MixTTE：用于可扩展、自适应行程时间估计的多层混合专家

Wenzhao Jiang 香港科技大学（广州） 中国广东广州
wjiang431@connect.hkust-gz.edu.cn

Jindong Han 香港科技大学 中国香港
jhanao@connect.ust.hk

Ruiqian Han
香港科技大学（广州） 中国广东广州
rhan464@connect.hkust-gz.edu.cn

Hao Liu*香港科技大学（广州） 中国广东广州 香港科技大学 中国香港 liuh@ust.hk

arXiv:2601.02943v1 [cs.LG] 6 Jan 2026

摘要

准确的行程时间估计（TTE）对于网约车平台至关重要，因为误差会直接影响用户体验和运营效率。尽管现有生产系统在整体路径级依赖建模方面表现出色，但难以捕捉城市规模交通动态和长尾场景，导致其在大型城市网络中的预测不够可靠。本文提出 MixTTE，这是一种可扩展且自适应的框架，协同融合链路级建模与工业级路径级 TTE 系统。具体而言，我们提出了一种时空外部注意力模块，以高效捕捉百万规模道路网络中的全局交通动态依赖关系。此外，我们构建了一个稳定的图混合专家网络，用于处理异构交通模式，同时保持推理效率。进一步地，我们设计了一种异步增量学习策略，使模型能够针对动态交通分布偏移进行实时且稳定的适应。在真实世界数据集上的实验表明，与七种基线方法相比，MixTTE 显著降低了预测误差。MixTTE 已部署于滴滴，大幅提升了 TTE 服务的准确性和稳定性。

CCS 概念

• **计算方法** → **机器学习**； •**应用计算** → **交通运输**； •**信息系统** → **时空系统**。

关键词

行程时间估计、外部注意力、混合专家、异步增量学习

*通讯作者。

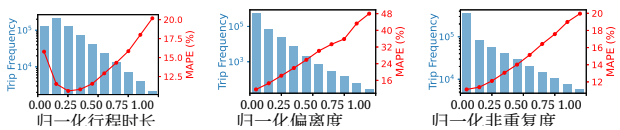
本作品依据知识共享署名4.0国际许可协议授权。*KDD '26*，济州岛，大韩民国
© 2026 版权所有者/作者保留版权。ACM ISBN 979-8-4007-2258-5/2026/08
https://doi.org/10.1145/3770854.3783940

ACM 参考文献格式：
Wenzhao Jiang、Jindong Han、Ruiqian Han 和 Hao Liu。2026 年。MixTTE：面向可扩展和自适应行程时间估计的多层混合专家。载于第*32*届*ACM SIGKDD*知识发现与数据挖掘大会论文集第*1*卷（*KDD '26*），*2026*年*8*月*9*日至*13*日，韩国济州岛。ACM，美国纽约州纽约市，12页。https://doi.org/10.1145/3770854.3783940

1 引言

行程时间估计（TTE）旨在预测给定行驶路线的耗时，该路线由按顺序排列的一系列道路链路组成，连接起点与目的地。现代网约车平台（如滴滴）每天处理数十亿次旅行时间查询，以支持路径规划、订单调度和动态定价等关键服务。平台在众多城市拥有大量活跃用户，其运营效率取决于行程时间估计（TTE）系统的准确性和可靠性；即使微小的改进，每年也能节省数百万成本，并显著提升用户满意度。

滴滴当前的生产系统采用**以路线为中心**的宽深循环架构（WDR） [50]，以捕捉复杂的路径特征交互，以及连续路线内链接之间的序列依赖关系。过去几年，滴滴持续改进特征工程、模型架构和训练策略 [13, 33]，以提升 TTE 系统的精度、效率和鲁棒性。虽然这种方法在常见场景中表现有效，但存在两项关键局限：（*1*）感受野有限。系统将路径视为彼此孤立的序列，无法考虑传播到路径中的更广泛交通动态（例如，拥堵从附近高速公路蔓延而来）。（*2*）长尾场景性能不佳。单体架构难以应对罕见但关键的模式（例如，活动交通、施工区域），导致长尾场景的错误率更高。在本研究中，我们旨在改进当前的 TTE 系统，整合**链路级**能力，明确捕捉路线内和路线外链接之间复杂的时空依赖关系，从而更充分地感知情境交通状况。



