An Empirical Evaluation of Cross-City POI Recommendation on a Large-Scale Benchmark

Peibo Li
University of New South Wales
Sydney, Australia
peibo.li@student.unsw.edu.au

Yang Song
University of New South Wales
Sydney, Australia
yang.song1@unsw.edu.au

Hao Xue
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
haoxue@hkust-gz.edu.cn

Maarten de Rijke
University of Amsterdam
The Netherlands
m.derijke@uva.nl

Flora D. Salim
University of New South Wales
Sydney, Australia
flora.salim@unsw.edu.au

Abstract

Cross-city point-of-interest (POI) recommendation is crucial for navigating unfamiliar urban environments, yet its progress has historically been constrained by data limitations. Using the recently proposed large-scale benchmark Trip World [2], we empirically re-examine whether conclusions drawn on small prior benchmarks still hold under worldwide coverage, low home–destination region overlap, and large, semantically rich POI inventories. Our evaluation surfaces three bottlenecks of representative state-of-the-art methods: (1) hometown-aware models appear to rely more on destination-region priors than on user-specific preference transfer; (2) their accuracy–efficiency trade-off degrades at this scale, where the simplest model is among the strongest; and (3) existing mechanisms for integrating semantic metadata yield little benefit. We further include a diagnostic pilot on agentic methods adapted from next-POI recommendation, finding that naive adaptation trails a simple popularity prior even though the relevant semantic signal is present in the data. These results highlight the need for task-specific designs that support cross-city preference transfer, semantic grounding, and scalable reasoning over unseen destination inventories.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Point-of-interest recommendation

ACM Reference Format:
Peibo Li, Yang Song, Hao Xue, Maarten de Rijke, and Flora D. Salim. 2026. An Empirical Evaluation of Cross-City POI Recommendation on a Large-Scale Benchmark. In *The 34th ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '26), November 03–06, 2026, Riverside, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3841645.3843403



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGSPATIAL '26, Riverside, CA, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2950-8/2026/11
https://doi.org/10.1145/3841645.3843403

1 Introduction

Point-of-interest (POI) recommendation underpins location-based services, helping users discover relevant places [21, 22]. Most work targets the in-city or next-POI setting, where users repeatedly interact with POIs in their home region [9, 16]. Many real travel scenarios break this assumption: a traveler arriving in an unfamiliar city has no prior interactions with the destination POIs, only a short and sparse trajectory, and preferences that shift from everyday routines to tourism. Cross-city POI recommendation therefore requires transferring user preferences from a home city to a destination with a largely unseen POI inventory [3, 9, 14], while contending with habitual histories that do not translate to travel behavior, geographically and semantically disjoint POIs, and limited sequential evidence [10, 17].

Progress on this setting has been limited by the lack of large-scale, realistic, and semantically rich benchmarks. Existing datasets are designed for single-city or in-city next-POI prediction [1, 15], or are small, geographically narrow, and metadata-poor [16, 18, 19]. Some provide only coordinates and coarse categories, while others offer rich reviews but rely on review-derived interactions that need not correspond to real visit sequences. These limitations make it hard to evaluate whether a method generalizes across cities, exploits semantic POI information, or handles the extreme cold start travelers face.

We conduct our evaluation on Trip World [2], a recently proposed large-scale benchmark that provides worldwide out-of-town check-ins together with enriched POI metadata and user reviews, to rigorously study cross-city POI recommendation. Its scale and coverage support a unified comparison of classical, neural, and recent agentic methods, and enable testing whether models leverage mobility history, destination context, POI semantics, and travel-specific behavior beyond simple proximity or popularity signals. Our contributions are:

- We evaluate representative cross-city POI recommendation methods on Trip World, spanning popularity-based, sequential, neural, semantic-aware, and large language model (LLM)-agentic approaches.
- We surface three bottlenecks: hometown-aware methods appear to rely more on destination-region priors than on transferable user preferences; the simplest model is among

Table 1: Statistics of Trip World after processing. CIs: check-ins; Avg $|c_o|$: mean number of POIs per out-of-town trip.

Dataset	#Users	#POIs	#Regions	#Home CIs	#Travel CIs	Avg $\lvert c_o \rvert$
Trip World	63,896	336,102	890	6,242,544	576,504	4.97

the strongest while the most compositional ranks lowest at much higher cost; and existing semantic-integration mechanisms yield little benefit.
- As a diagnostic pilot, we benchmark agentic methods adapted from next-POI recommendation and show that naive transfer is insufficient for cross-city-native designs.

