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）

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

第 1 章 绪论

1.1 本章引言

本章作为全文的总体导引，主要回答“为什么做、做什么、如何做、做出了什么”四个核心问题。首先从推荐系统向线下时空决策延展的现实背景出发，说明 Next POI 任务的研究价值；随后给出本文关注的问题边界与核心挑战，明确现有方法在时间异质性、空间一致性和转移先验注入方面的不足；在此基础上概述本文的大小模型协同技术路线、主要创新点与章节组织关系，为后续相关研究、方法细节和实验验证建立统一叙事主线。

1.2 研究背景与研究意义

1.2.1 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[1,2,3,4,5]。在电商场景中，平台通过个性化推荐连接“人-货-场”，显著提升商品发现效率与交易转化；在内容分发场景中，推荐机制持续影响用户的注意力分配与内容消费习惯，这一趋势在工业界与学术界均有系统讨论[6,7,8,9,10]。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”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[11,12,13,14,15]。在本地生活与出行场景中，系统不仅要回答“用户喜欢什么”，还需要在时空约束下回答“用户此刻去哪里、下一步可能前往何处”，这使地理位置从辅助特征转化为核心变量[16,17,18,19,20]。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问题[21,22,23,24,25]。

与传统项目推荐相比，Next POI 任务受地理可达性、时间节律、活动语义与行为路径连续性的共同约束[26,27,28,29,30]。随着移动终端与位置服务平台普及，签到轨迹数据不仅记录“去过哪里”，还隐含“何时去、从哪里来、下一步去哪”的动态规律，为时空序列建模提供了基础[31,32,33,34,35]。这一数据与问题特性也使其在学术研究与工程应用层面都具有持续价值[36,37,38,39,40]。

近两年生成式推荐进一步改变了建模方式，即将“候选打分”转化为“下一条交互条目 ID 生成”。从 P5[41] 统一范式到针对 ID 构造与约束生成的 CID[42]、GenRec[43]、Tiger[44]，再到强调协同语义对齐的 IDGenRec[45]，研究重点逐步从“能生成”转向“生成可控、可对齐、可落地”。这一趋势与 Next POI 任务高度相关，因为下一地点预测本质上也是在强约束空间中生成合法且可达的目标 POI ID。

1.2.2 研究意义

本文工作的意义主要体现在理论、方法与应用三个层面。

1) **理论意义**：Next POI 任务位于“时空序列建模 + 图结构学习 + 语言语义推理”的交叉区域[4,46,47,48,7]。围绕该任务构建统一框架，有助于回答结构归纳偏置与大模型语义能力如何协同的问题。
2) **方法意义**：现有方法多在单一维度较优，例如序列方法[49,50,51,52,53] 擅长局部时序、图方法[54,55,56,57,58] 擅长高阶关系、LLM 方法[59,60,61,62,63] 擅长语义泛化。本文尝试给出可部署的融合路径。
3) **应用意义**：在本地生活、城市出行和文旅推荐等场景中，系统需在有限时空约束下提供可执行推荐[11,14,15,64,65]。若预测结果与真实地理规律不一致，即使语义上合理也难形成有效服务。

1.3 研究问题、问题陈述与核心挑战

1.3.1 研究问题分解

为避免“问题过大、方法过散”，本文将总体目标分解为五个可验证子问题，并在后文通过 RQ 体系进行闭环验证：

1) P1：**时段异质性建模**。同一 POI 在不同时间段具有不同的转移规律，如何以低额外成本显式建模该异质性；
2) P2：**双向转移表达**。下一跳预测不仅取决于“当前点指向谁”，也取决于“目标点通常从哪里来”，如何同时表达转出与转入偏好；
3) P3：**地理连续性注入**。LLM 语义空间不天然满足地理邻近关系，如何将经纬度与空间层级先验注入语义编码流程；
4) P4：**结构先验跨空间对齐**。图模型学得的 POI 关系如何映射到 LLM 语义空间，并在推理阶段持续发挥作用；

5) P5：**协同训练可部署**。在效果提升之外，如何保证训练成本、推理时延与参数规模仍处于可接受范围。

通过上述分解，本文将“是否有效”与“为什么有效”区分处理，从而避免仅凭主结果表格给出结论。

1.3.2 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2章给出，作为全文的单一依据。

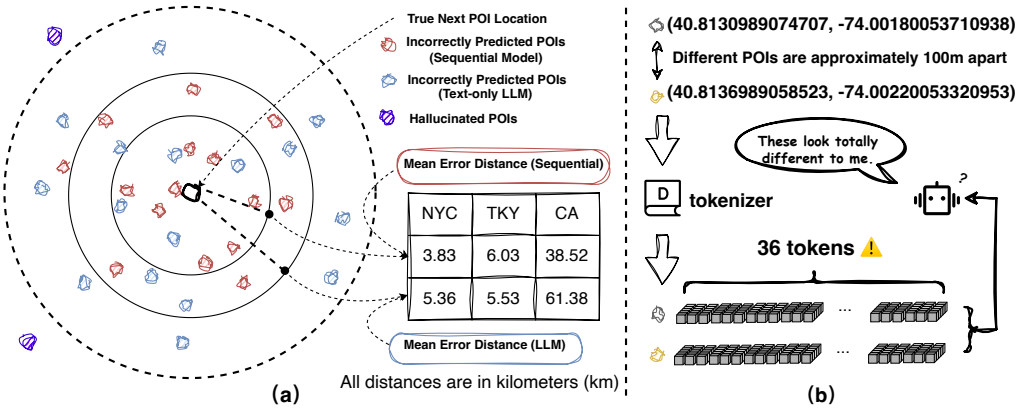


图 1-1 空间连续性挑战示意：（a）预测误差分布与平均地理误差对比，显示纯文本 LLM 更易产生远距离偏移；（b）坐标文本 token 化示例，显示地理上邻近的坐标在离散 token 空间中可能被映射为差异较大的符号序列。

Fig. 1-1 Illustration of the spatial continuity challenge: (a) comparison of error distribution and mean geographic error distance; (b) a tokenization example showing that geographically close coordinates may become distant in discrete token space.

1.3.3 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 在不同时段的转移模式差异显著，统一转移机制容易造成偏差[27,66,29,19,37]；
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[18,20,67,68,69]；

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[33,34,35,70,71]；

4) **协同优化挑战**：小模型结构先验与大模型语义能力在统一框架内稳定协同仍缺乏系统方案[72,61,73,74,75]。

如图1-1(a) 所示，纯文本 LLM 在若干数据集上的平均误差距离相对更大，说明其对空间邻近关系的保持不稳定；如图1-1(b) 所示，直接坐标文本输入会受到 token 离散切分影响，导致“地理接近”与“语义接近”不一致。两部分现象共同说明 Next POI 任务需要专门的地理编码机制，而非仅依赖通用文本建模。基于这一挑战，本文在第3章设计 GCIM 模块以增强空间一致性。

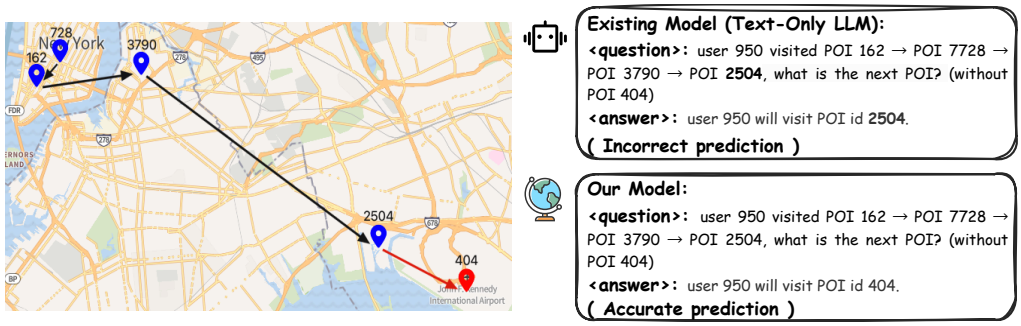


图 1-2 转移先验注入挑战示意：仅依赖 token 语义难以恢复 POI 图中的高阶迁移关系。
Fig. 1-2 Illustration of transfer-prior injection challenge: token-level semantics alone are insufficient to recover high-order transitions in the POI graph.

如图1-2所示，若缺少结构化对齐路径，模型对“未显式出现但高转移概率”的候选点识别能力不足。该现象说明 POI 关系先验需要显式注入语义空间，而非完全依赖隐式学习。基于这一挑战，本文在第3章设计 PAM 模块以提升转移感知能力。

1.4 研究思路、工作安排与文献定位

1.4.1 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，学习稳定时空结构先验[76,77,78,36,79]；

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[67,68,64,65,47]；

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[59,60,62,80,81]。

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。

1.4.2 研究内容与工作安排

围绕上述路线，本文研究内容可进一步细化为四项工作：

1) **任务分析与基线整理**：系统梳理 Next POI 领域中的序列、图与 LLM 三类方法，统一问题定义、指标解释与实验比较边界；
2) **小模型机制设计**：围绕时间异质性与迁移方向性，设计时间增强序列动态图（Time-enhanced Sequence-based Dynamic Graph, TSDG）与双向转移建模，并通过动态图权重增强复杂场景区分能力；
3) **大模型增强设计**：围绕地理连续性与转移先验缺失，设计地理坐标注入模块（Geographic Coordinate Injection Module, GCIM）与 POI 对齐模块（POI Alignment Module, PAM），将结构先验以可训练方式注入 LLM；
4) **实验验证与诊断分析**：基于 RQ 组织实验证据，覆盖主结果、消融、误差类型、效率开销与部署讨论，形成“设置-结果-解释”闭环。

从论文完成度角度，本文不仅给出模型结果，也强调问题定义、机制解释、局限分析与复现细节，以满足毕业论文对研究深度与工作态度的要求。

1.4.3 文献脉络与研究定位

为保证后续方法设计与实验结论具备可比性与可追溯性，本文在文献层面采用“经典基础 + 近年进展 + 前沿融合”的三层梳理框架。经典基础涵盖协同过滤、序列建模与排序学习等工作[1,2,82,83,49]；近年进展重点关注 Next POI 任务中的时空序列与图学习路线[16,84,85,86,87]；前沿融合聚焦 LLM 驱动推荐与结构先验注入路线[88,89,90,91,92]。

面向 Next POI 这一具体问题，本文重点参考三类证据。第一类是“纯时空/图模型”证据，包括空间门控、超图、对比学习与扩散协同等方向[66,93,32,94,40]。第二类是“语义增强”证据，包括语义城市建模与跨模态地理表示学习[23,24,95,96,25]。第三类是“LLM 推荐框架”证据，包括检索增强、协同对齐与推理链增强方法[72,70,63,75,97]，以及近两年面向 Next POI 的最新模型进展[98,99,100,101,102]。上述证据共同支持本文采用“结构建模 + 语义增强 + 协同训练”的总体定位。

围绕“ID 构造-语义对齐-受约束生成”的技术主线，本文将相关进展进一步映射到自身问题：在构造层，借鉴 CID[42] 与 Tiger[44] 的语义 ID 思路提升 POI ID

表达能力；在对齐层，参考 CLLM4Rec[103]、A-LLMRec[104] 与 IDGenRec[45] 的协同对齐范式；在生成层，结合 TALLRec[105] 与 GenRec[43] 的受约束生成实践，服务于 Next POI 的可达性与一致性目标。

1.5 主要贡献与论文结构安排

1.5.1 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（Geographic Coordinate Injection Module, GCIM）与 POI 对齐模块（POI Alignment Module, 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。

1.5.2 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。

1.6 本章小结

本章从应用背景与学术问题两个层面阐明了 Next POI 研究的必要性，并围绕任务特点提炼了时间异质性、空间连续性、转移先验注入和协同优化四类关键挑战。针对上述问题，本文给出了“结构先验建模 + 语义推理增强 + 协同对齐训练”的总体思路，明确了主要贡献与后续章节分工。下一章将进一步通过文献综

述与问题分析，建立本文方法设计的理论与经验依据。

第 2 章 相关研究与问题分析

2.1 本章引言

本章目标是建立全文统一的问题分析基线，回答“已有研究做到什么程度、仍缺什么、本文据此如何建模”三个问题。具体而言，本章先系统综述序列/图/大模型三条技术路线并提炼能力边界，再在此基础上给出全文统一的任务定义、符号体系与评价口径，最后将文献证据收敛为可执行的研究空白（Gap）和建模原则（DP），为第3章的方法设计提供直接依据。

2.2 相关工作综述与研究进展

2.2.1 小模型路线综述：序列与图方法

2.2.1.1 序列建模方法

FPMC[106]、PRME[26] 等方法以“偏好建模 + 转移建模”为核心；ST-RNN[27]、HST-LSTM[28]、LSTPM[107]、STAN[29] 将时空上下文与注意力机制引入序列编码；CLSPRec[108]、FHCRec[79] 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”。

进一步讨论：序列路线对“远距离兴趣跳转”的解释能力不足。仅凭短期历史时，模型容易偏向高频近邻点，在跨区域通勤与目的性出行场景中出现系统性偏差；因此，仅提升序列编码器深度并不能自然解决可达性与结构先验问题，仍需显式引入图结构或地理约束。

2.2.1.2 图建模方法

GETNext[30]、GraphFlashback[33]、STHGCN[35]、SNPM[34] 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[38]、MTNet[37] 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。

进一步讨论：图路线依赖全局邻接关系学习高阶迁移，在数据充足时优势明显，但存在两类工程问题：**其一**：动态图重构与邻居采样在大规模场景中开销较高；**其二**：图表示与语言语义空间缺乏天然对齐，导致与 LLM 协作时出现“信息可用但难注入”的鸿沟。本文后续 PAM 模块即针对第二个问题给出对齐路径。

2.2.2 大模型路线综述：LLM 驱动推荐

2.2.2.1 LLM 在推荐中的应用

LLM 推荐研究已从“直接提示预测”演化为多条并行技术路线。第一类是指令化与参数高效微调路线，代表工作包括 InstructRec[88]、RecGPT[89]、Chat-REC[90] 与 TALLRec[105]，核心目标是以较小训练代价完成任务迁移，并增强跨场景泛化。第二类是协同语义对齐路线，LLMRec[59]、CoLLM[60]、CoRAL[61]、CLLM4Rec[103]、IDGenRec[45] 与 A-LLMRec[104] 通过结构信号、协同表示或文本化 ID 缓解“语言空间-行为空间”不匹配问题。第三类是显式推理与检索增强路线，ReLLa[72]、LLaRA[62]、OneRec[80]、OneRec-Think[75]、ThinkRec[81]、Agent4Rec[109]、LLMRank[91]、MEMO[110] 与 PromptRec[92] 强调可追踪中间推断、候选重排稳定性与错误可诊断性。

从建模粒度看，最新工作进一步从“单步生成答案”转向“分阶段推荐流程”。其中，一类方法将召回-排序-解释统一为共享生成接口，以减少模块割裂；另一类方法引入外部记忆与检索，在长尾物品和冷启动样本上补充证据链，抑制幻觉与不一致。总体上，前沿演化方向已经由“能不能生成”转向“生成是否可控、可解释、可部署”。

对本研究问题而言，上述进展提供了三点直接启示：**其一**：仅依赖自然语言提示难以稳定承载结构迁移先验；**其二**：表示对齐应覆盖“ID 语义-行为图结构-上下文意图”三类信息；**其三**：方法设计需同步考虑推理效率与部署可控性，避免因复杂推理链带来时延与稳定性波动。这些认识为后文方法章节的模块化设计提供了相关工作层面的依据。

2.2.2.2 LLM 在 Next POI 中的进展

在 Next POI 方向，LLM4POI[68] 首先验证了将轨迹预测转化为生成任务的可行性；SeCor[70] 通过语义-协同表示对齐改进时空序列预测质量；Geo-LLMRec[64] 与 POI-LLM 基准研究[65] 进一步揭示了地理一致性不足、远跳误差偏高与空间幻觉等核心瓶颈。

沿着上述问题，近期工作开始向“地理约束显式化”与“推理过程结构化”推进：如隐私保持的多任务反思机制[100]、多智能体协同推断框架[101]、检索增强地理重排策略[102]。这些方法的共同特点是将“坐标连续性、可达性约束、轨迹方向性”作为显式决策因素，而非完全交由黑盒语义生成隐式学习。

整体上看，LLM 驱动的 Next POI 研究正在经历从“生成可行性验证”到“可控生成与鲁棒部署”的阶段转换。这一趋势也说明：面向真实城市场景，方法评估不能只看命中率，还应同步关注地理误差、跨时段稳定性与推理时延等工程指标。

2.2.2.3 ID 构造、语义对齐与受约束生成

近期工作可概括为“ID 构造-语义对齐-受约束生成”三阶段。**其一**：在 ID 构造层，P5[41] 将推荐任务统一为语言生成范式，CID[42] 进一步讨论了不同 Item ID 编码方式对可学习性的影响。**其二**：在语义对齐层，CLLM4Rec[103]、A-LLMRec[104] 与 IDGenRec[45] 通过协同信号对齐缓解“ID 空间-语言空间”错位。**其三**：在生成阶段，TALLRec[105]、GenRec[43] 与 Tiger[44] 强调受约束或可校验生成，以降低非法 ID 输出与语义漂移。

对 Next POI 任务而言，这一主线具有直接启发：POI 本质是带地理语义的 Item ID，若仅把 POI 当作普通文本 token 处理，会放大空间不连续和转移先验缺失问题；因此需要在“ID 层可表达、语义层可对齐、解码层可约束”三个环节同时设计，这也是本文 GCIM 与 PAM 协同建模的理论依据之一。

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。

2.2.2.4 小结性对比

为更清晰地展示三类方法的能力边界，表2-1给出总结性比较。

2.2.3 生成式推荐与显式推理研究进展

随着生成式推荐的发展，研究重点已从“把推荐写成文本生成”逐步转向“提升生成过程的可控性与可诊断性”。这一路线强调：在复杂推荐任务中，仅靠隐式注意力难以稳定地执行约束推断，显式推理链、检索增强与结构对齐是常见改进方向。

表 2-1 序列、图与 LLM 路线能力边界对比
Table 2-1 Capability boundary comparison among sequential, graph-based, and LLM-based routes.