（a）行程时长。 （b）途中交通状况偏差程度。 （c）途中交通状况偏非重复程度。

图 1：网约车行程数据在三个以路线为中心的测量指标上的长尾分布*w.r.t.*。一种最先进的以路线为中心的行程时间估计方法 [13] 在相当一部分长尾路径上的表现仍然不佳。

尽管针对链路级行程时间估计算法的研究已相当丰富 [4, 5, 9, 17, 53], , 但将其整合到工业级、以路线为中心的系统中，仍面临三个主要挑战，根源在于可扩展性、异质性和动态适应。首先，在大规模道路网络上建模全局时空依赖仍然效率低下。城市交通呈现长程相关性（例如，住宅区早间拥堵数小时后传播至商业区），也存在语义相似但时间相距较远的模式（例如，早晚交通高峰时段）。尽管近期研究 [17, 53] 改进了局部上下文建模，但在大规模 TTE 中，如何高效、有效地实现全局路线上下文感知仍缺乏充分研究。现有尝试全局建模的方法 [1, 32] 通常依赖链路两两相关性，复杂度为二次，因此难以应用于百万规模道路网络。其次，交通模式异质性造成了长尾场景下的行程时间估计精度瓶颈。如图 1 所示，城市道路网络会受到区域划分、时间和外生因素（例如天气、活动）的影响，呈现多样的动态变化，从而形成偏态分布，使罕见但关键的场景（例如体育场活动）难以得到妥善处理。单纯增加模型容量可能导致模型对高频模式过拟合，同时忽视长尾泛化，因为单体架构会在联合优化过程中将彼此无关的模式混为一谈。近期，混合专家（MoE）架构 [20, 24] 展现出处理不同模式的潜力，可通过条件计算和稀疏专家激活实现这一目标。然而，基础混合专家方案 [7, 43]缺乏针对噪声图结构交通数据的稳定机制，容易导致专家坍塌或相互干扰，削弱其在 TTE 中的效果。第三，若采用直接整合的方式，持续适应交通变化在计算上将难以实现。尽管增量学习（IL）已被证明在以路线为中心的系统中颇具成本效益 [13], , 但加入链路级建模会加剧两个瓶颈：（1）细粒度链路表示会造成参数爆炸，使得在延迟预算内频繁进行全模型更新变得不可行；（2）对瞬时链路级变化更加敏感，可能导致模型过拟合噪声或遗忘稳定的路线级模式。当前的增量学习策略 [8, 13] 缺乏模块化设计，无法解耦多层级行程组件的更新频率（即高频链路更新与低频路线更新），从而阻碍了高性价比的适应。

在本研究中，我们提出了 MixTTE，这是一种可扩展且自适应的多层行程时间框架，可将链路级建模以模块化方式集成到滴滴当前以路线为中心的系统中。首先，我们提出了一种时空外部注意力模块，以高效捕获百万规模道路网络中的全局依赖关系，从而突破链路表示的可扩展性瓶颈，增强其交通上下文感知能力。其次，我们构建了一个由外部稳定的图混合专家（MoE）模块，通过减少多样化异质交通模式之间的干扰，帮助模型应对长尾场景。最后，我们开发了一种异步增量学习策略，根据检测到的分布偏移，有选择地更新路线级和链路级模型参数，从而以可持续的计算成本实现实时适应。值得注意的是，MixTTE 无需重构现有数据流水线或以路线为中心的模型，即可无缝接入滴滴现有生产系统。自 2025年4月 起，MixTTE 已部署于滴滴网约车平台，显著提升了服务成效与用户体验。

我们的主要贡献总结如下：(1) 我们提出了 MixTTE，这是一种可扩展且自适应的框架，将链路级和路径级模型的进展协同整合到滴滴的 TTE 系统中。(2) 我们引入了时空外部衰减模块和外部稳定化图MoE，分别用于高效捕获全局时空依赖和适应细粒度异构模式。(3) 我们针对路径-链路混合模型更新设计了异步增量学习策略，从而能够高效、稳定地实时适应交通分布偏移。(4) 大量离线和在线实验表明，与最先进的方法相比，该方法在准确性、可扩展性和适应性方面均表现优越。我们还分享了真实世界部署中的经验，为未来的工业应用提供支持。