2 Experiments

2.1 Experimental setup

2.1.1 Task. Given a user's hometown check-in history H_u and an out-of-town trip specified by its origin o, destination d, target region r, and length L, the task is to predict the ordered intermediate POIs $c_o[1{:}{-}1]$ visited between o and d.

2.1.2 Baselines. We benchmark 7 baselines: 4 trip recommendation methods without hometown information (Popularity [1], GraphTrip [6], MatTrip [20], AR-Trip [13]) and 3 with hometown information (KDDC [11], PPROC [7], SPOT-Trip [10]), where KDDC and PPROC use the SPOT-Trip backbone. The **Popularity** baseline ranks POIs by their global training-set visit frequency.

2.1.3 Evaluation metric. We adopt F1 and Pairs-F1, following existing work [6, 8, 10, 13]: F1 measures the overlap between predicted and ground-truth intermediate POIs, while Pairs-F1 additionally rewards correct visiting order. Both are computed over the intermediate POIs only, excluding the fixed origin and destination.

2.1.4 Dataset processing. We follow [10] to process the dataset. We filter out users with fewer than three out-of-town check-in records, and exclude out-of-town trips whose travel durations are shorter than 1 hour or longer than 30 days. We partition by user into training, validation, and testing sets with an 80%/10%/10% ratio. The dataset stats are shown in Table 1.

2.2 Results and Discussion

We report baseline performance in Table 2. Using the published Foursquare/Yelp results as a contextual reference rather than a controlled re-run in our pipeline, we observe that the relative ranking on Trip World differs: there, the simplest model (MatTrip) is among the strongest while the most compositional, hometown-aware SPOT-Trip ranks lowest. We examine possible explanations below.

2.2.1 Transferable preferences, or region priors? On Trip World, methods that use hometown information do not outperform those relying solely on out-of-town signals, unlike on Foursquare and Yelp [10]. The standard framing casts these methods as learning user preferences from home history and transferring them to the destination. However, as shown in Table 3, only 30.9% of destination regions in Trip World also appear as home regions for other users, far below Foursquare and Yelp. This pattern suggests that, in this low-overlap regime, the methods rely more on region-conditional POI distributions (effectively per-region popularity priors) than on user preferences that transfer across regions. We test this directly on SPOT-Trip, the most complete hometown-aware model (KDDC and PPROC share its backbone): at test time we replace each trip's hometown history with that of a random other user, holding the trained model and decoder fixed. Across three seeds, this shuffling leaves F1 unchanged (relative change <0.55%, within the cross-seed standard deviation), so the model is invariant to *whose* home it sees and does not exploit user-specific hometown preference. We benchmark the most complete model; extending the ablation to every architecture is left to future work.

2.2.2 Accuracy–efficiency trade-off at scale. Trip World is an order of magnitude larger than previous benchmarks, raising the question of how the baselines behave at this scale. MatTrip, the simplest model (a Bi-LSTM with Bahdanau attention), attains the best accuracy, while the most compositional model, SPOT-Trip (a knowledge graph with an ODE decoder), ranks lowest while being the costliest to train (Table 4). This is consistent with the additional inductive biases adding variance rather than reducing it in a regime where the per-POI gradient signal is sparse (336,102 POIs across 890 regions). We cannot fully separate architecture from tuning or implementation here, and leave a controlled scaling study (e.g., training-size sweeps) to future work; nonetheless, the architectures that pay the most overhead also score lowest.

2.2.3 Integrating semantic metadata. Trip World provides rich semantic features. We probe one common integration mechanism: *SPOT-Trip full* adds a knowledge-graph relation encoding each POI's average rating, while *SPOT-Trip no-review* omits it. Counterintuitively, *SPOT-Trip no-review* outperforms *SPOT-Trip full*, indicating that this particular mechanism, a single review-derived rating relation, does not help. We stress that this probes *one* mechanism rather than semantic metadata in general; the richer review text and user-taste signals are examined in our agentic pilot (Section 2.3). Existing methods were not designed for such features, as they were unavailable in prior benchmarks, and we hope Trip World's semantics motivate models that can harness them.

2.3 Can Agentic Methods Transfer to Cross-City Recommendation?

As a diagnostic pilot, we test whether LLM-based *agentic* methods, which reason over textual context and natively consume semantic metadata, alleviate the bottlenecks above. We adapt two representative methods from next-point-of-interest (POI) prediction: LLM-Move [5], a direct LLM ranking prompt, and AgentMove [4], a full agent with memory and world-model reasoning, both backed by Claude Haiku 4.5 (`claude-haiku-4-5`).