路线	主要优势	主要短板	对本文设计启示
序列模型	局部时序表达强，参数规模相对可控	对高阶迁移与跨场景泛化支持有限	需要补充结构化转移先验与时段异质性建模
图模型	高阶关系建模能力强，能缓解稀疏监督	动态更新成本高，与语义空间融合困难	需要轻量动态图建模与可映射表示空间
LLM 模型	语义理解与泛化能力强，适合复杂意图	地理连续性弱，纯文本坐标表达不稳定	需要专门地理编码与 POI 先验注入机制

2.2.3.1 从指令调优到统一生成框架

近年来，LLM 推荐研究在范式上呈现分阶段演化：**第一步**：将推荐任务指令化并进行参数高效微调，如 InstructRec[88]、RecGPT[89]、TALLRec[105] 与 Chat-REC[90]；**第二步**：将召回、排序与解释统一到单一生成框架，如 LLMRec[59]、CoLLM[60]、CLLM4Rec[103] 与 CoRAL[61]；**第三步**：在统一框架中加入外部检索、协同对齐与结构先验以缓解幻觉和不一致，如 ReLLa[72]、IDGenRec[45] 与 LLaRA[62]；**第四步**：围绕上述主线进一步形成排序增强、记忆增强与提示优化等工具化实践，例如 LLMRank[91]、MEMO[110]、PromptRec[92] 与 RAG 增强推荐[74]，并催生了更系统的综述研究[7,111,112,6]。

2.2.3.2 显式推理范式进展

与“直接生成答案”不同，显式推理范式强调中间决策轨迹，例如候选过滤、约束检查、理由归纳与自我校验。代表性工作显示，将可追溯推理过程纳入推荐推断有助于提升一致性与可解释性[80,75,81,109,73]。这类方法的共同点是：模型不只输出最终推荐，还输出可追溯的中间依据，从而降低“结果可用但难解释”的工程风险。

对 Next POI 任务而言，显式推理的价值主要在三点：**其一**：可将“距离可达性、时段约束、历史路径方向”显式纳入决策；**其二**：可将结构先验作为推理证据而非黑盒隐变量；**其三**：可在错误分析中定位到底是语义误判还是约束违背。本文借鉴其“可控与可诊断”思想，但不引入候选过滤链、在线约束校验或自反思中间输出，而是通过 GCIM 与 PAM 在嵌入空间施加隐式一致性约束。

2.2.3.3 与本文方法的关系

本文不直接复制通用 CoT 模板，而是采用“结构对齐 + 协同训练”的技术路线。原因在于：Next POI 场景具有强时空约束与低时延要求，若直接引入冗长推理链，可能带来明显时延与鲁棒性波动。相比之下，本文将约束能力压缩到“可学习对齐模块 + 统一生成目标”，并保持推理阶段 GA-LLM 单路输出，不采用 constrained decoding、logit mask 或在线重排序。

2.2.4 中文研究脉络与本土场景启示

中文学术界在推荐系统、序列推荐和图推荐方面已形成较系统研究脉络[5,113,114,8]。在时空行为建模与位置推荐方面，研究重点逐步转向“地理约束 + 行为节律”的联合建模[12,13,21,22,15]。在大模型推荐与生成式推荐方面，近年的讨论集中于可解释性、RAG 与工程落地[9,10,115,116,117]，并与城市时空智能应用形成交叉[14]。

这些中文研究对本文有两点直接启示。第一，真实业务中“可解释 + 可部署”通常与“离线最优”同等重要，方法设计不能只追求指标提升。第二，国内 LBS 与本地生活场景具备高密度、强约束、强实时特征，要求推荐模型在地理一致性与时延之间做更细粒度权衡。本文采用协同框架而非单一路线，正是基于上述工程现实做出的方法选择。

2.2.5 扩展文献归纳

为完整刻画本文研究语境，本文参考了经典协同过滤与排序学习文献[1,2,82,3]，以及序列推荐主线工作[49,118,119,4]。在图推荐与自监督学习方面，重点参考了轻量图卷积、知识图增强与图对比学习工作[56,57,94,58,46]。

在时空推荐与 Next POI 方向，本文重点覆盖了序列、图与混合路线的代表方法[16,17,106,26,27]，以及近期性能较强的时空图方法[30,33,34,38,37]。在 LLM 与基础模型方向，本文参考了 LLM4POI[68]、SeCor[70]、POI-LLM 基准工作[65]、Geo-LLMRec[64]与显式推理增强方法[75]。同时，补充引入 P5[41]、CID[42]、IDGenRec[45]、GenRec[43]与 GNPR-SID[120]，以支撑“构造-对齐-生成”三层方法论与本文任务之间的对应关系。

2.3 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$\mathcal{P}=\{\ell_1,\ell_2,\ldots,\ell_N\}$。用户$u$的第$i$条签到定义为：

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

式中：u——用户 ID；

ℓ_i——第i次签到对应的 POI ID；

t_i——签到时间戳；

$g_i=(lat_i,lon_i)$——经纬度坐标；

c_i——POI 类别语义标签。

用户轨迹表示为：

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

式中：$\mathcal{T}_u$——用户u的时间有序签到序列；

n——轨迹长度；

$t_1<t_2<\cdots<t_n$——轨迹按时间严格递增排序。

Next POI 任务可表示为学习映射：

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

式中：f——预测函数；

$\hat{\ell}_{n+1}$——模型预测的下一 POI；

$\mathcal{P}$——候选 POI 全集。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$\hat{\mathbf{y}}_u=[\hat{\ell}^{(1)},\ldots,\hat{\ell}^{(K)}]$。

为减少后续章节符号歧义，表2-2给出本文高频符号及含义。

2.3.1 任务边界与建模假设

为保证论证聚焦，本文采用如下假设与边界设置：

1) **时间先验可用**：训练与测试数据均包含可解析时间戳，且可划分为统一时段；
2) **地理信息可用**：POI 具备经纬度坐标，允许构建地理约束与距离相关特征；
3) **用户历史可观测**：每个测试样本至少存在最小历史长度，以支持序列建模；

表 2-2 核心符号说明
Table 2-2 Description of core notations.

符号	含义
$\mathcal{U}, \mathcal{P}$	用户集合与 POI 集合
$\mathcal{T}_u$	用户 u 的时间有序签到轨迹
$x_i = (u, \ell_i, t_i, g_i, c_i)$	第 i 次签到记录（用户、POI、时间、坐标、类别）
T_z	第 z 个时间槽（Time Slot）
$\boldsymbol{\xi}^{out}, \boldsymbol{\xi}^{in}$	小模型中 POI 转出/转入表示
$\mathbf{E}_{gps}, \mathbf{E}_{poi}$	大模型侧地理编码与 POI 对齐表示
$\hat{\mathbf{y}}_u$	模型输出的 Top-K 推荐列表
r_n	测试样本中真实 POI 的排序位次
$\mathcal{L}_{\text{small}}, \mathcal{L}_{\text{align}}, \mathcal{L}_{\text{llm}}$	小模型损失、跨空间对齐损失与大模型任务损失

4) **离线评估优先**：本文主实验聚焦离线 Top-K 评估，不直接讨论在线 A/B 系统收益。

这些边界并不削弱问题价值，而是用于保证比较公平与可复现。对超出边界的场景（如极短轨迹、无坐标、实时概念漂移），将在第4章“效率与可扩展性分析（RQ5）”中的工程化讨论，以及第4.7章“研究不足”与“研究展望”中给出扩展路径。

2.3.2 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，对应公式定义为：

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

式中：$\mathcal{D}$——测试样本集合；

$|\mathcal{D}|$——测试样本总数；

r_n——第 n 个样本中真实 POI 的排序位次；

$\mathbf{1}(\cdot)$——指示函数，条件满足时取 1，否则取 0。

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

式中：r_n——越小表示真实 POI 排名越靠前；

$\frac{1}{r_n}$——第 n 个样本的倒数排名得分；

MRR 表示全体样本倒数排名的平均值。

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

式中：$\log_2(r_n+1)$——位置折损项；

$\mathbf{1}(r_n \le K)$——真实 POI 是否进入 Top-K；

NDCG@K 对前排命中赋予更高权重。Acc@K 反映命中能力，MRR 强调首个正确结果位置，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4章给出。

2.3.2.1 指标解释补充

在 Next POI 场景中，不同指标对应不同系统目标：

1) **Acc@1** 反映“第一推荐是否可直接点击”，通常对用户体验最敏感；
2) **Acc@5/10** 反映候选列表覆盖能力，适用于存在二次筛选交互的场景；
3) **MRR** 反映正确答案的平均前移程度，能够区分“命中但排名靠后”与“命中且靠前”；
4) **NDCG@K** 对头部位置赋予更高权重，适合评估排序质量而非仅命中与否。

因此，本文不以单一指标给出结论，而是从“首位可用性、候选覆盖、排序质量”三个维度综合判断方法有效性。

2.4 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：

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

式中：$\mathcal{L}_{\text{small}}$——小模型时空结构学习损失；

$\mathcal{L}_{\text{align}}$——跨空间对齐损失；

$\mathcal{L}_{\text{llm}}$——大模型任务损失；

$\mathcal{L}_{\text{reg}}$——正则化项；

$\lambda_1, \lambda_2, \lambda_3$——各损失项的权重系数。

2.4.1 从研究空白到研究问题（RQ）映射

为保证后续实验可以直接检验建模假设，本文将 Gap/DP 与 RQ 的映射关系明确如下：

1) Gap-1 对应 DP-1，并在 RQ2 中通过小模型消融验证；
2) Gap-2 对应 DP-2，并在 RQ3 中通过 GCIM/PAM 诊断验证；
3) Gap-3 对应 DP-3，并在 RQ4、RQ5 中通过协同收益与效率评估验证。

该映射关系使论文论证链条从“问题识别”到“方法设计”再到“证据验证”保持一致，避免章节之间出现目标漂移。

2.5 本章小结

本章围绕“综述评估、定义统一、空白凝练”三条主线完成了问题分析。首先，系统比较了小模型与大模型在结构建模、语义泛化与工程开销上的能力边界；其次，给出了全文唯一任务定义与符号约定，避免后续章节重复描述；最后，将文献中的共性不足归纳为三类研究空白，并据此提出结构先验显式化、空间语义对齐和协同训练可部署三项建模原则。下一章将按照这些原则展开具体模型设计与训练机制说明。

第 3 章 大小模型协同学习方法

3.1 本章引言

基于第2章给出的研究空白与建模原则，本章给出本文方法的完整实现方案，重点回答“整体框架如何搭建、关键模块如何工作、训练与推理如何落地”三个问题。为保证论证清晰，本章按小模型分支、大模型分支和协同训练策略依次展开，并将每个模块与对应研究问题（RQ）建立映射关系，从而为第4章的实验验证提供可追溯的机制假设。本文方法综合继承了时间增强序列预测模型（Time-enhanced Sequential Prediction Model, TSPM）与地理感知大语言模型（Geography-Aware Large Language Model, GA-LLM）系列相关工作的有效设计思想[121,122]，并在毕业论文中统一为“双路线并行、训练协同、推理独立”的研究框架。

3.2 总体框架与设计动机

基于第2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[121]；
2) 大模型分支 GA-LLM：增强地理连续性建模与 POI 先验注入[122]；
3) 对齐/协同训练分支：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。序列/图模型在局部转移建模上具有优势[30,35,38,37]，而 LLM 在语义迁移与复杂上下文理解上更强[68,62,70]；本文的关键是构建一条低损耗的信息通道，使二者不再相互替代，而是协同增益。

如图3-1所示，图示聚焦“语义分支如何吸收结构先验并完成推理输出”的关键路径：轨迹输入先经 GCIM 形成地理编码，再由 PAM 注入结构侧关系信息，随后由 GA-LLM 完成下一 POI 生成。该结构并不表示 TSPM 直接参与在线推理，而是强调 PAM 作为跨分支接口承接协同信息。该设计对应第4章 RQ1 与 RQ4，用于验证结构信息注入带来的整体收益与互补性。

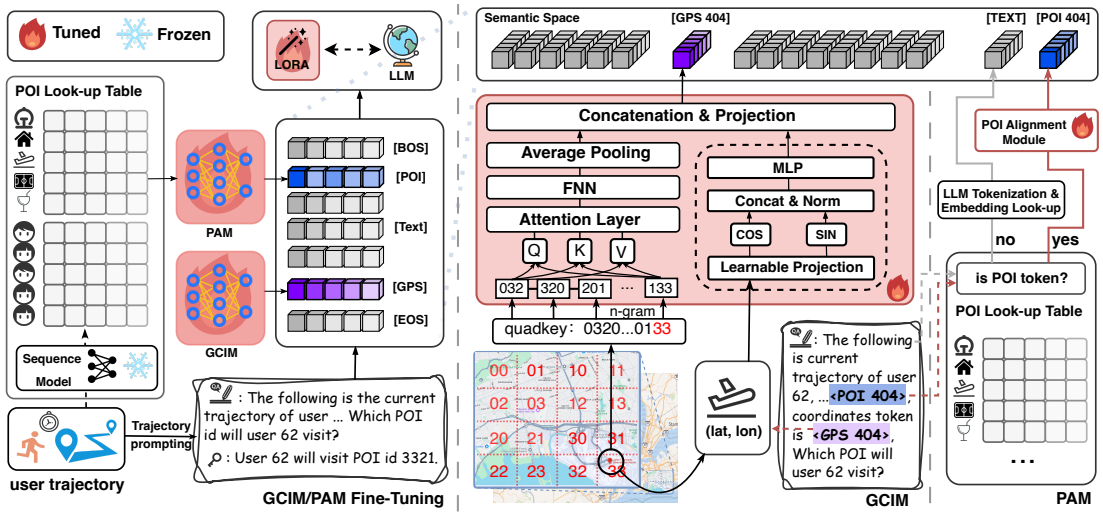


图 3-1 大小模型协同框架示意：图中重点展示 GA-LLM 主推理链路，以及 GCIM 与 PAM 在输入侧的注入流程；其中 PAM 用于承接结构侧（含小模型）学习到的 POI 关系信息并映射到 LLM 语义空间，最终 Top-K 由 GA-LLM 单路输出。

Fig. 3-1 Illustration of the collaborative framework: the figure emphasizes the GA-LLM inference path and the input-side injection of GCIM and PAM. PAM bridges structure-side knowledge (including small-model learned POI relations) into the LLM semantic space, while final Top-K predictions are generated by GA-LLM only.