2 预备知识

2.1 定义与问题陈述

本文重点研究道路网络中的网约车行程时间估计问题。首先介绍以下关键定义。

定义 1（交通网络）。交通网络被建模为有向加权图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$，其中每个节点 $v_i \in \mathcal{V}$表示一条道路链路，每条边 $e_{ij} \in \mathcal{E}$表示相邻链路 v_i与 v_j之间的连通性。在时间步 t，整个网络中的动态交通特征表示为 $\mathbf{X}^t \in \mathbb{R}^{N \times C}$，其中 $N := |\mathcal{V}|$，C表示动态交通特征的数量。我们将$\mathbf{X}_i^t \in \mathbb{R}^C$表示为时间步 t时链路 v_i的特征向量。

定义 2（交通切片）。链路 v_i 在时间步 τ 的交通切片定义为一组动态交通特征 $[\mathbf{X}_i^{\tau-T+1}, \ldots, \mathbf{X}_i^{\tau}] \in \mathbb{R}^{T \times C}$，其中 T 表示回溯窗口大小。

定义 3（路径）。路径 $\mathcal{R}$ 表示为一段连续的道路链路序列 $\{v_i : i = 1, \ldots, l, v_i \in \mathcal{V}\}$，其中 l是路径 $\mathcal{R}$中的道路链路总数。通常，一条路径包含数十到数百条链路，多个行程可以经过同一条路径。

接下来，我们正式定义本文关注的问题。

问题 1（行程时间估计）。给定一个行程查询 $q = (l_{ori}, l_{des}, \tau, \mathrm{R})$，从时间步 τ开始，我们旨在基于$\hat{y} = \mathrm{F}(q, \mathcal{G}, \mathbf{X}^{\tau-T+1:\tau})$，估计沿路径$\mathcal{R}$从起点 l_{ori}到目的地 l_{des}的旅行时间 y，其中$\mathrm{F}(\cdot) = \mathcal{F}_r \circ \mathcal{F}_l(\cdot)$是一个由以路线为中心和以路段为中心的映射函数构成的混合模型，我们旨在基于历史查询$\mathcal{Q} = \{q_1, q_2, \ldots\}$对其进行学习。

2.2 滴滴的 TTE 系统

作为领先的网约车平台之一，滴滴一直在持续改进行程时间估计（TTE）系统。本节将简要介绍滴滴当前的 TTE 系统，以便进一步讨论。

2.2.1 以路线为中心的骨干网络。 滴滴的 TTE 系统基于以路线为中心的宽深循环架构（WDR） [50] 构建，用于学习复杂的路线级特征交互与结构依赖关系。宽网络通过专门的特征工程和广义线性模型，记忆低阶特征交互。相比之下，深网络采用多层感知机（多层感知机）学习高阶非线性特征交互，以提升泛化能力。循环网络对路径中各道路链路之间的序列依赖关系进行建模。最终的旅行时间预测结果通过聚合三个模块的输出获得。尽管 WDR 于 2018 年推出，但凭借模块化维护和灵活升级能力，它至今仍是滴滴 TTE 系统的核心。依托平台丰富的动态与静态特征，宽模块持续进行特征工程优化，而循环模块则已升级为更先进的 Transformer架构 [13]。

2.2.2 增量更新。 尽管模型架构已高度优化，滴滴还部署了增量学习（IL）框架 iETA，以经济高效的方式每日调整模型，使其适应不断变化的交通环境 [13]。形式上，令 D_r表示在第 r. 天收集的数据集。第 s 天投入使用的模型，基于数据集 $\mathrm{D}^{\text{day}} = \mathrm{D}_{s-\delta_d} \cup \mathrm{D}_{s-7} \cup \cdots \cup \mathrm{D}_{s-7F}$，进行训练，其中 $\delta_d < 7$ 表示数据预处理的延迟， F 表示周期性数据回放所考虑的前几周数。无需完全重新训练，iETA 便能保持模型及时更新，同时巩固重复出现的交通模式，以确保稳定性。该框架还支持多种更新频率，以适应不同规模的交通分布偏移。

