SCOT: Multi-Source Cross-City Transfer with Optimal-Transport Soft-Correspondence Objectives

A Sinkhorn-based Alignment Framework for Urban Transfer Learning

Outline

Part I — Foundations

Part II — Extensions & Results

Motivation & Problem

Why Cross-City Transfer?

Urban computing needs region embeddings for:

- Regional GDP, population, CO_2 prediction
- Traffic / crowd flow forecasting
- Socio-economic analysis

The practical dilemma:

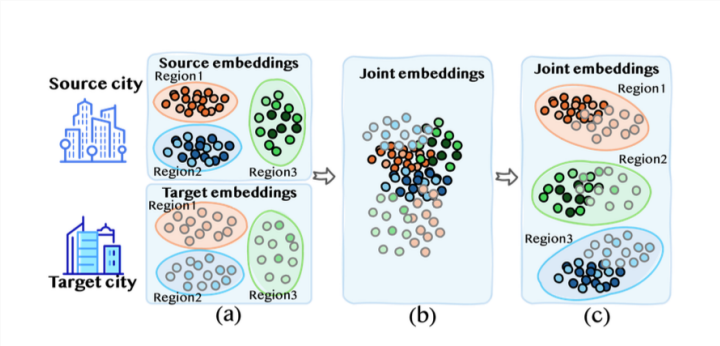
- Labels available only in a few *well-instrumented* cities
- Many cities remain label-scarce

$\Rightarrow$ Must **transfer** knowledge from labeled sources to a target

Why is this hard?

- Cities have **incompatible partitions**
- Unequal region counts: $n_s \neq n_t$
- No ground-truth region correspondences
- Regions are *nodes in mobility graphs*, not i.i.d. samples
- Only *part* of the semantics is transferable

The Central Bottleneck: *Alignment*



(a) Source and target embeddings come from *different* region sets

(b) Distribution matching (e.g., MMD) can **over-mix** clouds under heterogeneity

(c) We want **locally aligned yet structured** embeddings

Limitations of Existing Approaches

Distribution matching (MMD, Adv)

- Shrinks global distribution gaps
- But correspondences remain *implicit*
- Can over-mix under heterogeneity

Heuristic matching (RP, HBP, HSA)

- Anchor / nearest-neighbor pairing
- Sensitive to anchor choice
- Prone to hubness: many-to-one collapse

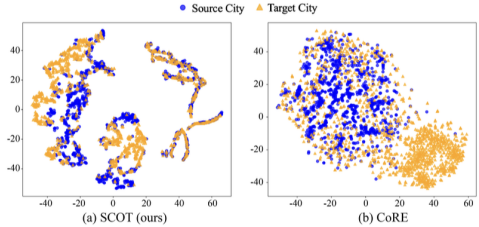


Figure 2: **t-SNE visualization for XA→BJ transfer.**

Fig. 2 — t-SNE for XA→BJ: CoRE (right) produces globally mixed embeddings that obscure functional structure.

Root cause: no *explicit, mass-controlled soft correspondence* between unequal region sets.

Problem Setup

Given two cities:

- Source $\mathcal{C}_s$ (labeled) with regions $V_s = \{1, \ldots, n_s\}$
- Target $\mathcal{C}_t$ (label-scarce) with regions $V_t = \{1, \ldots, n_t\}$, $n_s \neq n_t$

For each city we build:

- Undirected spatial adjacency graph $\mathrm{A}_s, \mathrm{A}_t$
- Directed mobility transition matrix from OD trips:

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

Goal: learn embeddings $\mathrm{z}_s \in \mathbb{R}^{n_s \times d}$, $\mathrm{z}_t \in \mathbb{R}^{n_t \times d}$ that

1. preserve *intra-city* mobility structure
2. are *comparable* across cities — without node correspondence

SCOT Framework

SCOT at a Glance

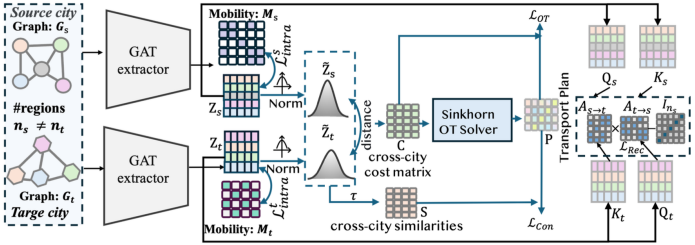


Figure 3: The pipeline of SCOT.