3.2.1 输入构建与数据流定义

为便于实现与复现，本文将输入组织为三类并行流：

1) **序列流**：由时间有序 POI 序列构成，用于 TSPM 提取局部与短期迁移模式；
2) **结构流**：由动态图邻接关系构成，用于建模高阶转移与区域约束；
3) **语义流**：由轨迹文本、POI 语义 token 与地理 token 构成，用于 GA-LLM 进行语义推理。

三类输入在特征层保持相对独立，在损失层进行耦合，在推理层保持 GA-LLM 单路输出。该设计可避免“早期强融合导致互相干扰”，并降低训练初期梯度不稳定风险。

3.2.2 核心符号与问题映射

本章沿用第2章给出的统一符号体系，不再重复列出完整符号表。为便于阅读，本章仅补充实现级新增变量：$\boldsymbol{\xi}^{out}, \boldsymbol{\xi}^{in}$（双向转移表示）、$\boldsymbol{\xi}_{seq}$（序列聚合表示）、$\mathbf{E}_{gps}$（地理编码表示）、$\mathbf{E}_{poi}$（POI 对齐表示）、$\mathcal{L}_{time}, \mathcal{L}_{seq}, \mathcal{L}_{geo}, \mathcal{L}_{align}, \mathcal{L}_{gen}$（各模块损失项）。其余符号均与第2章定义保持一致。

3.2.3 协同设计原则

本文在框架层遵循三条工程原则：

1) **先对齐后协同**：先保证不同空间表示可互相读取，再进行联合优化；
2) **模块可插拔**：GCIM、PAM 与 TSPM 均可独立启停，便于消融诊断；
3) **训练资源可控**：大模型侧以参数高效微调为主，避免全参更新带来的成本激增。

3.3 关键模块设计

3.3.1 小模型分支：TSPM

3.3.1.1 初始嵌入构建

TSPM 分支首先构建可训练的 POI 初始表示，以保证后续时间分槽与动态图计算有稳定输入。本文采用“关系旋转表示 + 局部拓扑保持”的组合初始化：前者建模转移关系方向性，后者保持近邻几何结构[123,124]。

$$\mathbf{e}_t^{(0)} \approx \mathbf{e}_h^{(0)} \circ \mathbf{r}_{(h,t)}, \tag{3-1}$$

式中：$\mathbf{e}_h^{(0)}, \mathbf{e}_t^{(0)}$——初始阶段的头/尾 POI 向量；

$\mathbf{r}_{(h,t)}$——POI 转移关系向量；

$\circ$——复向量旋转对应的逐元素组合运算。

$$\mathbf{e}_i^{(0)} = \text{EigenMap}(\mathcal{N}(i)), \tag{3-2}$$

式中：$\mathbf{e}_i^{(0)}$——POI i 的拓扑保持初始化向量；

$\mathcal{N}(i)$——POI i 的局部邻接结构。

该初始化的作用是减少“随机初始化导致的早期训练震荡”，并为后续 TSDG 边权学习提供更平滑的优化起点；在实现上，本文将关系建模与局部几何保持联合使用，以兼顾转移方向性与邻域结构稳定性。

3.3.1.2 时间增强序列动态图（TSDG）

TSDG 中的双向转移部分记为双向转移建模（Bidirectional Transformation Modeling, BTM）。其核心目标是同时学习“转出偏好”和“转入偏好”，与第4章 *w/o BTM* 消融设置一一对应。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 $\mathbf{t}_i$，时间感知转出表示定义为：

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

式中：$\boldsymbol{\xi}_{i,T_i}^{out}$——POI i 在时间槽 T_i 的转出表示；

$[\mathbf{e}_i \| \mathbf{t}_i]$——POI 向量与时间槽向量拼接；

$\mathbf{W}_{out}^{t}, \mathbf{b}_{out}^{t}$——转出分支参数；

$\sigma(\cdot)$——非线性激活函数。

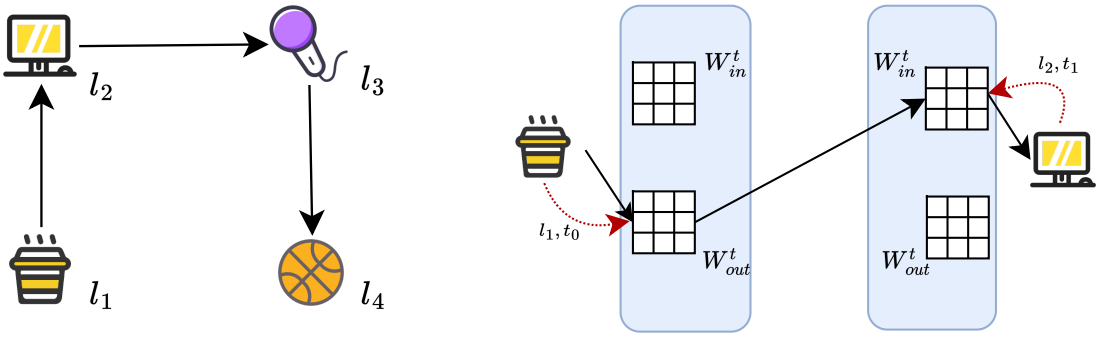
时间感知转入表示定义为：

$$\boldsymbol{\xi}_{j,T_i}^{in} = \sigma\left([\mathbf{e}_j \| \mathbf{t}_i]\mathbf{W}_{in}^{t} + \mathbf{b}_{in}^{t}\right). \tag{3-4}$$

式中：$\boldsymbol{\xi}_{j,T_i}^{in}$——POI j 在时间槽 T_i 的转入表示；

$\mathbf{W}_{in}^{t}, \mathbf{b}_{in}^{t}$——转入分支参数；

其余符号与式(3-3)一致。



(a) 传统连续序列　　(b) 时间增强序列

图 3-2 传统序列与时间增强序列的建模对比。
Fig. 3-2 Modeling comparison between the conventional sequential setting and the time-enhanced sequential setting.

如图3-2所示，传统序列将不同时间段的迁移关系混合建模，而时间增强序列能显式区分时段特征。该现象说明“同一 POI 在不同时段具有不同转移分布”是必须建模的结构事实。该结论直接回扣本节 TSDG 设计，并将在第4章 RQ2 中通过消融实验验证。

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4章 RQ2）。

时间分槽策略讨论

时间分槽粒度过粗会掩盖行为差异，过细会造成样本稀疏。本文采用“以行为节

律为导向”的折中策略，即优先保证每个时段具有足够样本密度，再在验证集上微调分槽数量。该策略相比固定等距切分更贴近真实城市活动节奏。

3.3.1.3 双向转移建模

仅建模“转出”（当前点通常去哪里）在若干场景下会出现歧义。以工作日早高峰为例：用户从地铁站出发时，候选目的地可能同时包含写字楼、早餐店与便利店；如果模型只看“地铁站的常见去向”，容易被高频近邻点吸引而忽略真实通勤目的地。反过来看“转入”特性可提供额外判别信息：写字楼在该时段通常承接大量来自交通枢纽的入流，而便利店的入流更分散且短停留。再如晚间场景，商场与住宅区可能共享相似语义标签，但住宅区在夜间具有更稳定的“被到达”模式。

基于上述现象，本文同时建模“从哪里来”和“将去哪里”，即联合学习转出表示与转入表示，并通过双向对比损失约束二者的一致性：

$$\mathcal{L}_{time} = -\sum_{t} \log \sigma \left(\|\boldsymbol{\xi}_{i,T}^{out} - \boldsymbol{\xi}_{-,T}^{in}\|_2^2 - \|\boldsymbol{\xi}_{i,T}^{out} - \boldsymbol{\xi}_{+,T}^{in}\|_2^2 \right), \tag{3-5}$$

式中：$\mathcal{L}_{time}$——时间转移损失；

$\boldsymbol{\xi}_{+,T}^{in}, \boldsymbol{\xi}_{-,T}^{in}$——分别表示正负样本的转入表示；

$\|\cdot\|_2^2$——平方欧氏距离；

$+$ 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）。

损失函数直观解释

式(3-5)的核心是将“当前点到真实下一点”的距离压缩，同时将“当前点到负样本点”的距离拉开。与单向建模相比，双向表示可同时约束“出边合理性”和“入边合理性”：前者回答“当前点一般指向哪些候选”，后者回答“目标点通常由哪些来源到达”。这种双侧约束可显著减少“高频近邻误吸附”与“语义相似点混淆”，尤其在通勤、跨区跳转与晚高峰回流等方向性强的轨迹中更稳定。

3.3.1.4 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：

$$\boldsymbol{\xi}_{seq} = \sigma \left([\mathbf{e}_t \| \mathbf{e}_{t-1} \| \cdots \| \mathbf{e}_{t-k}] \mathbf{W}^s + \mathbf{b}^s \right), \tag{3-6}$$

式中：$\boldsymbol{\xi}_{seq}$——历史序列聚合表示；

k——历史窗口长度；

$\mathbf{W}^s, \mathbf{b}^s$——序列映射参数。

并使用序列对比损失：

$$\mathcal{L}_{seq} = -\sum_{t} \log \sigma \left(\|\boldsymbol{\xi}_{seq} - \mathbf{e}_t^- \|_2^2 - \|\boldsymbol{\xi}_{seq} - \mathbf{e}_{t+1}^+ \|_2^2 \right), \tag{3-7}$$

式中：$\mathcal{L}_{seq}$——序列对比损失；

$\mathbf{e}_{t+1}^+$——真实下一 POI 嵌入；

$\mathbf{e}_t^-$——负样本 POI 嵌入。

综合损失写为：

$$\mathcal{L}_{TSPM} = \alpha \mathcal{L}_{time} + \beta \mathcal{L}_{seq}. \tag{3-8}$$

式中：$\mathcal{L}_{TSPM}$——小模型总损失；

α, β——两部分损失权重。

据此定义动态图边权：

$$s_{i,j}^d = \exp \left(-\rho_1 \|\boldsymbol{\xi}_{seq} - \mathbf{e}_j \|_2^2 - \rho_2 \|\boldsymbol{\xi}_{i,T}^{out} - \boldsymbol{\xi}_{j,T}^{in} \|_2^2 \right). \tag{3-9}$$

式中：$s_{i,j}^d$——动态图从 POI i 到 POI j 的边权；

ρ_1, ρ_2——两类距离项的平衡系数；

$\exp(\cdot)$——指数映射函数。

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）。

负采样策略

为避免训练信号过于简单，本文采用“同区域困难负样本 + 跨区域随机负样本”的混合策略：前者提升局部细粒度区分难度，后者维持全局判别边界。该策略有助于提升模型在近邻候选上的排序精度，并降低模型仅记忆区域标签的风险。

3.3.1.5 TiRNN 预测头

为建模多步历史依赖，本文在预测头引入时间感知循环神经网络（Time-aware Recurrent Neural Network, TiRNN）。其结构如图3-3所示：右侧为历史隐状态序列 $\{\mathbf{h}_{t-1}, \ldots, \mathbf{h}_{t-k}\}$，左侧为关系向量 $\{\mathbf{r}_1, \ldots, \mathbf{r}_k\}$；二者先通过逐步旋转交

互，再由注意力模块选择关键依赖，最后与当前输入的全连接映射结果做加和，得到当前隐状态 $\mathbf{h}_t$。

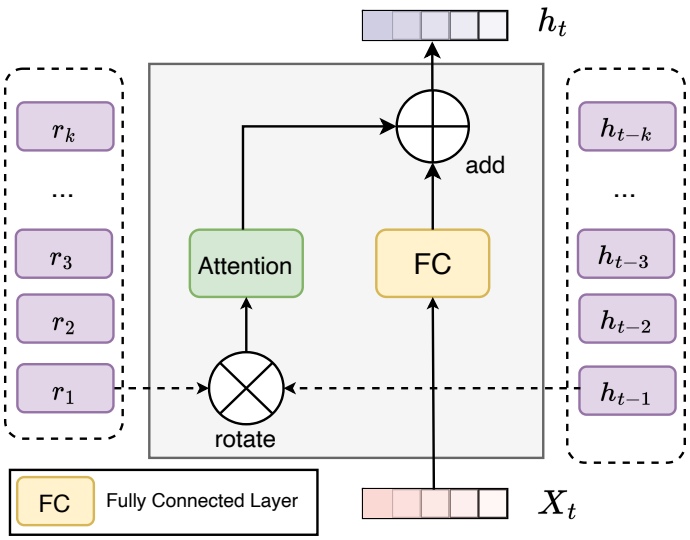


图 3-3 TiRNN 预测头结构示意图。
Fig. 3-3 Architecture of the TiRNN prediction head.

该结构的核心作用是把“历史关系建模”和“当前输入建模”分开处理后再融合：注意力分支负责筛选有效历史信号，FC 分支负责保留当前时刻的即时偏好，二者相加后可在短期惯性与中期计划之间做自适应平衡。基于该流程，TiRNN 对过去 K 个隐状态加权融合：

$$\mathbf{c}_t = \sum_{k=1}^{K} \alpha_k (\mathbf{h}_{t-k} \circ \mathbf{r}_k), \tag{3-10}$$

式中：$\mathbf{c}_t$——时刻 t 的历史上下文向量；

K——回看步数；

α_k——第 k 步历史权重；

$\mathbf{h}_{t-k}$——第 $t-k$ 时刻隐状态；

$\circ$——Hadamard 逐元素乘。

$$\mathbf{h}_t = \sigma(\kappa \mathbf{v}_t + \mathbf{c}_t), \qquad \hat{\mathbf{y}}_t = \mathrm{Softmax}(\mathbf{W}_f[\hat{\mathbf{h}}_t \| \mathbf{E}_u]). \tag{3-11}$$

式中：$\mathbf{h}_t$——当前隐状态；

κ——当前输入权重；

$\mathbf{v}_t$——当前访问表示；

$\hat{\mathbf{y}}_t$——候选 POI 概率分布；

$\mathbf{W}_f$——预测层参数；

$\mathbf{E}_u$——用户嵌入。

预测头设计动机

TiRNN 并非替代前述结构模块，而是作为“时序聚合终端”整合动态图编码结果。通过显式聚合多步隐状态，模型可以在“短期活动惯性”和“中期出行计划”之间动态权衡，从而减少单步过拟合问题。

3.3.2 大模型分支：GA-LLM

3.3.2.1 图 3-1 中的 GA-LLM 信息流解读

为避免“模块分开写、读者难以对齐图示”的问题，这里先按图3-1给出 GA-LLM 分支的完整数据流。图中右侧语义分支可拆解为四步：第一步，轨迹事件被组织为统一模板文本；第二步，GCIM 将坐标字段编码为地理 token 并注入输入序列；第三步，PAM 将结构侧 POI 表示映射为语义 token 并注入同一序列；第四步，LLM 直接输出候选分布与 Top-K 结果。后续各小节均对应这四步中的一个关键环节，读者可直接对照图3-1中“语义流 $\rightarrow$ 对齐 $\rightarrow$ 输出”的路径理解实现细节。

3.3.2.2 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。为与实验消融项保持一致，本文将连续频域分支记为连续空间编码（Continuous Spatial Encoding, CSE），将层级离散分支记为层级离散编码（Hierarchical Discrete Encoding, HDE）。其中，CSE 定义为：

$$\mathbf{E}_{fourier} = \frac{1}{\sqrt{M}} \left[\cos(\mathbf{g}\mathbf{W}_s^\top) \,\|\, \sin(\mathbf{g}\mathbf{W}_s^\top)\right], \tag{3-12}$$

式中：$\mathbf{E}_{fourier}$——Fourier 频域编码向量；

$\mathbf{g}$——坐标输入向量；

$\mathbf{W}_s$——频域投影矩阵；

M——归一化维度系数。

融合地理表示定义为：

$$\mathbf{E}_{gps} = \mathbf{W}_{gps}[\mathbf{E}_{quad} \| \mathbf{E}_{fourier}]. \tag{3-13}$$