Adaptation and setup. Open-vocabulary generation over the 336,102-POI inventory is infeasible for an LLM, a first finding in itself. We thus recast the task as *in-candidate* selection: each query receives a menu of 100 candidates guaranteed to contain the ground-truth intermediates, which the agent selects and orders. Negatives are sampled to *approximately* match ground-truth popularity (residual GT-vs-negative popularity AUC ≈0.55), so popularity cannot trivially separate the answer; the *in-candidate Popularity* baseline ranks candidates by leave-one-out training frequency. Because this

Table 2: Out-of-town trip recommendation performance, measured by F1 and Pairs-F1 over the intermediate POIs (higher is better, ↑). Foursquare and Yelp numbers are reproduced from Liu et al. [10] (single published values), *not* re-run in our pipeline; Trip World numbers are averaged over 3 random seeds (per-seed std ≤0.003 for F1 and ≤0.0013 for Pairs-F1, so the neural baselines differ by within roughly one standard deviation). Bold marks the best value per column. *SPOT-Trip no-review* is a Trip-World-specific ablation that removes the review-rating relation and has no counterpart in the original benchmark (—).

	Foursquare		Yelp		Trip World	
Architecture	F1 ↑	Pairs-F1 ↑	F1 ↑	Pairs-F1 ↑	F1 ↑	Pairs-F1 ↑
Popularity [1]	0.0261	0.0013	0.0257	0.0056	0.0222 ± 0.0012	0.0057 ± 0.0005
MatTrip [20]	0.0311	0.0037	0.0301	0.0119	**0.0601** ± 0.0019	**0.0166** ± 0.0010
GraphTrip [6]	0.0295	0.0048	0.0289	0.0126	0.0582 ± 0.0017	0.0164 ± 0.0005
AR-Trip [13]	0.0304	0.0045	0.0307	0.0153	0.0562 ± 0.0028	0.0151 ± 0.0013
KDDC [11]	0.0375	0.0079	0.0341	0.0156	0.0575 ± 0.0005	0.0159 ± 0.0001
PPROC [7]	0.0330	0.0071	0.0334	0.0159	0.0554 ± 0.0008	0.0148 ± 0.0008
SPOT-Trip no-review [10]	—	—	—	—	0.0572 ± 0.0004	0.0159 ± 0.0006
SPOT-Trip full [10]	**0.0400**	**0.0109**	**0.0399**	**0.0190**	0.0557 ± 0.0013	0.0151 ± 0.0010

Table 3: Region overlap between home regions ($\mathcal{R}_h$) and destination regions ($\mathcal{R}_o$) on the three benchmarks. *Dest-coverage* is the fraction of destination regions that also appear as a home region. *Dest-only regions* count regions that appear only as destinations and never as homes.

Dataset	Dest-coverage $\lvert\mathcal{R}_h \cap \mathcal{R}_o\rvert / \lvert\mathcal{R}_o\rvert$	Dest-only regions
Foursquare	95.2%	1
Yelp	69.9%	56
Trip World	30.9%	603

Table 4: Per-epoch training time on Trip World (one H200 GPU).

Architecture	s/epoch	Relative
MatTrip	58	1.0×
GraphTrip	178	3.1×
KDDC	137	2.4×
AR-Trip	304	5.2×
PPROC	330	5.7×
SPOT-Trip full	422	7.3×

restriction raises the chance level, absolute F1 here is *not* comparable to Table 2; the fair references are the in-candidate Popularity and Random baselines. Beyond trajectory context, we supply two review-grounded signals. The *review summary* is a short LLM-written description of what a POI is known for and whom it suits, distilled from its reviews and attached to every candidate. The *user-taste profile* is a concise LLM-written description of a user's preferred cuisines, venue types, vibe, and price level, distilled from the review summaries of the POIs the user visited at home, and prepended to the prompt.

Naive transfer underperforms a popularity prior. Table 5 reports the in-candidate results on data-rich destinations. Every agentic variant trails the simple in-candidate Popularity prior by a wide margin (e.g., 0.0718 vs. 0.1069 F1). Echoing the scaling result for neural baselines, the more elaborate AgentMove underperforms the lighter LLMMove prompt, indicating that additional agentic machinery again adds variance rather than signal. Strikingly, injecting the candidates' popularity into the prompt (LLMMove, + popularity) helps more than any semantic signal, yet still does not close the gap to the popularity baseline. Naive transfer of agentic methods is therefore *insufficient* for cross-city POI recommendation.

