produce expressive region embeddings, and within-city predictive quality is largely saturated by standard GAT-family architectures. What varies dramatically across methods is target-city performance under transfer—an asymmetry that points to the cross-city correspondence mechanism, not the encoder, as the locus of difficulty. We empirically confirm this in Section 6.2: backbone and readout substitutions shift performance only marginally, while different alignment designs cause far larger variation. These results suggest that, in this regime, methodological gains can be more efficiently obtained by investing effort in the alignment mechanism than in the encoder.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

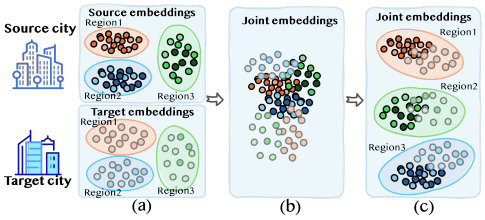


Figure 1: **Illustration of motivation.**

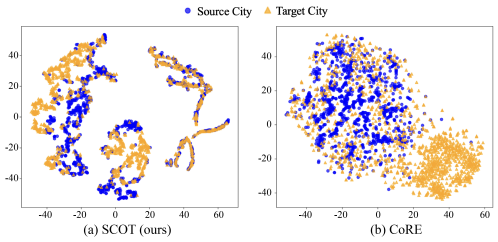


Figure 2: **t-SNE visualization for XA→BJ.**

Positioning relative to existing alignment paradigms. Prior work typically aligns cities along one of two axes [Wei et al., 2021, Liu et al., 2024, Yang et al., 2025c, Zhang et al., 2025b, Yang et al., 2025b, Yuan et al., 2025b, Zhang et al., 2025a], each with distinct trade-offs. *Distribution-matching* methods (e.g., MMD [Gretton et al., 2012], adversarial alignment [Ganin et al., 2016]) match embedding distributions without establishing explicit region-level correspondences, and even when augmented with region-level refinements, the distribution-level signal can encourage embeddings of heterogeneous cities to mix more than the data supports. *Anchor and nearest-neighbor matching* sits at the opposite end, constructing correspondences directly via hard one-to-one assignments, which can be sensitive to unequal region counts and to hubness [Lei et al., 2022, Tang et al., 2022, Zhao et al., 2023, Bao et al., 2022, Wang et al., 2021, Chen et al.]. Recent correspondence-learning approaches narrow this gap, though aspects of both behaviors can persist in cross-city transfer without targeted structural adaptation (Fig. 2 illustrates this for CoRE [Chen et al.]; quantitative diagnostics in Section 6.4). These two paradigms share a common underexplored ingredient: *explicit, mass-controlled soft correspondences between unequal region sets*—a primitive well-studied in optimal transport but not yet carefully adapted to the structural demands of cross-city transfer. What is needed is an alignment mechanism that (i) learns region-level correspondences without ground-truth matching, and (ii) scales to multiple sources without source domination or gradient conflict.

To address this unmet requirement, we propose **SCOT** (**S**emantic **C**orrespondence via **O**ptimal **T**ransport), which learns explicit soft correspondences for cross-city alignment (Fig. 3). **SCOT** rests on two design choices, each a principled response to a specific failure mode diagnosed above.

(1) **Optimal transport for capacity-controlled soft matching.** OT compares distributions via a minimum-cost coupling between point sets [Villani, 2021, Villani et al., 2008], and its entropic form [Cuturi, 2013] imposes marginal capacity constraints that address both failure modes at once: bounding per-region mass prevents the many-to-one collapse of anchor matching, while the resulting many-to-many soft coupling avoids the indiscriminate mixing of distribution-level objectives. Sinkhorn iterations [Peyré et al., 2019] scale this to urban settings, leveraging OT's established role in domain adaptation and representation learning [Courty et al., 2016, Chen et al., 2020, Li et al., 2024b].

(2) **OT-guided contrastive sharpening for semantic discriminability.** Geometric closeness alone does not yield semantically discriminative embeddings. We therefore couple OT with a contrastive objective whose soft positives are defined by the transport coupling itself: target candidates are weighted by transported mass, focusing similarity on transport-supported pairs while inheriting OT's capacity control [Genevay et al., 2018]. The result is a locally aligned yet non-collapsed embedding geometry that transfers more reliably (Fig. 2, left). A one-sided cycle reconstruction further stabilizes training under strong heterogeneity by enforcing transport-conditioned identity.

Contributions.

- (i) **Single-source alignment via OT soft correspondence.** We propose the *first* cross-city alignment framework that jointly enforces capacity-controlled correspondence, semantic discriminability, and rectangular-geometry stability, instantiated via Sinkhorn-based entropic OT, an OT-weighted contrastive objective, and a one-sided cycle regularizer.
- (ii) **Multi-source extension through a shared prototype hub.** We extend **SCOT** to multi-source transfer via a shared prototype hub guided by a target-induced prior and balanced OT, coordinating heterogeneous sources without explicit source-selection heuristics and reducing the impact of conflicting per-source signals through hub-mediated aggregation.

- (iii) **Empirical validation.** On GDP, population, and CO_2 prediction across cities and directions, **SCOT** outperforms strong baselines in both single- and multi-source settings, with multi-source **SCOT** consistently surpassing the strongest single-source transfer. Controlled substitution of encoders and regressors further confirms that gains stem from alignment design rather than encoder capacity or readout flexibility.

2 Related Work

2.1 Cross-City Transfer

Cross-city transfer uses well-instrumented source cities to compensate for label scarcity in a target city. Early work includes multimodal transfer for urban inference (FLORAL) [Wei et al., 2016], match-then-transfer via region correspondences (RegionTrans) [Wang et al., 2019], and meta-learning across cities (MetaST) [Yao et al., 2019]. Later methods address heterogeneity and negative transfer, e.g., by reweighting source regions (CrossTReS) [Jin et al., 2022]. Graph-based variants include spatio-temporal few-shot learning (ST-GFSL) [Lu et al., 2022], transferable graph structure learning (TransGTR) [Jin et al., 2023], parameter generation (CARPG) [Yang et al., 2023], and joint embedding and alignment (CoRE) [Chen et al.]. Many of these methods assume comparable spatial units such as grids. With irregular partitions, however, cities have unequal region counts and no natural one-to-one correspondence, which calls for local, selective alignment between region sets—the setting **SCOT** targets.

2.2 Spatio-Temporal Representation Learning

Spatio-temporal models learn urban embeddings for traffic forecasting, crowd flow and OD estimation [Mu et al., 2025, Yuan et al., 2025a], weather prediction [Gong et al., 2024], and regional attribute regression. On regular grids, ST-ResNet [Zhang et al., 2017], DeepSTN+ [Lin et al., 2019], and UrbanFM [Liang et al., 2019] model citywide flows. On graphs, STGNNs such as DCRNN [Li et al., 2017], STGCN [Yu et al., 2017], Graph WaveNet [Wu et al., 2019], and GMAN [Zheng et al., 2020] combine spatial message passing with temporal modules, and recent work adds self-supervised pretraining [Liu et al., 2022, Ji et al., 2023]. These models are designed for a single city or a fixed node set. Across cities with different partitions, a stronger encoder alone does not ensure transferability, so an explicit alignment mechanism is still needed.

2.3 Optimal Transport in Deep Learning

Optimal transport (OT) aligns distributions through a minimum-cost coupling; entropic regularization and the Sinkhorn algorithm make it scalable and differentiable [Cuturi, 2013, Peyré et al., 2019]. OT serves as a geometry-aware loss [Frogner et al., 2015, Genevay et al., 2018, Feydy et al., 2019] and as an alignment tool in domain adaptation, from OT-DA [Courty et al., 2016] to deep variants such as DeepJDOT [Damodaran et al., 2018].

OT variants. Unbalanced OT replaces hard marginals with soft penalties [Chizat et al., 2018, Fatras et al., 2021], partial OT transports only part of the mass [Caffarelli and McCann, 2010, Chapel et al., 2020], and quadratic regularization yields sparse couplings [Blondel et al., 2018]. **SCOT** uses balanced entropic OT because its strict marginals are what prevent many-to-one concentration (Appendix F.4).

Graph-structural OT. Gromov–Wasserstein [Mémoli, 2011, Peyré et al., 2016] and fused GW [Titouan et al., 2019] align graphs by preserving intra-domain distances. We instead apply OT to GAT embeddings, which already encode graph structure; this avoids GW's non-convex optimization and higher cost (Appendix F.6).

3 Problem Setup

We study cross-city transfer between a labeled source $\mathcal{C}_s$ and a label-scarce target $\mathcal{C}_t$, with region sets $V_s = \{1, \ldots, n_s\}$ and $V_t = \{1, \ldots, n_t\}$. Each city c is equipped with an undirected spatial

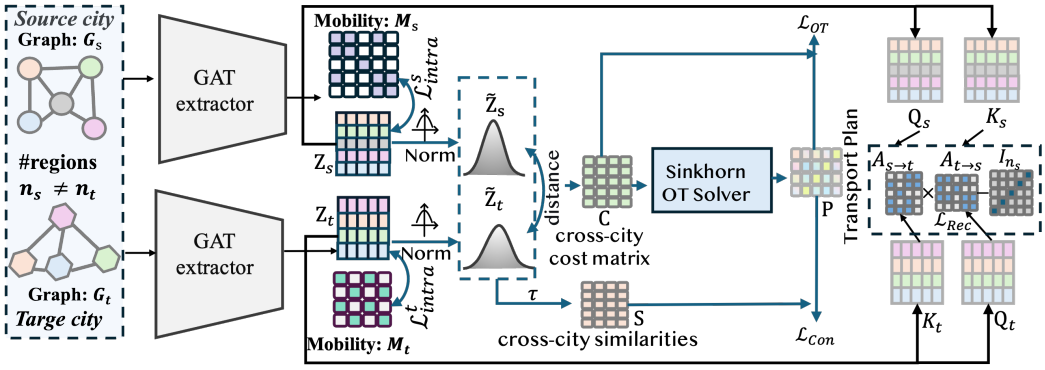


Figure 3: The pipeline of SCOT.

adjacency $\mathbf{A}_c$ and a row-stochastic mobility matrix $\mathbf{M}_c$ derived from OD trips:

$$M_{ij} = \frac{\text{count}(i \to j)}{\sum_k \text{count}(i \to k)}, \tag{1}$$

so $M_{i\cdot}$ encodes the destination distribution from region i. Our goal is to learn embeddings $\mathbf{z}_s \in \mathbb{R}^{n_s \times d}$, $\mathbf{z}_t \in \mathbb{R}^{n_t \times d}$ that preserve intra-city mobility and remain comparable across cities despite $n_s \neq n_t$, without any node correspondence (Section 4).

4 Method

SCOT jointly learns mobility-preserving region embeddings and a soft correspondence between two unequal region sets, without node-level correspondence (Fig. 3, Alg. 1).

Encoder and intra-city objective. Because performance depends far more on the alignment design than on the encoder (Section 6.2), we use a standard GAT encoder [Veličković et al., 2017]. For each city $c \in \{s, t\}$, learnable initial embeddings $\mathbf{H}_c^{(0)} \in \mathbb{R}^{n_c \times d}$ pass through L GAT layers over the spatial adjacency $\mathbf{A}_c$, giving $\mathbf{z}_c = \mathbf{H}_c^{(L)}$. We model the destination distribution of each origin i as

$$\hat{P}_{ij}^{(c)} = \frac{\exp(\mathbf{z}_{c,i}^\top \mathbf{z}_{c,j})}{\sum_{k=1}^{n_c} \exp(\mathbf{z}_{c,i}^\top \mathbf{z}_{c,k})}, \tag{2}$$

and fit it to the empirical transition matrix $\mathbf{M}_c$:

$$\mathcal{L}_{\text{intra}}^{(c)} = -\sum_{i,j} (\mathbf{M}_c)_{ij} \log \hat{P}_{ij}^{(c)}. \tag{3}$$

4.1 Alignment via Soft Transport-Guided Matching

Cross-city alignment has no ground-truth correspondence, unequal region counts, and only partially shared semantics. One-to-one anchor matching is fragile under unequal partitions, and distribution-level alignment can mix regions that should stay distinct. We therefore model the correspondence as an OT coupling $\mathbf{P} \in \mathbb{R}_+^{n_s \times n_t}$, whose marginal constraints limit how much mass any region can send or receive. Unlike prior uses of OT as a discrepancy, matching step, or regularizer [Courty et al., 2016, Damodaran et al., 2018, Genevay et al., 2018], **SCOT** feeds $\mathbf{P}$ directly into a contrastive loss, so the coupling shapes representation learning.

4.1.1 Sinkhorn-Based Soft Correspondence

We ℓ_2-normalize the embeddings, $\tilde{\mathbf{z}}_i^s = \mathbf{z}_i^s / \|\mathbf{z}_i^s\|_2$ and $\tilde{\mathbf{z}}_j^t = \mathbf{z}_j^t / \|\mathbf{z}_j^t\|_2$, so that the cost reflects direction rather than city-specific scale, and set

$$C_{ij} = \left\|\tilde{\mathbf{z}}_i^s - \tilde{\mathbf{z}}_j^t\right\|_2, \qquad \mathbf{C} \in \mathbb{R}^{n_s \times n_t}. \tag{4}$$

With Gibbs kernel $\mathbf{K} = \exp(-\mathbf{C}/\varepsilon)$, uniform marginals $\mathbf{a} = \frac{1}{n_s}\mathbf{1}$ and $\mathbf{b} = \frac{1}{n_t}\mathbf{1}$, and $\mathbf{u}^{(0)} = \mathbf{v}^{(0)} = \mathbf{1}$, we run T Sinkhorn iterations [Cuturi, 2013]:

$$\mathbf{u}^{(k+1)} = \mathbf{a} \oslash \left(\mathbf{K}\mathbf{v}^{(k)}\right), \qquad \mathbf{v}^{(k+1)} = \mathbf{b} \oslash \left(\mathbf{K}^\top \mathbf{u}^{(k+1)}\right), \tag{5}$$

where $\oslash$ is elementwise division, and obtain the differentiable coupling

$$\mathbf{P} = \operatorname{diag}\big(\mathbf{u}^{(T)}\big)\,\mathbf{K}\,\operatorname{diag}\big(\mathbf{v}^{(T)}\big). \tag{6}$$

Smaller ε gives sharper but less stable couplings; we use $\varepsilon = 0.15$ (Section 6.5). The OT loss is the expected transport cost:

$$\mathcal{L}_{\mathrm{OT}} = \frac{1}{\min(n_s, n_t)} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} P_{ij} C_{ij}. \tag{7}$$

4.1.2 OT-Weighted Contrastive Alignment

$\mathcal{L}_{\mathrm{OT}}$ makes coupled pairs close in absolute terms, but a target region may be equally close to several sources. We therefore add a contrastive loss that makes pairs supported by $\mathbf{P}$ more similar than competing pairs. With similarities $S_{ij} = \tilde{\mathbf{z}}_i^{s\top} \tilde{\mathbf{z}}_j^t / \tau$ and row i of $\mathbf{P}$ as soft positive weights,

$$\mathcal{L}_{\mathrm{Con}} = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log \frac{\sum_{j=1}^{n_t} P_{ij} \exp(S_{ij})}{\sum_{j=1}^{n_t} \exp(S_{ij})}. \tag{8}$$

This pulls each source region toward its high-mass targets and away from the rest. The following transfer bound links $\mathcal{L}_{\mathrm{Con}}$ to target error.