One-stage framework with three coordinated objectives:

1. **Intra-city** — GAT encoder + mobility-preserving softmax loss
2. **Cross-city alignment** — Sinkhorn entropic OT + OT-weighted contrastive
3. **Stabilization** — Cycle reconstruction regularizer

Intra-City Objective

Backbone: L-layer GAT over spatial adjacency $\quad \mathrm{z}_c = \mathrm{GAT}(\mathrm{A}_c),\ c \in \{s, t\}$

Mobility-preserving softmax models the destination distribution from origin i:

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

Negative log-likelihood weighted by true mobility M_c:

$$\mathcal{L}_{\mathrm{intra}} = - \sum_{c \in \{s,t\}} \sum_{i,j} (\mathrm{M}_c)_{ij} \log \hat{P}_{ij}^{(c)}$$

- Forces intra-city embeddings to encode *who flows where*
- Applied symmetrically in both cities

Sinkhorn-Based Soft Correspondence

Why Optimal Transport?

We need a soft region-to-region correspondence $\mathrm{P} \in \mathbb{R}_{+}^{n_s \times n_t}$ where P_{ij} = association between source region i and target region j.

Optimal Transport: find the minimum-cost coupling that moves mass between two point sets.

Entropic OT (Sinkhorn):

- **Marginal constraints** control how much mass each region sends/receives
  $\Rightarrow$ discourages many-to-one shortcuts
- Yields **structured many-to-many** correspondence
- Fast Sinkhorn iterations $\Rightarrow$ practical at urban scale
- Fully **differentiable** $\Rightarrow$ end-to-end training

Sinkhorn Soft Correspondence — Mechanics

Step 1. ℓ_2-normalize embeddings *(unit sphere)*

$$\tilde{z}_i^s = \frac{z_i^s}{\|z_i^s\|_2}, \qquad \tilde{z}_j^t = \frac{z_j^t}{\|z_j^t\|_2}$$

Step 2. Cross-city cost matrix — *directional similarity, not scale*

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

Step 3. Gibbs kernel + Sinkhorn scaling *(T iterations)*

$$K = \exp(-C/\varepsilon), \quad u^{(k+1)} = 1 \oslash (Kv^{(k)}), \quad v^{(k+1)} = 1 \oslash (K^\top u^{(k+1)})$$

Step 4. Soft matching matrix $\Longrightarrow$ OT alignment loss

$$\boxed{P = \mathrm{diag}(u^{(T)})\, K\, \mathrm{diag}(v^{(T)})} \quad \Longrightarrow \quad \mathcal{L}_{\mathrm{OT}} = \frac{1}{\min(n_s, n_t)} \sum_{i,j} P_{ij}\, C_{ij}$$

$\varepsilon = 0.15$ *balances sharpness and diffusion — too small → noisy coupling; too large → mushy alignment.*

OT-Weighted Contrastive Learning

From Geometric to Semantic Alignment

Issue: minimizing $\mathcal{L}_{\mathrm{OT}}$ enforces *geometric closeness* but does not guarantee discriminative embeddings.

Idea: use the Sinkhorn coupling P_{ij} as a **soft positive weight** in a contrastive loss.

$$S_{ij} = \frac{\tilde{z}_i^{s\top} \tilde{z}_j^{t}}{\tau} \qquad \text{(cross-city similarity, temperature } \tau\text{)}$$

Sinkhorn-weighted contrastive loss:

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

- Pulls source region i toward target regions with high transported mass
- Pushes i away from unmatched targets
- Concentrates similarity on *transport-supported* pairs — no brittle NN matches

Combined alignment loss: $\mathcal{L}_{\mathrm{Align}} = \mathcal{L}_{\mathrm{OT}} + \eta \, \mathcal{L}_{\mathrm{Con}}$

Theoretical Guarantee (Theorem 3.1)

Target MAE bound via OT-weighted contrastive alignment
For unit embeddings, couplings P with marginals a, b, and Lipschitz regressors g, h:

$$\mathcal{R}_t^b(h) \;\le\; \mathcal{R}_s^a(h) \;+\; (L_h + L_g)\sqrt{2 - 2m}$$

where

$$m = \max\{-1,\; \tau \log n_t + \tau H(a) - \tau\, \mathcal{L}_{\mathrm{Con}}(P) - 1 - \tfrac{1}{2\tau}\}$$

Interpretation:

- Target MAE $\le$ source MAE + a transfer gap
- Smaller $\mathcal{L}_{\mathrm{Con}} \Rightarrow$ larger $m \Rightarrow$ **tighter bound**
- Explicitly links our alignment objective to downstream error

(Empirically, $\mathcal{L}_{\mathrm{Con}}$ shows standardized coefficient 0.77 and partial correlation 0.95 with target error — Table 9.)

