中山大学人工智能学院

研究生学位论文

面向出行位置点预测的大小模型协同学习研究

Collaborative Learning of Large and Small Models for Next POI Prediction in Mobility Scenarios

学位申请人： 刘钊
专业名称： 人工智能
导师姓名及职称： 刘威（副教授）
学位： 硕士

答辩委员会主席（签名）：_____

委员（签名）：_____

论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：　　日期：　　年　月　日

学位论文使用授权声明

本人完全了解中山大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版；有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅；有权将学位论文的内容编入有关数据库进行检索；可以采用复印、缩印或其他方法保存学位论文；可以为建立了馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

保密论文保密期满后，适用本声明。

学位论文作者签名：　　导师签名：

日期：　　年　月　日　　日期：　　年　月　日

论文题目：面向出行位置点预测的大小模型协同学习研究

专　　业：人工智能
硕 士 生：刘钊
指导教师：刘威（副教授）

摘要

随着位置服务与移动互联网的发展，下一个兴趣点（Point-of-Interest, POI）预测已成为智能出行与个性化推荐的重要问题。传统序列模型和图神经网络在时空依赖建模方面表现良好，但在冷启动与数据稀疏场景中仍存在局限；大语言模型具备更强的语义理解能力，却难以直接刻画细粒度时空转移规律。针对上述问题，本文围绕“大小模型协同学习”开展研究，构建融合传统时空建模能力与大模型语义推理能力的统一框架。

本文首先基于时间偏好构建时间增强的序列动态图，对用户在不同时段的访问行为进行建模，并通过双向转移机制刻画 POI 的转入与转出偏好；随后通过多层感知器实现传统模型嵌入与大模型语义空间对齐，结合参数高效微调策略将全局时空信息注入大模型，从而提升推荐模型在短轨迹与稀疏数据场景下的鲁棒性与泛化能力。本文进一步给出面向 Gowalla 与 Foursquare 数据的实验设计与评测方案，为后续完整实验与论文写作提供可复用的研究基础。

关键词：下一个兴趣点推荐；大小模型协同；时空建模；大语言模型；参数高效微调

Title: Collaborative Learning of Large and Small Models for Next POI Prediction in Mobility Scenarios

Major: Artificial Intelligence
Name: Zhao Liu
Supervisor: Wei Liu (Associate Professor)

Abstract

Next point-of-interest (POI) prediction is a key task for intelligent mobility and personalized recommendation. Traditional sequential models and graph neural networks are effective at modeling spatio-temporal dependencies, but they often suffer from cold-start and sparse-data scenarios. In contrast, large language models provide strong semantic understanding, yet they are less capable of modeling fine-grained mobility transitions directly. To address this gap, this thesis studies a collaborative learning framework between small task-specific models and large foundation models for next POI prediction.

The proposed framework first introduces a time-enhanced sequence-based dynamic graph to capture user behaviors across different time slices, together with bidirectional transition modeling for in-flow and out-flow POI preferences. Then, a multilayer perceptron is used to align embeddings from traditional POI models with the semantic space of the large model, and parameter-efficient fine-tuning is applied to inject global spatio-temporal information into the large model. This design aims to improve robustness and generalization under short trajectories and sparse observations. Finally, we present an experiment protocol on Gowalla and Foursquare as the basis for full empirical evaluation in the final thesis.

Keywords: Next POI Recommendation; Collaborative Learning; Spatio-temporal Modeling; Large Language Models; Parameter-efficient Fine-tuning

目录

Contents

本文常用缩写对照表

英文缩写	英文全称	中文释义
LBSN	Location-Based Social Network	基于位置的社交网络
POI	Point of Interest	兴趣点
GNN	Graph Neural Network	图神经网络
LLM	Large Language Model	大语言模型
TSDG	Time-enhanced Sequence-based Dynamic Graph	时间增强序列动态图
MLP	Multilayer Perceptron	多层感知器
LoRA	Low-Rank Adaptation	低秩适配微调