3 方法

如图 2 所示，为构建 MixTTE，我们着力解决两项主要任务：i）设计一个表达能力强且可扩展的链路表示学习模块，为以路线为中心的模型补充丰富的交通上下文；ii）为混合模型定制优化策略，使其能够高效地适应频繁变化的交通状况。针对第一项任务，我们提出时空外部注意力（STEA）模块，以捕获全局时空相关性；随后引入外部稳定图混合专家层（ESGMoE），以学习更细粒度但具有异质性的交通模式。这两个模块输出的路段表示随后与路线内原始链路特征拼接，作为下游以路线为中心的模型的增强输入。针对第二项任务，我们提出异步增量学习（ASIL），根据检测到的分布偏移，高效且自适应地更新混合模型。

3.1 时空外部注意力

如前所述，现实世界中的城市道路网络包含数百万条道路链路，因此，实现可扩展的全局相关性建模是一项重大挑战。受近期进展的启发 [11, 28]，我们提出了一种时空外部注意力（STEA）模块，该模块学习一小组外部记忆单元$\mathbf{M} = [\mathbf{M}_1, \ldots, \mathbf{M}_{U_{ex}}] \in \mathbb{R}^{U_{ex} \times d}$, 其中 $U_{ex} \ll N$和 d 为嵌入维度，用于调节交通切片之间代价高昂的成对特征交互。

具体而言，对于交通切片 $\mathbf{X}^{t-T+1:t}$，我们首先利用轻量级多层感知机，将时间依赖关系编码为高维表示 $\mathbf{H} \in \mathbb{R}^{N \times d}$。然后，我们通过交叉注意力机制从记忆单元 $\mathbf{M}$ 中进行外部知识检索（EKR），以增强 $\mathbf{H}$ 的独特性，即，

$$\begin{aligned} \mathrm{EKR}(\mathbf{H}, \mathbf{M}) &= \text{Cross-Attention}(\mathbf{H}, \mathbf{M}) \\ &= \mathrm{Softmax}\left((\mathbf{H}\mathbf{W}_Q)(\mathbf{M}\mathbf{W}_K)^\top / \tau\right)(\mathbf{M}\mathbf{W}_V) \end{aligned} \tag{1}$$

其中，$\mathbf{W}_Q$，$\mathbf{W}_K$，$\mathbf{W}_V \in \mathbb{R}^{d \times d}$均为可学习的线性变换， τ为温度超参数。作为另一种方案，我们可以用两个不同的可学习外部记忆单元 $\mathbf{M}_K$，$\mathbf{M}_V \in \mathbb{R}^{U_{ex} \times d}$直接替换$\mathbf{M}\mathbf{W}_K$，$\mathbf{M}\mathbf{W}_V$，以提升模型容量。最后，整体 STEA 机制可形式化表示为：

$$\mathbf{H}^{ex} = \mathrm{STEA}(\mathbf{H}, \mathbf{M}) = \mathrm{LN}\left(\mathrm{EKR}(\mathbf{H}, \mathbf{M}) + \mathbf{H}\right), \tag{2}$$

其中，LN(·)表示层归一化算子。通过端到端训练，记忆单元会根据每个交通切片产生的梯度持续更新，从而逐渐积累整个数据集中的原型交通模式。因此，每个 EKR 步骤都使一个交通切片能够间接地与时空中的所有交通切片交互。相似的交通模式会检索到相似的知识，从而使潜在空间中的表示更加集中。值得注意的是，EKR 步骤的复杂度为线性复杂度 $O(N \cdot U_{ex})$，不仅相比成对建模方法的二次复杂度实现了更具可扩展性的时段内全局建模，还填补了融合时间上相距较远的交通切片这一方面的空白。

在实践中，同一时间的交通切片可能呈现出更显著的语义相关性，例如城市尺度极端拥堵。为此，我们将 EKR 与空间层次模型无缝集成 [14, 52]，以强化同一时间的全局相关性建模。我们首先对 $\mathbf{H}$进行软聚类，得到交通语义令牌 $\mathbf{H}^{se} \in \mathbb{R}^{U_{in} \times d}$，其中 $U_{in} \ll N$.。随后，我们在语义层面执行 EKR 步骤，以增强语义令牌之间同一时间的自注意力（SA）。最后，我们根据软聚类权重，将 $\mathbf{H}^{se,ex}$广播回来，得到路段表示 $\mathbf{H}^{ex}$。上述过程可形式化为

$$\begin{aligned} \mathbf{H}^{ex} &= \mathrm{STEA}(\mathbf{H}, \mathbf{M}) \\ &= \mathrm{LN}\left(\mathrm{BC}\left(\text{Self-Attention}\left(\mathbf{H}^{se}\right) + \mathrm{EKR}\left(\mathbf{H}^{se}, \mathbf{M}\right)\right) + \mathbf{H}\right). \end{aligned} \tag{3}$$