Cycle Reconstruction Regularizer

Stabilizing Training: One-Sided Cycle Reconstruction

Motivation: a source region mapped to the target should be *recoverable* from its matched target counterpart.

Cross-attention maps (shared W_q, W_k):

$$A_{s\to t} = \mathrm{softmax}\Big(\frac{Q_s K_t^\top}{\sqrt{d}}\Big), \qquad A_{t\to s} = \mathrm{softmax}\Big(\frac{Q_t K_s^\top}{\sqrt{d}}\Big)$$

One-sided cycle loss (handles rectangular $n_s \neq n_t$):

$$\mathcal{L}_{\mathrm{cyc}} = \|A_{s\to t} A_{t\to s} - I_{n_s}\|_F^2$$

Entropy penalty to avoid overly diffuse attention:

$$\mathcal{R}_{\mathrm{ent}} = -\frac{1}{n_s} \sum_{i,j} A_{s\to t}(i,j) \log\big(A_{s\to t}(i,j) + \delta\big)$$

Final stabilizer: $\mathcal{L}_{\mathrm{Rec}} = \mathcal{L}_{\mathrm{cyc}} + \beta\, \mathcal{R}_{\mathrm{ent}}$

Total Loss for Single-Source SCOT

End-to-end joint objective

$$\mathcal{L}_{\text{Total}} = \underbrace{\mathcal{L}_{\text{intra}}^{(s)} + \mathcal{L}_{\text{intra}}^{(t)}}_{\text{intra-city}} + \lambda_{\text{align}} \underbrace{\mathcal{L}_{\text{Align}}}_{\text{cross-city}} + \lambda_{\text{rec}} \underbrace{\mathcal{L}_{\text{Rec}}}_{\text{stabilization}}$$

Three roles, one optimization:

- Intra-city: encode *within*-city structure
- Cross-city: establish *between*-city soft correspondence (capacity-controlled)
- Stabilization: regularize training and discourage degenerate attention

Default hyperparameters: $\lambda_{\text{align}} = 1$, $\lambda_{\text{rec}} = 0.5$, $\eta = 0.5$, $\beta = 0.05$, $\tau = 0.1$, $\varepsilon = 0.15$, $T = 30$ Sinkhorn steps.

Multi-Source Hub Alignment

Multi-Source Transfer is Harder

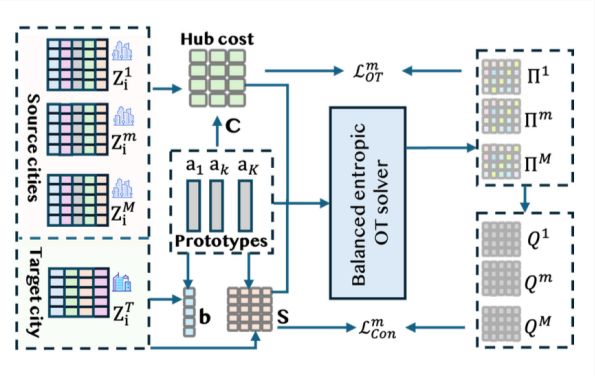
New difficulties beyond single-source:

- Different sources may induce conflicting correspondences to the same target
- Independent pairwise alignments can be dominated by one source
- Gradients from multiple sources can interfere

SCOT's answer: a shared *semantic hub*

- Introduce K **learnable prototypes** $\{a_k\}_{k=1}^{K}$ as a common anchor space
- Align *all* cities (sources & target) to the hub via **balanced entropic OT**
- Use a **target-induced prototype prior** to emphasize target-relevant semantics

Multi-Source Pipeline



For each city $m \in \mathcal{S} \cup \{t\}$**:**

- Hub cost $C_{ik}^m = \|\tilde{z}_i^m - \tilde{a}_k\|_2$
- Balanced entropic OT Π^m with marginals (a^m, b)
- Row-normalize $\Pi^m \to Q^m \Rightarrow$ soft positive weights for contrastive loss

Target-Induced Prototype Prior

Idea: let the target city *tell us* which prototypes matter.