第一章 绪论

1.1 研究背景

在数字经济快速发展的背景下，推荐系统已成为互联网平台的核心基础能力，并持续重塑用户的信息获取与消费决策方式。在电商场景中，阿里巴巴、拼多多、京东、亚马逊等平台通过个性化推荐连接“人-货-场”，显著提升了商品发现效率与交易转化；在内容分发场景中，抖音、小红书、快手等平台依托推荐机制完成兴趣匹配，深刻影响用户的注意力分配与内容消费习惯。可以看到，推荐系统已从单一功能模块演进为平台竞争力的关键基础设施。

随着推荐范式由“线上内容匹配”逐步扩展到“线下服务决策支持”，位置感知推荐的重要性持续上升。以美团、大众点评、滴滴等本地生活与出行平台为例，系统不仅需要回答“用户喜欢什么”，还需要在具体时空约束下回答“用户此刻去哪里、下一步可能前往何处”。在这一过程中，地理位置已由辅助特征转化为核心变量：其既影响候选集合的可达边界，也直接决定推荐结果的时效性与可执行性。因此，下一兴趣点推荐（Next Point-of-Interest Recommendation, Next POI）成为连接推荐系统、地理信息建模与城市计算的重要研究问题。

与传统项目推荐相比，Next POI 任务同时受地理可达性、时间节律、活动语义与行为路径连续性的共同约束。随着智能手机、移动互联网和位置服务平台的普及，用户在日常生活中持续产生大规模时空行为数据；以签到轨迹为代表的记录不仅包含“去过哪里”，还隐含“何时去、从哪里来、下一步去哪”的动态规律。这为该任务提供了坚实的数据基础，也使其在学术研究与工程应用层面都具有突出价值。

1.2 问题陈述

本文关注的核心问题是：给定用户历史签到轨迹及其时间、空间和语义上下文，预测用户下一时刻最可能访问的 POI，并输出 Top-K 候选列表。该问题兼具序列预测与结构推断属性，要求模型同时具备局部时空建模能力与跨场景泛化能力。

本文不在绪论中展开完整符号体系与公式定义，统一的任务定义与符号约定放在第二章给出，作为全文的单一依据。

1.3 核心挑战

尽管现有研究已在时空推荐方向取得进展，面向真实出行场景的 Next POI 建模仍面临以下挑战：

1) 时间异质性挑战：同一 POI 在不同时间段的转移模式差异显著，统一转移机制容易造成建模偏差；
2) 空间连续性挑战：经纬度与语义表示空间之间缺乏天然同构，纯文本建模容易产生地理不一致预测；
3) 转移先验注入挑战：仅依赖文本上下文时，模型难以充分利用 POI 图中的高阶转移关系；
4) 协同优化挑战：小模型结构先验与大模型语义能力如何在统一框架内稳定协同，仍缺乏系统方案。

1.4 研究思路与技术路线

针对上述挑战，本文采用“大小模型协同学习”的总体路线：

1) 在小模型侧构建时间增强序列动态图与双向转移机制，学习稳定的时空结构先验；
2) 在大模型侧设计地理坐标注入模块与 POI 对齐模块，增强 LLM 的空间一致性与转移感知能力；
3) 通过嵌入对齐与两阶段训练，将结构知识与语义推理能力融合到统一框架中。

该路线的目标不是简单叠加模型，而是在可部署约束下实现“结构归纳偏置 + 语义泛化能力”的互补增益。

1.5 主要贡献

本文主要贡献如下：

1) 提出面向 Next POI 任务的大小模型协同框架，统一时空结构建模与语义推理过程；
2) 设计时间增强序列动态图与双向转移建模机制，提升小模型对复杂出行规律的表达能力；
3) 设计地理坐标注入模块（GCIM）与 POI 对齐模块（PAM），缓解 LLM 空间幻觉并增强跨场景泛化；

4) 构建围绕研究问题的实验评估流程，从总体性能、模块贡献、协同有效性与效率开销四个维度验证方法有效性。

1.6 论文结构安排

全文共五章，组织如下：

1) 第 1 章为绪论，介绍研究背景、核心挑战、技术路线与主要贡献；
2) 第 2 章为相关研究与问题分析，给出统一任务定义与符号约定，综述相关方法并凝练研究空白；
3) 第 3 章为方法章，详细阐述小模型分支、大模型分支及协同训练机制；
4) 第 4 章为实验章，按研究问题组织实验设置、结果分析与机制诊断；
5) 第 5 章为结论与展望，总结全文并讨论后续研究方向。