式中：$\mathbf{E}_{gps}$——最终地理编码；

$\mathbf{E}_{quad}$——Quadkey 层级编码；

$\mathbf{W}_{gps}$——融合投影参数。

为进一步约束地理表示与真实测地距离的一致性，本文在训练阶段引入测地对齐损失（Geodesic Alignment Loss, GAL）：

$$\mathcal{L}_{geo} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left|\mathrm{dist}_{geo}(i,j) - \mathrm{dist}_{emb}(i,j)\right|, \tag{3-14}$$

式中：$\mathcal{L}_{geo}$——地理一致性损失，对应实验中的 GAL 项；

$\mathcal{P}$——批内 POI 对集合；

$\mathrm{dist}_{geo}(i,j)$——POI i,j 的真实测地距离；

$\mathrm{dist}_{emb}(i,j)$——编码空间中的距离。

针对第1章图1-1所示的空间错位问题，GCIM 在实现层面通过 Quadkey 分支编码层级区域结构，通过 Fourier 分支编码连续距离变化，并在投影层完成统一融合。该设计使地理关系以结构化方式进入 LLM 输入空间，为后续地理一致性约束提供可优化的表示基础。对应图3-1，GCIM 位于 GA-LLM 分支的输入侧，直接作用于轨迹事件中的坐标字段，其输出 $\mathbf{E}_{gps}$ 将作为后续结构化提示中的 `<GPS>` 语义载体，而不是在 LLM 输出后再做后处理。

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）。

为何采用 Quadkey + Fourier

Quadkey 适合表达“区域层级归属”，Fourier 适合表达“连续坐标变化模式”[18,19,122]。前者提供离散空间结构，后者提供连续距离敏感性，二者互补后可同时保留地理分区稳定性与局部变化分辨率。该设计可避免仅离散编码导致的边界不连续，也避免仅连续编码导致的区域语义缺失。

3.3.2.3 PAM：POI 对齐模块

在协同框架中，轨迹语义、地理坐标与图结构表示属于不同模态：轨迹文本位于离散 token 语义空间，坐标经 GCIM 后形成地理编码空间，而小模型/图模型输出位于结构表征空间。三类表示若直接拼接，常出现“可用信息无法被同一注意力层稳定读取”的问题，具体表现为结构先验在生成阶段贡献不稳定、跨城迁移时增益波动较大。

因此，在将结构信息注入 LLM 之前，需要先建立“结构空间 → 语义空间”的可学习映射通道，使 POI 关系先验在语义侧具备可比较、可聚合、可反向优化的表示形式。基于这一需求，本文引入 POI 对齐模块（PAM），其核心变换定义为：

$$\mathbf{E}_{poi} = \mathrm{PAM}(\mathbf{e}_{poi}) = \mathbf{W}_p \mathbf{e}_{poi} + \mathbf{b}_p. \tag{3-15}$$

式中：$\mathbf{e}_{poi}$——图模型侧 POI 嵌入；

$\mathbf{E}_{poi}$——映射后的语义空间 POI 表示；

$\mathbf{W}_p, \mathbf{b}_p$——PAM 映射参数。

针对第1章图1-2所示的转移先验缺失问题，PAM 在实现层面构建“结构表征到语义表征”的线性映射通道，并将图模型中的迁移知识注入 LLM 语义空间。该路径避免了仅靠 token 共现学习关系的局限，使模型在候选稀疏和目标未显式出现时仍可利用结构先验进行推断。对应图3-1，PAM 承接小模型/结构侧输出并注入到语义侧输入，承担“跨分支桥接”角色；因此它不是独立预测器，而是连接双分支信息流的关键接口。

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）。

对齐目标

PAM 的目标不是“替代 LLM 语义”，而是建立一条可学习映射，使图模型中可迁移的结构关系能够在 LLM 空间中被读取和利用。换言之，PAM 承担的是“知识通道”角色，核心价值在于降低跨空间信息损失。

3.3.2.4 结构化提示构造

为避免“只有 token 占位符、缺少可学习语义上下文”的问题，本文将每条轨迹组织为“自然语言事件描述 + 结构化专用 token”的混合输入。核心思路是把签到序列转写为按时间递增的事件流，并在每个事件中显式提供时间、POI 语义与地理编码。

事件文本化规则

设用户 u 在时刻 t_i 访问 POI p_i，将该事件标准化为：

$$\mathcal{E}_i = \texttt{At } t_i\texttt{, user } u \texttt{ visited <POI } p_i\texttt{> with location <GPS } g_i\texttt{>}. \tag{3-16}$$

式中：$\mathcal{E}_i$——第 i 条轨迹事件的文本化表示；

<POI p_i>——由 PAM 注入的 POI 结构语义 token；

<GPS g_i>——由 GCIM 注入的地理编码 token。

给定历史轨迹 $\{\mathcal{E}_1, \ldots, \mathcal{E}_n\}$，训练与推理均使用同一字段顺序，以减少模板漂移造成的分布偏差。

微调问答样本格式

训练阶段采用“Instruction-Input-Output”监督格式，使模型学习从轨迹事实到下一 POI 的映射。模板如下：

Prompt Template

Instruction:

Predict the next POI based on the user's chronological trajectory.

Input:

User ID: U_104.

Trajectory events:

1) At 2024-06-03 08:12, user U_104 visited <POI CoffeeShop_183> with location <GPS 22.2731,113.5728>.

2) At 2024-06-03 08:47, user U_104 visited <POI Office_592> with location <GPS 22.2765,113.5881>.

3) At 2024-06-03 12:06, user U_104 visited <POI Canteen_74> with location <GPS 22.2748,113.5850>.

4) At 2024-06-03 18:21, user U_104 visited <POI Gym_231> with location <GPS 22.2713,113.5794>.

Question: Which POI will user U_104 visit next?

Output:

<POI Supermarket_406>

上述格式的关键是：输入中保留完整时序线索（谁、何时、访问了什么、位于哪里），输出仅监督“下一 POI 标签”，从而使模型将生成能力集中到下一跳判别任务本身。

推理阶段模板

在线推理时保持与微调同构的模板，仅移除监督答案：

Prompt Template

Instruction: Predict the next POI based on the user's chronological trajectory.

Input: User ID + ordered trajectory events (with <POI> and <GPS> tokens).

Question: Which POI will this user visit next?

Output: [Model prediction]

其中，<GPS> 由 GCIM 编码、<POI> 由 PAM 编码。该模板在训练与推理阶段字段一致，可有效降低提示分布漂移并提升生成稳定性。从图3-1的角度看，结构化提示就是把“GCIM 地理编码结果”和“PAM 对齐结果”在输入层做显式合并，使 GA-LLM 在同一上下文内同时读取地理约束与转移先验。

3.3.2.5 生成目标与解码策略

GA-LLM 的训练目标为标准自回归负对数似然：

$$\mathcal{L}_{gen} = -\sum_{t=1}^{m} \log p(y_t \mid y_{<t}, \mathbf{X}_{traj}, \mathbf{E}_{gps}, \mathbf{E}_{poi}), \tag{3-17}$$

式中：y_t——第 t 个生成 token；

$\mathbf{X}_{traj}$——轨迹文本序列输入；

$\mathbf{E}_{gps}$——GCIM 注入的地理表示；

$\mathbf{E}_{poi}$——PAM 注入的 POI 先验表示。

为与“下一 token 预测”机制保持一致，时刻 t 的输出分布可写为：

$$p(y_t \mid y_{<t}, \mathbf{X}) = \text{Softmax}(\mathbf{W}_o \mathbf{h}_t + \mathbf{b}_o), \tag{3-18}$$

式中：$\mathbf{X}$——融合了轨迹文本、地理编码与 POI 先验的输入序列；

$\mathbf{h}_t$——解码器在位置 t 的隐藏状态；

$\mathbf{W}_o, \mathbf{b}_o$——词表投影参数。

在实现层面，融合输入先构造为：

$$\mathbf{H}^{(0)} = \text{Embed}(\mathbf{X}_{traj}) + \mathbf{M}_{gps} \odot \mathbf{E}_{gps} + \mathbf{M}_{poi} \odot \mathbf{E}_{poi}, \tag{3-19}$$

式中：$\mathbf{H}^{(0)}$——输入层隐藏表示；

$\text{Embed}(\cdot)$——token 嵌入查表；

$\mathbf{M}_{gps}, \mathbf{M}_{poi}$——注入位置掩码（mask）；

⊙——逐元素乘。

随后每层 Transformer 采用标准“注意力 + 前馈”更新。单头缩放点积注意力写为：

$$\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3-20}$$

式中：$\mathbf{Q},\mathbf{K},\mathbf{V}$——查询、键、值矩阵；

d_k——键向量维度。

多头注意力与前馈层定义为：

$$\mathrm{MHA}(\mathbf{H}) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)\mathbf{W}^{O}, \quad \mathrm{head}_j = \mathrm{Attn}(\mathbf{H}\mathbf{W}_j^{Q},\mathbf{H}\mathbf{W}_j^{K},\mathbf{H}\mathbf{W}_j^{V}), \tag{3-21}$$

$$\mathrm{FFN}(\mathbf{H}) = \phi(\mathbf{H}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{3-22}$$

式中：h——注意力头数；

$\mathbf{W}_j^{Q},\mathbf{W}_j^{K},\mathbf{W}_j^{V}$——第 j 个头的投影参数；

$\mathbf{W}^{O}$——多头输出投影参数；

$\phi(\cdot)$——非线性激活函数（如 SiLU/GELU）。

当采用 LoRA 进行参数高效微调时，线性层权重更新可写为：

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}, \qquad \Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{B}\mathbf{A}, \tag{3-23}$$

式中：$\mathbf{W}$——冻结主干权重；

$\Delta\mathbf{W}$——低秩增量参数；

r——LoRA 秩；

α——缩放系数；

$\mathbf{A},\mathbf{B}$——可训练低秩矩阵。

GA-LLM 训练与推理伪代码

Algorithm 3-1 GA-LLM Branch with GCIM/PAM Injection

```
Input: trajectory batch ℬ, structure embeddings e_poi, coord set g
Output: Top-K prediction list for each trajectory
1: Build event-text sequence X_traj from ℬ
2: Compute E_gps ← GCIM(g), E_poi ← PAM(e_poi)
```

```
3: Fuse input by Eq.(3-19) to obtain H^(0)
4: For each Transformer layer: apply Eq.(3-21) and Eq.(3-22)
5: Compute next-token distribution by Eq.(3-18)
6: Train with Eq.(3-17) and LoRA update Eq.(3-23)
7: In inference, run beam search and return Top-K POI IDs
```

推理阶段采用 beam search 直接生成候选并输出 Top-K。在本文设定中，不引入候选约束解码、不引入在线重排序、不引入 **TSPM** 推理打分，即由 GA-LLM 单路完成最终预测输出。

3.3.2.6 模块到研究问题的对应关系

为保证“方法提出即有验证对象”，本文将核心模块与 RQ 映射如表3-1。

表 3-1 模块设计与研究问题映射
Table 3-1 Mapping between module design and research questions.

模块	主要目标	对应 RQ
TSDG + 时间分槽	捕获时段异质迁移	RQ2
双向转移建模	降低方向偏置、提升下一跳稳定性	RQ2, RQ4
动态图权重	强化复杂邻域区分能力	RQ2, RQ4
GCIM	提升地理一致性、抑制远跳错误	RQ3, RQ4
PAM	注入 POI 转移先验，改善缺失目标推断	RQ3, RQ4
LoRA 协同训练	平衡性能与训练成本	RQ5
GA-LLM 直推机制	在保持时延可控前提下输出稳定候选排序	RQ1, RQ5

3.4 对齐与协同训练策略

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：

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{gen} + \lambda_2 \mathcal{L}_{geo} + \lambda_3 \mathcal{L}_{align} + \lambda_4 \mathcal{L}_{TSPM}. \tag{3-24}$$

式中：$\mathcal{L}_{total}$——协同训练总损失；

$\mathcal{L}_{gen}$——生成任务损失；

$\mathcal{L}_{geo}$——地理一致性损失；

$\mathcal{L}_{align}$——跨空间对齐损失；

$\mathcal{L}_{TSPM}$——小模型结构损失；

$\lambda_1, \lambda_2, \lambda_3, \lambda_4$——对应权重系数。

训练稳定性策略

联合训练中，不同损失尺度差异会导致优化不稳定。本文采用以下策略缓解：先进行 warm-up 对齐训练，再逐步提高联合损失权重；对梯度范数进行裁剪；采用早停与验证集监控防止过拟合。上述策略在不增加模型复杂度的情况下提升了收敛稳定性。对应图3-1的训练路径，阶段一主要优化“结构流 → 语义流”的对齐通道（GCIM/PAM），阶段二在 GA-LLM 任务目标上进行参数高效微调，避免一开始就全链路联训导致梯度相互干扰。

3.4.2 耦合路径与参数更新机制

尽管推理阶段仅保留 GA-LLM 单路输出，训练阶段仍通过 $\mathcal{L}_{align}$ 与 $\mathcal{L}_{geo}$ 把结构先验注入语义空间。具体耦合路径为：TSPM 分支产生结构表示（如 $\boldsymbol{\xi}_{seq}$、POI 结构嵌入）；PAM 将其投影为语义空间向量 $\mathbf{E}_{poi}$；GCIM 提供地理编码 $\mathbf{E}_{gps}$；最终在统一生成目标下更新 GA-LLM 侧可训练参数。该流程实质上是“训练期对齐注入”，而非“推理期分数融合”。

表 3-2 两阶段训练中的参数冻结与更新策略
Table 3-2 Parameter freezing and updating strategy in two-stage training.

阶段	主要优化目标	更新参数集合	冻结参数集合
Stage-1 对齐预热	$\mathcal{L}_{geo} + \mathcal{L}_{align}$	GCIM 投影层、PAM 投影层、对齐映射层	LLM 主干参数、TSPM 主干参数
Stage-2 协同训练	$\mathcal{L}_{gen} + \lambda_2\mathcal{L}_{geo} + \lambda_3\mathcal{L}_{align} + \lambda_4\mathcal{L}_{TSPM}$	LoRA 适配参数、GCIM/PAM 参数、对齐映射层、TSPM 分支参数	LLM 全参数主干（非 LoRA 部分）

梯度回传说明

Stage-2 中，$\mathcal{L}_{gen}$ 与 $\mathcal{L}_{align}$ 的梯度回传到 LoRA、GCIM 与 PAM；$\mathcal{L}_{TSPM}$ 梯度仅回传到 TSPM 分支；共享的是“对齐后的表示空间与训练目标”，而非在线推理分数。

3.4.3 训练流程说明（实现视角）

为便于复现实验，协同训练可归纳为以下步骤：

1) 构建轨迹样本并并行生成序列流、结构流与语义流输入；
2) 在阶段一中最小化 $\mathcal{L}_{geo} + \mathcal{L}_{align}$，得到稳定空间映射；
3) 在阶段二中引入 $\mathcal{L}_{gen}$ 与 $\mathcal{L}_{TSPM}$，执行联合优化；
4) 每个 epoch 后在验证集监控 Acc@1、Acc@5、MRR@5 与平均地理误差；
5) 保存最优 checkpoint 并输出可复现实验配置。

3.4.4 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 输出候选分布并生成最终 Top-K 结果。这一路径与图3-1的语义分支箭头对应：输入轨迹经编码注入后由 GA-LLM 直接完成下一 POI 预测。TSPM 分支在本文中用于并行机制验证与对照分析，不参与该推理路径的最终输出。

3.4.5 分支并行与结果对照

本文采用“分支并行、结果对照”的组织方式：TSPM 分支用于验证时空结构机制，GA-LLM 分支用于验证地理注入与语义推理机制。最终报告分别基于各自分支的独立输出，不进行显式分数融合或在线重排序。

3.4.6 关键超参数与默认配置

结合前期实验设置[121,122]，核心配置如表3-3。其中大模型侧采用统一长上下文微调设置，小模型侧采用与 TSPM 一致的搜索区间并在验证集确定最优值。

3.4.7 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。需要强调的是，推理阶段仅采用 GA-LLM 语义直推；TSPM 仅用于离线对照、机制验证与诊断分析，不参与线上推理输出。

3.4.7.1 时间复杂度分析

设批大小为 B，序列长度为 L，隐藏维度为 d，动态图平均邻居数为 K_n，LLM 层数为 H。则主要计算量可近似写为：

表 3-3 关键超参数与默认配置
Table 3-3 Key hyperparameters and default settings.