Proposition 4.1 (Contrastive alignment as a transfer surrogate)**.** *Let $\{u_i\}_{i=1}^{n_s}, \{v_j\}_{j=1}^{n_t} \subset \mathbb{S}^{d-1}$ be the normalized source and target embeddings, with marginals $a \in \Delta^{n_s}$, $b \in \Delta^{n_t}$, and coupling $P \in \mathbb{R}_+^{n_s \times n_t}$ satisfying $P\mathbf{1} = a$ and $P^\top \mathbf{1} = b$. Let g be the unknown region-level task function and h a source-trained predictor applied to the target. Assume that g and h are L_g- and L_h-Lipschitz, and define*

$$\mathcal{R}_s^a(h) := \sum_i a_i |h(u_i) - g(u_i)|, \qquad \mathcal{R}_t^b(h) := \sum_j b_j |h(v_j) - g(v_j)|.$$

Then

$$\mathcal{R}_t^b(h) \le \underbrace{\mathcal{R}_s^a(h)}_{\textit{source risk}} + \underbrace{(L_h + L_g)\sqrt{2 - 2\underline{m}}}_{\textit{transfer gap}},$$

where

$$\underline{m} := \max\left\{-1,\, \tau \log n_t + \tau H(a) - \tau \mathcal{L}_{\mathrm{Con}}(P) - 1 - \frac{1}{2\tau}\right\}, \qquad H(a) := -\sum_i a_i \log a_i.$$

With uniform marginals, the two risks are the source and target MAE. A smaller $\mathcal{L}_{\mathrm{Con}}$ raises $\underline{m}$ and tightens the transfer gap. The bound motivates the contrastive term; it does not characterize the full **SCOT** objective. The proof is in Appendix B, and an empirical check against target MAE is in Appendix E.4.

4.1.3 Alignment Loss

We combine the two terms as

$$\mathcal{L}_{\mathrm{Align}} = \mathcal{L}_{\mathrm{OT}} + \eta\, \mathcal{L}_{\mathrm{Con}}. \tag{9}$$

$\mathcal{L}_{\mathrm{OT}}$ controls global mass allocation, and $\mathcal{L}_{\mathrm{Con}}$ sharpens local discrimination among candidates.

4.2 Cycle Reconstruction Regularization

We regularize the correspondence with a one-sided cross-attention cycle. Its goal is not an exact inverse between unequal region sets, but to limit distortion when representations are mapped to the other city and back. With shared projections $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d \times d}$, the cross-attention maps are

$$\mathbf{A}_{s \to t} = \operatorname{softmax}\left(\frac{(\mathbf{Z}_s \mathbf{W}_q^\top)(\mathbf{Z}_t \mathbf{W}_k^\top)^\top}{\sqrt{d}}\right), \qquad \mathbf{A}_{t \to s} = \operatorname{softmax}\left(\frac{(\mathbf{Z}_t \mathbf{W}_q^\top)(\mathbf{Z}_s \mathbf{W}_k^\top)^\top}{\sqrt{d}}\right), \tag{10}$$

and we penalize the source-returning cycle:

$$\mathcal{L}_{\text{cyc}} = \|\mathbf{A}_{s\to t}\mathbf{A}_{t\to s} - \mathbf{I}_{n_s}\|_F^2 . \tag{11}$$

We cycle back to the source because labels and the downstream predictor live there; the target-returning cycle does not improve target MAE (Appendix F.5). An entropy penalty discourages diffuse forward attention:

$$\mathcal{R}_{\text{ent}} = -\frac{1}{n_s}\sum_{i,j} A_{s\to t}(i,j)\log(A_{s\to t}(i,j)+\delta), \qquad \mathcal{L}_{\text{Rec}} = \mathcal{L}_{\text{cyc}} + \beta\,\mathcal{R}_{\text{ent}}, \tag{12}$$

with $\delta = 10^{-8}$.

4.3 Model Training

SCOT is trained end-to-end with

$$\mathcal{L}_{\text{Total}} = \underbrace{\mathcal{L}_{\text{intra}}^{(s)} + \mathcal{L}_{\text{intra}}^{(t)}}_{\text{intra-city reconstruction}} + \lambda_{\text{align}} \underbrace{\mathcal{L}_{\text{Align}}}_{\text{cross-city alignment}} + \lambda_{\text{rec}} \underbrace{\mathcal{L}_{\text{Rec}}}_{\text{cycle regularization}} . \tag{13}$$

The intra-city terms preserve each city's mobility structure, $\mathcal{L}_{\text{Align}}$ learns the cross-city correspondence, and $\mathcal{L}_{\text{Rec}}$ regularizes it. All weights are fixed across city pairs and tasks (Section 6).

5 Multi-Source Hub Alignment

Multi-source transfer is fundamentally harder than the single-source case: different sources induce *conflicting correspondences* to the same target, and independent pairwise alignments are easily destabilized or dominated by one source. We resolve both pathologies through a **shared semantic hub**— a set of K learnable prototypes that serve as intermediate anchors in a common alignment space (Fig. 9, Alg. 2).

Instead of aligning each source to the target separately, all cities (sources and target) are aligned to the hub via balanced entropic OT, yielding a *coordinated many-to-hub matching*. A target-induced prototype marginal controls hub capacity and steers prototypes toward target-relevant semantics, simultaneously stabilizing optimization and preventing source domination. The resulting transport plans $\{\Pi^{(m)}\}$ jointly supervise $\mathcal{L}_{\text{OT}}^m$ and $\mathcal{L}_{\text{Con}}^m$, enabling scalable alignment that avoids the source-domination and conflict pathologies observed under independent pairwise OT.

Balanced entropic OT to the hub. Let $\mathcal{S} = \{s_1, \ldots, s_M\}$ denote the source cities and t the target. For each city $m \in \mathcal{S} \cup \{t\}$, we ℓ_2-normalize region embeddings and prototypes ($\tilde{\mathbf{z}}_i^m = \mathbf{z}_i^m/\|\mathbf{z}_i^m\|_2$, $\tilde{\mathbf{a}}_k = \mathbf{a}_k/\|\mathbf{a}_k\|_2$) and form the hub cost $C_{ik}^m = \|\tilde{\mathbf{z}}_i^m - \tilde{\mathbf{a}}_k\|_2$, $\mathbf{C}^m \in \mathbb{R}^{n_m\times K}$.

Target-induced prototype marginal. The shared prototype marginal $\mathbf{b} \in \Delta^{K-1}$ is constructed from target–prototype affinities:

$$\bar{s}_k = \tfrac{1}{n_t}\sum_{j=1}^{n_t} \tilde{\mathbf{z}}_j^{t\top}\tilde{\mathbf{a}}_k, \qquad b_k \propto \max\{\exp(\bar{s}_k/\tau_b),\, \epsilon_b\}, \tag{14}$$

normalized to $\sum_k b_k = 1$, where $\tau_b > 0$ controls sharpness and $\epsilon_b > 0$ prevents dead prototypes (Appendix F.3). City-side marginals are uniform: $\mathbf{a}^m = \frac{1}{n_m}\mathbf{1}$.

Balanced entropic coupling. For each $m \in \mathcal{S} \cup \{t\}$, we solve

$$\mathbf{\Pi}^m \in \arg\min_{\mathbf{P}\geq 0} \langle \mathbf{P}, \mathbf{C}^m\rangle + \varepsilon \textstyle\sum_{i,k} P_{ik}(\log P_{ik} - 1) \quad \text{s.t.} \quad \mathbf{P}\mathbf{1} = \mathbf{a}^m,\ \mathbf{P}^\top\mathbf{1} = \mathbf{b}, \tag{15}$$

via T Sinkhorn iterations, and obtain row-normalized assignments $Q_{ik}^m = \Pi_{ik}^m / \sum_{k'} \Pi_{ik'}^m$.

OT-guided contrastive alignment to the hub. For each $m \in \mathcal{S}\cup\{t\}$, we compute region–prototype similarities $S_{ik}^m = \tilde{\mathbf{z}}_i^{m\top}\tilde{\mathbf{a}}_k/\tau$ and use the OT-induced assignments Q_{ik}^m as soft positive weights:

$$\mathcal{L}_{\text{Con}}^m = -\frac{1}{n_m}\sum_{i=1}^{n_m} \log \frac{\sum_k Q_{ik}^m \exp(S_{ik}^m)}{\sum_k \exp(S_{ik}^m)}. \tag{16}$$

With the transport cost $\mathcal{L}_{\text{OT}}^m = \langle \mathbf{\Pi}^m, \mathbf{C}^m\rangle$, the per-city alignment loss is $\mathcal{L}_{\text{Align}}^m = \mathcal{L}_{\text{OT}}^m + \lambda_c \mathcal{L}_{\text{Con}}^m$.

Table 1: Single-source transfer results on XA↔BJ. Lower is better. Red: best; Blue: runner-up. Gain row reports SCOT's relative improvement over the strongest baseline.

Method	XA → BJ						BJ → XA					
	GDP		Population		CO_2		GDP		Population		CO_2	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Non-Alignment	264.30	12.08	981.07	8.56	288.41	6.40	252.91	5.84	946.93	8.53	270.42	8.66
RP	189.84	9.07	684.50	6.55	196.05	4.70	181.08	4.58	670.44	6.46	191.85	6.42
HBP	177.69	8.40	665.46	5.95	188.72	4.18	196.20	3.10	627.46	4.37	176.35	4.25
HSA	201.27	6.83	619.33	4.00	176.40	3.20	188.40	2.11	636.33	5.03	182.91	5.02
MMD	183.32	5.71	588.34	3.17	165.70	2.54	180.73	1.99	499.94	1.85	141.57	1.83
Adv	192.59	8.72	702.19	6.78	199.23	4.83	199.21	6.32	805.01	9.16	203.27	7.21
CrossTReS	207.43	7.39	633.25	4.42	179.75	3.50	170.27	4.23	639.44	5.62	182.72	5.55
CoRE	157.83	5.46	611.18	4.05	166.28	2.95	162.19	1.91	547.74	2.17	153.63	2.09
SCOT (Ours)	115.33	3.17	528.50	2.13	149.42	1.79	154.92	1.60	452.67	1.58	128.74	1.63
Δ vs. best baseline	**+26.9%**	**+41.9%**	+10.2%	**+32.8%**	+9.8%	**+29.5%**	+4.5%	+15.7%	+9.5%	+14.6%	+9.1%	+10.9%

Hub-cycle stabilization. We extend the one-sided cycle of Section 4.1 to the city-to-hub setting. With shared $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d\times d}$, we form cross-attention maps between region embeddings $\mathbf{Z}^m \in \mathbb{R}^{n_m\times d}$ and prototypes $\mathbf{A} \in \mathbb{R}^{K\times d}$:

$$\mathbf{A}_{m\to h} = \mathrm{softmax}\Big(\frac{(\mathbf{Z}^m\mathbf{W}_q)(\mathbf{A}\mathbf{W}_k)^\top}{\sqrt{d}}\Big), \quad \mathbf{A}_{h\to m} = \mathrm{softmax}\Big(\frac{(\mathbf{A}\mathbf{W}_q)(\mathbf{Z}^m\mathbf{W}_k)^\top}{\sqrt{d}}\Big). \tag{17}$$

The hub-cycle loss and entropy penalty (with floor $\delta = 10^{-8}$) are

$$\mathcal{L}^m_{\mathrm{cyc}} = \tfrac{1}{n_m^2}\big\|\mathbf{A}_{m\to h}\mathbf{A}_{h\to m} - \mathbf{I}_{n_m}\big\|_F^2, \quad \mathcal{R}^m_{\mathrm{ent}} = -\tfrac{1}{n_m}\sum_{i,k}\mathbf{A}_{m\to h}(i,k)\log\big(\mathbf{A}_{m\to h}(i,k)+\delta\big), \tag{18}$$

giving $\mathcal{L}^m_{\mathrm{Rec}} = \mathcal{L}^m_{\mathrm{cyc}} + \beta\,\mathcal{R}^m_{\mathrm{ent}}$.

Objective. Multi-source SCOT is trained end-to-end by minimizing

$$\mathcal{L} = \sum_{m\in\mathcal{S}\cup\{t\}}\mathcal{L}^m_{\mathrm{intra}} + \frac{1}{|\mathcal{S}|+1}\sum_{m\in\mathcal{S}\cup\{t\}}\big(\lambda_{\mathrm{align}}\mathcal{L}^m_{\mathrm{Align}} + \lambda_{\mathrm{rec}}\mathcal{L}^m_{\mathrm{Rec}}\big). \tag{19}$$

6 Experiments

Setup, tasks, and metrics. We evaluate SCOT on mobility data from three Chinese cities (Beijing, Xi'an, Chengdu), aggregating anonymized OD trips into region-level mobility graphs and evaluating transfer on all ordered city pairs. Cross-country generalization with New York City is reported in Appendix L. For each pair $X \to Y$, a downstream regressor is fit on $(\mathbf{Z}^X, \mathbf{y}^X)$ and applied directly to $\mathbf{Z}^Y$ to predict $\mathbf{y}^Y$; we use ridge by default and verify in Section 6.2 that Lasso, SVR, and Elastic Net yield comparable results. For two-source experiments, each baseline adaptively weights the two transfer directions via softmax (with MMD and Adv additionally using a joint-mixture variant), and a single predictor is trained on the union of labeled source regions. We report MAE and MAPE on GDP, population, and CO_2 (lower is better).

Implementation. A two-layer GAT encoder ($d = 128$, $H = 8$) with PReLU and a linear output layer, trained end-to-end with Adam (lr $= 10^{-3}$). Hyperparameters are fixed to a single configuration across all city pairs and tasks: $\lambda_{\mathrm{align}} = 1.0$, $\lambda_{\mathrm{rec}} = 0.5$, $\eta = 0.5$, $\beta = 0.05$, $\tau = 0.1$, $\varepsilon = 0.15$. Multi-source additionally uses $K = 32$ prototypes, $\tau_b = 0.5$, and probability floor 10^{-3}.

Baselines. We compare SCOT against three paradigms: a *non-alignment* baseline (intra-city objectives only); *correspondence-based* alignment via surrogate matches (RP, HBP, HSA); and *correspondence-free* transfer via distributional or relational alignment (MMD [Gretton et al., 2012], Adv [Ganin et al., 2016], CrossTReS [Jin et al., 2022], CoRE [Chen et al.]). Details in Appendix A.1.

Table 2: Multi-source cross-city transfer results (two targets shown; the remaining target is reported in Appendix). Lower is better. Red: best; Blue: runner-up. Gain row reports SCOT's relative improvement over the strongest baseline.

Target: Beijing (BJ)

Method	GDP		Population		CO_2	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
RP	172.76	7.64	679.86	6.98	166.80	2.69
HBP	164.25	7.13	662.60	6.53	165.53	2.57
HSA	156.40	6.62	644.89	5.53	160.05	2.59
MMD	127.45	4.93	605.81	4.11	160.52	1.29
Adv	196.76	9.90	717.96	6.34	189.41	3.19
CrossTReS	151.17	6.41	666.74	5.03	187.59	2.32
CoRE	152.88	5.86	620.34	4.30	152.24	1.99
SCOT (Ours)	104.16	2.57	525.10	1.87	143.53	1.16
Δ vs. best	**+18.3%**	**+47.9%**	**+13.3%**	**+54.5%**	+5.7%	+10.1%

Target: Xi'an (XA)

Method	GDP		Population		CO_2	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
RP	195.91	2.89	642.01	3.67	181.16	2.88
HBP	200.04	4.24	670.41	3.80	150.16	2.09
HSA	183.55	2.42	648.67	3.79	155.20	2.45
MMD	163.78	2.18	506.22	3.05	144.61	3.18
Adv	221.31	4.84	731.92	5.43	184.73	3.46
CrossTReS	179.01	4.94	625.63	5.52	151.22	3.48
CoRE	173.72	5.48	549.37	3.89	134.19	1.97
SCOT (Ours)	156.94	1.71	446.13	1.86	127.66	1.26
Δ vs. best	+4.2%	**+21.6%**	**+11.9%**	**+39.0%**	+4.9%	**+36.0%**

6.1 Experimental Results

(i) **Single-source transfer.** SCOT achieves the best results across all target cities and tasks under both MAE and MAPE (Table 1, Appendix E.1). (ii) **Multi-source transfer.** In the two-source setting (Table 2), SCOT again attains the best performance, with gains from the shared hub stabilizing multi-source aggregation. (iii) **Multi-source vs. single-source.** Multi-source SCOT consistently outperforms its best single-source counterpart on most target–task pairs (Fig. 4), reflecting complementary signals aggregated through the hub rather than reliance on a single closest source. The marginal Xi'an GDP degradation is mechanistically diagnosed via hub statistics in Appendix K.