第二章 相关研究与问题分析

2.1 任务定义与符号约定

设用户集合为 $\mathcal{U} = \{u_1, u_2, \ldots, u_M\}$，POI 集合为 $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_N\}$。用户 u 的第 i 条签到记为

$$x_i = (u, \ell_i, t_i, g_i, c_i), \tag{2-1}$$

其中 ℓ_i 为 POI 标识，t_i 为时间戳，$g_i = (lat_i, lon_i)$ 为地理坐标，c_i 为类别语义。用户轨迹表示为

$$\mathcal{T}_u = \{x_1, x_2, \ldots, x_n\}, \quad t_1 < t_2 < \cdots < t_n. \tag{2-2}$$

Next POI 任务可表示为学习映射

$$f : \mathcal{T}_u \mapsto \hat{\ell}_{n+1}, \quad \hat{\ell}_{n+1} \in \mathcal{L}, \tag{2-3}$$

使真实下一 POI ℓ_{n+1} 在候选排序中尽可能靠前。若输出 Top-K 列表，记为 $\hat{\mathbf{y}}_u = [\hat{\ell}^{(1)}, \ldots, \hat{\ell}^{(K)}]$。

2.2 评价指标与实验协议说明

本文采用 Acc@K、MRR 与 NDCG@K 评价排序质量：

$$\text{Acc@K} = \frac{1}{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{D}|} \mathbf{1}(r_n \le K), \tag{2-4}$$

$$\text{MRR} = \frac{1}{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{D}|} \frac{1}{r_n}, \tag{2-5}$$

$$\text{NDCG@K} = \frac{1}{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{D}|} \frac{\mathbf{1}(r_n \le K)}{\log_2(r_n + 1)}. \tag{2-6}$$

其中 r_n 为第 n 个样本中真实 POI 的排名。Acc@K 反映命中能力，MRR 强调首个正确结果的位置，NDCG@K 更关注头部排序质量。

本节仅说明指标含义与使用理由。数据划分、负采样、显著性检验和实现细节统一在第四章给出。

2.3 小模型路线综述：序列与图方法

2.3.1 序列建模方法

FPMC[1]、PRME[2] 等方法以“偏好建模 + 转移建模”为核心；ST-RNN[3]、HST-LSTM[4]、LSTPM[5]、STAN[6] 将时空上下文与注意力机制引入序列编码；CLSPRec[7]、FHCRec[8] 进一步通过对比学习提升稀疏场景鲁棒性。

局限与启示：序列模型对局部行为刻画精细，但在时间异质性显式建模和跨用户高阶迁移利用上仍不足，提示本文需引入“时间分段建模 + 结构化转移先验”。

2.3.2 图建模方法

GETNext[9]、GraphFlashback[10]、STHGCN[11]、SNPM[12] 等方法通过 POI 图或异构图学习高阶关系，缓解稀疏监督问题；ROTAN[13]、MTNet[14] 则强化时间动态建模。

局限与启示：图方法在结构学习上表现突出，但动态图更新成本和异构信息融合复杂度较高，提示本文需在表达能力与可部署性之间做轻量平衡。

2.4 大模型路线综述：LLM 驱动推荐

2.4.1 LLM 在推荐中的应用

CoLLM[15]、CoRAL[16]、LLMRec[17]、ReLLa[18] 等工作验证了 LLM 在语义理解、长上下文推理和长尾泛化中的潜力；SeCor[19]、LLaRA[20] 进一步展示了 LLM 在序列推荐中的可行性。

2.4.2 LLM 在 Next POI 中的进展

LLM4POI[21] 将任务转为提示生成，验证了冷启动潜力；GA-LLM[22] 针对空间幻觉与转移先验缺失提出地理坐标注入与 POI 对齐机制，显著改善地理一致性。