关于层次建模的更多细节，请参见附录A.1。

3.2 外部稳定化图MoE

除了 STEA 的全局建模之外，我们还提出了外部稳定化图混合专家（ESGMoE）层，以高效捕捉具有长尾分布的异质局部交通模式。本节首先概述 ESGMoE 层，随后详细介绍路由和专家设计。

*3.2.1 ESGMoE层概述。*总体而言，ESGMoE利用一组 N_e 图专家来捕捉异质交通模式的全貌，同时通过稀疏

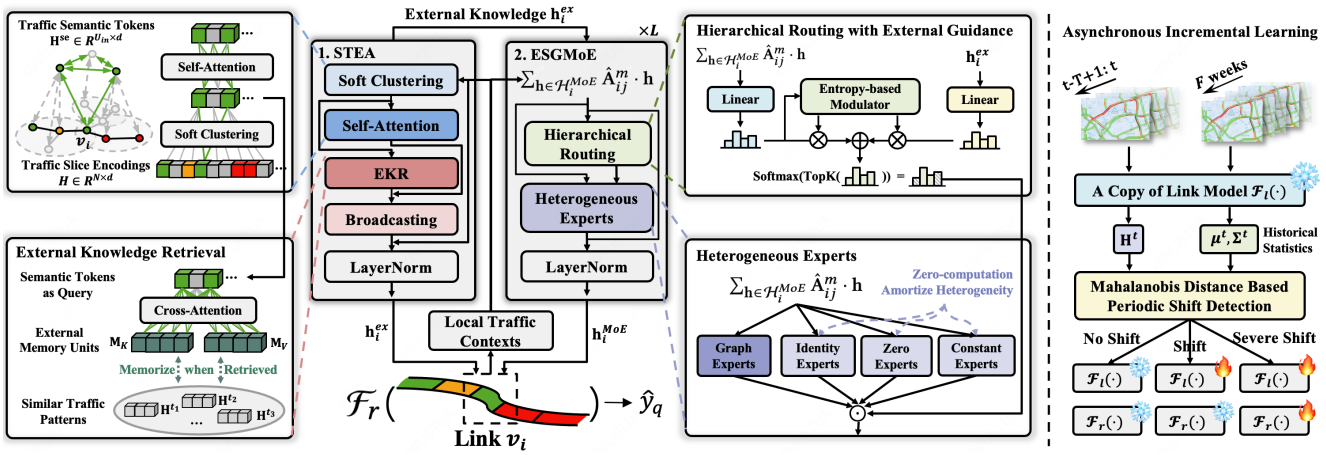


图 2：MIXTTE的整体框架。左侧展示了从特定道路链路视角出发的多层级 TTE 流程。右侧描绘了每个 IL 更新步骤开始时采用的异步策略。

激活保持较低的推理成本。形式化地，ESGMoE层定义为

$$\mathbf{h}_i^{MoE'} = LN\left(\sum_{n=1}^{N_e} G_n\left(\mathcal{H}_i^{MoE}, \mathbf{h}_i^{ex}\right)\cdot E_n\left(\mathcal{H}_i^{MoE}\right) + \mathbf{h}_i^{MoE}\right), \quad (4)$$

该模块将每个链路 m 跳邻域 $\mathrm{N}_m(i)$ 内的局部交通上下文 $\mathrm{H}_i^{MoE} = \{\mathbf{h}_j^{MoE}: j \in \mathrm{N}_m(i) \cup \{i\}\}$ 编码为紧凑表示 $\mathbf{h}_i^{MoE'} \in \mathbb{R}^d$. 其中，$\mathbf{h}_j^{MoE}$表示初始交通特征（与 STEA 模块的输入相同），或前一个 ESGMoE 层针对链路 v_j 输出的隐藏表示。该表示在保留关键时间模式的同时，避免了在每个时间步进行开销高昂的图建模。门控函数 G(·) 和图专家 $E_n(\cdot)$ 是其设计的核心。具体而言，$G_n(\cdot)$ 是稀疏门控函数 G(·) 的第 n 个元素，它根据局部交通上下文和从 STEA 模块检索到的外部知识 $\mathbf{h}_i^{ex}$，决定将链路 v_i 路由至第 n 个专家 $E_n(\cdot)$ 的概率。专家设计各不相同。大多数专家采用简化的图卷积网络实现，计算开销较低 [51]，，即 $E_n(\mathrm{H}_i^{MoE})$ = 多层感知机$_n(\sum_{\mathbf{h}\in\mathrm{H}_i^{MoE}} \hat{\mathrm{A}}_{ij}^m \cdot \mathbf{h})$，其中 $\hat{\mathrm{A}}^m$ 是归一化邻接矩阵 A 的第 m 个幂，可用于预计算以实现推理加速。其余专家是零计算专家，旨在分摊常见交通模式下的专家激活开销。