6.2 Robustness to Encoder and Readout Choice

A natural concern is whether SCOT's gains stem from its alignment design or from incidental choices in the encoder and downstream regressor. To isolate this, we hold the alignment module fixed and vary the GNN encoder (GAT, GATv2, SuperGAT) and the regressor (Ridge, Lasso, SVR, Elastic Net) on BJ→XA (Table 3). The asymmetry is striking: **encoder/readout substitution <3%** vs. **alignment-design gap 25–40%** —roughly an order of magnitude larger. This provides strong evidence that SCOT's gains stem primarily from its *alignment design* rather than from encoder capacity or readout flexibility, and justifies treating both as *interchangeable components*.

Table 3: Robustness to encoder and readout choice on BJ→XA (mean ± std). Red: best per column.

(a) Encoder ablation

Encoder	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
GAT [Veličković et al., 2017] *(default)*	160.21 ± 3.53	1.87 ± 0.18	450.14 ± 2.81	1.73 ± 0.12	127.79 ± 1.08	1.78 ± 0.10
GATv2 [Brody et al., 2021]	162.60 ± 2.27	1.74 ± 0.22	455.36 ± 4.56	1.74 ± 0.13	128.83 ± 1.29	1.82 ± 0.11
SuperGAT [Kim and Oh, 2022]	164.42 ± 5.29	1.49 ± 0.13	461.45 ± 8.37	1.95 ± 0.21	132.31 ± 3.05	1.86 ± 0.21

(b) Readout ablation

Readout	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
Ridge [Hoerl and Kennard, 1970] *(default)*	160.21 ± 3.53	1.87 ± 0.18	450.14 ± 2.81	1.73 ± 0.12	127.79 ± 1.08	1.78 ± 0.10
Lasso [Tibshirani, 1996]	158.66 ± 4.07	2.03 ± 0.12	455.68 ± 5.66	1.96 ± 0.02	131.40 ± 3.26	1.99 ± 0.05
Linear SVR [Drucker et al., 1996]	162.23 ± 2.00	1.61 ± 0.09	456.35 ± 3.62	1.94 ± 0.19	128.62 ± 0.78	1.86 ± 0.18
Elastic Net [Zou and Hastie, 2005]	164.60 ± 2.80	1.57 ± 0.19	459.46 ± 3.93	1.89 ± 0.34	129.90 ± 1.83	1.82 ± 0.31

6.3 Ablation Study

Single-source components. Removing $\mathcal{L}_{\text{Con}}$, $\mathcal{L}_{\text{OT}}$, or $\mathcal{L}_{\text{Rec}}$ consistently degrades transfer performance (Figs. 6 and 10). We retain the t-SNE plots only as qualitative views of the embedding geometry; direct diagnostics are reported in Appendix F.1. The three terms exhibit distinct failure

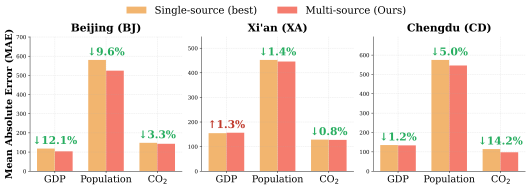


Figure 4: Best single-source SCOT (orange) vs. multi-source SCOT (red). Labels show relative MAE change Δ (blue: improvement; red: degradation).

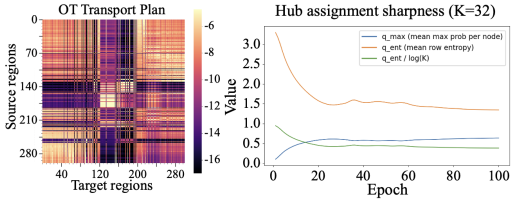


Figure 5: Diagnostics: (left) entropic OT coupling (XA→BJ, epoch 100); (right) hub assignment sharpness for $K = 32$.

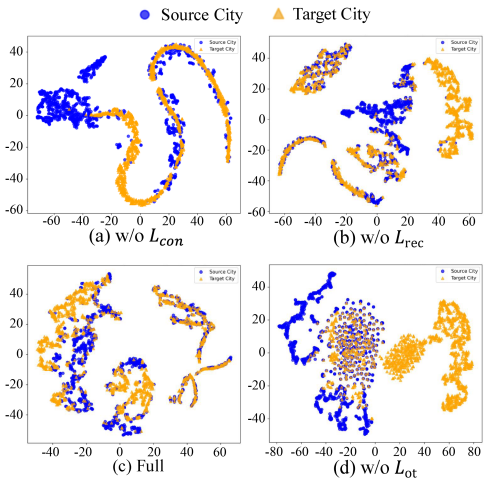


Figure 6: **Ablation t-SNE of SCOT (XA→BJ).** (a) w/o contrastive, (b) w/o reconstruction, (c) full, (d) w/o OT. The full model is the best.

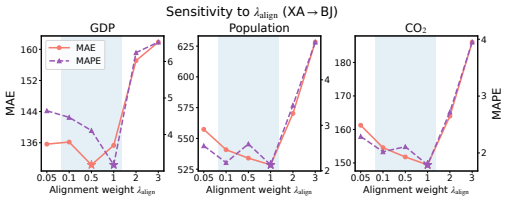


Figure 7: Sensitivity to λ_{align}.

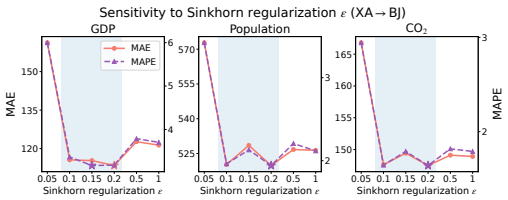


Figure 8: Sensitivity to OT regularization ε.

modes. Without $\mathcal{L}_{\text{Con}}$, the row-normalized transport becomes nearly uniform, showing insufficient local discrimination. Without $\mathcal{L}_{\text{OT}}$, the coupling develops a dominant entry together with a broad diffuse tail, indicating that contrastive learning alone does not recover structured global mass allocation. Removing $\mathcal{L}_{\text{Rec}}$ increases cycle error and cross-seed variation. Thus, $\mathcal{L}_{\text{OT}}$, $\mathcal{L}_{\text{Con}}$, and $\mathcal{L}_{\text{Rec}}$ respectively contribute global allocation, local relative discrimination, and correspondence stability; the full model achieves the lowest MAE/MAPE.

Multi-source and additional ablations. SCOT's multi-source design rests on three deliberate choices (Appendix F): *hub mediation* (vs. pairwise OT), *target-induced prior* (vs. uniform), and *balanced OT* (vs. unbalanced). **Each contributes meaningfully**; one-sided cycle wins under $n_s \neq n_t$, and alignment preserves within-city quality.

6.4 Diagnostics

(i) **OT coupling structure.** The learned coupling $\mathbf{P}$ for XA→BJ shows clear block structure after barycentric reordering (Fig. 5, left), confirming *selective many-to-many correspondences* without hubness or indiscriminate spreading.

(ii) **Hub assignment sharpness.** The normalized entropy $q_{\text{ent}}/\log K \approx 0.4$ ($K = 32$) corresponds to roughly four active prototypes per region—*stable specialization rather than uniform pooling*, indicating the target-induced prior steers prototypes toward task-relevant semantics (App. E.5).

(iii) **Semantic coherence of transport blocks.** Using two region attributes excluded from training, FUNCTION and FORM_TYPE, the top-k targets of the source region in Fig. 5 are more semantically coherent than size-matched random subsets, with purity gains of 10.2–26.9% and 10.4–22.1%, respectively. A row-shuffled control preserves each transport profile but breaks its source-region association, reducing same-FUNCTION mass and worsening FORM_TYPE similarity. This indicates source-dependent semantic concentration rather than global block reuse (App. H).

6.5 Hyperparameter Sensitivity

SCOT remains stable across broad parameter ranges, staying competitive with baselines even at off-recommended settings—valuable for label-scarce targets where per-target tuning is infeasible. We show three cases; remaining sweeps (η, τ, τ_b) are in Appendix G.

Sensitivity to λ_{align}. Sweeping $\lambda_{\text{align}} \in \{0.05, 0.1, 0.5, 1, 2, 3\}$ (Fig. 7), performance is best for $\lambda_{\text{align}} \in [0.1, 1]$ and degrades only at $\lambda_{\text{align}} \geq 2$ (over-alignment). Default: 1.

Sensitivity to ε. Sweeping $\varepsilon \in \{0.05, 0.1, 0.15, 0.2, 0.5, 1\}$ (Fig. 8), performance peaks at $\varepsilon \in [0.1, 0.2]$; smaller values give noisy couplings, larger values diffuse matching. Default: 0.15.

Hub size K. On CD,BJ→XA, performance is stable for $K \in [4, 32]$ (smaller K underfits, $K \geq 64$ over-resolves; App. G.3). Default: 32. For $M > 2$ sources, $K \approx \min(32 \log_2(M+1),\ n_t/2)$ serves as a starting heuristic; hub-conflict diagnostics (App. K) provide an unsupervised adjustment signal.

6.6 Generality beyond mobility graphs.

SCOT's cross-city alignment operates on normalized region embeddings and does not require shared edges, node identities, equal region counts, or comparable partitions. It can therefore be paired with different graph encoders and input signals. This is supported by our encoder and readout ablations (Section 6.3) and by the cross-country experiments with NYC (Appendix L). We further verify the alignment module in a controlled study with mobility, functional-similarity, and spatial-functional graphs, and under increasing source–target heterogeneity (Appendix I).

7 Conclusion

We proposed **SCOT**, a framework for cross-city region transfer that addresses two complementary failure modes of existing alignment paradigms—hubness under unequal partitions and over-mixing under heterogeneity—through Sinkhorn-based entropic OT paired with an OT-guided contrastive objective. A shared prototype hub with target-induced prior extends this to multi-source transfer, coordinating heterogeneous sources without gradient conflict or explicit source selection. Experiments on GDP, population, and CO_2 across multiple cities and directions show consistent gains over strong baselines. Future work includes uncertainty-aware mechanisms for selectively integrating sources under severe heterogeneity.

8 Acknowledgments

The work of the authors affiliated with Shandong University was supported in part by the National Natural Science Foundation of China (Grant No. 62476154), the Major Basic Research Project of the Shandong Provincial Natural Science Foundation (Grant No. ZR2024ZD03), and the Shandong University Qilu Young Scholars Program.

References

Yang An, Zhibin Li, Xiaoyu Li, Wei Liu, Xinghao Yang, Haoliang Sun, Meng Chen, Yu Zheng, and Yongshun Gong. Spatio-temporal multivariate probabilistic modeling for traffic prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Han Bao, Xun Zhou, Yiqun Xie, Yanhua Li, and Xiaowei Jia. Storm-gan: spatio-temporal meta-gan for cross-city estimation of human mobility responses to covid-19. In *Proceedings of the 2022 IEEE International Conference on Data Mining, ICDM 2022, Orlando, FL, USA, November 28 - December 1*, pages 1–10, 2022.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pages 880–889. PMLR, 2018.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pages 673–730, 2010.

Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020.

Meng Chen, Hongwei Jia, Zechen Li, Wenzhen Jia, Kai Zhao, Hongjun Dai, and Weiming Huang. Cross-city latent space alignment for consistency region embedding. In *Forty-second International Conference on Machine Learning*.

Meng Chen, Zechen Li, Weiming Huang, Yongshun Gong, and Yilong Yin. Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 319–328, 2024.

Meng Chen, Zechen Li, Hongwei Jia, Xin Shao, Jun Zhao, Qiang Gao, Min Yang, and Yilong Yin. Mgrl4re: A multi-graph representation learning approach for urban region embedding. *ACM Transactions on Intelligent Systems and Technology*, 16(2):1–23, 2025.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018.

Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9, 1996.

Ziquan Fang, Dongen Wu, Lu Pan, Lu Chen, and Yunjun Gao. When transfer learning meets cross-city urban flow prediction: Spatio-temporal adaptation matters. In *IJCAI*, volume 22, pages 2030–2036, 2022.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International conference on machine learning*, pages 3186–3197. PMLR, 2021.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 2681–2690. PMLR, 2019.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

Yongshun Gong, Tiantian He, Meng Chen, Bin Wang, Liqiang Nie, and Yilong Yin. Spatio-temporal enhanced contrastive and contextual learning for weather forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):4260–4274, 2024.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Junbo Zhang, and Yu Zheng. Spatio-temporal self-supervised learning for traffic flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 4356–4364, 2023.

Yilun Jin, Kai Chen, and Qiang Yang. 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*, pages 731–741, 2022.

Yilun Jin, Kai Chen, and Qiang Yang. Transferable graph structure learning for graph-based traffic forecasting across cities. In *Proceedings of the Twenty-Ninth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10*, pages 1032–1043, 2023.

Dongkwan Kim and Alice Oh. How to find your friendly neighborhood: Graph attention design with self-supervision. *arXiv preprint arXiv:2204.04879*, 2022.

Xiaoliang Lei, Hao Mei, Bin Shi, and Hua Wei. Modeling network-level traffic flow transitions on sparse data. In *Proceedings of the Twenty-Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2022, Washington DC Convention Center, USA, August 14-18*, pages 835–845, 2022.

Mingxi Li, Yihong Tang, and Wei Ma. Few-sample traffic prediction with graph networks using locale as relational inductive biases. *IEEE Transactions on Intelligent Transportation Systems*, 24 (2):1894–1908, 2022.

Xiaoyu Li, Yongshun Gong, Wei Liu, Yilong Yin, Yu Zheng, and Liqiang Nie. Dual-track spatio-temporal learning for urban flow prediction with adaptive normalization. *Artificial Intelligence*, 328:104065, 2024a.

Xiaoyu Li, Yitian Zhang, Guodong Long, Yupeng Hu, Wenpeng Lu, Meng Chen, Chengqi Zhang, and Yongshun Gong. Adaptive traffic forecasting on daily basis: A spatio-temporal context learning approach. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Xinhang Li, Xiangyu Zhao, Zihao Wang, Yang Duan, Yong Zhang, and Chunxiao Xing. Optimal transport enhanced cross-city site recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1441–1451, 2024b.

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

Yuxuan Liang, Kun Ouyang, Lin Jing, Sijie Ruan, Ye Liu, Junbo Zhang, David S Rosenblum, and Yu Zheng. Urbanfm: Inferring fine-grained urban flows. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3132–3142, 2019.

Ziqian Lin, Jie Feng, Ziyang Lu, Yong Li, and Depeng Jin. Deepstn+: Context-aware spatial-temporal neural network for crowd flow prediction in metropolis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1020–1027, 2019.

Xu Liu, Yuxuan Liang, Chao Huang, Yu Zheng, Bryan Hooi, and Roger Zimmermann. When do contrastive learning signals help spatio-temporal graph forecasting? In *Proceedings of the 30th international conference on advances in geographic information systems*, pages 1–12, 2022.

Zhanyu Liu, Jianrong Ding, and Guanjie Zheng. Frequency enhanced pre-training for cross-city few-shot traffic forecasting. *arXiv preprint arXiv:2406.02614*, 2024.

Bin Lu, Xiaoying Gan, Weinan Zhang, Huaxiu Yao, Luoyi Fu, and Xinbing Wang. Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proceedings of the Twenty-Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2022, Washington DC Convention Center, USA, August 14-18*, pages 1162–1172, 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21, 2008.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Wenqian Mu, Jiyuan Liu, Yongshun Gong, Ji Zhong, Wei Liu, Haoliang Sun, Xiushan Nie, Yilong Yin, and Yu Zheng. Gem: Gaussian embeddings with multi-hop graph transfer for next poi recommendation. *Neural Networks*, 186:107290, 2025.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Shiliang Sun, Honglei Shi, and Yuanbin Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24:84–92, 2015.