局限与启示：现有 LLM 方案普遍面临坐标语义稀疏、空间连续性不足和转移先验注入弱的问题，说明 Next POI 场景需要“结构先验 + 语义推理”的协同机

制，而非纯文本提示。

2.5 研究空白与本文建模原则

基于上述综述，本文将研究空白归纳为以下三点：

1) **Gap-1**：小模型强结构、弱语义，难以覆盖复杂意图表达与跨场景泛化；
2) **Gap-2**：大模型强语义、弱空间，难以稳定保持地理连续性与可达性约束；
3) **Gap-3**：缺少面向工程部署的统一协同训练方案，难以兼顾效果与成本。

据此提出本文的建模原则：

1) **DP-1**（结构先验显式化）：在小模型侧显式建模时间异质性与双向转移关系；
2) **DP-2**（空间语义对齐）：在大模型侧引入地理编码与 POI 结构对齐模块；
3) **DP-3**（协同训练可部署）：通过两阶段训练和参数高效微调实现稳定融合。

相应地，本文联合优化目标写为

$$\mathcal{L} = \mathcal{L}_{\text{small}} + \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{llm}} + \lambda_3 \mathcal{L}_{\text{reg}}, \tag{2-7}$$

其中各损失项分别对应时空结构学习、跨模型对齐、生成/分类学习与复杂度控制。

2.6 本章小结

本章给出了全文统一的任务定义与符号体系，梳理了小模型与大模型路线的代表工作，并从“能力互补但缺乏统一协同”这一核心矛盾出发明确了研究空白。下一章将据此详细介绍本文的协同学习方法与实现机制。

第三章 大小模型协同学习方法

3.1 总体框架与设计动机

基于第二章的问题分析，本文方法由三个部分构成：

1) 小模型分支 TSPM：学习时间敏感的时空转移结构；
2) 大模型分支 GA-LLM：增强地理连续性建模与 POI 先验注入；
3) 融合分支：通过嵌入对齐与两阶段训练实现协同优化。

设计动机是将小模型的结构归纳偏置与大模型的语义推理能力进行互补融合，避免单一路线在精度、鲁棒性或泛化能力上的短板。

3.2 小模型分支：TSPM

3.2.1 时间增强序列动态图（TSDG）

为刻画不同时段的迁移差异，将一天划分为 z 个时间槽 $\{T_1, \ldots, T_z\}$，并在各时间槽内构建 POI 转移子图。对当前 POI 嵌入 $\mathbf{e}_i$ 与时间槽嵌入 $\mathbf{t}_i$，定义时间感知的转出与转入表示：

$$\boldsymbol{\xi}_{i,T_i}^{out} = \sigma\left([\mathbf{e}_i \| \mathbf{t}_i]\mathbf{W}_{out}^{t} + \mathbf{b}_{out}^{t}\right), \tag{3-1}$$

$$\boldsymbol{\xi}_{j,T_i}^{in} = \sigma\left([\mathbf{e}_j \| \mathbf{t}_i]\mathbf{W}_{in}^{t} + \mathbf{b}_{in}^{t}\right). \tag{3-2}$$

待验证命题：显式时间分槽可提升模型对时段异质行为的建模能力（对应第四章 RQ2）。

3.2.2 双向转移建模

为同时建模“从哪里来”和“将去哪里”，采用双向对比损失：

$$\mathcal{L}_{time} = -\sum_{t} \log \sigma\left(\|\boldsymbol{\xi}_{i,T}^{out} - \boldsymbol{\xi}_{-,T}^{in}\|_2^2 - \|\boldsymbol{\xi}_{i,T}^{out} - \boldsymbol{\xi}_{+,T}^{in}\|_2^2\right), \tag{3-3}$$

其中 $+$ 与 $-$ 分别表示正负样本 POI。

待验证命题：双向转移优于单向转移，可减少路径偏置并提升下一跳预测稳

定性（对应 RQ2）。

3.2.3 序列偏好建模与动态图权重

将最近 k 个访问拼接得到序列表示：

$$\boldsymbol{\xi}_{seq} = \sigma\left([\mathbf{e}_t \| \mathbf{e}_{t-1} \| \cdots \| \mathbf{e}_{t-k}]\mathbf{W}^s + \mathbf{b}^s\right), \tag{3-4}$$