3.2.2 采用自适应外部引导的分层路由。在混合专家中，成熟的路由机制对于实现稳定的专家专门化至关重要 [6]。现有时空混合专家方法通常将短期交通切片 [24,26] 或可捕获长期模式的可学习链路和时间嵌入 [20], 输入路由门控，但前者缺乏判别能力，后者则限制了门控对细粒度短期差异的学习。为此，我们提出一种基于熵的层级路由机制，利用 STEA 模块积累的外部知识，自适应地增强门控输入的区分度。

具体而言，该路由机制采用经典的 Top-k门控函数 [7] 来生成图专家的路由概率，即，G(·) = Softmax(Topk(g(·)))，其中 g(·) 通常实现为一个将输入特征映射到路由逻辑值的线性层，而 k 表示每个样本所激活的专家数量。此外，还引入了负载均衡损失，以防止稀疏门控机制坍缩到少数占主导地位的专家上，具体内容将在第 3.4 节中介绍。不同之处在于，我们将 g(·) 设计为具有如下层次结构：

$$\mathrm{g}(\mathcal{H}_i^{MoE}, \mathbf{h}_i^{ex}) = \alpha_i \cdot \mathrm{g}^{ex}(\mathbf{h}_i^{ex}) + (1-\alpha_i)\cdot \mathrm{g}^{loc}(\mathcal{H}_i^{MoE}), \quad (5)$$

其中，$\mathrm{g}^{ex}(\cdot)$ 和 $\mathrm{g}^{loc}(\cdot)$ 是两个独立的前馈神经网络，分别利用经过外部增强的链路表示和局部交通上下文，计算链路 v_i 的路由逻辑值。自适应权重 α_i根据局部路由分布的熵 S_i 计算得出：

$$\begin{aligned} \alpha_i &= \mathrm{Sigmoid}\left(\gamma S_i + \mu\right), \\ S_i &= -\bar{\mathrm{g}}_i^{loc}\cdot \log \bar{\mathrm{g}}_i^{loc}, \quad \bar{\mathrm{g}}_i^{loc} = \mathrm{Softmax}\left(\mathrm{g}^{loc}(\mathcal{H}_i^{MoE})\right) \end{aligned} \quad (6)$$

其中 γ，μ为可学习参数。这样，当路由器对局部交通上下文不确定时，α_i 会增大，从而在路由时更多地依赖外部引导，以提高路由稳定性，进而实现更稳健的专家专门化。

3.2.3 借助零计算专家摊销异质性。尽管训练专门化专家以捕捉异质性交通模式具有潜力，但传统的稀疏门控混合专家设计在面对长尾模式时存在固有局限：（i）由于负载均衡正则化，旨在处理罕见模式的专家往往会在常见模式上被过度激活，而单纯增加专家数量则可能导致容量过剩和训练不足；（ii）罕见模式应分配更多专家，但 top-k 稀疏激活缺乏灵活性。单纯提高 k 会增加计算量，并有过拟合常见模式的风险。为此，我们进一步引入不执行图建模的零计算专家，以摊销常见模式的激活，并隐式地允许罕见模式激活更多专家。

受已有研究 [21],的启发，我们引入三类零计算专家：i）恒等专家，仅返回输入特征；ii）常量专家，输出一个可学习常量向量；iii）空值

专家，输出一个零向量。这些专家的激活由第 3.2.2 节介绍的同一路由机制与图专家同时确定。直观而言，三类零计算专家在处理某些常见交通模式时各有优势：i）恒等专家适用于平稳/自由流或近线性状态，此时当前表示已具备预测能力；ii）常量专家通过学习偏置记忆高频、模板化的模式；iii）空专家在信号嘈杂或边际增益微乎其微时，跳过图建模。因此，引入这些专家可以有效分担频繁出现但边际增益较低的交通模式，而不会过度占用图专家容量。