Yihong Tang, Ao Qu, Andy HF Chow, William HK Lam, Sze Chun Wong, and Wei Ma. Domain adversarial spatial-temporal network: A transferable framework for short-term traffic forecasting across cities. In *Proceedings of the Thirty-First ACM International Conference on Information and Knowledge Management, CIKM 2022, Atlanta, GA, USA, October 17-21*, pages 1905–1915, 2022.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR, 2019.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Leye Wang, Xu Geng, Xiaojuan Ma, Feng Liu, and Qiang Yang. Cross-city transfer learning for deep spatio-temporal prediction. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16*, pages 1893–1899, 2019.

Senzhang Wang, Hao Miao, Jiyue Li, and Jiannong Cao. Spatio-temporal knowledge transfer for urban crowd flow prediction via deep attentive adaptation networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(5):4695–4705, 2021.

Xiaohui Wei, Tao Guo, Hongmei Yu, Zijian Li, Hao Guo, and Xiang Li. Areatransfer: A cross-city crowd flow prediction framework based on transfer learning. In *Proceedings of the International Conference on Smart Computing and Communications, ICSCC 2021, New York, USA, December 29*, pages 238–253. Springer, 2021.

Ying Wei, Yu Zheng, and Qiang Yang. Transfer knowledge between cities. In *Proceedings of the Twenty-Second ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, San Francisco, California, USA, August 13-17*, pages 1905–1914, 2016.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, and Satish V Ukkusuri. City2city: Translating place representations across cities. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 412–415, 2019.

Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, and Satish V Ukkusuri. Unsupervised translation via hierarchical anchoring: functional mapping of places across cities. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2841–2851, 2020.

Guang Yang, Yuequn Zhang, Jinquan Hang, Xinyue Feng, Zejun Xie, Desheng Zhang, and Yu Yang. Carpg: Cross-city knowledge transfer for traffic accident prediction via attentive region-level parameter generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2939–2948, 2023.

Min Yang, Yang An, Jinliang Deng, Xiaoyu Li, Bin Xu, Ji Zhong, Xiankai Lu, and Yongshun Gong. Can-st: Clustering adaptive normalization for spatio-temporal ood learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 3543–3551, 2025a.

Min Yang, Xiaoyu Li, Bin Xu, Xiushan Nie, Muming Zhao, Chengqi Zhang, Yu Zheng, and Yongshun Gong. Stda: Spatio-temporal deviation alignment learning for cross-city fine-grained urban flow inference. *IEEE Transactions on Knowledge and Data Engineering*, 2025b.

Ying Yang, Jiahao Zhan, Yang Liu, and Qi Wang. Cross-city transfer learning: Applications and challenges for smart cities and sustainable transportation. *Communications in Transportation Research*, 5:100206, 2025c.

Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *Proceedings of The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17*, pages 2181–2191, 2019.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

Shilu Yuan, Xiaoyu Li, Wenqian Mu, Ji Zhong, Meng Chen, Haoliang Sun, and Yongshun Gong. Spatio-temporal prototype-based hierarchical learning for od demand prediction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 3597–3605, 2025a.

Xiaoming Yuan, Zhenyu Luo, Ning Zhang, Ge Guo, Lin Wang, Changle Li, and Dusit Niyato. Federated transfer learning for privacy-preserved cross-city traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2025b.

Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Qianru Zhang, Chao Huang, Lianghao Xia, Zheng Wang, Siu Ming Yiu, and Ruihua Han. Spatial-temporal graph learning with adversarial contrastive adaptation. In *International Conference on Machine Learning*, pages 41151–41163. PMLR, 2023.

Xijun Zhang, Guangyu Wan, and Hong Zhang. Transfer learning for cross-city traffic prediction to solve data scarcity. *Transportation Research Record*, page 03611981241283013, 2025a.

Yudong Zhang, Xu Wang, Xuan Yu, Zhaoyang Sun, Kai Wang, and Yang Wang. Drawing informative gradients from sources: A one-stage transfer learning framework for cross-city spatiotemporal forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1147–1155, 2025b.

Yifan Zhao, Tong Zhang, Jia Li, and Yonghong Tian. Dual adaptive representation alignment for cross-domain few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11720–11732, 2023.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1234–1241, 2020.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A Experimental Details

A.1 Data

We use datasets from Xi'an (XA), Chengdu (CD), and Beijing (BJ). Each city is partitioned into irregular road-network-based regions, with one month of anonymized taxi OD trips mapped to regions to form a directed mobility graph. We evaluate three region-level targets (GDP, population, and carbon emissions) aggregated from public gridded/raster products by assigning grid cells to polygons and summing within each region.

Table 4: Dataset summary.

City	# Regions	# Trips	Targets
XA	1306	559,729	GDP / Pop / CO_2
CD	1056	384,618	GDP / Pop / CO_2
BJ	1311	78,945	GDP / Pop / CO_2

A.2 Baselines.

Baselines. We compare **SCOT** with the following methods. **Non-Alignment** trains on the source and directly applies to the target without adaptation. **RP** [Yabe et al., 2019] forms one-to-one anchors by rank-matching regions and aligns them via Procrustes [Schönemann, 1966]. **HBP** [Yabe et al., 2020] uses level-wise prototype (mean) vectors as anchors under a hierarchical partition, aligned by Procrustes. **HSA** [Yabe et al., 2020] samples anchors within each hierarchical level and fits an unconstrained affine map for more flexible alignment. **MMD** [Gretton et al., 2012] is an RKHS-based correspondence-free loss matching source and target embedding distributions. **Adv** (DANN) [Ganin et al., 2016] learns domain-invariant embeddings via gradient reversal against a domain discriminator. **CrossTReS** [Jin et al., 2022] is a selective fine-tuning framework that meta-learns region weights to prioritize source regions most helpful for the target. **CoRE** [Chen et al.] jointly learns region embeddings and aligns the two latent spaces both globally and at the region level.

B Proof of Proposition 4.1

Proposition 4.1 (Contrastive alignment as a transfer surrogate)**.** *Let $\{u_i\}_{i=1}^{n_s}, \{v_j\}_{j=1}^{n_t} \subset \mathbb{S}^{d-1}$ be the normalized source and target embeddings, with marginals $a \in \Delta^{n_s}$, $b \in \Delta^{n_t}$, and coupling $P \in \mathbb{R}_+^{n_s \times n_t}$ satisfying $P\mathbf{1} = a$ and $P^\top \mathbf{1} = b$. Let g be the unknown region-level task function and h a source-trained predictor applied to the target. Assume that g and h are L_g- and L_h-Lipschitz, and define*

$$\mathcal{R}_s^a(h) := \sum_i a_i |h(u_i) - g(u_i)|, \qquad \mathcal{R}_t^b(h) := \sum_j b_j |h(v_j) - g(v_j)|.$$

Then

$$\mathcal{R}_t^b(h) \le \underbrace{\mathcal{R}_s^a(h)}_{\text{source risk}} + \underbrace{(L_h + L_g)\sqrt{2 - 2\underline{m}}}_{\text{transfer gap}},$$

where

$$\underline{m} := \max\left\{-1,\, \tau \log n_t + \tau H(a) - \tau \mathcal{L}_{\mathrm{Con}}(P) - 1 - \frac{1}{2\tau}\right\}, \qquad H(a) := -\sum_i a_i \log a_i.$$

Proof. Throughout, let $Z_i := \sum_k e^{\langle u_i, v_k\rangle/\tau}$, $q_i(j) := e^{\langle u_i, v_j\rangle/\tau}/Z_i$, and

$$\mathcal{L}_{\mathrm{Con}}(P) := -\sum_i a_i \log \sum_j P_{ij}\, q_i(j),$$

which coincides with Eq. (8) for uniform a. Rows with $a_i = 0$ contribute nothing and are dropped.

Let $\phi := |h - g|$. For any $x, x' \in \mathbb{S}^{d-1}$, $|\phi(x) - \phi(x')| \le |(h-g)(x) - (h-g)(x')| \le (L_h + L_g)\|x - x'\|_2$. Let $(I, J) \sim P$; since $P\mathbf{1} = a$ and $P^\top \mathbf{1} = b$, we have $I \sim a$ and $J \sim b$. Hence

$$\mathcal{R}_t^b(h) - \mathcal{R}_s^a(h) = \mathbb{E}\big[\phi(v_J) - \phi(u_I)\big] \le (L_h + L_g)\,\mathbb{E}\|u_I - v_J\|_2.$$

Since $\|u_i\|_2 = \|v_j\|_2 = 1$, we have $\|u_i - v_j\|_2^2 = 2 - 2\langle u_i, v_j\rangle$, and Jensen's inequality gives

$$\mathbb{E}\|u_I - v_J\|_2 \le \sqrt{\mathbb{E}\|u_I - v_J\|_2^2} = \sqrt{2 - 2\,\mathbb{E}\langle u_I, v_J\rangle}.$$

Fix i and let $p_i(j) := P_{ij}/a_i$ and $\ell_i := -\log \sum_j p_i(j) q_i(j)$. Using $\langle u_i, v_k\rangle \ge -1$,

$$\mathbb{E}_{J\sim p_i} e^{\langle u_i, v_J\rangle/\tau} = Z_i e^{-\ell_i} \ge n_t\, e^{-1/\tau - \ell_i}.$$

For $X := \langle u_i, v_J\rangle \in [-1, 1]$, Hoeffding's lemma gives $\log \mathbb{E} e^{\lambda X} \le \lambda\,\mathbb{E}X + \lambda^2/2$. Taking $\lambda = 1/\tau$,

$$\mathbb{E}_{J\sim p_i}\langle u_i, v_J\rangle \ge \tau \log \mathbb{E}_{J\sim p_i} e^{X/\tau} - \frac{1}{2\tau} \ge \tau \log n_t - 1 - \tau\ell_i - \frac{1}{2\tau}.$$

Since $\log \sum_j P_{ij} q_i(j) = \log a_i - \ell_i$, we have $\sum_i a_i \ell_i = \mathcal{L}_{\mathrm{Con}}(P) - H(a)$. Averaging over $I \sim a$ therefore yields

$$\mathbb{E}\langle u_I, v_J\rangle \ge \tau \log n_t + \tau H(a) - \tau \mathcal{L}_{\mathrm{Con}}(P) - 1 - \frac{1}{2\tau}.$$

Together with $\langle u_I, v_J\rangle \ge -1$, this gives $\mathbb{E}\langle u_I, v_J\rangle \ge \underline{m}$. Note also $\underline{m} \le \mathbb{E}\langle u_I, v_J\rangle \le 1$, so the square root below is well defined. Combining the three steps,

$$\mathcal{R}_t^b(h) - \mathcal{R}_s^a(h) \le (L_h + L_g)\sqrt{2 - 2\,\underline{m}}. \qquad \square$$

Remark B.1. The constants come from worst-case steps ($Z_i \ge n_t e^{-1/\tau}$ and Hoeffding's lemma), so the bound is loose and can be vacuous ($\underline{m} = -1$) at small τ. We use it qualitatively: the transfer gap is non-increasing in $\mathcal{L}_{\mathrm{Con}}$, a relation we verify empirically in Appendix E.4.

C Training Algorithms

We provide pseudocode for the single-source (Algorithm 1) and multi-source (Algorithm 2) variants of **SCOT**.

C.1 Single-source SCOT Training Algorithm

Algorithm 1 Single-source **SCOT** training

Input: Graphs $\mathcal{G}_s, \mathcal{G}_t$; mobility $\mathbf{M}_s, \mathbf{M}_t$; hyperparameters $\tau, \varepsilon, T, \eta, \beta, \lambda_{\mathrm{align}}, \lambda_{\mathrm{rec}}$; step size α
Output: Trained parameters Θ

▶ **Stage 1.** Encode regions & compute intra-city objective
for epoch $= 1, 2, \ldots$ **do**
  $\mathbf{z}_c \leftarrow \mathrm{GAT}(\mathcal{G}_c; \Theta)$, $\mathcal{L}_{\mathrm{intra}}^c \leftarrow \mathcal{L}_{\mathrm{intra}}(\mathbf{z}_c; \mathbf{M}_c)$ for $c \in \{s, t\}$
▶ **Stage 2.** Sinkhorn entropic OT coupling
  $\tilde{\mathbf{z}}_c \leftarrow \mathrm{RowNorm}(\mathbf{z}_c)$, $C_{ij} \leftarrow \|\tilde{\mathbf{z}}_i^s - \tilde{\mathbf{z}}_j^t\|_2$, $\mathbf{K} \leftarrow \exp(-\mathbf{C}/\varepsilon)$
  Initialize $\mathbf{u}, \mathbf{v} \leftarrow \mathbf{1}$
  for $k = 1, \ldots, T$ **do**
    $\mathbf{u} \leftarrow \mathbf{1} \oslash (\mathbf{K}\mathbf{v})$, $\mathbf{v} \leftarrow \mathbf{1} \oslash (\mathbf{K}^\top \mathbf{u})$
  end for
  $\mathbf{P} \leftarrow \mathrm{diag}(\mathbf{u})\,\mathbf{K}\,\mathrm{diag}(\mathbf{v})$ *// soft correspondence*
▶ **Stage 3.** Alignment & reconstruction losses
  $\mathcal{L}_{\mathrm{OT}} \leftarrow \langle \mathbf{P}, \mathbf{C}\rangle / \min(n_s, n_t)$, $\mathcal{L}_{\mathrm{Con}}$ via Eq. (8)
  $\mathcal{L}_{\mathrm{Align}} \leftarrow \mathcal{L}_{\mathrm{OT}} + \eta\,\mathcal{L}_{\mathrm{Con}}$, $\mathcal{L}_{\mathrm{Rec}} \leftarrow \mathcal{L}_{\mathrm{cyc}} + \beta\,\mathcal{R}_{\mathrm{ent}}$
▶ **Stage 4.** Total objective & parameter update
  $\mathcal{L} \leftarrow \underbrace{\mathcal{L}_{\mathrm{intra}}^s + \mathcal{L}_{\mathrm{intra}}^t}_{\text{intra-city}} + \lambda_{\mathrm{align}} \underbrace{\mathcal{L}_{\mathrm{Align}}}_{\text{cross-city}} + \lambda_{\mathrm{rec}} \underbrace{\mathcal{L}_{\mathrm{Rec}}}_{\text{stabilization}}$
  $\Theta \leftarrow \Theta - \alpha\,\nabla_\Theta \mathcal{L}$
end for

C.2 Multi-source hub alignment (illustration)

Figure 9 provides an illustration of the proposed multi-source hub alignment mechanism. Given region embeddings from multiple source cities $\{Z^{(1)}, \ldots, Z^{(M)}\}$ and a target city $Z^{(T)}$, we introduce a set of shared prototypes (hubs) that serve as intermediate anchors to mediate cross-city alignment.

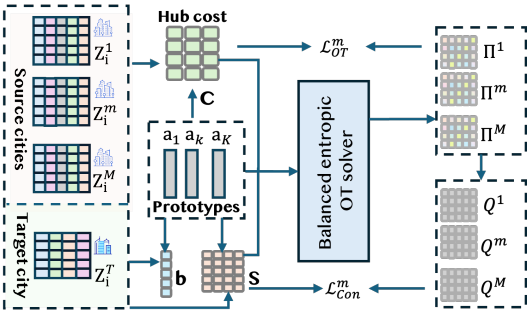


Figure 9: **Illustration of multi-source SCOT.**

Specifically, each city is softly assigned to the prototypes via learned assignment distributions, inducing hub-level representations that summarize transferable structure across sources. These hub representations are then aligned with the target city through a balanced entropic OT solver, producing transport plans $\{\Pi^{(m)}\}$ that guide both the OT alignment loss $\mathcal{L}^m_{\mathrm{OT}}$ and the OT-weighted contrastive objective $\mathcal{L}^m_{\mathrm{Con}}$. This design enables structured many-to-many alignment across cities while avoiding brittle pairwise correspondences and mitigating hub collapse, yielding a scalable and robust extension of **SCOT** to the multi-source setting.