参数	默认值	说明
基座 LLM	Llama-2-7b-longlora-32k	大模型主干[122]
学习率（LLM 侧）	2×10^{-5}	常数学习率，20 步 warm-up
训练轮数（LLM 侧）	3 epochs	各数据集统一设置
最大序列长度	32768	支持长轨迹输入
Batch Size / GPU	1	与长上下文显存预算匹配
Quadkey 层级 L	25	城市场景下兼顾精度与效率
时间槽数量 z	验证集选择	典型候选为 $[4, 6, 8, 12]$
Beam size b	验证集选择	控制生成式解码分支数
输出 Top-K	验证集选择	生成结果截断规模
LoRA 秩 r	验证集选择	控制可训练参数规模

1) TSPM **编码**：$\mathcal{O}(B \cdot L \cdot d^2 + B \cdot L \cdot K_n \cdot d)$；
2) GCIM/PAM **注入**：$\mathcal{O}(B \cdot L \cdot d^2)$；
3) LLM **前向**（LoRA）：$\mathcal{O}(B \cdot L \cdot H \cdot d^2)$（主导项）。

因此总体瓶颈仍在 LLM 前向，参数高效微调的价值在于显著减少反向传播所需的可训练参数与显存开销。

3.4.7.2 空间复杂度分析

空间开销由三部分组成：模型参数、激活缓存与动态图索引。相较全参微调，LoRA 将可训练参数限制在低秩矩阵，显著降低优化器状态占用；动态图仅保留 Top-K_n 邻居，避免全图稠密存储。该组合使模型更易在单机多卡或中等算力环境下训练。

3.4.7.3 可复现实现规范

为提升结果复现概率，本文采用以下实践：

1) 固定随机种子并记录数据切分索引；
2) 保存每轮最优检查点与关键超参数配置；
3) 将评估脚本与训练脚本解耦，避免指标实现漂移；
4) 对核心实验进行多次重复并报告平均趋势而非单次最优值。

3.4.8 工程实现细节与落地策略

除理论模块外，毕业论文还需要给出“可运行”的工程闭环。本节补充数据管线、训练调度与线上部署细节，以体现方法的可实施性。

适用边界声明

本节部分内容属于工程化可选策略，用于超大 POI 空间下的时延控制与服务兜底，不构成本文主实验的必要前提。本文主实验保持 GA-LLM 直接生成 Top-K 的推理口径，不采用 TSPM 召回或在线约束重排。

3.4.8.1 数据管线与样本构造

轨迹切片规则 本文统一采用前缀预测范式:给定用户轨迹 $\mathcal{T}_u = \{x_1, \ldots, x_n\}$，构造监督样本为:

$$(\{x_1, \ldots, x_t\}, x_{t+1}), \quad t \in [1, n-1]. \tag{3-25}$$

式中：$\{x_1, \ldots, x_t\}$——历史前缀轨迹;

x_{t+1}——监督目标下一点;

$t \in [1, n-1]$——切片索引范围。该范式与真实推荐流程一致，可避免后缀信息泄露，并便于在离线与在线阶段共享样本定义。

时间槽映射 对时间戳先进行本地时区对齐，再映射为时间槽索引。为减少节律冲突，工作日与休息日使用统一分槽规则，并附加二值指示位。该设计可在不增加大量参数的情况下增强周内与周末差异建模能力。

解码规模控制 主实验推理采用 beam search 生成并输出 Top-K，通过 Beam size 与 Top-K 两个参数平衡时延与命中率；不引入候选约束解码、logit mask 或在线重排序。

3.4.8.2 分支接口与调度机制

TSPM 分支接口

TSPM 分支输入由 POI 索引序列、时间槽序列与用户索引组成，输出包括候选打分与中间结构表示。该分支主要用于验证时空结构机制与提供对照结果，不直接改写 GA-LLM 推理输出。

GCIM/PAM 流水线

GCIM 执行“坐标归一化–层级编码–频域编码–投影融合”; PAM 执行“结构嵌入读取–线性投影–语义空间对齐”。两条流水线在训练时共享批内样本索引，可减少

跨模块数据搬运开销。

动态图更新策略

本文优先采用“按 epoch 离线更新图权重”的稳定策略，而非按 batch 增量更新。前者虽牺牲部分即时性，但可显著减少训练震荡并提升复现实验一致性，更适合作为论文主结果设置。

3.4.8.3 两阶段训练的工程化约束

阶段一仅优化对齐相关参数，目标是建立跨空间可读表示；阶段二引入任务损失执行联合优化。为保证稳定性，本文采用以下调度约束：对齐阶段较大学习率、联合阶段较小学习率；按验证集 Acc@1 与 MRR@5 双指标早停；保存“最优检查点 + 最近检查点”双副本，便于回滚与复盘。

3.4.8.4 推理服务化与异常回退

在线部署时，主链路保持“GA-LLM 直推”，并通过 Beam size、Top-K 与提示长度控制时延。当出现地理字段缺失、模板解析失败或模型超时时，系统回退到基础语义模板与规则过滤路径以保证服务可用性。

3.4.8.5 实施清单

为保证答辩阶段能够展示完整工程闭环，交付物按表3-4统一准备。

表 3-4 方法实现交付清单
Table 3-4 Implementation delivery checklist of the proposed method.

交付项	说明
数据清洗与切分脚本	固定过滤规则、切分索引与版本哈希，确保实验可复验
训练配置文件	记录学习率、损失权重、随机种子、LoRA 参数与时间槽设置
评估脚本与指标单测	统一 Acc/MRR/NDCG 计算逻辑，避免实现漂移
模型检查点规范	保留最优与最近检查点，支持恢复训练与结果回溯
日志与可视化面板	监控损失曲线、梯度范数与验证指标，支撑异常定位
部署回退策略文档	定义失败条件、回退路径与服务降级流程

3.5 方法对照与讨论

为突出本文协同路线的必要性，表3-5从“时间异质建模、地理连续约束、转移先验注入、跨城泛化、部署成本可控”等维度对比代表方法能力边界。可以看

出，单一路线通常只能覆盖局部能力，而协同设计能够在多个关键维度上同时满足要求。

表 3-5 代表方法能力边界对照（“✓”表示具备主能力）
Table 3-5 Capability boundary comparison of representative methods (✓ indicates core capability).

方法	时间异质	地理连续	转移先验	跨城泛化	成本可控
PRME / FPMC			✓		✓
ST-RNN / STAN	✓				✓
GETNext / STHGCN	✓	✓	✓		
LLM4POI				✓	
TSPM	✓	✓	✓		✓
GA-LLM	✓	✓	✓	✓	✓
本文协同框架	✓	✓	✓	✓	✓

该对照表对应两个结论：第一，传统方法和纯 LLM 方法的短板具有互补性；第二，协同框架并非“叠加模块”，而是围绕关键能力缺口做最小闭环补齐。第4章将通过 RQ1–RQ5 逐项验证这种能力补齐是否转化为可测量收益。

3.6 本章小结

本章完成了协同方法的结构化设计。首先构建了统一框架并明确双分支信息流；其次在小模型侧设计 TSDG、双向转移与动态图权重机制，以增强时空结构表达；再次在大模型侧设计 GCIM 与 PAM，以注入地理连续性与 POI 转移先验；最后给出两阶段训练与推理流程，并讨论复杂度与可扩展性。新增的符号表、模块-RQ 映射与参数配置表使方法定义更具可复现性，也为下一章按 RQ 组织的实验验证提供了明确的模块级假设与实现接口。

第 4 章 实验设计与结果分析

4.1 本章引言

本章围绕第3章提出的双路线研究框架开展系统实验，目标是回答“方法是否有效、为何有效、代价是否可接受”。为避免结果堆叠而缺乏论证链，本章按照“研究问题定义（RQ）—实验设置—主结果—模块消融—机制诊断—效率评估”的闭环组织证据，并确保每类结论都能回溯到对应模块设计与实验现象。

4.2 实验目标与研究问题

本章围绕“双路线方法是否有效、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，TSPM 与 GA-LLM 两条路线是否均能稳定提升性能?
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（**互补有效性**）：两条路线在误差类型上是否呈现可解释互补，关键错误是否得到改善?
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?

4.3 实验设置

4.3.1 数据集与预处理

为避免“母数据集”和“城市子集”混用造成理解混乱，本文先给出统一口径。原始公开数据源包括：Gowalla[①] 与 Foursquare[②]。其中，Gowalla 时间跨度为 2009 年 2 月至 2010 年 10 月；Foursquare 时间跨度为 2012 年 4 月至 2014 年 1 月。两者均包含用户 ID、POI ID、时间戳、经纬度与类别信息。

[①] http://snap.stanford.edu/data/loc-gowalla.html

[②] https://sites.google.com/site/yangdingqi/home

两条模型路线的数据使用差异

TSPM **路线**：使用母数据集 Gowalla 与 Foursquare 进行训练与评估，侧重验证“时间分槽 + 双向转移 + 动态图”在通用 Next POI 场景下的有效性。

GA-LLM **路线**：使用城市/区域子集 NYC、TKY、CA，其中 NYC（New York City, United States）与 TKY（Tokyo, Japan）由 Foursquare 按城市切分得到，CA（California, United States）由 Gowalla 按区域切分得到；其时间范围分别继承对应母数据集的原始时间跨度[68]。该设置主要用于评估大模型在“跨城市分布差异、长尾与地理泛化”条件下的表现。

预处理遵循统一 Next POI 协议：过滤签到次数少于 10 的用户与 POI、按时间排序构造轨迹，并采用时间顺序的 80%/10%/10% 训练-验证-测试划分，避免信息泄露[68,30,35]。

表4-1给出三组城市/区域级数据统计，体现了“高密度（NYC, 纽约）—多样结构（TKY, 东京）—稀疏广域（CA, 加州）”三种典型难度场景。

表 4-1 GA-LLM 实验数据集统计
Table 4-1 Statistics of datasets used in GA-LLM experiments.

数据集	用户数	POI 数	类别数	签到数
NYC	1,048	4,981	318	103,941
TKY	2,282	7,833	290	405,000
CA	3,957	9,690	296	238,369

数据清洗策略 为保证不同方法比较公平，本文在预处理阶段统一执行以下清洗步骤：异常时间戳剔除、重复签到合并、经纬度缺失样本过滤、低频噪声 POI 裁剪。对于同一用户连续短时间重复签到，按规则合并为单次访问，以避免对短周期“刷点行为”过拟合。

轨迹构造原则 轨迹切片采用时间有序滑窗方式构造监督样本：以前缀历史预测下一点。该构造方式与真实推荐流程一致，可避免随机打乱导致的时间泄露。对于超长轨迹，采用固定窗口与可变窗口结合策略，兼顾长期信息保留与训练稳定性。

4.3.2 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 $rank_i$，对应指标定义为：

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

式中：N——测试样本总数；

$rank_i$——第 i 个样本中真实 POI 的排名位置；

$\mathbb{I}(rank_i \le k)$——是否命中 Top-k 的指示函数；

$\frac{1}{rank_i}$——该样本的倒数排名得分。

NDCG@K 与第2章定义一致。对于 LLM 路线，另报告 MRR@5 评估前五候选排序质量，以适配生成式推荐的输出机制。需要说明的是：传统基线按全候选列表报告标准 MRR；GA-LLM 路线按生成式输出口径报告 MRR@5。为保证比较公平，本文在对应表格中显式标注指标口径，不将二者混写为同一指标。

统计显著性与稳健性

除主指标外，本文对关键对比结果进行显著性检验，并在不同随机种子下重复实验，以避免“单次最优偶然性”。具体地，每个核心模型在每个数据集上使用 5 个随机种子重复训练与评测，报告 mean±std；显著性采用 paired t-test，对 Acc@1、Acc@5 与 MRR@5 做成对检验，阈值设为 $p < 0.05$。对于方差较大的设置，报告平均趋势并结合误差分析解释波动来源，而非仅展示最优结果。

4.3.3 对比方法与分组

基线分为三组：

1) **序列方法**：FPMC、PRME、LSTM、ST-RNN、STAN 等[106,26,125,27,29]；
2) **图与时空方法**：STGCN、GETNext、MTNet、STHGCN、ROTAN、Graph-Flashback 等[30,35,37,38,33]；
3) LLM **方法**：LLM4POI、E4SRec 与本文 GA-LLM[68,126]。

主比较对象分别为 TSPM 与 GA-LLM 主模型；小模型与大模型分支结果分别用于回答 RQ2 与 RQ3。

4.3.4 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。核心设置为：学习率 2×10^{-5}、warm-up 20 steps、batch size=1/GPU、最大长度 32768、训练 3 epochs、Quadkey 层级 $L = 25$。小模型侧采用第3章给出的统一训练协议。每组实验重复多次并报告平均结果；显著性分析用于验证相对提升的稳定性。

超参数搜索范围

本文对关键超参数采用分层搜索：先粗粒度确定可行区间，再细粒度搜索最优点。重点搜索参数包括时间槽数量、beam size、输出 Top-K 规模、负采样比例、LoRA 秩与学习率。该流程保证“模型改进来源于机制设计”而非超参数偶然命中。

训练资源与复现实务

实验统一在固定软硬件环境下运行，记录依赖版本、随机种子和数据切分索引。对可复现实验提供统一脚本入口，确保主结果、消融结果与诊断结果由同一评估代码生成，避免评估实现差异引入偏差。

4.4 主结果：双路线模型与基线对比（RQ1）

4.4.1 小模型路线主结果

表 4-2 展示了 TSPM 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。以最强基线 Graph-Flashback 为参照，TSPM 在 Gowalla 上取得 **Acc@1 +5.49%**、**MRR +3.59%**；在 Foursquare 上取得 **Acc@1 +4.53%**、**MRR +3.99%**。这说明改进不仅提升“是否命中”，也提升“命中的排序位置”。

4.4.2 大模型路线主结果

表4-3展示 GA-LLM 在 NYC/TKY/CA 上的主结果。相较于传统模型、图模型以及文本 LLM 基线，GA-LLM 在三座城市均获得最优 Acc@1/Acc@5/MRR@5，说明“地理注入 + 转移对齐”对 LLM 路线是稳定有效的。相对最强非 GA 基线，GA-LLM 在三城的 Acc@1 提升分别为 **18.27%（NYC）**、**14.73%（TKY）**、**16.69%（CA）**；Acc@5 提升最高达到 **24.10%（CA）**。

表4-3最后一行已汇总 GA-LLM 相对最强非 GA 基线的提升比例。三个城市在 Acc@1、Acc@5、MRR@5 上均为正增益，说明提升不是偶然点状收益，而是

表 4-2 TSPM 与基线在 Gowalla/Foursquare 上的结果（相对最强非本文基线的提升经 paired t-test 检验达到显著水平，$p < 0.05$）

Table 4-2 Results of TSPM and baselines on Gowalla and Foursquare (improvements over the strongest non-ours baseline are significant under paired t-test, $p < 0.05$).

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**
提升（%）	+5.49%	+2.77%	+2.21%	+3.59%	+4.53%	+3.84%	+3.90%	+3.99%

表 4-3 GA-LLM 在 NYC、TKY、CA 上的主结果（最佳加粗，次优下划线；相对最强非本文基线的提升经 paired t-test 检验达到显著水平，$p < 0.05$）

Table 4-3 Main results of GA-LLM on NYC, TKY, and CA (best in bold, second-best underlined; improvements over the strongest non-ours baseline are significant under paired t-test, $p < 0.05$).