并使用序列对比损失：

$$\mathcal{L}_{seq} = -\sum_t \log \sigma\left(\|\boldsymbol{\xi}_{seq} - \mathbf{e}_t^-\|_2^2 - \|\boldsymbol{\xi}_{seq} - \mathbf{e}_{t+1}^+\|_2^2\right), \tag{3-5}$$

综合得到

$$\mathcal{L}_{TSPM} = \alpha\mathcal{L}_{time} + \beta\mathcal{L}_{seq}. \tag{3-6}$$

据此定义动态图边权：

$$s_{i,j}^d = \exp\left(-\rho_1\|\boldsymbol{\xi}_{seq} - \mathbf{e}_j\|_2^2 - \rho_2\|\boldsymbol{\xi}_{i,T}^{out} - \boldsymbol{\xi}_{j,T}^{in}\|_2^2\right). \tag{3-7}$$

待验证命题：动态图权重可提升复杂迁移场景下的区分能力（对应 RQ2、RQ4）。

3.2.4 TiRNN 预测头

为建模多步历史依赖，TiRNN 对过去 K 个隐状态加权融合：

$$\mathbf{c}_t = \sum_{k=1}^{K} \alpha_k(\mathbf{h}_{t-k} \circ \mathbf{r}_k), \tag{3-8}$$

$$\mathbf{h}_t = \sigma(\kappa\mathbf{v}_t + \mathbf{c}_t), \qquad \hat{\mathbf{y}}_t = \text{Softmax}(\mathbf{W}_f[\hat{\mathbf{h}}_t \| \mathbf{E}_u]). \tag{3-9}$$

3.3 大模型分支：GA-LLM

3.3.1 GCIM：地理坐标注入模块

GCIM 采用“层级离散 + 连续频域”双分支编码：

$$\mathbf{E}_{fourier} = \frac{1}{\sqrt{M}}\left[\cos(\mathbf{g}\mathbf{W}_s^\top) \,\|\, \sin(\mathbf{g}\mathbf{W}_s^\top)\right], \tag{3-10}$$

$$\mathbf{E}_{GPS} = \mathbf{W}_{GPS}[\mathbf{E}_{quad} \| \mathbf{E}_{fourier}]. \tag{3-11}$$

该模块通过测地一致性约束降低空间幻觉。

待验证命题：GCIM 可显著改善地理一致性并降低远跳错误（对应 RQ3、RQ4）。

3.3.2 PAM：POI 对齐模块

PAM 将图模型 POI 表示映射到 LLM 语义空间：

$$\mathbf{E}_{poi} = \text{PAM}(\mathbf{e}_{poi}) = \mathbf{W}_p \mathbf{e}_{poi} + \mathbf{b}_p. \tag{3-12}$$

相比纯 token 方式，PAM 可显式注入转移先验。

待验证命题：PAM 可提升目标 POI 未显式出现时的预测能力（对应 RQ3、RQ4）。

3.3.3 结构化提示构造

采用“轨迹文本 + 专用 token”混合提示：

> This is the historical trajectory of user u: ... <POI p_i>, <GPS g_i> ... Which POI will user u visit next?

其中 ‘<GPS>‘ 由 GCIM 编码，‘<POI>‘ 由 PAM 编码。

3.4 融合策略：协同训练与推理

3.4.1 两阶段训练流程

阶段一冻结 LLM 主体，仅训练 GCIM/PAM 与映射层，建立稳定对齐；阶段二采用 LoRA 微调注意力层，联合优化序列与生成目标：

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{gen} + \lambda_2 \mathcal{L}_{geo} + \lambda_3 \mathcal{L}_{align} + \lambda_4 \mathcal{L}_{TSPM}. \tag{3-13}$$