C.3 Algorithm: Multi-source SCOT training

Algorithm 2 Multi-source **SCOT** training (hub entropic OT)

Input: $\{\mathcal{G}_m, \mathbf{M}_m\}_{m\in\mathcal{S}}, \mathcal{G}_t, \mathbf{M}_t$; hub size K; hyperparameters $\tau, \varepsilon, T, \lambda_{\mathrm{align}}, \lambda_{\mathrm{rec}}, \lambda_c, \lambda_{\mathrm{hub}}, \beta$; step size α

Output: Trained parameters Θ (including learnable prototypes $\mathbf{a}$)

▶ **Stage 1.** Encode all cities & compute intra-city losses

for epoch $= 1, 2, \ldots$ **do**

  $\mathbf{z}_m \leftarrow \mathrm{GAT}_m(\mathcal{G}_m; \Theta), \quad \mathcal{L}^m_{\mathrm{intra}} \leftarrow \mathcal{L}_{\mathrm{intra}}(\mathbf{z}_m; \mathbf{M}_m) \quad$ for $m \in \mathcal{S} \cup \{t\}$

▶ **Stage 2.** City-to-hub coupling via balanced entropic OT

  $\tilde{\mathbf{z}}_m \leftarrow \mathrm{RowNorm}(\mathbf{z}_m)$ for $m \in \mathcal{S} \cup \{t\}, \quad \tilde{\mathbf{a}} \leftarrow \mathrm{RowNorm}(\mathbf{a})$

  for all $m \in \mathcal{S} \cup \{t\}$ **do**

    $\mathbf{C}^m_{ik} \leftarrow \|\tilde{\mathbf{z}}^m_i - \tilde{\mathbf{a}}_k\|_2$ *// city-to-hub cost*

    $\mathbf{\Pi}^m \leftarrow \arg\min_{\mathbf{P}\in\Pi(\mathbf{a}^m,\mathbf{b})} \langle \mathbf{P}, \mathbf{C}^m \rangle - \varepsilon\, H(\mathbf{P})$ *// Sinkhorn, T iters*

    $\mathbf{Q}^m \leftarrow \mathrm{RowNorm}(\mathbf{\Pi}^m)$

  end for

▶ **Stage 3.** Per-city alignment & hub-balance losses

  for all $m \in \mathcal{S} \cup \{t\}$ **do**

    $\mathcal{L}^m_{\mathrm{OT}} \leftarrow \langle \mathbf{\Pi}^m, \mathbf{C}^m \rangle / \min(n_m, K)$

    $\mathcal{L}^m_{\mathrm{Con}} \leftarrow -\frac{1}{n_m} \sum_i \log\Big(\frac{\sum_k Q^m_{ik} \exp(\tilde{\mathbf{z}}_i^{m\top} \tilde{\mathbf{a}}_k/\tau)}{\sum_k \exp(\tilde{\mathbf{z}}_i^{m\top} \tilde{\mathbf{a}}_k/\tau)} \Big)$

    $\mathcal{L}^m_{\mathrm{align}} \leftarrow \mathcal{L}^m_{\mathrm{OT}} + \lambda_c\, \mathcal{L}^m_{\mathrm{Con}}$

    $\mathbf{p}^m \leftarrow (\mathbf{\Pi}^m)^\top \mathbf{1}, \quad \mathcal{L}^m_{\mathrm{hub}} \leftarrow \mathrm{KL}\big(\mathbf{p}^m \,\|\, \frac{1}{K}\mathbf{1}\big)$

  end for

▶ **Stage 4.** Aggregate & update

  $\mathcal{L}_{\mathrm{align}} \leftarrow \frac{1}{|\mathcal{S}|+1} \sum_m \mathcal{L}^m_{\mathrm{align}} + \lambda_{\mathrm{hub}} \cdot \frac{1}{|\mathcal{S}|+1} \sum_m \mathcal{L}^m_{\mathrm{hub}}$

  $\mathcal{L}_{\mathrm{rec}} \leftarrow \mathcal{L}_{\mathrm{cycle}}(\{\mathbf{z}_m\}_{m\in\mathcal{S}}, \mathbf{z}_t) + \beta\, \mathrm{Ent}(\mathbf{A}_{\cdot\to t})$

  $\mathcal{L} \leftarrow \underbrace{\sum_{m\in\mathcal{S}} \mathcal{L}^m_{\mathrm{intra}} + \mathcal{L}^t_{\mathrm{intra}}}_{\text{intra-city}} + \lambda_{\mathrm{align}} \underbrace{\mathcal{L}_{\mathrm{align}}}_{\text{hub alignment}} + \lambda_{\mathrm{rec}} \underbrace{\mathcal{L}_{\mathrm{rec}}}_{\text{stabilization}}$

  $\Theta \leftarrow \Theta - \alpha \nabla_\Theta \mathcal{L}$

end for

Table 6: Single-source transfer results on XA↔CD. Lower is better. Red: best; Blue: runner-up. Gain row reports **SCOT**'s relative improvement over the strongest baseline.

Method	XA → CD						CD → XA					
	GDP		Population		CO_2		GDP		Population		CO_2	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Non-Alignment	323.90	11.49	908.18	4.73	237.73	12.91	232.86	3.74	748.16	5.22	179.25	1.67
RP	191.11	11.64	644.22	3.32	144.87	9.18	194.67	6.57	671.30	6.41	145.51	2.30
HBP	188.61	13.42	660.01	4.26	135.14	11.59	199.79	7.06	697.44	6.93	146.76	2.47
HSA	184.81	11.95	619.72	3.57	137.29	9.92	202.01	7.27	708.80	7.14	147.27	2.54
MMD	201.60	10.49	645.42	3.15	162.21	8.83	176.00	5.26	608.58	5.15	139.15	1.83
Adv	194.71	13.68	645.59	3.98	124.98	10.01	215.47	8.54	785.78	8.50	150.10	2.99
CrossTReS	183.22	12.82	712.74	4.55	159.86	9.26	180.31	5.09	630.66	5.37	143.58	1.92
CoRE	174.28	10.55	615.97	2.96	128.77	8.82	174.52	5.22	600.45	4.98	138.12	1.78
SCOT (Ours)	165.88	7.67	575.43	2.35	114.68	7.83	158.95	3.12	538.23	3.37	130.04	1.24
Δ vs. best baseline	+4.8%	**+26.9%**	+6.6%	**+20.6%**	**+8.2%**	**+11.2%**	**+8.9%**	**+16.6%**	**+10.4%**	**+32.3%**	+5.8%	**+25.7%**

D Intra-city Prediction with and without Alignment

Cross-city gains should not come at the cost of within-city accuracy. We therefore compare full **SCOT** with an alignment-free variant on intra-city prediction (XA→XA, BJ→BJ). The differences are small: MAE changes by less than 1% in every case, and MAPE by at most 0.2 points (Table 5). Alignment thus leaves within-city quality largely intact.

Table 5: Intra-city prediction with and without alignment (4 seeds, mean ± std). MAE differs by less than 1% in all cases. Lower is better.

Direction / Variant	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
▶ **TARGET: XI'AN (XA)**						
XA→XA *(w/o Alignment)*	155.31 ± 1.92	3.63 ± 0.06	467.03 ± 4.45	2.50 ± 0.04	130.35 ± 0.97	2.60 ± 0.05
XA→XA *(Full **SCOT**)*	156.44 ± 2.29	3.67 ± 0.11	467.96 ± 8.19	2.51 ± 0.05	130.55 ± 1.93	2.59 ± 0.06
▶ **TARGET: BEIJING (BJ)**						
BJ→BJ *(w/o Alignment)*	95.41 ± 1.19	2.48 ± 0.06	531.27 ± 10.10	2.85 ± 0.06	154.57 ± 2.24	2.35 ± 0.28
BJ→BJ *(Full **SCOT**)*	96.05 ± 1.65	2.45 ± 0.07	534.79 ± 13.17	2.93 ± 0.07	155.58 ± 1.71	2.55 ± 0.06

E Additional Experiment Details

E.1 Additional Single-Source Results

Tables 6 and 7 report the remaining single-source directions (XA↔CD and BJ↔CD). **SCOT** is best on every task, metric, and direction.

E.2 Additional Results on XA↔BJ (4 Random Seeds)

To verify that **SCOT**'s gains are not artifacts of seed selection, we report single-source transfer in both directions (XA→BJ and BJ→XA) under *four random seeds*, with mean ± standard deviation of MAE and MAPE for GDP, population, and CO_2 (Tables 8, 9). **SCOT** achieves **the lowest mean error on every metric in both directions**, with *the smallest standard deviation* on most entries, indicating that the improvements are both consistent and statistically robust rather than driven by favorable seeds.

Table 7: Single-source transfer results on BJ↔CD. Lower is better. Red: best; Blue: runner-up. Gain row reports SCOT's relative improvement over the strongest baseline.

Method	BJ → CD						CD → BJ					
	GDP		Population		CO_2		GDP		Population		CO_2	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Non-Alignment	191.36	5.31	889.37	3.97	192.64	15.77	163.93	7.04	805.78	5.07	192.96	1.98
RP	158.57	8.71	726.42	5.39	175.21	14.49	150.56	6.63	697.53	6.64	156.97	2.03
HBP	155.05	8.30	698.75	4.80	172.65	13.22	166.30	7.29	715.27	6.84	154.77	1.79
HSA	147.74	6.87	667.95	4.20	160.10	11.55	148.87	6.59	648.28	5.75	166.55	2.64
MMD	154.96	8.29	706.31	4.98	170.85	12.15	130.41	4.94	632.71	5.08	151.57	1.75
Adv	178.84	11.91	861.43	7.44	226.49	19.88	162.20	7.42	701.22	6.18	184.74	3.20
CrossTReS	159.73	6.86	654.14	3.45	137.63	9.83	140.37	5.58	686.19	5.66	198.34	2.39
CoRE	150.51	7.57	680.81	4.34	126.66	9.66	159.00	6.81	673.17	5.96	163.39	2.51
SCOT (Ours)	135.63	3.55	597.80	2.38	121.21	8.94	118.48	3.41	580.95	2.74	148.50	1.54
Δ *vs. best baseline*	**+8.2%**	**+33.1%**	**+8.6%**	**+31.0%**	+4.3%	+7.5%	**+9.1%**	**+31.0%**	**+8.2%**	**+46.0%**	+2.0%	**+12.0%**

Table 8: XA→BJ results averaged over 4 random seeds (mean ± std). Lower is better. Red: best; Blue: runner-up. Gain row reports SCOT's relative improvement over the strongest baseline.

Method	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
Non-Alignment	276.33 ± 10.63	9.46 ± 1.98	958.52 ± 31.66	5.99 ± 1.82	278.25 ± 9.95	5.39 ± 0.75
RP	192.05 ± 26.62	6.65 ± 1.48	676.21 ± 28.24	4.00 ± 0.66	194.70 ± 11.88	3.71 ± 0.22
HBP	186.21 ± 10.14	8.17 ± 0.88	664.15 ± 24.83	4.68 ± 0.89	188.14 ± 5.39	3.99 ± 0.20
HSA	179.96 ± 17.62	7.30 ± 1.00	631.69 ± 27.93	4.64 ± 1.19	180.21 ± 8.29	3.57 ± 0.64
MMD	162.63 ± 16.27	5.93 ± 0.90	596.60 ± 21.41	3.63 ± 0.66	169.99 ± 7.14	2.91 ± 0.46
Adv	200.33 ± 13.90	8.98 ± 0.60	694.64 ± 9.39	6.15 ± 1.04	199.99 ± 2.61	4.63 ± 0.37
CrossTReS	194.87 ± 28.96	7.28 ± 0.90	629.37 ± 22.46	4.29 ± 0.18	182.88 ± 9.74	3.59 ± 0.28
CoRE	159.53 ± 14.64	6.19 ± 1.65	607.79 ± 39.24	4.19 ± 1.13	170.55 ± 11.99	3.12 ± 0.67
SCOT (Ours)	120.25 ± 7.30	3.59 ± 0.48	527.04 ± 6.38	2.17 ± 0.23	149.20 ± 1.58	1.80 ± 0.17
Δ *vs. best baseline*	**+24.6%**	**+39.5%**	**+11.7%**	**+40.2%**	**+12.2%**	**+38.1%**

Table 9: BJ→XA results averaged over 4 random seeds (mean ± std). Lower is better. Red: best; Blue: runner-up. Gain row reports SCOT's relative improvement over the strongest baseline.

Method	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
Non-Alignment	220.31 ± 22.10	7.91 ± 1.43	975.83 ± 65.55	9.63 ± 1.27	276.84 ± 18.56	9.70 ± 1.14
RP	175.21 ± 4.15	4.60 ± 0.53	671.08 ± 11.95	6.37 ± 0.22	191.56 ± 3.33	6.30 ± 0.24
HBP	180.60 ± 13.16	2.97 ± 0.82	629.09 ± 5.07	4.95 ± 0.39	179.17 ± 2.76	4.91 ± 0.45
HSA	176.42 ± 10.52	3.66 ± 1.35	649.08 ± 16.82	5.66 ± 0.62	185.72 ± 4.29	5.62 ± 0.57
MMD	180.71 ± 5.39	2.25 ± 0.22	500.12 ± 27.35	1.93 ± 0.09	141.24 ± 4.96	1.91 ± 0.08
Adv	192.06 ± 6.28	6.15 ± 0.34	778.54 ± 30.34	8.53 ± 0.62	212.72 ± 9.74	7.84 ± 0.62
CrossTReS	165.18 ± 3.45	3.98 ± 0.22	627.96 ± 8.88	5.18 ± 0.32	179.42 ± 2.30	5.16 ± 0.29
CoRE	162.64 ± 5.07	2.80 ± 0.80	576.95 ± 25.66	3.43 ± 1.15	164.26 ± 8.68	3.44 ± 1.18
SCOT (Ours)	160.21 ± 3.53	1.87 ± 0.18	450.14 ± 2.81	1.73 ± 0.12	127.79 ± 1.08	1.78 ± 0.10
Δ *vs. best baseline*	+1.8%	**+16.9%**	**+10.0%**	**+10.4%**	**+9.5%**	+6.8%

E.3 Additional Multi-Source Results (Target: CD)

Table 10 reports multi-source transfer with Chengdu (CD) as target. SCOT achieves the **best performance on all three tasks** under both MAE and MAPE, with the largest gains on *population and CO_2*. When existing single-source alignment and distribution-matching strategies are naively extended to multiple sources, they yield only limited improvements. This confirms that effective

multi-source transfer is not a matter of summing per-source losses, but requires the *coordinated aggregation* provided by **SCOT**'s shared hub.

Table 10: Multi-source transfer results (Target: CD). Lower is better. Red: best; Blue: runner-up. Gain row reports **SCOT**'s relative improvement over the strongest baseline.

Method	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
RP	187.13	13.31	630.44	3.62	167.76	14.05
HBP	176.35	11.11	631.26	3.69	125.97	9.96
HSA	180.71	8.18	651.06	3.41	159.86	11.79
MMD	163.82	5.07	639.32	3.93	145.74	10.83
Adv	183.26	12.43	687.72	4.73	156.37	12.92
CrossTReS	167.53	10.26	668.00	4.37	139.22	11.84
CoRE	156.10	4.40	621.11	3.44	121.33	9.57
SCOT (Ours)	133.94	3.82	546.82	2.43	98.43	5.10
Δ *vs. best baseline*	**+14.1%**	**+13.1%**	**+11.9%**	**+28.7%**	**+18.8%**	**+46.7%**

E.4 Empirical Check of the Proposition

To complement Proposition 4.1, we test its qualitative mechanism by relating target error y to both alignment terms in a *joint regression*:

$$y \;=\; \beta_0 + \beta_1\,\mathcal{L}_{\mathrm{Con}} + \beta_2\,\mathcal{L}_{\mathrm{OT}} + \varepsilon.$$

This isolates the effect of $\mathcal{L}_{\mathrm{Con}}$ on target error *after controlling for* $\mathcal{L}_{\mathrm{OT}}$, addressing the concern that the two alignment terms might be confounded. As shown in Table 11, the standardized coefficient on $\mathcal{L}_{\mathrm{Con}}$ is **0.77**, and its partial Pearson/Spearman correlations remain at **0.95/0.93** after controlling for $\mathcal{L}_{\mathrm{OT}}$. This empirically supports the proposition's qualitative message: *stronger contrastive alignment is closely associated with lower target error, independently of OT alignment quality*.