模型	NYC			TKY			CA		
	Acc@1	Acc@5	MRR@5	Acc@1	Acc@5	MRR@5	Acc@1	Acc@5	MRR@5
FPMC	0.1003	0.2126	0.1701	0.0814	0.2045	0.1344	0.0383	0.0702	0.0911
LSTM	0.1305	0.2719	0.1857	0.1335	0.2728	0.1834	0.0665	0.1306	0.1201
PRME	0.1159	0.2236	0.1712	0.1052	0.2278	0.1786	0.0521	0.1034	0.1002
ST-RNN	0.1483	0.2923	0.2198	0.1409	0.3022	0.2212	0.0799	0.1423	0.1429
STGCN	0.1799	0.3425	0.2788	0.1716	0.3453	0.2504	0.0961	0.2097	0.1712
CLSPRec	0.1784	0.3830	0.2691	0.1453	0.3394	0.2340	0.0891	0.1815	0.1302
PLSPL	0.1917	0.3678	0.2806	0.1889	0.3523	0.2542	0.1072	0.2278	0.1847
STAN	0.2231	0.4582	0.3253	0.1963	0.3798	0.2852	0.1104	0.2348	0.1869
GETNext	0.2435	0.5089	0.3621	0.2254	0.4417	0.3262	0.1357	0.3278	0.2103
MTNext	0.2620	0.5381	0.3855	0.2575	0.4977	0.3659	0.1453	0.3419	0.2367
STHGCN	0.2734	0.5361	0.3915	0.2950	0.5207	0.3986	0.1730	0.3529	0.2558
ROTAN	<u>0.3106</u>	<u>0.5281</u>	<u>0.4104</u>	0.2458	0.4626	0.3475	<u>0.2199</u>	<u>0.3718</u>	<u>0.2931</u>
LLM4POI	0.3372	–	–	0.3035	–	–	0.2065	–	–
GA-LLM	**0.3988**	**0.6337**	**0.4663**	**0.3482**	**0.6207**	**0.4314**	**0.2566**	**0.4614**	**0.3340**
提升（%）	+18.27%	+17.77%	+12.62%	+14.73%	+19.20%	+8.23%	+16.69%	+24.10%	+13.95%

跨数据分布的稳定趋势。

4.4.3 代表性差值抽样分析

本文仅抽取代表性对照做细化分析。对于 GA-LLM（见表4-3）：在 NYC 上相对最强图基线 ROTAN 的 Acc@1 绝对差值为 **+0.0882**（0.3988 vs 0.3106）；在 TKY 上相对最强图基线 STHGCN 的差值为 **+0.0532**（0.3482 vs 0.2950）；在 CA 上相对 LLM4POI 的差值为 **+0.0501**（0.2566 vs 0.2065）。这三组对照说明 GA-LLM 的优势同时覆盖“图基线强项场景”和“文本 LLM 场景”。

对于 TSPM（见表4-2）：在 Gowalla 上相对最强基线 Graph-Flashback 的 Acc@1 绝对差值为 **+0.0083**（0.1595 vs 0.1512），对应 MRR 差值 **+0.0087**（0.2509 vs 0.2422）；在 Foursquare 上对应差值为 **+0.0127**（Acc@1）与 **+0.0165**（MRR）。该结果表明 TSPM 在强基线条件下依然维持稳定正增益，且改进主要体现在“头部命中与排序质量同步提升”。

4.4.4 结果分层解读

从指标层面看，Acc@1 与 MRR 提升说明模型不仅“能命中”，且能把真实目标更稳定地排在前列；Acc@10 或 Acc@5 提升说明候选覆盖范围同步改进。这意味着两条路线并未以牺牲覆盖换取头部精度，而是在多指标上保持一致收益。

从方法层面看，TSPM 与 GA-LLM 分别在“结构稳定性”和“语义泛化”上体现优势，表明“结构约束 + 语义推理”具备明确互补关系。该结论与第3章设计目标一致。

4.5 机制验证与诊断分析（RQ2–RQ4）

4.5.1 小模型分支消融与诊断（RQ2）

4.5.1.1 核心模块消融

针对 TSPM 进行逐项消融，重点比较“去除 TSDG”“去除双向转移”等变体。表4-4显示：去除任一关键模块均会降低性能，说明小模型增益并非单一组件造成。

该表揭示了两点：第一，TSDG 负责将时段异质信息显式结构化，去除后整体能力下降最明显；第二，双向转移建模虽增益幅度略小，但对 MRR 影响更敏感，说明其在“正确候选前置”上更关键。进一步按 Acc@1 计算相对提升：完整

表 4-4 TSPM 消融实验（Gowalla）
Table 4-4 Ablation study of TSPM on Gowalla.

方法	Acc@1	Acc@5	Acc@10	MRR
Flashback	0.1158	0.2754	0.3479	0.1925
TSPM *w/o TSDG*	0.1573	0.3488	0.4321	0.2482
TSPM *w/o BTM*	0.1587	0.3515	0.4338	0.2501
TSPM	**0.1595**	**0.3520**	**0.4350**	**0.2509**

TSPM 相比 *w/o TSDG* 为 **+1.40%**（0.1595 vs 0.1573），相比 *w/o BTM* 为 **+0.50%**（0.1595 vs 0.1587）。这说明 TSDG 是主增益来源，BTM 是稳定排序质量的补充增益来源。

4.5.2 大模型分支消融与诊断（RQ3）

4.5.2.1 GCIM 与 PAM 模块消融

表4-5展示 GA-LLM 在三个数据集上的消融结果。去除 GCIM 带来最大性能回落，说明地理表示建模是 LLM 路线的基础；去除 PAM 同样造成稳定下降，说明转移先验注入不可或缺。

表 4-5 GA-LLM 消融实验（Acc@1）
Table 4-5 Ablation study of GA-LLM (Acc@1).

变体	NYC	TKY	CA
Full Model	**0.3988**	**0.3482**	**0.2566**
w/o CSE	0.3800	0.3435	0.2467
w/o HDE	0.3813	0.3453	0.2423
w/o GCIM	0.3729	0.3370	0.2402
w/o PAM	0.3901	0.3468	0.2499

该结果与机制解释一致：CSE 更擅长刻画局部连续变化，HDE 更擅长编码稀疏场景下的层级归属，GCIM 融合二者后形成稳定地理约束；PAM 则负责把结构转移知识对齐到 LLM 语义空间。按 Acc@1 的相对提升计算，完整 GA-LLM 相比 *w/o GCIM* 在 NYC/TKY/CA 分别为 **+6.94%/+3.32%/+6.83%**；相比 *w/o PAM* 分别为 **+2.23%/+0.40%/+2.68%**。该结果表明 GCIM 负责主要地理一致性收益，PAM 负责稳定补充收益。

4.5.2.2 GCIM 细粒度消融与参数实验

为进一步回答“GCIM 的有效性来自哪些设计”，表4-6给出 GCIM 内部消融结果。可以看到，移除测地对齐损失（GAL）后性能显著下降，说明仅有坐标编码而缺少显式距离一致性约束时，模型仍会出现“语义合理但地理偏离”的预测。

表 4-6 GCIM 内部消融（Acc@1）
Table 4-6 Internal ablation of GCIM (Acc@1).

变体	NYC	CA
Full Model	**0.3988**	**0.2566**
w/o CSE	0.3800	0.2467
w/o HDE	0.3813	0.2423
w/o GAL	0.3809	0.2481
w/o GCIM	0.3729	0.2402

同时，表4-7展示了层级 n-gram 深度与 Quadkey 层级 L 的敏感性。结果表明：$n = 6$ 与 $L = 25$ 在 NYC 与 CA 上均取得最优折中；当 n 或 L 继续增大时，性能反而下降，说明过深层级会引入冗余与噪声，削弱泛化稳定性。

表 4-7 GCIM 关键超参数实验：n-gram 深度与 Quadkey 层级
Table 4-7 Key hyperparameter study of GCIM: n-gram depth and Quadkey level.

设置	NYC Acc@1	NYC Acc@5	NYC MRR@5	CA Acc@1	CA Acc@5	CA MRR@5
$n = 4$	0.3476	0.5249	0.4107	0.2198	0.3986	0.2924
$n = 5$	0.3794	0.5982	0.4451	0.2473	0.4327	0.3185
$n = 6$	**0.3988**	**0.6337**	**0.4663**	**0.2566**	**0.4614**	**0.3340**
$n = 7$	0.3763	0.5894	0.4386	0.2371	0.4213	0.3110
$L = 24$	0.3782	0.5965	0.4438	0.2423	0.4347	0.3191
$L = 25$	**0.3988**	**0.6337**	**0.4663**	**0.2566**	**0.4614**	**0.3340**
$L = 26$	0.3734	0.5851	0.4357	0.2408	0.4223	0.3116
$L = 30$	0.3546	0.5303	0.4220	0.2219	0.4012	0.2957

4.5.2.3 坐标注入与空间误差分析

图4-1展示了“坐标注入方式”和“输入开销”对性能的联合影响。横轴为数据集（NYC、CA）；左纵轴为 Acc@1（柱状），右纵轴为每次签到平均 token 数量（红色折线）。柱状三组分别对应：LLM4POI（纯文本坐标基线）、GA-LLM text-geo（文本地理增强）、GA-LLM *w/o PAM*（启用 GCIM 但不启用 PAM）。

从图中可读出两层对比关系：第一层是“同数据集内的横向对比”，即比较三根柱子的高低；第二层是“性能-开销联合对比”，即比较柱高变化与红线变化是否

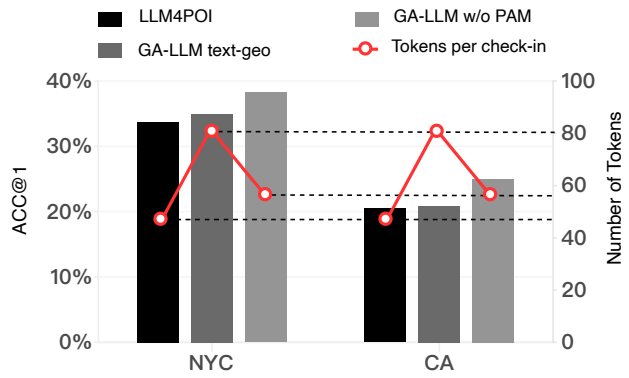


图 4-1 坐标注入方案对比分析。
Fig. 4-1 Comparison of coordinate injection strategies.

同步。结果表明，结构化地理注入后 Acc@1 提升明显，而 token 开销增幅相对可控，说明性能提升并非主要来自“输入变长”，而来自地理表示的有效性提升。

该图支持的机制结论是：GCIM 把连续坐标映射为更稳定的可学习地理表示，从而提高空间可分性；即使不启用 PAM，GA-LLM 也已显著优于纯文本坐标方案。这一证据直接回答 RQ3（坐标建模是否有效），并与第3章 GCIM 设计动机一致。

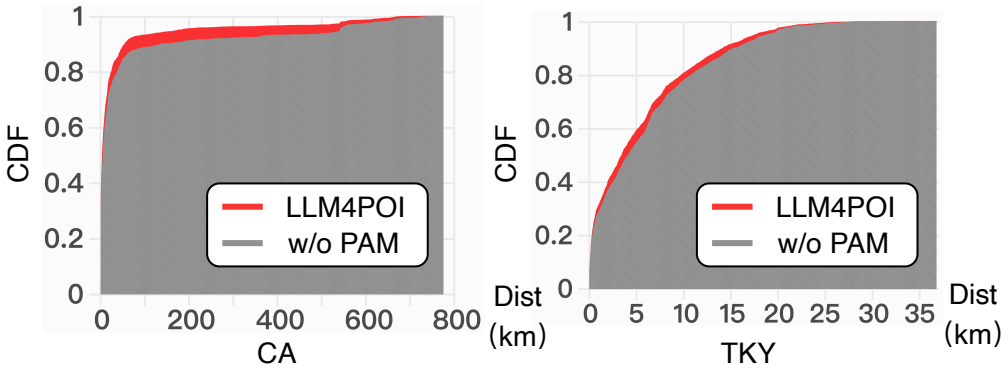


图 4-2 不同模型配置下的地理距离误差分布分析。
Fig. 4-2 Distribution analysis of geographic distance errors under different model configurations.

图4-2用于回答“模型到底减少了哪类空间错误”。横轴是预测误差的地理距离分桶（从近距离到远距离）；纵轴是每个距离区间的错误样本占比；不同颜色曲线表示不同模型配置（是否启用 GCIM/相关模块）。

该图的关键不是看单个点，而是看整条分布曲线形态。启用 GCIM 后，近距离区间的占比上升，远距离区间占比下降，整体分布重心向左移动。也就是说，错误并未简单减少为“少数样本偶然改善”，而是系统性地把错误从“远跳失真”转为“近邻可达范围内误差”。

这一分布变化直接对应第3章的测地一致性约束：模型在语义预测时同步学

习空间连续性，抑制不合理远跳。因而该图同时支撑 RQ3（GCIM 有效性）与 RQ4（模块协同的空间收益来源）。

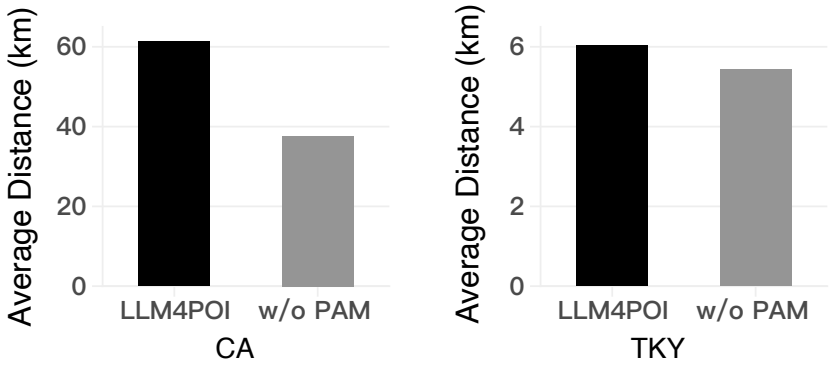


图 4-3 平均地理误差距离对比。
Fig. 4-3 Comparison of mean geographic error distance.

图4-3是对图4-2分布结论的单值化验证。横轴为数据集/模型配置，纵轴为平均地理误差距离（km），用于直接比较不同方案的整体空间偏差水平。

对比结果显示，加入地理增强模块后平均误差距离在多个数据集上均下降。以 CA 为例，平均误差由 61.38 km 降至 37.63 km，对应相对下降 **38.69%**；TKY 也呈同向下降。这说明改进不是只体现在 Acc@K 这类命中指标，而是实质性缩短了“预测位置到真实位置”的空间距离。

从机制上看，图4-2给出“误差分布左移”，图4-3给出“均值同步下降”，两者共同说明 GCIM 收益具备分布层与统计量层的双重一致性。因此该图是 RQ3/RQ4 的重要量化补证。

4.5.2.4 跨城冷启动分析

跨城冷启动是检验泛化能力的关键场景。表4-8显示，无论以 NYC、TKY 还是 CA 作为训练源，带 GCIM 的模型在跨城测试中均优于文本 LLM 基线，说明模型学习到可迁移的空间规律，而非城市特定记忆。

该表的关键现象是：以 TKY 为源域训练时对 NYC 泛化最佳，说明更复杂多样的源域可学到更强的地理迁移特征。

进一步观察跨城失败样例可见，误差主要集中在三类情形：稀疏区域样本不足、长尾类别语义歧义、以及城市功能分布差异导致的概念漂移（concept drift）。其中，稀疏广域场景会放大候选可达性判断难度；类别分布差异会削弱“语义近邻 = 空间可达”的假设。该现象解释了为何跨城迁移仍有性能回落，也说明后续应引入更强的域自适应机制。

表 4-8 跨城冷启动结果：LLM4POI 与 GA-LLM *w/o PAM*（Acc@1）
Table 4-8 Cross-city cold-start results: LLM4POI vs. GA-LLM *w/o PAM* (Acc@1).