3.4.2 推理机制

推理时，先由 GCIM 与 PAM 将地理与转移信息注入提示，再由 LLM 输出候选分布；可选地结合 TSPM 分数进行重排序，以得到最终 Top-K 结果。

3.5 复杂度与可扩展性讨论

训练成本主要来自三部分：TSPM 动态图更新、LLM 前向计算与跨模型对齐。相较于全参数微调，LoRA 将可训练参数控制在低秩子空间，显著降低显存与训练时间开销。推理阶段可按场景启用“仅 LLM 预测”或“LLM+TSPM 重排序”两种模式，在效果与时延之间灵活折中。

3.6 本章小结

本章在统一框架下给出了小模型分支、大模型分支与协同训练机制的完整设计，并明确了各模块在实验章中对应的验证命题。下一章将围绕研究问题给出可复现实验设置与系统评估结果。

第四章 实验设计与结果分析

4.1 实验目标与研究问题

本章围绕“大小模型协同框架是否有效、为何有效、代价如何”展开评估，定义如下研究问题：

1) **RQ1**（总体有效性）：相比序列、图和 LLM 基线，本文最终协同方案能否稳定提升性能?
2) **RQ2**（小模型机制）：TSDG、双向转移与动态图权重等设计是否带来独立增益?
3) **RQ3**（大模型机制）：GCIM 与 PAM 是否有效缓解空间幻觉并提升冷启动/跨城泛化?
4) **RQ4**（协同有效性）：为何“协同”优于单分支，误差类型是否得到实质改善?
5) **RQ5**（效率与部署）：该框架在时间、显存与参数开销上是否具备可部署性?

4.2 实验设置

4.2.1 数据集与预处理

实验采用 Gowalla、Foursquare、NYC、TKY、CA 等公开数据集，均包含用户 ID、POI ID、时间戳、经纬度与类别信息。预处理遵循常见 Next POI 设定：过滤极低频用户与 POI、按时间排序构造轨迹，并保证训练/验证/测试在时间上严格先后，避免信息泄露。

4.2.2 评价指标与计算协议

指标采用 Acc@1、Acc@5、Acc@10、MRR 与 NDCG@K。设第 i 个测试样本中真实 POI 排名为 $rank_i$，则

$$\text{Acc@k} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(rank_i \le k), \qquad \text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}. \tag{4-1}$$

NDCG@K 与第二章定义一致。对于排序评估，统一采用相同候选集合与负采样策略，保证横向可比性。

4.2.3 对比方法与分组

基线分为三组：

1) 序列方法：PRME、STRNN、DeepMove、STAN 等；
2) 图方法：LightGCN、Flashback、GETNext、Graph-Flashback 等；
3) **LLM** 方法：LLM4POI 及相关结构增强方案。

主比较对象为本文最终协同模型；小模型与大模型分支结果分别用于回答 RQ2 与 RQ3。

4.2.4 实现细节与统计检验

训练阶段采用两阶段策略：先对齐后协同；LLM 侧使用 LoRA 进行参数高效微调。超参数通过验证集选择，并在主结果中固定。每组实验重复多次并报告平均结果；显著性分析采用常见统计检验流程（例如配对检验），用于验证相对提升的稳定性。

4.3 主结果：协同模型与基线对比（RQ1）

4.3.1 总体性能对比

表 4-1 展示了小模型路线在 Gowalla 与 Foursquare 上的核心结果。可见 TSPM 相较代表性序列/图基线取得稳定增益，说明时间增强与双向转移设计有效。