Table 5: Agentic methods on cross-city POI recommendation (*in-candidate* setting, data-rich destinations, Claude Haiku 4.5). Values are not comparable to Table 2; the references are the in-candidate Popularity and Random baselines. Bold marks the best value per column.

Method	F1 ↑	Pairs-F1 ↑
Random (in-candidate)	0.0361	0.0040
AgentMove [4], no semantics	0.0627	0.0102
AgentMove [4], + semantics	0.0658	0.0116
LLMMove [5], + semantics	0.0718	0.0146
LLMMove [5], + popularity	0.0820	0.0189
Popularity (in-candidate)	**0.1069**	**0.0234**

Table 6: Disentangling the two semantic signals on *cold-start* destinations (≤ 30 training trips), LLMMove with Claude Haiku 4.5 (in-candidate). Removing either signal *improves* F1; the variant with no semantics is best among the LLMs, and all trail the Popularity prior.

Variant	F1 ↑	Pairs-F1 ↑
Random (in-candidate)	0.0324	0.0089
+ taste + review	0.0490	0.0083
+ taste −review	0.0536	0.0097
−taste + review	0.0524	0.0125
−taste −review	0.0591	0.0144
Popularity (in-candidate)	**0.0686**	**0.0176**

Semantics are not exploited, but the signal is present. To probe the semantic bottleneck, we disentangle the user-taste and review signals on the cold-start slice (destinations with ≤ 30 training trips), where transfer matters most. Table 6 shows that *both* signals independently hurt: removing either improves F1, and the variant with no semantics is the strongest LLM configuration. This is not a data-quality artifact, cold-start POI summaries are in fact well sourced (a median of 342 reviews per summarized POI). Rather, the agent cannot convert the metadata into correct selections and is instead distracted away from the popularity prior. Crucially, the information is *present*: the ground-truth intermediates are measurably

aligned with the user's hometown taste, matching the user's home POI categories 40.6% of the time versus 25.3% for the popularity-matched negatives. The limitation is therefore one of *method*, not of available signal.

Where the headroom lies. The dominance of popularity is itself partly an artifact of the benchmark and metric. Popularity still wins overall only because the candidate matching is approximate and most queries have popular ground truth: stratifying by ground-truth popularity, the prior is strong where the answer is itself popular, but on the long-tail third of queries it falls to near the Random floor, where the LLM agents *exceed* it. This is where semantic reasoning should help, pointing to directions for cross-city-native methods: (i) anchor on the popularity prior and use semantics only to *refine* among comparable candidates, rather than letting free-form generation override a reliable prior; (ii) align POI semantics to the user-taste vocabulary, since current review summaries describe service sentiment largely orthogonal to the preferences that distinguish the ground truth; and (iii) develop scalable retrieval and grounding to relax the in-candidate restriction toward open-vocabulary cross-city recommendation. These mirror the three bottlenecks found for neural baselines, preference transfer, semantic grounding, and scale.

3 Conclusion

Using Trip World, a large-scale, worldwide, and semantically rich benchmark, we re-examined state-of-the-art methods for cross-city POI recommendation. Our evaluation surfaces three bottlenecks. First, under Trip World's low home–destination region overlap, hometown-aware methods rely on per-region popularity priors rather than transferable user preferences: replacing a user's hometown history with a random other user's leaves SPOT-Trip—the most complete such model—essentially unchanged, confirming directly that it does not exploit user-specific home signal. Second, their accuracy–efficiency trade-off degrades at this scale: the simplest model (MatTrip) is among the strongest, while the most compositional (SPOT-Trip) ranks lowest at the highest cost. Third, a common mechanism for integrating semantic metadata yields no benefit. A diagnostic pilot on agentic LLM methods fares no better: they trail a simple in-candidate popularity prior, even though the relevant semantic signal is present in the data. Overall, progress will require task-specific designs for cross-city preference transfer, semantic grounding, and scalable reasoning over unseen destination inventories.

Acknowledgments

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005), and funded by the Australian Government through the Australian Research Council. In addition, this research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government, and the computational cluster Katana [12] supported by Research Technology Services at UNSW Sydney.