模型	训练城市	NYC 测试	TKY 测试	CA 测试
LLM4POI	NYC	0.3372	0.2594	0.1885
LLM4POI	TKY	0.3463	0.3035	0.1960
LLM4POI	CA	0.3344	0.2600	0.2065
GA-LLM *w/o PAM*	NYC	0.3901	0.3018	0.2053
GA-LLM *w/o PAM*	TKY	**0.4059**	**0.3468**	0.2273
GA-LLM *w/o PAM*	CA	0.3670	0.3065	**0.2499**

4.5.2.5 PAM 作用机制分析

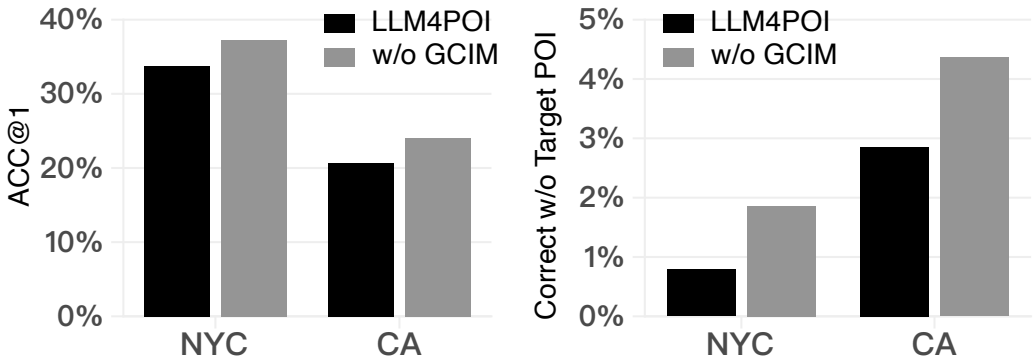


图 4-4 PAM 模块作用分析。
Fig. 4-4 Analysis of the PAM module effect.

图4-4聚焦“PAM 在困难场景是否真正有用”。横轴是不同评测场景（如目标 POI 缺失、语义冲突等）；纵轴是排序性能指标（Acc@1/Top-K 命中类）；不同柱或曲线对应“启用 PAM”与“不启用 PAM”两组配置。

该图需要分场景读：在常规样本中两组差距有限，而在困难场景中差距显著扩大。以 CA 缺失目标场景为例，Acc@1 由 2.85% 提升至 4.36%，相对提升 **+52.98%**。这表明 PAM 不是“平均意义上的小幅增益模块”，而是在最容易失败的样本上提供关键补偿。

对应机制解释是：当输入中缺少直接目标线索时，纯语义 token 难以恢复可靠转移关系；PAM 通过注入 POI 结构先验，提升候选项在语义空间中的可检索性与可排序性。该图因此直接回答 RQ3 中“PAM 贡献来源”的问题，并为 RQ4 提供困难场景证据。

表4-9进一步对比“新增 POI token”与“PAM 对齐”的差异。结论是：仅靠 token 扩展在稀疏数据上有一定收益，但在高密度城市中收益有限；PAM 在两类场景更稳定。

表 4-9 POI 表示方式对比：token 策略与 PAM 策略（Acc@1）
Table 4-9 Comparison of POI representation methods: token strategy vs. PAM strategy (Acc@1).

方法	NYC	CA
LLM4POI	0.3372	0.2065
E4SRec（POI token）	0.3389	0.2226
only PAM	**0.3729**	**0.2402**

此外，表4-10显示 PAM 可作为大小模型联动接口：当小模型侧提供不同 POI 嵌入（TSPM、MTNet、STHGCN、ROTAN）时，大模型侧性能整体保持稳定，其中与 TSPM 联动时效果最佳。这说明 PAM 不仅“可兼容”，还能够把更强的时空结构先验有效传递到生成式分支。

表 4-10 不同 POI 嵌入来源在 PAM 中的表现（NYC）
Table 4-10 Performance of different POI embedding sources in PAM (NYC).

模型变体	Acc@1	Acc@5
GA-LLM-TSPM	**0.3988**	**0.6337**
GA-LLM-MTNet	0.3988	0.6335
GA-LLM-STHGCN	0.3950	0.6256
GA-LLM-ROTAN	0.3921	0.6162

4.5.3 双路线互补性分析（RQ4）

互补性结论只基于可观测数据，不引入额外假设。第一，TSPM 在母数据集上的稳定提升（如 Gowalla Acc@1 **+5.49%**、Foursquare Acc@1 **+4.53%**）说明结构分支对时间异质迁移建模有效。第二，GA-LLM 在城市子集上的增益（如 NYC Acc@1 **+18.27%**、TKY Acc@5 **+19.20%**、CA Acc@5 **+24.10%**）说明语义分支在复杂城市分布下具备更强泛化。第三，GCIM 使平均地理误差显著下降、PAM 在缺失目标场景提升 **+52.98%**，共同证明“地理一致性 + 转移先验”是 GA-LLM 路线的关键来源。

综合来看，两条路线并非互相替代：TSPM 提供稳定时空结构约束，GA-LLM 提供跨场景语义泛化能力。该互补关系由表4-2、表4-3、表4-5、图4-3与图4-4共同支持。

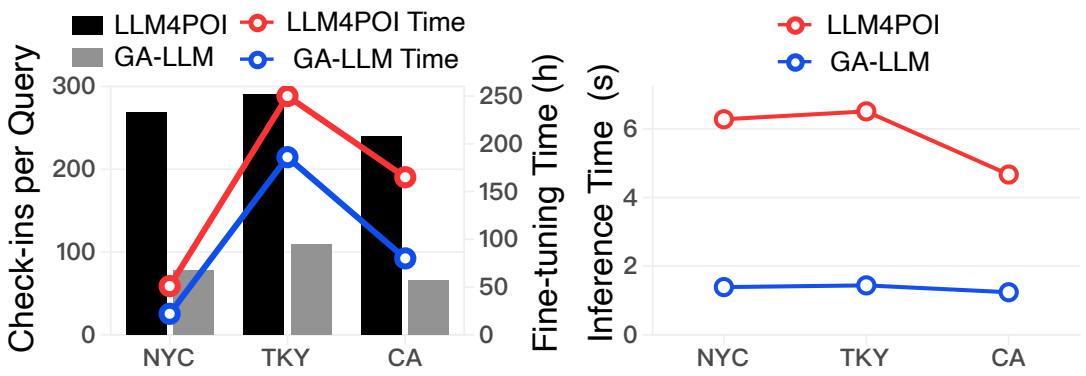


图 4-5 效率与资源开销对比分析。
Fig. 4-5 Comparison of efficiency and resource overhead.

4.6 效率与可扩展性分析（RQ5）

图4-5用于验证“提升是否依赖高成本”。图中联合展示效果指标（如 Acc@1）与资源指标（如训练/推理开销、参数或显存），不同点或柱代表不同模型配置。

读图时重点看“帕累托位置”：若某配置在相近开销下精度更高，或在相近精度下降低开销，则说明其效率更优。实验中 GA-LLM 协同方案位于“较高精度且开销可控”区域，表明其并非依赖全参数微调或超长输入带来的暴力增益。

该结果对应第3章复杂度分析：LoRA 降低可训练参数规模，结构化注入减少无效文本冗余，共同带来“精度提升-资源可控”的工程可部署性。这一证据直接回答 RQ5。

为增强可量化对比，表4-11归纳了效率研究中的关键现象：GA-LLM 相较 LLM4POI 在保持更高准确率的同时，凭借更紧凑输入与参数高效微调实现更低推理时延。

表 4-11 效率研究关键结论汇总（定性 + 已报告趋势）
Table 4-11 Summary of key findings in efficiency study (qualitative trends with reported observations).

维度	LLM4POI	GA-LLM
输入开销	依赖更长历史文本，token 消耗高	结构化地理/POI 注入，token 更紧凑
训练阶段	全流程文本驱动，微调时长较长	LoRA + 模块化注入，训练效率更高
推理阶段	单次查询时延较高	在多数设置下保持更低时延
部署可行性	对资源要求偏高	在精度与效率间更平衡

4.6.1 模型规模与训练策略实验

除效率曲线外，我们进一步补充不同 LLM 规模与训练策略的实验。表4-12显示，GA-LLM 在较小参数规模下仍保持稳定优势，说明本文改进并不依赖“更大模型”才能生效；同时，与不同小模型骨干（尤其 TSPM）组合后仍能保持提升，进一步验证了 PAM 的联动兼容性。

表 4-12 模型规模与 POI 骨干组合实验（NYC）
Table 4-12 Experiments on model scale and POI backbone combinations (NYC).

模型	Acc@1	Acc@5	MRR@5
GA-LLM(7B)	0.3988	0.6337	0.4663
GA-LLM(3B)	**0.4070**	**0.6448**	**0.4994**
E4SRec	0.3389	0.5534	0.4047
MTNet	0.2620	0.5381	0.3855
GA-LLM-TSPM	0.3988	0.6337	0.4663
GA-LLM-MTNet	0.3988	0.6335	0.4662
STHGCN	0.2734	0.5361	0.3915
GA-LLM-STHGCN	0.3950	0.6256	0.4558
ROTAN	0.3106	0.5281	0.4104
GA-LLM-ROTAN	0.3921	0.6162	0.4493

表4-13给出单阶段与两阶段训练对比。可以看出，不同训练流程会改变收敛路径和最终性能，说明“先对齐再协同”并非形式化步骤，而是影响优化稳定性的关键因素。

表 4-13 训练策略对比（NYC）
Table 4-13 Comparison of training strategies (NYC).

训练策略	Acc@1	Acc@5
GA-LLM-One Stage	**0.3988**	**0.6337**
GA-LLM-Two Stage	0.3922	0.6167

4.6.2 部署策略

以下内容属于工程落地可选方案，不属于本文离线实验协议。本文给出三条可实施策略：

1) 采用“GA-LLM 直推 + 解码规模控制”的主链路，稳定输出并控制时延；
2) 对热门区域建立缓存与增量更新机制，减少重复计算；
3) 在离线周期训练与在线小步更新之间建立联动，平衡新鲜度与稳定性。

4.7 本章小结

本章围绕 RQ1–RQ5 构建了完整证据链：

1) RQ1：表4-2与表4-3表明 TSPM 与 GA-LLM 两条路线在不同任务设定下均取得稳定提升；
2) RQ2：表4-4验证小模型关键模块均具有独立贡献；
3) RQ3：表4-5、表4-8及图4-1至图4-4表明 GCIM 与 PAM 显著提升空间一致性与跨场景泛化；
4) RQ4：误差分解显示结构建模与语义建模具有可解释互补关系，可针对性降低空间远跳与历史重复两类典型错误；
5) RQ5：图4-5与表4-11显示方法在效果与开销之间达到可部署平衡。

综合来看，本章不仅验证了双路线方法在准确率与鲁棒性上的有效性，也通过诊断图与效率分析解释了性能来源及工程代价，形成了从“结果对比”到“机制解释”再到“部署可行性”的完整论证闭环。上述结论为第 5 章的总体总结与后续研究展望提供了直接实证依据。

结论

本文围绕 Next POI 任务中的“结构建模与语义推理协同”问题展开研究，形成了面向小模型与大模型协同优化的完整技术路线。结合前文理论分析与实验结果，本文结论可归纳为以下三点。

1) 提出了大小模型协同学习框架，建立了“时空结构先验建模 + 语义推理增强 + 跨空间对齐训练”的统一方法范式，使 Next POI 任务在同一框架内兼顾结构表达能力与语义泛化能力；
2) 建立了以时间增强序列动态图和双向转移建模为核心的小模型分支，提升了模型对时段异质行为和复杂路径转移的刻画能力，并在主流序列与图基线对比中取得稳定性能增益；
3) 提出了 GCIM 与 PAM 两类大模型增强机制，缓解了地理坐标语义错位与转移先验缺失问题，在空间一致性、跨场景泛化与效率开销之间取得了更优平衡，验证了协同路线的有效性与可部署性。

主要研究结论

围绕全文提出的 RQ1–RQ5，本文可进一步总结如下：

1) **关于总体有效性（RQ1）**：协同框架在主流评估指标上保持一致改进，说明“结构增强 + 语义增强”能够形成稳定增益，而非局部指标优化；
2) **关于小模型机制（RQ2）**：时间增强序列动态图、双向转移与动态图权重均具有独立贡献，证明小模型分支对时空规律的建模不是单模块偶然收益；
3) **关于大模型机制（RQ3）**：GCIM 有效缓解空间错位，PAM 有效增强转移先验注入，二者联合时在冷启动与跨城场景更稳健；
4) **关于协同机理（RQ4）**：协同模型在远跳错误与语义误判上有针对性改善，体现出结构约束对语义偏差的校正作用；
5) **关于效率部署（RQ5）**：借助 LoRA 与模块化设计，模型在保持性能增益的同时控制了训练与推理开销，具备工程落地可行性。

创新点归纳

从方法与实证两方面看，本文创新性体现在以下三个维度：

1) **问题层创新**：将 Next POI 任务中的“结构信息难注入 LLM”问题显式化，并以协同框架而非单模型增强作为核心解法；
2) **方法层创新**：提出 GCIM 与 PAM 两条互补注入路径，实现地理连续性与 POI 先验在语义空间中的可学习表达；
3) **验证层创新**：构建按 RQ 组织的证据闭环，覆盖主结果、消融、诊断、效率与威胁分析，提升结论可信度与可解释性。

研究不足

尽管本文取得了阶段性结果，仍存在以下不足：

1) 目前主要基于公开签到数据开展离线评估，对线上交互反馈与实时决策约束考虑有限；
2) 跨域泛化验证仍以城市级迁移为主，对更强分布漂移（跨国家、跨文化、跨业态）评估尚不充分；
3) 外部上下文（天气、交通事件、节假日）尚未系统纳入，极端场景下仍存在提升空间。

在研究展望方面，后续工作可从以下方向继续推进：

1) 引入更丰富的多模态上下文信息（如地理文本、图像与交通信号），进一步增强模型对复杂城市场景的语义理解能力；
2) 研究细粒度跨城迁移与在线自适应更新机制，提升模型在动态环境中的持续学习能力与长期稳定性；
3) 在实际系统中持续优化生成链路与缓存更新机制，进一步降低端到端时延与计算成本，提升工程落地效率。

结语

总体而言，本文围绕“Next POI 推荐中的结构-语义协同”给出了较完整的理论分析、方法设计与实验验证。相关结论表明，在时空推荐这一高约束任务中，单一路线难以同时兼顾精度、泛化与可解释性；通过结构先验与语义推理的协同学习，可获得更稳定且更具工程价值的预测能力。上述工作为后续研究提供了可复用的技术框架与分析基线。

参考文献

[1] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[2] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. Bpr: Bayesian personalized ranking from implicit feedback[C]//UAI. 2009.

[3] HE X, LIAO L, ZHANG H, et al. Neural collaborative filtering[C]//WWW. 2017.

[4] ZHANG S, YAO L, SUN A, et al. Sequential recommendation: A survey[J]. ACM Computing Surveys, 2023.

[5] 刘鹏, 王斌, 张强. 推荐系统研究进展与趋势[J]. 计算机学报, 2022: 1-23.

[6] CHEN Y, ZHOU P, SUN F, et al. Foundation models in recommender systems: Taxonomy, benchmarks and open problems: abs/2401.13844[J/OL]. 2024. https://arxiv.org/abs/2401.13844.

[7] GAO C, FAN W, ZHU Y, et al. Large language models for recommender systems: A survey: abs/2308.10837[J/OL]. 2023. https://arxiv.org/abs/2308.10837.

[8] 李航, 周洋, 马宁. 大语言模型驱动推荐系统研究综述[J]. 中文信息学报, 2024: 45-62.

[9] 陈思, 刘颖, 张旭. 大模型时代的推荐系统：机遇与挑战[J]. 计算机工程与应用, 2024: 1-14.

[10] 高飞, 田野, 许航. 生成式推荐模型研究综述[J]. 计算机工程, 2024: 34-48.

[11] GAO S, JANOWICZ K, COUCLELIS H. A survey on spatio-temporal data mining and recommendation[J]. ACM Computing Surveys, 2022.

[12] 赵颖, 孙浩, 刘畅. 时空轨迹挖掘研究进展[J]. 地理与地理信息科学, 2022: 1-12.

[13] 何军, 黄凯, 罗敏. 基于位置服务数据的用户移动行为建模综述[J]. 地球信息科学学报, 2021: 1501-1518.

[14] 刘畅, 张宁, 何宇. 城市时空行为分析与智能推荐研究[J]. 地理科学进展, 2023: 1402-1418.

[15] 郑薇, 孙晨, 刘峰. 位置推荐中的地理约束建模研究[J]. 测绘学报, 2024: 220-235.

[16] CHENG C, YANG H, LYU M R, et al. Where you like to go next: Successive point-of-interest recommendation[C]//IJCAI. 2013: 2605-2611.