Step 1. Aggregate target–prototype similarity

$$\bar{s}_k = \frac{1}{n_t} \sum_{j=1}^{n_t} \tilde{z}_j^{t\top} \tilde{a}_k$$

Step 2. Temperature-softmax with probability floor

$$b_k \;\propto\; \max\!\big\{ \exp(\bar{s}_k/\tau_b),\, \epsilon_b \big\}, \qquad \sum_k b_k = 1$$

- Prevents dead prototypes (via floor ϵ_b)
- Emphasizes target-relevant semantics during multi-source aggregation
- Avoids source domination: no source can pull the hub away from the target

Default: $K = 32$, $\tau_b = 0.5$, $\epsilon_b = 10^{-3}$

Multi-Source Objective

For each city $m \in \mathcal{S} \cup \{t\}$:

$$\mathcal{L}_{\mathrm{OT}}^{m} = \langle \Pi^{m}, \mathsf{C}^{m} \rangle$$

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

$$\mathcal{L}_{\mathrm{align}}^{m} = \mathcal{L}_{\mathrm{OT}}^{m} + \lambda_c \, \mathcal{L}_{\mathrm{Con}}^{m}$$

Final training objective:

$$\mathcal{L} = \sum_{m} \mathcal{L}_{\mathrm{intra}}^{m} \; + \; \lambda_{\mathrm{align}} \cdot \frac{1}{|\mathcal{S}|+1} \sum_{m} \mathcal{L}_{\mathrm{align}}^{m} \; + \; \lambda_{\mathrm{rec}} \cdot \frac{1}{|\mathcal{S}|+1} \sum_{m} \mathcal{L}_{\mathrm{rec}}^{m}$$

Complexity: $O(TK \sum_m n_m)$ — far cheaper than pairwise $O(T \sum_m n_m n_t)$ when $K \ll n_t$.

Experiments

Experimental Setup

Datasets — three Chinese cities:

City	# Regions	# Trips	Targets
Xi'an (XA)	1,306	559,729	GDP / Pop / CO_2
Chengdu (CD)	1,056	384,618	GDP / Pop / CO_2
Beijing (BJ)	1,311	78,945	GDP / Pop / CO_2

Evaluation protocol:

- All ordered city pairs: BJ↔XA, BJ↔CD, XA↔CD
- Fit ridge regressor on source labels, apply directly to target embeddings
- Report **MAE** and **MAPE** (lower is better)

Baselines:

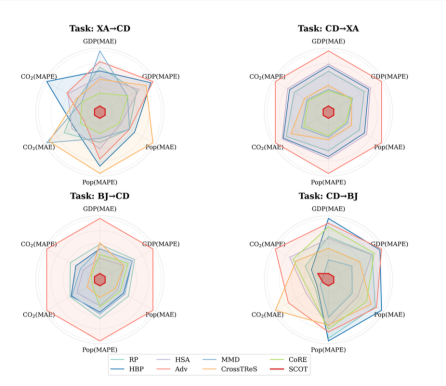
- **Heuristic correspondence**: RP, HBP, HSA
- **Distribution / relational**: MMD, Adv (DANN), CrossTReS, CoRE

Single-Source Results: XA $\leftrightarrow$ BJ

	XA $\rightarrow$ BJ						BJ $\rightarrow$ XA					
	GDP		Pop		CO_2		GDP		Pop		CO_2	
Method	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Non-Align	264.3	12.08	981.1	8.56	288.4	6.40	252.9	5.84	946.9	8.53	270.4	8.66
RP	189.8	9.07	684.5	6.55	196.1	4.70	181.1	4.58	670.4	6.46	191.9	6.42
HBP	177.7	8.40	665.5	5.95	188.7	4.18	196.2	3.10	627.5	4.37	176.4	4.25
HSA	201.3	6.83	619.3	4.00	176.4	3.20	188.4	2.11	636.3	5.03	182.9	5.02
MMD	183.3	5.71	588.3	3.17	165.7	2.54	180.7	1.99	499.9	1.85	141.6	1.83
Adv	192.6	8.72	702.2	6.78	199.2	4.83	199.2	6.32	805.0	9.16	203.3	7.21
CrossTReS	207.4	7.39	633.3	4.42	179.8	3.50	170.3	4.23	639.4	5.62	182.7	5.55
CoRE	157.8	5.46	611.2	4.05	166.3	2.95	162.2	1.91	547.7	2.17	153.6	2.09
SCOT (Ours)	115.3	3.17	528.5	2.13	149.4	1.79	154.9	1.60	452.7	1.58	128.7	1.63
Gain(%)	+26.9	+41.9	+10.2	+32.8	+9.8	+29.5	+4.5	+15.7	+9.5	+14.6	+9.1	+10.9