Table 11: Empirical verification of Proposition 4.1. The standardized coefficient on $\mathcal{L}_{\mathrm{Con}}$ remains **strongly positive after controlling for** $\mathcal{L}_{\mathrm{OT}}$, and partial Pearson/Spearman correlations both exceed 0.93 with $p < 0.001$.

Joint OLS: $y \sim \mathcal{L}_{\mathrm{Con}} + \mathcal{L}_{\mathrm{OT}}$		**Partial correlation with** $\mathcal{L}_{\mathrm{Con}}$ *(controlling for* $\mathcal{L}_{\mathrm{OT}}$*)*		
Std. coef. on $\mathcal{L}_{\mathrm{Con}}$	Adj. R^2	Pearson r	Spearman ρ	p-value
0.77	**0.94**	**0.95**	**0.93**	<0.001

E.5 Additional OT Coupling Diagnostics

The log-scale plot in Fig. 5 visually exaggerates tiny tail entries, so we quantify the coupling directly using row-normalized profiles $q(j \mid i) = P_{ij} / \sum_{j'} P_{ij'}$. We ask two questions: is each row's mass concentrated, and do different source regions share the same target pattern?

Mass is concentrated. On average, 35, 76, and 112 of the 1,311 Beijing regions cover 50%, 80%, and 90% of a row's mass; the representative region in Fig. 5 behaves similarly (top-20 targets hold 50.8%).

Rows are region-specific. Pairwise cosine similarity between rows is near zero (mean 0.082, median 0.002) and the mean JS divergence is 0.587, so source regions do not collapse onto a shared target pattern.

In short, the coupling is dense in support but selective in mass.

Table 12: Selectivity and region specificity of the XA→BJ coupling. Coverage is the number of target regions required to contain the specified fraction of transported mass.

Representative source region		Average over all source regions		Across-row diversity	
Diagnostic	**Result**	**Diagnostic**	**Result**	**Diagnostic**	**Result**
Mass in top-10 targets	**32.5%**	Targets covering 50% mass	**35**	Mean cosine similarity	**0.082**
Mass in top-20 targets	**50.8%**	Targets covering 80% mass	**76**	Median cosine similarity	**0.002**
Targets covering 90% mass	**108**	Targets covering 90% mass	**112**	Mean JS divergence	**0.587**

F Ablation Study

F.1 Ablation Study for Single-Source SCOT

We remove each of $\mathcal{L}_{\mathrm{OT}}$, $\mathcal{L}_{\mathrm{Con}}$, and $\mathcal{L}_{\mathrm{Rec}}$ in turn.

Transfer performance. Across six directions (Fig. 10), removing $\mathcal{L}_{\mathrm{OT}}$ causes the largest drop, removing $\mathcal{L}_{\mathrm{Con}}$ consistently hurts, and removing $\mathcal{L}_{\mathrm{Rec}}$ has a smaller but noticeable effect.

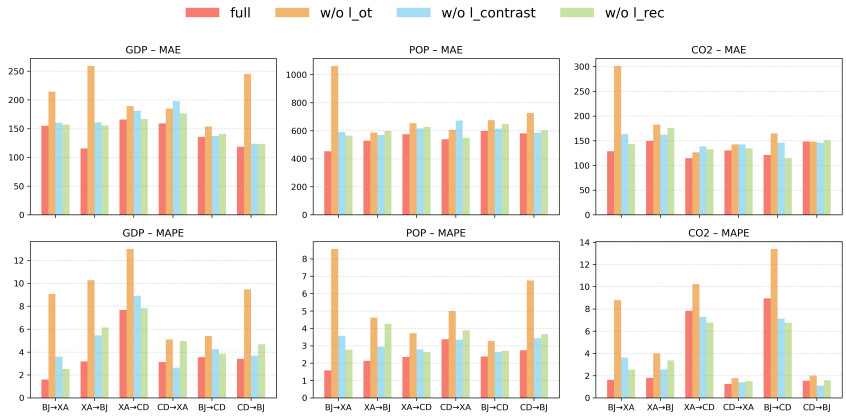


Figure 10: **Single-source ablation results.** MAE (top) and MAPE (bottom) are reported for GDP, population, and CO_2 prediction across six transfer directions: BJ→XA, XA→BJ, XA→CD, CD→XA, BJ→CD, and CD→BJ.

Effect on the coupling. To see how $\mathcal{L}_{\mathrm{OT}}$ and $\mathcal{L}_{\mathrm{Con}}$ shape the coupling, we measure each row of $q(j \mid i) = P_{ij} / \sum_{j'} P_{ij'}$ by its *top-1 mass* (largest entry) and *normalized entropy* (spread), averaged over source regions (Table 13).

Table 13: Coupling diagnostics for XA→BJ (mean ± std over ablation runs). Values in parentheses are ratios relative to full **SCOT**.

Variant	Top-1 mass	Normalized entropy
SCOT	0.015 ± 0.002	0.745 ± 0.015
w/o $\mathcal{L}_{\mathrm{Con}}$	0.0011 ± 0.0001 $(0.07\times)$	0.9999 ± 0.0001 $(1.34\times)$
w/o $\mathcal{L}_{\mathrm{OT}}$	0.205 ± 0.025 $(13.7\times)$	0.815 ± 0.020 $(1.09\times)$

Without $\mathcal{L}_{\mathrm{Con}}$, rows become nearly uniform (entropy ≈ 1): OT alone gives little local discrimination. Without $\mathcal{L}_{\mathrm{OT}}$, each row puts large mass on one target but keeps a diffuse tail (entropy 0.815): contrastive learning alone does not yield a structured global allocation. The two terms are thus complementary.

F.2 Ablation: Hub vs. Pairwise OT (Multi-source)

We compare our hub, which aligns all cities to K shared prototypes, with a pairwise variant that aligns each source to the target separately and combines the two objectives with a learnable global gate (Table 14).

Table 14: Multi-source ablation: shared-prototype **Hub** vs. pairwise OT with global gating (**No Hub**), across three targets (mean $\pm$ std over 4 seeds). Red: better in each pair. Lower is better.

Target / Variant	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
▶ **TARGET: XI'AN (XA)**						
Ours (Hub)	154.49 ± 2.10	2.12 ± 0.31	467.54 ± 17.42	2.23 ± 0.25	133.58 ± 4.64	1.51 ± 0.18
No Hub (Pairwise)	157.45 ± 3.31	2.52 ± 0.48	511.37 ± 23.34	2.98 ± 0.64	135.56 ± 6.90	1.71 ± 0.35
▶ **TARGET: CHENGDU (CD)**						
Ours (Hub)	143.91 ± 6.84	4.54 ± 0.49	565.65 ± 14.55	2.38 ± 0.09	102.20 ± 11.20	5.92 ± 0.67
No Hub (Pairwise)	146.72 ± 7.48	6.00 ± 1.03	585.49 ± 15.77	2.17 ± 0.20	105.96 ± 2.38	5.87 ± 0.47
▶ **TARGET: BEIJING (BJ)**						
Ours (Hub)	110.40 ± 5.34	3.30 ± 0.50	533.11 ± 10.52	2.98 ± 0.74	145.72 ± 3.58	1.46 ± 0.21
No Hub (Pairwise)	140.86 ± 15.09	5.15 ± 1.18	580.37 ± 20.59	4.20 ± 0.55	152.83 ± 4.38	1.92 ± 0.27

The hub is better on 16 of 18 metrics and wins every MAE, with the largest gain on Beijing GDP (110.40 vs. 140.86). The pairwise variant is slightly better only on two Chengdu MAPE entries.

F.3 Ablation: Target-Induced Prototype Prior

We compare three choices of the prototype marginal $\mathbf{b}$ (Eq. (14)): **Uniform** ($1/K$, no target signal), **Frozen** (computed from early target embeddings and then fixed), and **Adaptive** (ours, updated during training).

Table 15: Ablation on the target-induced prototype marginal $\mathbf{b}$ (XA as target, 4 seeds, mean $\pm$ std). The **adaptive** prior wins on every metric. Lower is better.

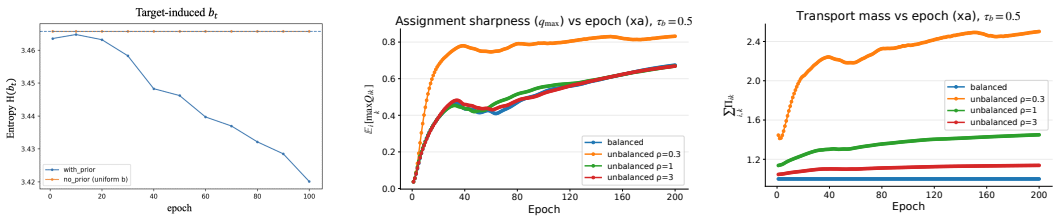
Prior Variant	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
Uniform *(no target)*	186.33 ± 3.93	3.42 ± 0.14	575.81 ± 17.09	3.27 ± 0.33	156.56 ± 10.13	1.85 ± 0.18
Frozen *(fixed)*	181.47 ± 3.16	4.79 ± 0.19	553.05 ± 8.68	3.81 ± 0.10	148.89 ± 0.67	2.83 ± 0.02
Adaptive (Ours)	154.49 ± 2.10	2.12 ± 0.31	467.54 ± 17.42	2.23 ± 0.25	133.58 ± 4.64	1.51 ± 0.18
Δ *vs. Uniform*	**+17.1%**	**+38.0%**	**+18.8%**	**+31.8%**	**+14.7%**	**+18.4%**

The adaptive prior is best on every metric (Table 15). Fig. 11a shows why: the uniform prior's entropy stays at $\log K$, while the adaptive prior's entropy decreases steadily as prototypes specialize toward target semantics. Fixing $\mathbf{b}$ early lowers MAE relative to Uniform but raises MAPE, so the prior must evolve with the target encoder.

F.4 OT Formulation

We test whether relaxing balanced OT helps, first in single-source alignment and then in the multi-source hub.

Single source. We replace balanced OT with unbalanced OT [Chizat et al., 2018], partial OT [Chapel et al., 2020], or the Sinkhorn divergence [Genevay et al., 2018, Feydy et al., 2019], keeping everything



(a) Entropy of $\mathbf{b}$ (b) Assignment sharpness $q_{\max}$ (c) Transport mass $\sum_{i,k} \Pi_{ik}$

Figure 11: Hub diagnostics on CD,BJ→XA. (a) Entropy of the prototype marginal $\mathbf{b}$ during training: it stays at $\log K$ under the uniform prior and decreases under the adaptive prior. (b, c) Unbalanced OT with small ρ gives sharper assignments only by inflating the total mass above one ($\tau_b = 0.5$).

else fixed (Table 16). Balanced OT gives the lowest MAE in every case, and unbalanced OT, the only variant that relaxes the marginals, is worst on most metrics.

Table 16: OT variant comparison on XA↔BJ. Balanced entropic OT (our default) gives the lowest MAE on all tasks and directions. Red: best per column. Lower is better.

OT variant in SCOT	**XA → BJ**						**BJ → XA**					
	GDP		Pop		CO_2		GDP		Pop		CO_2	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Balanced (Ours)	115.33	3.17	528.50	2.13	149.42	1.79	154.92	1.60	452.67	1.58	128.74	1.63
Partial (frac= 0.7)	137.18	4.54	556.49	3.11	156.70	2.34	172.49	2.01	462.15	1.56	131.25	1.60
Sinkhorn divergence	141.78	3.90	562.76	2.23	160.57	2.06	167.83	2.62	476.30	2.54	136.74	2.50
Unbalanced (KL marginal)	145.04	5.77	627.52	5.15	177.08	3.67	172.09	2.42	487.21	2.62	139.69	2.52

Multi-source hub. In the city-to-hub coupling, we vary the KL penalty ρ of unbalanced OT. Small ρ gives sharper early assignments (Fig. 11b), but only by inflating the total mass far above one (Fig. 11c), i.e., by duplicating mass rather than matching better. Balanced OT gives the lowest MAE on all three tasks, while unbalanced OT is sensitive to ρ and only approaches it at $\rho = 3$ (Table 17).

Table 17: Balanced vs. unbalanced OT in the hub on CD,BJ→XA (4 seeds, mean ± std). Balanced OT gives the lowest MAE on all tasks. Red: best; Blue: runner-up. Lower is better.

OT Variant	**GDP**		**Population**		**CO_2**	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
Unbalanced OT ($\rho = 0.3$)	212.96 ± 27.46	7.22 ± 1.16	721.83 ± 95.63	4.80 ± 0.69	174.77 ± 41.48	3.17 ± 1.29
Unbalanced OT ($\rho = 1$)	173.59 ± 13.68	4.16 ± 1.54	563.30 ± 33.74	3.49 ± 1.09	149.69 ± 9.89	1.90 ± 0.48
Unbalanced OT ($\rho = 3$)	165.98 ± 11.66	2.06 ± 0.90	530.68 ± 13.40	2.27 ± 0.47	147.91 ± 8.99	1.81 ± 0.14
Balanced OT (Ours)	154.49 ± 2.10	2.12 ± 0.31	467.54 ± 17.42	2.23 ± 0.25	133.58 ± 4.64	1.51 ± 0.18

F.5 Cycle Design

We examine two choices in the cycle regularizer: whether to enforce one or both cycle directions, and, if one, which.

One-sided vs. two-sided. The two-sided variant adds the reverse target→source→target cycle to our source cycle. Because $n_s \neq n_t$, the two cities are not related by an invertible map, so requiring both cycles over-constrains the matching. The one-sided design is better on 23 of 24 metrics across four real-city directions (Table 18).

Table 18: One-sided vs. two-sided cycle reconstruction across four transfer directions (4 seeds, mean ± std). The one-sided design wins on 23 of 24 metrics. Red: better in each pair. Lower is better.

Variant	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
▶ **DIRECTION: BJ → XA**						
One-sided (Ours)	160.21 ± 3.53	1.87 ± 0.18	450.14 ± 2.81	1.73 ± 0.12	127.79 ± 1.08	1.78 ± 0.10
Two-sided	160.92 ± 2.94	1.65 ± 0.20	503.85 ± 14.49	2.51 ± 0.35	143.18 ± 2.82	2.47 ± 0.30
▶ **DIRECTION: XA → BJ**						
One-sided (Ours)	120.25 ± 7.30	3.59 ± 0.48	527.04 ± 6.38	2.17 ± 0.23	149.20 ± 1.58	1.80 ± 0.17
Two-sided	164.88 ± 12.64	6.38 ± 0.44	583.17 ± 18.17	3.92 ± 0.13	167.51 ± 5.62	2.98 ± 0.15
▶ **DIRECTION: XA → CD**						
One-sided (Ours)	154.60 ± 2.33	4.88 ± 0.57	558.56 ± 11.01	2.19 ± 0.21	114.71 ± 4.22	6.48 ± 0.78
Two-sided	172.93 ± 8.88	7.66 ± 1.58	582.95 ± 7.16	2.48 ± 0.08	130.86 ± 6.30	7.11 ± 0.35
▶ **DIRECTION: CD → XA**						
One-sided (Ours)	159.61 ± 0.71	3.17 ± 0.29	531.00 ± 12.57	3.29 ± 0.10	131.10 ± 2.00	1.24 ± 0.02
Two-sided	172.65 ± 5.98	4.38 ± 0.15	572.33 ± 13.30	4.10 ± 0.31	135.88 ± 1.98	1.49 ± 0.13

Which direction. We cycle back to the source because labels and the downstream predictor live there. When $n_s > n_t$, however, the target-returning cycle is better posed: $\mathrm{rank}(\mathbf{A}_{s\to t}\mathbf{A}_{t\to s}) \le n_t < n_s$ prevents the source cycle from reaching the identity. To test whether this matters, we use paired synthetic cities with $n_s = 480$ and $n_t = 240$ and vary only the cycle direction (four seeds, 200 training epochs).