References

[1] Dawei Chen, Cheng Soon Ong, and Lexing Xie. 2016. Learning points and routes to recommend trajectories. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 2227–2232.
[2] CRUISE Research Group. 2026. TripWorld: A Large-Scale Benchmark for Cross-City POI Recommendation. Hugging Face dataset. https://huggingface.co/datasets/CRUISEResearchGroup/TripWorld.
[3] Jingtao Ding, Guanghui Yu, Yong Li, Depeng Jin, and Hui Gao. 2019. Learning from hometown and current city: Cross-city POI recommendation via interest drift and transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–28.
[4] Jie Feng, Yuwei Du, Jie Zhao, and Yong Li. 2025. Agentmove: A large language model based agentic framework for zero-shot next location prediction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1322–1338.
[5] Shanshan Feng, Haoming Lyu, Fan Li, Zhu Sun, and Caishun Chen. 2024. Where to move next: Zero-shot generalization of llms for next poi recommendation. In *2024 ieee conference on artificial intelligence (cai)*. IEEE, 1530–1535.
[6] Qiang Gao, Wei Wang, Li Huang, Xin Yang, Tianrui Li, and Hamido Fujita. 2023. Dual-grained human mobility learning for location-aware trip recommendation with spatial–temporal graph knowledge fusion. *Information Fusion* 92 (2023), 46–63.
[7] Valerii Iakovlev and Harri Lähdesmäki. 2024. Learning spatiotemporal dynamical systems from point process observations. *arXiv preprint arXiv:2406.00368* (2024).
[8] Ai-Te Kuo, Haiquan Chen, and Wei-Shinn Ku. 2023. BERT-Trip: effective and scalable trip representation using attentive contrast learning. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, 612–623.
[9] Dichao Li and Zhiguo Gong. 2020. A deep neural network for crossing-city POI recommendations. *IEEE Transactions on Knowledge and Data Engineering* 34, 8 (2020), 3536–3548.
[10] Yinghui Liu, Hao Miao, Guojiang Shen, Yan Zhao, Xiangjie Kong, and Ivan Lee. 2025. SPOT-Trip: Dual-preference driven out-of-town trip recommendation. *arXiv preprint arXiv:2506.01705* (2025).
[11] Yinghui Liu, Guojiang Shen, Chengyong Cui, Zhenzhen Zhao, Xiao Han, Jiaxin Du, Xiangyu Zhao, and Xiangjie Kong. 2024. KDDC: Knowledge-Driven Disentangled Causal Metric Learning for Pre-Travel Out-of-Town Recommendation.. In *IJCAI*. 2207–2215.
[12] Research Technology Services, UNSW Sydney. 2010. Katana. https://doi.org/10.26190/669X-A286
[13] Wenzheng Shu, Kangqi Xu, Wenxin Tai, Ting Zhong, Yong Wang, and Fan Zhou. 2024. Analyzing and mitigating repetitions in trip recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2276–2280.
[14] Ke Sun, Chenliang Li, and Tieyun Qian. 2024. City matters! A dual-target cross-city sequential poi recommendation model. *ACM Transactions on Information Systems* 42, 6 (2024), 1–27.
[15] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.
[16] Hao Wang, Yanmei Fu, Qinyong Wang, Hongzhi Yin, Changying Du, and Hui Xiong. 2017. A location-sentiment-aware recommender system for both home-town and out-of-town users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1143.
[17] Haoran Xin, Xinjiang Lu, Nengjun Zhu, Tong Xu, Dejing Dou, and Hui Xiong. 2022. Captor: A crowd-aware pre-travel recommender system for out-of-town users. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1174–1184.
[18] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. Nation-Telescope: Monitoring and visualizing large-scale collective behavior in LBSNS. *Journal of Network and Computer Applications* 55 (2015), 170–180.
[19] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 3 (2016), 1–23.
[20] Jiale Zhang, Mingqian Ma, Xiaofeng Gao, and Guihai Chen. 2024. Encoder-decoder based route generation model for flexible travel recommendation. *IEEE Transactions on Services Computing* 17, 3 (2024), 905–920.
[21] Qianru Zhang, Peng Yang, Junliang Yu, Haixin Wang, Xingwei He, Siu-Ming Yiu, and Hongzhi Yin. 2025. A survey on point-of-interest recommendation: Models, architectures, and security. *IEEE Transactions on Knowledge and Data Engineering* (2025).
[22] Shenglin Zhao, Irwin King, and Michael R Lyu. 2016. A survey of point-of-interest recommendation in location-based social networks. *arXiv preprint arXiv:1607.00647* (2016).