SCOT wins on *all* 12 cells, with especially large gains on GDP MAPE (+42%) and Population MAPE (+33%).

Radar View: All City Pairs, All Tasks



The SCOT polygon (red) is *consistently the smallest* across all four transfer directions and all six metrics — uniform improvements, not a task-specific artifact.

Multi-Source Results

	Target: BJ						Target: XA					
	GDP		Pop		CO_2		GDP		Pop		CO_2	
Method	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
RP	172.8	7.64	679.9	6.98	166.8	2.69	195.9	2.89	642.0	3.67	181.2	2.88
HBP	164.3	7.13	662.6	6.53	165.5	2.57	200.0	4.24	670.4	3.80	150.2	2.09
HSA	156.4	6.62	644.9	5.53	160.1	2.59	183.6	2.42	648.7	3.79	155.2	2.45
MMD	127.5	4.93	605.8	4.11	160.5	1.29	163.8	2.18	506.2	3.05	144.6	3.18
Adv	196.8	9.90	718.0	6.34	189.4	3.19	221.3	4.84	731.9	5.43	184.7	3.46
CrossTReS	151.2	6.41	666.7	5.03	187.6	2.32	179.0	4.94	625.6	5.52	151.2	3.48
CoRE	152.9	5.86	620.3	4.30	152.2	1.99	173.7	5.48	549.4	3.89	134.2	1.97
SCOT	**104.2**	**2.57**	**525.1**	**1.87**	**143.5**	**1.16**	**156.9**	**1.71**	**446.1**	**1.86**	**127.7**	**1.26**
Gain(%)	+18.3	+47.9	+13.3	+54.5	+5.7	+10.1	+4.2	+21.6	+11.9	+39.0	+4.9	+36.0

SCOT wins all 12 multi-source cells — the *shared hub* successfully aggregates complementary signals without source domination.

Single vs. Multi-Source SCOT

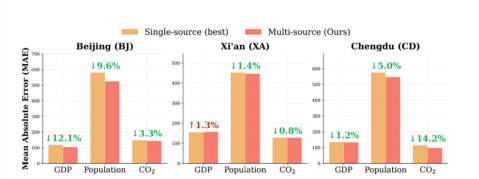


Figure 7: Best single-source SCOT (orange) vs. multi-source SCOT (red). Labels show relative MAE change Δ (green: improvement; red: degradation).

Multi-source SCOT beats the *best single-source* SCOT on most target-task pairs:

- BJ target: -19.6% on GDP, -12.1% on Pop
- CD target: -1.2% / -14.2% on GDP/CO_2
- XA: mostly flat (mild GDP trade-off from source conflict — see Appendix J)

$\Rightarrow$ Gains come from *complementary* signals across cities, not a single closest source.

Ablation & Diagnostics

Ablation: All Three Components Matter

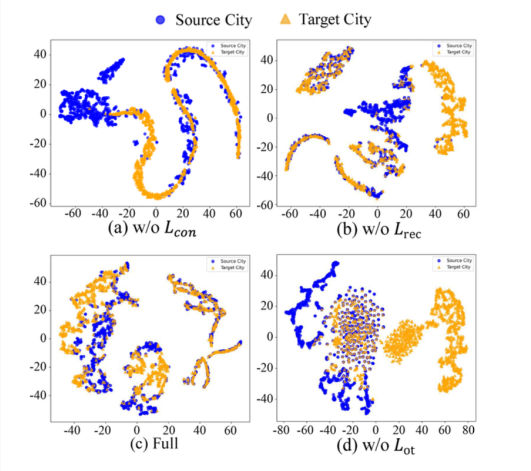


Fig. 6 — t-SNE for XA→BJ under each ablation

w/o $\mathcal{L}_{\mathrm{Con}}$

Embeddings remain largely **city-specific** — limited mixing.

w/o $\mathcal{L}_{\mathrm{Rec}}$

Training becomes **less stable** — noisier optimization.

w/o $\mathcal{L}_{\mathrm{OT}}$

Target-side branches persist — unresolved mismatches.