Table 19: Target MAE under source-returning and target-returning cycles at $n_s = 480$ and $n_t = 240$ (mean ± std over four seeds). Lower is better.

Graph	Source-returning	Target-returning
Mobility	1.158 ± 0.198	1.218 ± 0.150
Feature similarity	1.025 ± 0.177	1.037 ± 0.138
Spatial-functional	0.945 ± 0.233	0.979 ± 0.213

The better-posed target-returning cycle does not improve prediction: all differences are within one standard deviation, and the source-returning cycle has slightly lower mean MAE in every case. Since the cycle acts as a regularizer rather than an exact inverse, we keep the source-returning cycle as the default.

F.6 Comparison to Structure-Aware Cost Formulations

We compare our cost (ℓ_2 distance between normalized embeddings) with three structure-aware alternatives: a **graph-aware cost**, $C_{ij} = \alpha\|\tilde{u}_i - \tilde{v}_j\|_2 + (1-\alpha)|\sigma_s(i) - \sigma_t(j)|$ with $\sigma(\cdot) = \|L\tilde{u}(\cdot)\|_2$ and $\alpha = 0.7$; a **mobility-profile cost**, which adds OD volume and entropy statistics ($\alpha = 0.5$); and **Gromov–Wasserstein** [Peyré et al., 2016, Mémoli, 2011], which matches intra-city distance structure. Our cost is best on all 12 metrics (Table 20). Since the GAT encoder already encodes graph and mobility structure, adding it again through the cost brings no benefit.

Table 20: Cost formulation comparison on XA ↔ BJ. ℓ_2 on ℓ_2-normalized embeddings (**SCOT**'s default) achieves the best result on *all* 12 metrics. Red: best per cell. Lower is better.

Cost variant in SCOT	XA → BJ						BJ → XA					
	GDP		Pop		CO_2		GDP		Pop		CO_2	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
ℓ_2 **on embeddings (Ours)**	115.33	3.17	528.50	2.13	149.42	1.79	154.92	1.60	452.67	1.58	128.74	1.63
Graph-aware ($\alpha = 0.7$)	152.93	5.73	588.56	3.74	168.17	2.96	166.51	2.43	458.53	2.27	131.21	2.24
Mobility-profile ($\alpha = 0.5$)	175.20	7.49	598.49	4.58	176.77	3.67	172.54	2.15	468.85	1.93	132.97	1.88
Gromov-Wasserstein	177.10	6.62	607.10	3.54	174.19	3.03	167.70	2.53	501.56	2.57	142.65	2.65

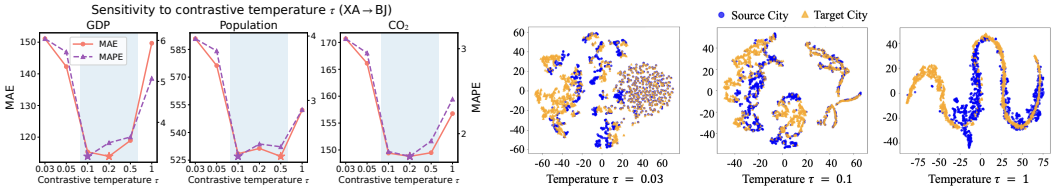
G Hyperparameter Sensitivity

The main paper reports quantitative sensitivity for λ_{align} and ε. Here we provide the remaining sweeps and visualizations: τ, hub size K, contrastive weight η, and target-prior temperature τ_b, together with t-SNE visualizations that show how each hyperparameter shapes the cross-city embedding geometry. Across all six hyperparameters, **SCOT** exhibits broad stable regions and only degrades at extreme values, supporting the main-paper claim that the gains stem from the OT-based alignment framework rather than fine-grained tuning.

G.1 Sensitivity to Contrastive Temperature τ

We sweep $\tau \in \{0.03, 0.05, 0.1, 0.2, 0.5, 1\}$ on XA→BJ (Fig. 12), revealing a characteristic *U-shape* in target performance and a matching geometric pattern in embedding space.

(1) **Quantitative U-shape.** Very small τ (0.03–0.05) over-sharpens similarity weighting and amplifies noise; large $\tau = 1$ dilutes the discriminative signal. The *sweet spot* $\tau \in [0.1, 0.5]$ balances sharpness against smoothness, and we adopt $\tau = 0.1$ as the default (Fig. 12 (a)). (2) **Geometric mirror in embedding space.** The t-SNE visualization (Fig. 12 (b)) reveals the same trade-off geometrically: small τ leaves the two cities *weakly interleaved*, moderate τ produces *clean interleaving with preserved clusters*, and large τ *over- smooths* the embeddings into indistinguishable mixtures. The embedding geometry tracks the metric U-shape, providing visual confirmation of the sweet spot.



(a) **Metrics.** MAE / MAPE vs. τ.

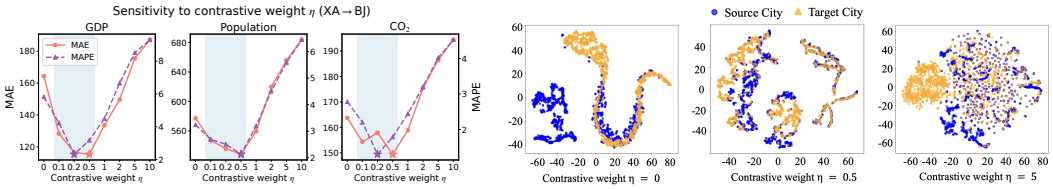
(b) **Embeddings.** t-SNE at $\tau = 0.03,\ 0.1,\ 1$.

Figure 12: **Sensitivity to τ on XA→BJ.** (a) U-shape with sweet spot $\tau \in [0.1, 0.5]$. (b) Geometry mirrors metrics: extreme τ yields weak interleaving or over-smoothing; $\tau = 0.1$ aligns cleanly without collapsing clusters.

G.2 Sensitivity to Contrastive Weight η

The contrastive weight η balances OT-based geometric correspondence against contrastive discriminative sharpening. We sweep $\eta \in \{0, 0.1, 0.2, 0.5, 1, 2, 5, 10\}$ on XA→BJ (Fig. 13).

(1) **Moderate η wins; both extremes fail.** $\eta \in [0.1, 0.5]$ consistently yields the best or near-best performance (Fig. 13 (a)). Too small η under-uses contrastive sharpening; too large $\eta \geq 2$ causes sharp error increases. (2) **The geometry confirms the failure modes.** t-SNE (Fig. 13 (b)) shows that $\eta = 0$ leaves the two cities *barely interleaved*, $\eta = 0.5$ produces *clean alignment with preserved clusters*, and $\eta = 5$ drives *near-complete feature mixing*. OT and contrastive alignment thus act complementarily, with **SCOT** robust over a practical mid-range.



(a) **Metrics.** MAE / MAPE vs. η.

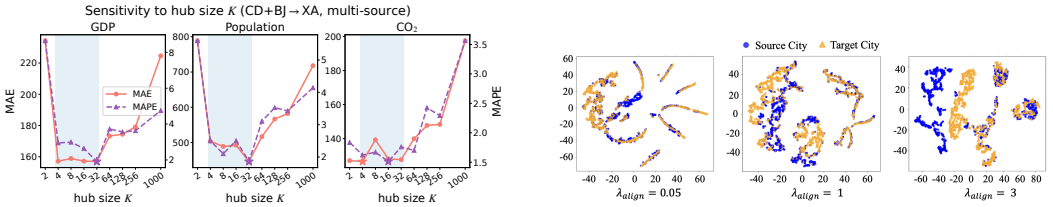
(b) **Embeddings.** t-SNE at $\eta = 0,\ 0.5,\ 5$. Blue $\circ$: source; orange $\triangle$: target.

Figure 13: **Sensitivity to η on XA→BJ.** (a) Best in $\eta \in [0.1, 0.5]$; large η degrades sharply. (b) $\eta = 0$ under-aligns; $\eta = 0.5$ aligns cleanly; $\eta = 5$ over-mixes the two cities.

G.3 Sensitivity to Hub Size K and λ_{align}

Hub size K. We sweep $K \in \{2, 4, 8, 16, 32, 64, 128, 256, 1000\}$ on CD,BJ→XA (Fig. 14a). Performance is stable for $K \in [4, 32]$. At $K = 2$, heterogeneous regions must share too few prototypes and performance drops sharply; for $K \geq 64$, the hub becomes too fine to act as a compact bridge. We use $K = 32$.

Effect of λ_{align} on geometry. Fig. 14b complements the sweep in Fig. 7. At $\lambda_{\text{align}} = 1$, the two cities interleave while keeping their cluster structure. At 0.05 they remain largely separate (alignment too weak), and at 3 they are over-mixed (alignment overrides intra-city structure).



(a) Hub size K (CD,BJ→XA)

(b) t-SNE at $\lambda_{\text{align}} = 0.05,\ 1,\ 3$ (XA→BJ)

Figure 14: (a) MAE/MAPE vs. hub size K; performance is stable for $K \in [4, 32]$. (b) Embeddings under varying λ_{align}; blue $\circ$: source, orange $\triangle$: target.

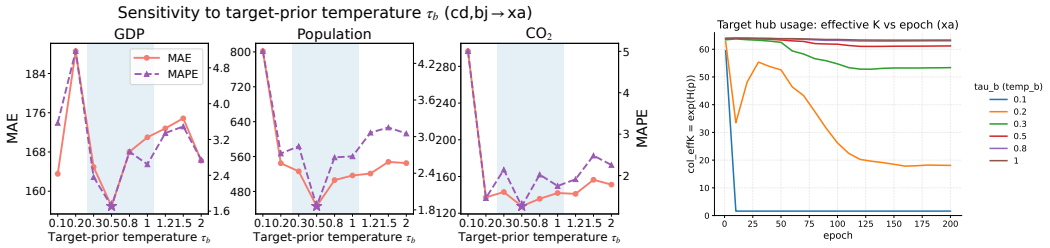
G.4 Sensitivity to Target-Prior Temperature τ_b

The target-prior temperature τ_b controls the sharpness of the target-induced prototype marginal $\mathbf{b}$: smaller τ_b peaks $\mathbf{b}$, larger τ_b flattens it. We sweep $\tau_b \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1.0, 2.0\}$ on multi-source CD,BJ→XA, jointly with hub-usage diagnostics that reveal the underlying mechanism.

(1) **The recommended range $\tau_b \in [0.3, 0.8]$ delivers the best performance.** Within this range, MAE and MAPE are stably minimized across GDP, Population, and CO_2, with $\tau_b = 0.5$ as the operating midpoint (Fig. 15 (a)). The plateau spans nearly an order of magnitude, eliminating the need for per-target tuning.

(2) **Hub diagnostics explain why.** The OT column marginal $p_k = \sum_i \Pi_{ik}$ reports how the target city distributes mass across prototypes. Within the recommended range, the normalized entropy $H(p)/\log K$ stabilizes around 0.4, corresponding to $\exp(H(p)) \approx 4$ *effective prototypes* per region—a selective, non-collapsed hub usage that aligns with strong transfer (Fig. 15 (b)).

(3) **Boundary behavior is interpretable.** Departures from the recommended range deviate predictably: small $\tau_b \leq 0.2$ peaks $\mathbf{b}$ onto few prototypes (low entropy, prototype *specialization without diversity*), while large $\tau_b \geq 1.0$ flattens $\mathbf{b}$ toward uniform ($\exp(H(p)) \to K$, *diffuse usage with weakened target guidance*). Both extremes serve only to bracket the boundary behavior of the target-induced prior.



(a) **Metrics.** MAE (solid) and MAPE (dashed) vs. τ_b. (b) **Hub diagnostics.**

Figure 15: **Sensitivity to τ_b on multi-source CD,BJ→XA.** (a) Performance plateau across $\tau_b \in [0.3, 0.8]$. (b) Normalized entropy $H(p)/\log K$ and effective prototype count $\exp(H(p))$ on XA: hub usage stabilizes at ~ 4 effective prototypes within the same range, providing direct mechanistic evidence for the metric plateau.

H Semantic Analysis of the Learned Transport

We test whether high-mass patterns in the learned coupling reflect semantic correspondence rather than arbitrary block concentration, at three levels: a single source region, blocks of the full coupling, and synthetic cities with known correspondence. For real cities, we use two attributes never seen during training as coarse semantic proxies: FUNCTION (dominant urban function) and FORM_TYPE (building height and layout). The real-city results should therefore be read as evidence of semantic enrichment, not exact matching.

H.1 Single Source Region

For the source region in Fig. 5, we compare the mass-weighted purity of its top-k targets against size-matched random subsets. We also compare against a row-shuffled control, which keeps every transport profile but assigns it to a different source region (Table 21).

Table 21: Semantic diagnostics for the high-mass target regions in Fig. 5. Purity gains are measured against size-matched random subsets. The row-shuffled control preserves each transport profile but breaks its association with the original source region.

Analysis	**Subset / diagnostic**	**Semantic consistency**	
		FUNCTION	**FORM_TYPE**
Top-k purity gain	Top-10	+10.2%	+20.9%
Top-k purity gain	Top-20	+13.6%	+22.1%
Top-k purity gain	Top-50	+26.9%	+10.4%
SCOT vs. row shuffle	**Semantic consistency**	**Same-FUNCTION mass:** +18.1%	**Weighted distance:** −8.4%

The top targets are purer than random subsets, and agreement drops after row shuffling. The selected targets thus depend on the source region rather than being reused globally.

H.2 Blocks of the Full Coupling

For the full XA→BJ coupling, we represent each source region by its outgoing profile $P_{i,:}$ and each target region by its normalized incoming profile $\widetilde{P}_{:,j} = P_{:,j}/\sum_i P_{ij}$. We build cosine kNN graphs over these profiles and cluster each side independently with spectral clustering ($k \in \{20, 50, 80\}$, $C \in \{8, 10, 12\}$, four seeds; 36 partitions per side). FUNCTION is used only for evaluation.

Within-side enrichment. The shuffled control keeps profiles, block sizes, and citywide FUNCTION frequencies, but breaks the region–block association (Table 22).

Table 22: FUNCTION structure of transport-profile blocks. Purity is averaged over all neighborhood sizes, cluster counts, and seeds. Representative enriched blocks use $k = 50$, $C = 10$, and seed 42. Higher purity and enrichment indicate stronger semantic structure.

Side	Average block purity			Representative enriched block				
	SCOT	Shuffled	Relative gain	Block	Size	Dominant category	Block share	Enrichment
Source	**0.370**	0.276	+34%	S3	107	Category 0	0.607	2.26×
Source				S4	45	Category 3	0.378	1.82×
Source				S2	96	Category 3	0.354	1.71×
Target	**0.398**	0.338	+18%	T1	49	Category 0	0.633	1.87×
Target				T8	161	Category 3	0.224	2.04×
Target				T0	302	Category 4	0.219	1.79×

Both sides are purer than the control. For a representative partition, none of 5,000 label permutations reaches the observed purity ($p < 0.001$); NMI and ARI show the same pattern. Enriched blocks include the less frequent Categories 3 and 4, so the blocks are functionally enriched rather than dominated by the majority category. Partitions are stable across seeds (ARI/NMI: $0.84/0.93$ source, $0.79/0.90$ target).

Cross-city block alignment. We aggregate the coupling into block-level mass $M_{ab} = \sum_{i \in S_a} \sum_{j \in T_b} P_{ij}$ and test whether mass concentrates between blocks with similar FUNCTION compositions. The null shuffles target FUNCTION labels with blocks fixed; a lift of one denotes the null expectation. Overlap compares which functions two blocks contain, while JS similarity also compares their proportions (Table 23).