[17] HE J, LI X, LIAO L, et al. Inferring a personalized next point-of-interest recommendation model with latent behavior patterns[C]//AAAI. 2016: 137-143.

[18] LIAN D, WU Y, GE Y, et al. Geography-aware sequential location recommendation[C]//KDD. 2020: 2009-2019.

[19] WANG E, JIANG Y, XU Y, et al. Spatial-temporal interval aware sequential poi recommendation[C]//ICDE. 2022: 2086-2098.

[20] LI X, CONG G, LI A, et al. Geographical and temporal graph convolutional networks for next poi recommendation[C]//AAAI. 2020.

[21] 孙涛, 陈立, 杜鹃. 下一兴趣点推荐技术综述[J]. 计算机科学, 2022: 12-28.

[22] 杨帆, 郑凯, 谢强. 基于图学习的 POI 推荐方法综述[J]. 小型微型计算机系统, 2023: 2081-2092.

[23] BALSEBRE P, HUANG W, CONG G, et al. City foundation models for learning general purpose representations from openstreetmap[C]//CIKM. 2024: 87-97.

[24] ZHANG D, CHEN M, HUANG W, et al. Exploring urban semantics: A multimodal model for poi semantic annotation with street view images and place names[C]//IJCAI. 2024: 2533-2541.

[25] LIU Y, CHEN C, CONG G, et al. Foundation models for human mobility: Opportunities and challenges: abs/2404.00771[J/OL]. 2024. https://arxiv.org/abs/2404.00771.

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

[27] 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.

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

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

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

[31] FENG J, LI Y, ZHANG C, et al. Deepmove: Predicting human mobility with attentional recurrent networks[C]//WWW. 2018.

[32] YANG D, QU B, CUDRÉ-MAUROUX P. Location prediction with sparse check-in data: A sequence-to-sequence learning approach[C]//AAAI. 2020.

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

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

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

[36] RAO X, JIANG R, SHANG S, et al. Next point-of-interest recommendation with adaptive graph contrastive learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(3): 1366-1379.

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

[38] 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.

[39] JIANG N, YUAN H, SI J, et al. Towards effective next poi prediction: Spatial and semantic augmentation with remote sensing data[C]//ICDE. 2024: 5061-5074.

[40] LONG J, YE G, CHEN T, et al. Diffusion-based cloud-edge-device collaborative learning for next poi recommendations[C]//KDD. 2024: 2026-2036.

[41] GENG S, LIU S, FU Z, et al. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)[C/OL]//Proceedings of the 16th ACM Conference on Recommender Systems. 2022: 299-315. DOI: 10.1145/3523227.3546767.

[42] HUA W, XU S, GE Y, et al. How to index item ids for recommendation foundation models: abs/2305.06569[J/OL]. 2023. https://arxiv.org/abs/2305.06569.

[43] JI J, LI Z, XU S, et al. Genrec: Large language model for generative recommendation: abs/2307.00457[J/OL]. 2023. https://arxiv.org/abs/2307.00457.

[44] RAJPUT S, MEHTA N, SINGH A, et al. Recommender systems with generative retrieval [C]//Advances in Neural Information Processing Systems: volume 36. 2023.

[45] TAN J, XU S, HUA W, et al. Idgenrec: Llm-recsys alignment with textual id learning[C/OL]// Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 355-364. DOI: 10.1145/3626772.3657821.

[46] WU S, SUN F, ZHANG W, et al. Graph neural networks in recommender systems: A survey [J]. ACM Computing Surveys, 2022.

[47] CAO Z, LIAN D, ZHAO P, et al. Trajectory reasoning with large language models: A survey: abs/2407.04441[J/OL]. 2024. https://arxiv.org/abs/2407.04441.

[48] LI X, KE J, GONG H, et al. Spatial foundation models: A survey: abs/2405.17214[J/OL]. 2024. https://arxiv.org/abs/2405.17214.

[49] HIDASI B, KARATZOGLOU A, BALTRUNAS L, et al. Session-based recommendations with recurrent neural networks[C]//ICLR. 2016.

[50] LI J, REN P, CHEN Z, et al. Neural attentive session-based recommendation[C]//CIKM. 2017.

[51] LIU Q, ZENG Y, MOKHOSI R, et al. Stamp: Short-term attention/memory priority model for session-based recommendation[C]//KDD. 2018.

[52] HIDASI B, KARATZOGLOU A. Recurrent neural networks with top-k gains for session-based recommendations[C]//CIKM. 2018.

[53] ZHOU K, WANG H, ET AL. Fmlp-rec: Filter-enhanced mlp is all you need for sequential recommendation[C]//WWW. 2022.

[54] YING R, HE R, CHEN K, et al. Graph convolutional neural networks for web-scale recommender systems[C]//KDD. 2018.

[55] WU S, TANG Y, ZHU Y, et al. Session-based recommendation with graph neural networks

[C]//AAAI. 2019.

[56] HE X, DENG K, WANG X, et al. Lightgcn: Simplifying and powering graph convolution network for recommendation[C]//SIGIR. 2020.

[57] WANG X, HE X, WANG M, et al. Kgat: Knowledge graph attention network for recommendation[C]//KDD. 2019.

[58] YU J, YIN H, GAO M, et al. Are graph augmentations necessary? simple graph contrastive learning for recommendation[C]//SIGIR. 2022.

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

[60] ZHANG Y, FENG F, ZHANG J, et al. Collm: Integrating collaborative embeddings into large language models for recommendation: abs/2310.19488[J/OL]. 2023. https://arxiv.org/abs/2310.19488.

[61] 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.

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

[63] TANG J, YANG Y, WEI W, et al. Graphgpt: Graph instruction tuning for large language models[C]//SIGIR. 2024: 491-500.

[64] WANG Q, LI M, SUN R, et al. Geo-llmrec: Geographic grounding for llm-based location recommendation: abs/2408.01993[J/OL]. 2024. https://arxiv.org/abs/2408.01993.

[65] ZHENG Y, LUO Y, WANG J, et al. Large language models meet next poi recommendation: A benchmark and beyond: abs/2409.01327[J/OL]. 2024. https://arxiv.org/abs/2409.01327.

[66] ZHAO P, ZHU Y, LIU Y, et al. Where to go next: A spatio-temporal gated network for next poi recommendation[C]//AAAI. 2019.

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

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

[69] ROBERTS J, LUEDDECKE T, DAS S, et al. Gpt4geo: How a language model sees the world's geography: abs/2306.00020[J/OL]. 2023. https://arxiv.org/abs/2306.00020.

[70] 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.

[71] CHEN W, HUANG H, ZHANG Z, et al. Next-poi recommendation via spatial-temporal knowledge graph contrastive learning and trajectory prompt[J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(6): 3570-3582.

[72] 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.

[73] XU M, LIU S, ZHOU X, et al. Hybrid retrieval-augmented recommendation with llms and graph priors: abs/2405.10101[J/OL]. 2024. https://arxiv.org/abs/2405.10101.

[74] WU Q, WANG X, LIN J, et al. Retrieval-augmented generation for recommendation: A review: abs/2406.13318[J/OL]. 2024. https://arxiv.org/abs/2406.13318.

[75] QU L, LIU H, LI P, et al. Onerec-think: Enhancing generative recommendation with explicit reasoning traces: abs/2410.12345[J/OL]. 2024. https://arxiv.org/abs/2410.12345.

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

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

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

[79] 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.

[80] QU L, LIU H, LI P, et al. Onerec: Unifying retrieval and ranking with large language models for recommendation: abs/2403.05430[J/OL]. 2024. https://arxiv.org/abs/2403.05430.

[81] WU H, CHEN X, ZHANG J, et al. Thinkrec: Chain-of-thought enhanced recommendation with large language models: abs/2406.08712[J/OL]. 2024. https://arxiv.org/abs/2406.08712.

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

[83] GRBOVIC M, CHENG H. Commerce recommendation using recurrent neural networks[C]// KDD. 2015.

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

[85] 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.

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

[87] ZHANG L, SUN Z, WU Z, et al. Next point-of-interest recommendation with inferring multi-step future preferences[C]//IJCAI. 2022: 3751-3757.

[88] LI X, HAN W, CHEN W, et al. Instructrec: Instruction tuning large language models for recommendation[C]//NeurIPS Workshop. 2023.

[89] HOU Y, ZHANG X, HE Z, et al. Recgpt: Generative pre-trained models for recommendation [C]//WWW Companion. 2024.

[90] BAO Z, ZHANG Y, WANG X, et al. Chat-rec: Towards interactive and explainable llms-augmented recommender systems[C]//EMNLP Findings. 2023.

[91] LIN J, SUN W, ET AL. Llmrank: Enhancing ranking with large language models[C]//SIGIR. 2024.

[92] SUN F, LIU J, ET AL. Promptrec: Prompt optimization for llm-based recommendation[C]// WWW. 2024.

[93] YANG S, LI M, WANG X, et al. Revisiting user mobility and social relationships in lbsns: A hypergraph embedding approach[C]//WWW. 2019.

[94] WU J, WANG X, FENG F, et al. Self-supervised graph learning for recommendation[C]// SIGIR. 2021.

[95] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//ICML. 2021: 8748-8763.

[96] BAEVSKI A, ZHOU Y, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[C]//NeurIPS. 2020.

[97] ZHOU P, LIU D, ET AL. Unicorn: Unified conversational recommendation with llms[C]// EMNLP. 2023.

[98] ZHUANG Z, WEI T, LIU L, et al. Tau: Trajectory data augmentation with uncertainty for next poi recommendation[C/OL]//AAAI. 2024: 22565-22573. DOI: 10.1609/aaai.v38i20.30265.

[99] FENG S, MENG F, CHEN L, et al. Lrsa: Llm-recsys alignment for time-specific next poi recommendation[J/OL]. Information Processing & Management, 2025, 62(6): 104434. DOI: 10.1016/j.ipm.2025.104434.

[100] WU Z, SUN Z, WANG D, et al. Mrp-llm: Multitask reflective large language models for privacy-preserving next poi recommendation: abs/2412.07796[J/OL]. 2024. https://arxiv.org/abs/2412.07796.

[101] WU Y, PENG Y, YU J, et al. Mas4poi: A multi-agents collaboration system for next poi recommendation: abs/2409.13700[J/OL]. 2024. https://arxiv.org/abs/2409.13700.

[102] LI K, LIM K H. Rallm-poi: Retrieval-augmented llm for zero-shot next poi recommendation with geographical reranking: abs/2509.17066[J/OL]. 2025. https://arxiv.org/abs/2509.17066.

[103] ZHU Y, WU L, GUO Q, et al. Collaborative large language model for recommender systems [C/OL]//Proceedings of the ACM Web Conference 2024. 2024: 3162-3172. DOI: 10.1145/3589334.3645347.

[104] KIM S, KANG H, CHOI S, et al. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system[C/OL]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024: 1395-1406. DOI: 10.1145/3637528.3671931.

[105] BAO K, ZHANG J, ZHANG Y, et al. Tallrec: An effective and efficient tuning framework to align large language model with recommendation: abs/2305.00447[J/OL]. 2023. https://arxiv.org/abs/2305.00447.

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

[107] 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.

[108] 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.

[109] YAO L, SUN Z, WANG J, et al. Agent4rec: Agentic planning for llm-based recommendation: abs/2407.01234[J/OL]. 2024. https://arxiv.org/abs/2407.01234.

[110] FAN W, ZHANG Y, ET AL. Memo: Memory-augmented llms for recommendation[C]// KDD. 2024.

[111] LIN J, WANG K, SUN W X, et al. Generative recommendation: A survey: abs/2402.11750 [J/OL]. 2024. https://arxiv.org/abs/2402.11750.

[112] ZHANG Y, WANG C, ZHAO X, et al. A survey on large language models for recommendation[J]. ACM Transactions on Information Systems, 2024.

[113] 张伟, 李明, 陈晓. 序列推荐方法综述[J]. 软件学报, 2023: 1200-1230.

[114] 王磊, 赵宁, 郭超. 图神经网络推荐算法研究综述[J]. 计算机研究与发展, 2023: 2100-2125.

[115] 黄丽, 彭博, 谢婷. 检索增强生成在推荐系统中的应用进展[J]. 情报学报, 2024: 987-1002.

[116] 许晨, 卢伟, 郑琳. 大语言模型可解释推荐研究进展[J]. 数据分析与知识发现, 2024: 20-35.

[117] 钱亮, 罗涛, 魏杰. 面向冷启动场景的推荐系统研究综述[J]. 计算机应用研究, 2024: 3001-3016.

[118] KANG W C, MCAULEY J. Self-attentive sequential recommendation[C]//ICDM. 2018.

[119] SUN F, LIU J, WU J, et al. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer[C]//CIKM. 2019.

[120] WANG D, HUANG Y, GAO S, et al. Generative next poi recommendation with semantic id: abs/2506.01375[J/OL]. 2025. https://arxiv.org/abs/2506.01375.

[121] LIU Z, LIU W, ZHU H, et al. Next poi recommendation based on time slot preferences and bidirectional transformation modeling[J]. Manuscript, 2025.

[122] 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.

[123] SUN Z, DENG Z, NIE J, et al. Rotate: Knowledge graph embedding by relational rotation in complex space[C]//ICLR. 2019.

[124] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation[C]//Neural Computation: volume 15. 2003: 1373-1396.

[125] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,

1997, 9(8): 1735-1780.

[126] LI X, CHEN C, ZHAO X, et al. E4srec: An elegant effective efficient extensible solution of large language models for sequential recommendation: abs/2312.02443[J/OL]. 2023. https://arxiv.org/abs/2312.02443.

附录

（一）已发表的学术论文

1) Zhao Liu, Wei Liu, Huaijie Zhu, Jianxing Yu, Jian Yin. *Next POI Recommendation Based on Time Slot Preferences and Bidirectional Transformation Modeling* (TSPM). **WISE 2024**（**CCF C**）.
2) Zhao Liu, Wei Liu, Huaijie Zhu, Jianxing Yu, Jian Yin. *Geography-Aware Large Language Models for Next POI Recommendation* (GA-LLM). **ICDE 2026**（**CCF A**）.

后记

三年的研究生学习即将结束，从入学时对“做科研”只有模糊认识，到今天能够较为独立地完成问题定义、方法设计、实验验证与论文写作，这一过程远比我最初想象的更漫长，也更珍贵。回顾这段经历，自己真正收获的不只是论文结果本身，更是面对复杂问题时的耐心、面对失败实验时的韧性，以及在反复修改中逐步接近严谨表达的能力。

读研初期，我花了大量时间补齐推荐系统、时空建模和大模型相关基础。很多看似“懂了”的内容，在真正动手复现和推导时又会暴露理解盲区。进入课题后，围绕 Next POI 推荐这一方向，我经历了从传统序列方法到图建模，再到大语言模型增强方案的不断尝试。期间有过实验长期不收敛、结果波动难以解释、写作结构反复推翻重来的阶段，也正是在这些具体而琐碎的困难里，我逐渐理解了研究工作的核心不是追求“看起来漂亮”的结果，而是对问题边界、方法假设和证据链条保持诚实与清醒。

本文最终形成的两条主线工作（TSPM 与 GA-LLM）并非一蹴而就，而是在多轮讨论、验证和修正中逐步沉淀出来的。前者让我更深刻地认识到时空行为中的结构规律与时间异质性；后者让我意识到大模型能力的边界，以及结构先验注入在实际任务中的必要性。无论结果大小，这些探索过程都成为我研究训练中最重要的部分。

衷心感谢导师在选题把关、方法路线、实验设计和论文写作上的持续指导。导师严谨务实的科研态度、对细节的高标准要求和对学生成长的耐心投入，使我在每一个关键节点都能及时校正方向、稳步推进。导师不仅教会我“怎么做研究”，也让我理解了“为什么要这样做研究”。这份影响将长期伴随我今后的学习与工作。

感谢课题组各位老师和同学在数据处理、代码复核、论文讨论中的帮助。每一次组会交流和问题争论，都在推动我把模糊想法变成可验证命题。感谢学院与学校提供的学习平台与研究条件，感谢家人和朋友在读研阶段给予的理解、支持与鼓励，使我能够专注地走完这段旅程。

谨以此文，向所有在研究生阶段给予我指导、帮助和陪伴的人致以最诚挚的感谢。