Alignment Diagnostics

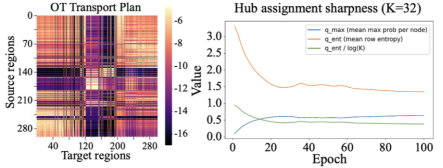


Figure 8: Diagnostics: (left) entropic OT coupling(XA→BJ, epoch 100), subsample after reordering; (right) hub assignment sharpness for $K = 32$ ($q_{\max}$, q_{ent}, and $q_{\text{ent}}/\log K$).

OT coupling (left):

- Clear block structure after barycentric reordering
- Selective many-to-many correspondences
- Limited hubness

Hub sharpness (right), $K = 32$**:**

- $q_{\max}$ rises, $q_{\text{ent}}/\log K \approx 0.4$
- $\exp(q_{\text{ent}}) \approx 4$ active prototypes per region
- Specialization, not pooling

Hyperparameter Sensitivity: Robust Across Broad Ranges

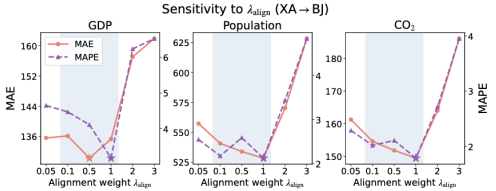


Figure 9: Sensitivity to λ_{align} (XA→BJ).

Sensitivity to λ_{align}

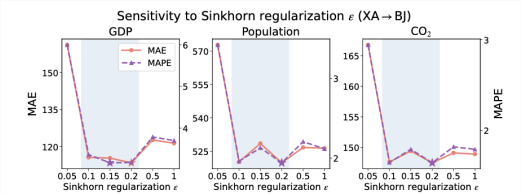


Figure 10: Sensitivity to Sinkhorn regularization ε (XA→BJ).

Sensitivity to Sinkhorn ε

Stable sweet spots:

- $\lambda_{\text{align}} \in [0.1, 1]$ — over-alignment only above 2
- $\varepsilon \in [0.1, 0.2]$ — too sharp below 0.05, too diffuse above 0.5
- $\tau \in [0.1, 0.5]$, hub size $K \in [4, 32]$

$\Rightarrow$ A *single globally fixed* configuration works across all city pairs.

Robustness to Backbone and Downstream Head

Are the gains from alignment design or from the encoder/evaluator?

Swap the GNN backbone (BJ→XA):

Backbone	GDP MAE	Pop MAE
GAT	160.2	450.1
GATv2	162.6	455.4
SuperGAT	164.4	461.5

Swap the downstream regressor:

Readout	GDP MAE	Pop MAE
Ridge	160.2	450.1
Lasso	158.7	455.7
Linear SVR	162.2	456.4
Elastic Net	164.6	459.5

Performance is essentially flat across both dimensions.

Gains are attributable to the *alignment design*, not to the encoder or evaluator.

Conclusion

Takeaways — Contributions

Key Contributions

1. Problem framing

Identified **explicit soft correspondence under unequal partitions** as the central bottleneck of cross-city transfer.

2. SCOT framework

Sinkhorn entropic OT + OT-weighted contrastive + cycle regularizer, jointly controlling *capacity*, *discriminability*, and *stability*.

3. Multi-source hub

Shared prototypes with a **target-induced prior** prevent source domination and conflicting gradients.

4. Empirical wins

Consistent gains on **GDP** / **Pop** / **CO_2** across BJ, XA, CD; robust to backbones and regressors.

Takeaways — Insights & Outlook

Why it works

Every component directly addresses an **alignment failure mode**:

- **Transport** $\longrightarrow$ capacity control, no many-to-one collapse
- **Contrastive** $\longrightarrow$ sharpen transport-supported pairs
- **Hub** $\longrightarrow$ aggregate complementary sources without domination

Future work

- **Uncertainty-aware** source weighting
- Severe cross-city **heterogeneity**
- **Finer** spatial resolutions
- Non-comparable urban modalities