Table 23: Cross-city semantic consistency over 36 independently extracted source–target partitions. Higher lift indicates that transport concentrates between blocks with more similar FUNCTION compositions.

Metric	Observed lift	Permuted null	p
FUNCTION distribution overlap	**1.089**	1.001 ± 0.010	< 0.001
FUNCTION JS similarity	**1.048**	1.000 ± 0.005	< 0.001
*Same-dominant-*FUNCTION *transport mass*	**63.1%**	**49.3%**	—

Both lifts are significantly above the null, and 63.1% of transport mass falls between blocks with the same dominant FUNCTION, versus 49.3% under permutation. Because source and target blocks are extracted independently, this agreement is not imposed by the clustering.

H.3 Synthetic Cities with Known Correspondence

Real cities lack ground-truth correspondence, so we use paired synthetic cities in which each region's latent type is known. We compare **SCOT** with marginal mixing (no selectivity), MMD+OT, and Non-aligned+OT (OT on embeddings trained without a cross-city objective). Lifts are normalized by the agreement expected from source and target type frequencies alone (lift $= 1$: no selectivity). *Fixed-type* lift matches type indices directly; *permutation-invariant* lift allows labels to be reordered across cities; *overlap* and *JS* lifts compare the full type mixtures of transported blocks.

SCOT places 33.1% more mass than expected between regions of the same type, rising to 49.4% when labels may be reordered. All OT-based controls stay near one, so the recovered correspondence comes from the **SCOT** alignment objective rather than from OT alone.

I Controlled Generality Across Graph Constructions and City Heterogeneity

Public urban data rarely offer consistent partitions, graphs, and labels across cities and modalities. We therefore use synthetic cities in which the latent regions, labels, and task are fixed, and vary only (i) the graph construction or (ii) the source–target heterogeneity.

Table 24: Recovery of ground-truth semantic correspondence in controlled synthetic cities. A lift of 1 indicates no selectivity beyond the independent marginal expectation. Red: best.

Method	Fixed-type lift	Permutation-invariant lift	Shared overlap lift	Shared JS lift
Marginal mixing	1.000	1.000	1.000	1.000
MMD+OT	0.969	1.040	0.964	0.963
Non-aligned+OT	0.980	1.007	0.980	0.981
SCOT (Ours)	1.331	1.494	1.317	1.330

Setup. We generate $K = 10$ latent region types with centers $\boldsymbol{\mu}_k \in \mathbb{R}^{16}$, 75% of which are shared by the two cities. Source features are $\mathbf{x}_i^s = \boldsymbol{\mu}_{c_i^s} + \boldsymbol{\epsilon}_i^s$, where c_i is the latent type of region i. Shared types are shifted in the target as $\boldsymbol{\mu}_k^t = \boldsymbol{\mu}_k \mathbf{R} + 0.45h\,\boldsymbol{\delta}_k$, where $\mathbf{R}$ is a rotation, $\boldsymbol{\delta}_k$ a random perturbation, and h the heterogeneity level. Labels follow $y_i = \boldsymbol{\mu}_{c_i}^\top \mathbf{w} + \alpha_{c_i} + \epsilon_i$. We use $n_s = 300$, $n_t = 450$, top-12 row-normalized graphs, 100 training epochs, and four seeds.

Graph constructions. From the same regions and labels, we build three graphs that differ only in the edge rule: mobility interaction, functional similarity, and spatial-functional.

Table 25: Target MAE under different graph constructions (mean $\pm$ std over four seeds). Lower is better. Gain is SCOT's relative improvement over the strongest baseline.

Graph construction	Non-align	MMD	SCOT	Gain
Mobility interaction	4.051 ± 3.228	1.289 ± 0.515	$\mathbf{1.198 \pm 0.189}$	$\mathbf{+7.1\%}$
Functional similarity	1.536 ± 0.654	1.222 ± 0.159	$\mathbf{1.156 \pm 0.167}$	$\mathbf{+5.4\%}$
Spatial-functional	1.659 ± 0.394	1.426 ± 0.126	$\mathbf{1.311 \pm 0.126}$	$\mathbf{+8.1\%}$

SCOT is best under all three graphs, so the alignment is not tied to mobility-derived edges.

Heterogeneity. With the functional-similarity graph fixed, we vary $h \in \{0.2, 0.5, 0.8\}$, which scales the target-side shift and noise ($h = 0.5$ is the setting above).

Table 26: Target MAE under increasing source–target heterogeneity (mean $\pm$ std over four seeds). Lower is better. Gain is SCOT's relative improvement over the strongest baseline.

h	Non-align	MMD	SCOT	Gain
0.2	1.415 ± 0.627	1.164 ± 0.247	$\mathbf{1.057 \pm 0.235}$	$\mathbf{+9.2\%}$
0.5	1.536 ± 0.654	1.222 ± 0.159	$\mathbf{1.156 \pm 0.167}$	$\mathbf{+5.4\%}$
0.8	1.511 ± 0.619	1.227 ± 0.071	$\mathbf{1.202 \pm 0.155}$	$\mathbf{+2.0\%}$

SCOT stays best at every level. Its MAE rises gradually from 1.057 to 1.202, while its margin over MMD narrows from 9.2% to 2.0% as the cities become less alike.

J Complexity Analysis

Table 27 lists the cost of computing the OT couplings with T Sinkhorn iterations; the contrastive term reuses the same matrices at no extra asymptotic cost. With M sources and cities of about n regions, the hub replaces M problems of size $n \times n$ with $M + 1$ problems of size $n \times K$, a speedup of $\frac{M}{M+1} \cdot \frac{n}{K}$ (about $20\times$ for $M = 2$, $n \approx 1000$, $K = 32$). The cycle term adds $O(n_s^2 n_t)$ in the single-source case and $O(n^2 K)$ per city with the hub.

Table 27: Cost of the OT couplings. n: city size; T: Sinkhorn iterations; M: number of sources; K: hub size.

Variant	Time	Memory
Single-source SCOT	$O(Tn_sn_t)$	$O(n_sn_t)$
Multi-source, pairwise OT	$O(TMn^2)$	$O(Mn^2)$
Multi-source SCOT (hub)	$O(T(M+1)nK)$	$O((M+1)nK)$

K Multi-Source vs. Best Single Source

Choosing the best single source normally requires target labels, which are scarce in our setting [Mansour et al., 2008, Sun et al., 2015, Pan and Yang, 2009]. SCOT avoids explicit selection by aggregating all sources through the hub. Table 28 compares it with an *oracle* that picks the best single source separately for each target, task, and metric.

Table 28: Multi-source SCOT vs. the best single-source SCOT (oracle, chosen per target, task, and metric). Red: better in each pair. Lower is better.

Target / Variant	GDP		Population		CO_2	
	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
▶ TARGET: BEIJING (BJ)						
Best single-source *(oracle)*	115.33	3.17	528.50	2.13	148.50	1.54
Multi-source (Ours)	104.16	2.57	525.10	1.87	143.53	1.16
▶ TARGET: XI'AN (XA)						
Best single-source *(oracle)*	154.92	1.60	452.67	1.58	128.74	1.24
Multi-source (Ours)	156.94	1.71	446.13	1.86	127.66	1.26
▶ TARGET: CHENGDU (CD)						
Best single-source *(oracle)*	135.63	3.55	575.43	2.35	114.68	7.83
Multi-source (Ours)	133.94	3.82	546.82	2.43	98.43	5.10

Multi-source SCOT beats the oracle on 8 of 9 MAE entries and on 4 of 9 MAPE entries. It wins every metric on Beijing, while its shortfalls lie on Xi'an and on Chengdu's GDP and population MAPE.

Label-free conflict signal. Table 29 reports, for each target, the gap Δ between the two sources on three hub statistics, all computed without target labels. Beijing has near-zero gaps, Chengdu intermediate ones, and Xi'an the largest on all three. This ordering matches where aggregation helps least, suggesting that source disagreement in the hub can flag cases where a single source may be preferable. Using this signal for conflict-aware gating is left to future work.

Table 29: Hub statistics near convergence for the two sources (S1, S2) of each target. Δ: absolute gap between sources; larger Δ indicates stronger source disagreement. No target labels are used.

Target	$\mathcal{L}_{\mathrm{OT}}$			$\mathcal{L}_{\mathrm{Con}}$			Transport mass		
	S1	S2	Δ	S1	S2	Δ	S1	S2	Δ
Beijing (BJ)	0.558	0.557	0.001	0.999	0.987	0.012	0.431	0.447	0.016
Chengdu (CD)	0.435	0.403	0.032	1.075	0.839	0.237	0.489	0.515	0.026
Xi'an (XA)	0.470	0.420	**0.050**	1.080	0.660	**0.420**	0.480	0.580	**0.100**

L Cross-Country Generalization: Transfer to and from New York City

L.1 Single-Source Transfer to and from New York City

We extend our evaluation to **New York City (NYC)**, a metropolis with substantially different urban form, mobility patterns, and POI distributions from the Chinese cities. We test both *transfer to NYC* from each Chinese source and *transfer from NYC* to each Chinese target (Table 30). Population prediction is the downstream task; we report MAE here.

(1) **SCOT wins on all six NYC-involving directions.** Relative MAE reductions over the strongest baseline range from 5.4% to 31.8%, showing that mass-controlled soft correspondence remains effective even when source and target diverge substantially in urban form and partition granularity.

(2) **Failure modes mirror the within-China patterns.** Distribution-matching (MMD, Adv) and anchor-based (RP, HBP, HSA) methods both struggle on these cross-country directions; CoRE remains the strongest non-SCOT method but is consistently outperformed. The failure modes SCOT addresses are thus not specific to any geographic context.

These results indicate that SCOT's gains are not idiosyncratic to within-China transfer or our specific partitions—the framework generalizes across distinct urban contexts regardless of NYC's role as source or target.

Table 30: Cross-country transfer with NYC (Population MAE; lower is better). SCOT wins on all six directions, with larger gains for transfer *to* NYC (21.0–31.8%) than *from* NYC (5.4–9.7%). Red: best; Blue: runner-up.

▶ Direction: Source → NYC				▶ Direction: NYC → Target			
Method	**Population MAE ↓**			**Method**	**Population MAE ↓**		
	BJ → NYC	CD → NYC	XA → NYC		NYC → BJ	NYC → CD	NYC → XA
Non-Alignment	468.84	502.65	531.28	Non-Alignment	1086.45	1038.38	934.76
RP	382.72	421.94	472.61	RP	952.24	934.93	842.21
HBP	294.10	317.57	386.93	HBP	843.73	859.02	754.40
HSA	354.46	389.83	418.72	HSA	874.36	862.91	772.58
MMD	331.55	372.64	356.38	MMD	821.49	812.82	724.33
Adv	421.29	456.72	489.16	Adv	1008.65	972.74	886.52
CrossTReS	346.91	381.27	369.85	CrossTReS	838.72	826.44	738.86
CoRE	238.82	285.36	251.54	CoRE	739.26	707.28	632.57
SCOT (Ours)	188.69	194.72	196.76	**SCOT (Ours)**	667.39	657.15	598.45
Δ *vs. best*	**+21.0%**	**+31.8%**	**+21.8%**	Δ *vs. best*	**+9.7%**	**+7.1%**	**+5.4%**

L.2 Scaling to Three Sources: Multi-Source Transfer to NYC

To verify that SCOT's hub scales beyond $M = 2$, we test the most challenging scenario: three Chinese cities (BJ, CD, XA) jointly transferring to NYC—a cross-country target with substantial divergence from all sources (Table 31).

(1) **Hub still outperforms all baselines.** SCOT-Hub achieves MAE 181.36, beating the strongest baseline (CoRE) by 21.0%. The margin over distribution-matching (MMD, Adv) and anchor-based (RP, HBP, HSA) methods is larger than at $M = 2$, suggesting hub coordination becomes *more valuable as M grows* and gradient conflict intensifies.

(2) **Hub beats best-single-source aggregation.** SCOT-Hub outperforms the best single-source variant (BJ→NYC: 188.69) by 3.9%, extracting complementary signal from three sources rather than collapsing to the dominant one—validating graceful scaling with M.

Table 31: Multi-source transfer to NYC with $M=3$ sources (BJ, CD, XA). Population MAE; lower is better. SCOT-Hub wins over both the strongest baseline (+21.0%) and the best single-source variant of SCOT (+3.9%), confirming that the hub design scales effectively beyond $M=2$. Red: best.

Method	BJ+CD+XA → NYC
Non-Alignment	487.36
RP	426.58
HBP	339.84
HSA	356.27
MMD	327.45
Adv	462.91
CrossTReS	318.76
CoRE	229.64
SCOT-Hub (Ours)	181.36
Δ *vs. best baseline (CoRE)*	**+21.0%**
Δ *vs. best single-source SCOT (BJ→NYC)*	**+3.9%**

L.3 Scaling to Three Sources: Adding NYC to Chinese Source Pools

We further test whether SCOT's hub design continues to extract complementary signal as a structurally divergent source (NYC) is added to the source pool. For each Chinese target, we compare multi-source SCOT at $M=2$ (two Chinese sources) against $M=3$ (two Chinese sources plus NYC), reporting Population MAE in Table 32.

(1) **Adding NYC as a third source improves over $M=2$.** On both targets, $M=3$ further reduces MAE relative to $M=2$ (-2.0% on BJ, -1.8% on XA), even though NYC is structurally divergent from Chinese cities. The hub design's ability to absorb heterogeneous sources without conflict prevents the dilution that naive aggregation typically suffers from when sources differ widely.
(2) **The gap over best-single-source widens.** Compared to the best single-source variant of SCOT for each target, $M=3$ achieves $+2.6\%$ on BJ and $+3.2\%$ on XA, exceeding the gains at $M=2$ ($+0.5\%$ and $+1.4\%$ respectively). This confirms that hub-based coordination becomes *increasingly valuable as more sources are added*, even when the additional source is heterogeneous.

Table 32: Multi-source scaling: SCOT-Hub at $M=3$ (adding NYC as a third source) vs. $M=2$ (Chinese sources only). Population MAE; lower is better. Adding a heterogeneous source (NYC) further improves performance, demonstrating that the hub absorbs additional sources without dilution. Red: best per target.

Variant	Population MAE ↓	
	Target: BJ	Target: XA
Best single-source SCOT	528.50	452.67
SCOT-Hub ($M=2$, Chinese sources)	525.10	446.13
SCOT-Hub ($M=3$, +NYC)	514.76	438.14
Δ *vs. best single-source*	**+2.6%**	**+3.2%**
Δ *vs. $M=2$ Hub*	**+2.0%**	**+1.8%**

M Limitations.

Although SCOT provides interpretable couplings and hub assignments, these diagnostics are not guarantees of causal correctness and should be used with domain knowledge in practical deployments. Besides, our experiments focus on mobility-derived region graphs and aggregated socioeconomic targets; extending the framework to finer-grained spatial resolutions or highly non-comparable urban modalities may require additional modeling assumptions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The scope and contributions of this work were included in the abstract and introduction, particularly highlighted at the end of the Introduction section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and future directions of the work were included in the Appendix section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Full assumptions and proof were included in the theory part and the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The proposed algorithms and implementation details of experiments were described in the Experiments Section and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The real data used in this work are publicly accessible and properly cited. We also include the code and descriptions in the supplementary materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The details were discussed in the Experiments section and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the Experiments section and Appendix, the error bars, standard deviation, and number of replicates, were reported and clearly stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed descriptions of the computational resources used for each experiment in the Appendix. Furthermore, our open-sourced code includes configuration files and parameter settings to replicate the experiments efficiently.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work of this paper is conducted with NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Societal impacts of the work were included in the Introduction section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risk.

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing assets were cited and credited throughout this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Documentation was provided with the supplementary code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not conduct research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not conduct research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methods presented in this research do not utilize Large LLMs as an essential, original, or non-standard component. LLMs were not used in any part of the model design, experimental evaluation, or optimization processes. Additionally, LLMs were not employed for data augmentation or pre-processing in our experiments.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.