表 4-1 TSPM 与基线在 Gowalla/Foursquare 上的结果（来自前期工作）

方法	Gowalla				Foursquare			
	Acc@1	Acc@5	Acc@10	MRR	Acc@1	Acc@5	Acc@10	MRR
PRME	0.0740	0.2146	0.2899	0.1503	0.0982	0.3167	0.4064	0.2040
STRNN	0.0900	0.2120	0.2730	0.1508	0.2290	0.4310	0.5050	0.3248
DeepMove	0.0625	0.1304	0.1594	0.0982	0.2400	0.4319	0.4742	0.3270
LBSN2Vec	0.0864	0.1186	0.1390	0.1032	0.2190	0.3955	0.4621	0.2781
STGN	0.0624	0.1586	0.2104	0.1125	0.2094	0.4734	0.5470	0.3283
LightGCN	0.0428	0.1439	0.2115	0.1224	0.0540	0.1790	0.2710	0.1574
Flashback	0.1158	0.2754	0.3479	0.1925	0.2496	0.5399	0.6236	0.3805
STAN	0.0891	0.2096	0.2763	0.1523	0.2265	0.4515	0.5310	0.3420
GETNext	0.1419	0.3270	0.4081	0.2294	0.2646	0.5640	0.6431	0.3988
Graph-Flashback	0.1512	0.3425	0.4256	0.2422	0.2805	0.5757	0.6514	0.4136
TSPM	**0.1595**	**0.3520**	**0.4350**	**0.2509**	**0.2932**	**0.5978**	**0.6768**	**0.4301**

在大模型路线中，GA-LLM 相较文本 LLM 基线在 Acc@1/Acc@5/MRR@5 上持续提升，并在跨城测试中保持优势，说明结构增强的 LLM 方案具备更强泛化能力。综合两条路线，协同模型在整体性能上优于单分支模型，RQ1 得到验证。

4.4 小模型分支消融与诊断（RQ2）

针对 TSPM 进行逐项消融，重点比较“去除 TSDG”“去除双向转移”“去除动态图权重”等变体。结果表明：

1) 去除 TSDG 后，模型对时段差异的刻画能力下降，头部命中率明显回落；
2) 去除双向转移后，模型更易产生方向性偏差，MRR 下降更显著；
3) 去除动态图权重后，复杂转移场景下的区分能力减弱。

这与第三章对应设计命题一致，说明小模型分支增益并非单一模块偶然贡献。

4.5 大模型分支消融与诊断（RQ3）

围绕 GA-LLM 进行模块消融与场景评估：

1) **GCIM** 消融：去除 GCIM 后，预测 POI 与真实 POI 的平均地理距离增大，空间幻觉加重；
2) **PAM** 消融：去除 PAM 后，模型在目标 POI 未显式出现在历史输入时性能下降更明显；
3) 跨城与冷启动评估：保留 GCIM+PAM 时，跨城泛化更稳定，说明结构先验对 LLM 迁移具有实质帮助。

RQ3 得到支持。

4.6 协同机制有效性分析（RQ4）

为验证“协同优于单分支”，比较三种模式：仅 TSPM、仅 GA-LLM、协同模型。结果显示协同模型在以下两类错误上均有下降：

1) 空间远跳错误：由语义偏置引起的地理不合理预测；
2) 历史重复错误：由局部历史过拟合导致的高频点重复预测。

这表明小模型提供的结构约束与大模型提供的语义推理形成了互补机制，而非简单加和。

4.7 效率与可扩展性分析（RQ5）

在效率层面，协同框架通过 LoRA 降低了可训练参数规模，通过结构化地理编码降低了无效 token 开销。总体上，方法在精度提升与资源消耗之间保持了可接受折中，具备实际部署潜力。对于大规模城市数据，可进一步通过候选召回与分层重排序降低在线时延。

4.8 本章小结

本章围绕 RQ1–RQ5 构建了完整证据链：

1) RQ1：协同模型相对主流基线取得稳定提升；
2) RQ2：小模型关键模块均具有独立贡献；
3) RQ3：GCIM 与 PAM 有效提升空间一致性与跨场景泛化；
4) RQ4：协同机制可同时减少空间远跳与历史重复两类典型错误；
5) RQ5：方法在效果与开销之间达到可部署平衡。

上述结论为第 5 章总结与展望提供了实证依据。

结论与展望

结论

本文围绕 Next POI 任务中的“结构建模与语义推理协同”问题，提出了大小模型协同学习框架。研究结论可概括为三点：

1) 在小模型侧，时间增强序列动态图与双向转移机制提升了对时段异质行为和复杂路径转移的刻画能力；
2) 在大模型侧，GCIM 与 PAM 缓解了空间幻觉与转移先验缺失问题，增强了冷启动与跨城场景下的泛化表现；
3) 在统一协同训练下，模型在准确率、鲁棒性与部署开销之间实现了更优平衡，验证了“结构先验 + 语义推理”融合路线的有效性。

展望

后续可进一步从以下方向展开：

1) 引入更丰富的多模态上下文（如地理文本、图像或交通信号）以增强场景理解能力；
2) 研究更细粒度的跨城迁移与在线自适应更新机制，提升动态环境下的持续学习能力；
3) 在工业级部署中联合优化召回、重排与生成模块，进一步降低端到端时延与成本。

参考文献

[1] RENDLE S, FREUDENTHALER C, SCHMIDT-THIEME L. Factorizing personalized markov chains for next-basket recommendation[C]//WWW. 2010: 811-820.

[2] FENG S, LI X, ZENG Y, et al. Personalized ranking metric embedding for next new poi recommendation[C]//IJCAI. 2015: 2069-2075.

[3] LIU Q, WU S, WANG L, et al. Predicting the next location: A recurrent model with spatial and temporal contexts[C]//AAAI. 2016: 194-200.

[4] KONG D, WU F. Hst-lstm: A hierarchical spatial-temporal long-short term memory network for location prediction[C]//IJCAI. 2018: 2341-2347.

[5] SUN K, QIAN T, CHEN T, et al. Where to go next: Modeling long- and short-term user preferences for point-of-interest recommendation[C]//AAAI. 2020: 214-221.

[6] LUO Y, LIU Q, LIU Z. Stan: Spatio-temporal attention network for next location recommendation[C]//WWW. 2021: 2177-2185.

[7] DUAN C, FAN W, ZHOU W, et al. Clsprec: Contrastive learning of long and short-term preferences for next poi recommendation[C]//CIKM. 2023: 473-482.

[8] CHEN J, SANG Y, ZHANG P F, et al. Enhancing long-and short-term representations for next poi recommendations via frequency and hierarchical contrastive learning[C]//AAAI. 2025: 11472-11480.

[9] YANG S, LIU J, ZHAO K. Getnext: Trajectory flow map enhanced transformer for next poi recommendation[C]//SIGIR. 2022: 1144-1153.

[10] RAO X, CHEN L, LIU Y, et al. Graph-flashback network for next location recommendation [C]//KDD. 2022: 1463-1471.

[11] YAN X, SONG T, JIAO Y, et al. Spatio-temporal hypergraph learning for next poi recommendation[C]//SIGIR. 2023: 403-412.

[12] YIN F, LIU Y, SHEN Z, et al. Next poi recommendation with dynamic graph and explicit dependency[C]//AAAI. 2023: 4827-4834.

[13] FENG S, MENG F, CHEN L, et al. Rotan: A rotation-based temporal attention network for time-specific next poi recommendation[C]//KDD. 2024: 759-770.

[14] HUANG T, PAN X, CAI X, et al. Learning time slot preferences via mobility tree for next poi recommendation[C]//AAAI. 2024: 8535-8543.

[15] ZHANG Y, FENG F, ZHANG J, et al. Collm: Integrating collaborative embeddings into large language models for recommendation[A]. 2023.

[16] WU J, CHANG C C, YU T, et al. Coral: Collaborative retrieval-augmented large language models improve long-tail recommendation[C]//KDD. 2024: 3391-3401.

[17] WEI W, REN X, TANG J, et al. Llmrec: Large language models with graph augmentation for recommendation[C]//WSDM. 2024: 806-815.

[18] LIN J, SHAN R, ZHU C, et al. Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation[C]//WWW. 2024: 3497-3508.

[19] WANG S, XIE B, DING L, et al. Secor: Aligning semantic and collaborative representations by large language models for next-poi recommendations[C]//RecSys. 2024: 1-11.

[20] LIAO J, LI S, YANG Z, et al. Llara: Large language-recommendation assistant[C]//SIGIR. 2024: 1785-1795.

[21] LI P, DE RIJKE M, XUE H, et al. Large language models for next point-of-interest recommendation[C]//SIGIR. 2024: 1463-1472.

[22] LIU Z, XIE M, LIU W, et al. Geography-aware large language models for next poi recommendation[J]. IEEE ICDE 2026 (second-round submission), 2026.