3.3 异步增量学习

持续适应对于工业级旅行时间估计系统在不断变化的交通环境中保持稳定至关重要。然而，以路段为中心的模型使频繁进行全参数更新的计算成本高昂，也更容易对多样的路段级分布偏移产生过拟合。为此，我们提出一种异步增量学习（ASIL）策略，主动检测分布偏移，从而实现选择性参数更新。

首先，由于模型已充分学习周期性交通模式，我们仅在检测到周期性分布偏移时才启动更新。具体而言，我们维护每条链路 v_i的表示历史统计量，包括均值 $\boldsymbol{\mu}_i^t$和协方差Σ_i^t，这些统计量根据过去 F周内相同时间步的链路表示计算得出。更新开始时，我们针对每条链路计算其当前表示 $\mathbf{h}_i^t$与历史统计量之间的马氏距离（MD），以量化周期性偏移程度：

$$d_i^t = \sqrt{(\mathbf{h}_i^t - \boldsymbol{\mu}_i^t)^\top (\Sigma_i^t)^{-1} (\mathbf{h}_i^t - \boldsymbol{\mu}_i^t)}. \tag{7}$$

当一条路段的 d_i^t超过预设阈值 δ_d时，该路段被视为异常路段，即正在经历周期性偏移。只有当城市内异常路段的比例超过其历史分布的 δ_l 分位数时，我们才会执行 IL 更新步骤。此外，考虑到以路线为中心的模型在输入链路级上下文时，自然会更多地关注相对稳定的结构依赖关系，因此我们对路线侧参数更新施加更严格的限制。只有当异常比例超过其历史分布的δ_r分位数时，我们才会同时触发链路侧和路线侧更新，其中 $\delta_r > \delta_l$。

然而，随着路段编码器不断增量更新，不同模型版本中的路段表示潜在空间会发生变化，这可能影响 MD 的可靠性。为此，我们使用每日更新中的以路段为中心的模型的冻结副本来生成用于漂移检测的表示，从而确保潜在空间保持一致，同时将额外推理成本控制在较低水平。

3.4 优化目标

MixTTE 的优化目标由两部分组成。第一部分是行程时间回归损失，用于以端到端的方式联合优化以路径和路段为中心的模型，即，

$$\mathcal{L}_{reg} = \sum_{q \in Q} |\hat{y}_q - y_q|. \tag{8}$$

第二部分是专家负载均衡正则项，用于防止 ESGMoE 层出现模型坍塌问题，即，

$$\mathcal{L}_{load} = N_e \cdot \sum_{n=1}^{N_e} \xi_n \cdot \left(\frac{1}{N} \sum_{i=1}^{N} G_{n,i} \right) \cdot \left(\frac{C_n}{N \cdot k} \right) \tag{9}$$

其中， C_n表示在 top-k 选择下路由到专家 n 的词元数量。 ξ_n表示专家类型权重，对于图专家，该权重设为 1；对于零计算专家，该权重设为 δ，其中 δ是用于控制摊销程度的超参数。

总体而言，我们通过联合优化目标函数 $\mathrm{L} = \mathrm{L}_{reg} + \alpha \sum_{l=1}^{L} \mathrm{L}_{load}^{(l)}$，来训练 MixTTE，其中 $\mathrm{L}_{load}^{(l)}$是第 l 个 ESGMoE 层的负载均衡损失， α是用于控制专家均衡程度的超参数。

4 系统部署

MixTTE 已部署于滴滴网约车平台，用于升级现有的以路线为中心的 TTE 系统。本节概述系统部署策略。

4.1 离线模型训练

每个城市的离线模型训练均在 Linux 服务器上使用 TensorFlow 完成，该服务器配备 4 个 Intel Xeon E5-2630 v4 CPU（90 GB）和 1 个 NVIDIA Tesla P40 GPU（24 GB）。模型按照两级时间层级进行增量更新。在日级尺度上，更新流程与 iETA [13] 类似。在小时级尺度上，则基于过去一小时记录的已完成行程数据采用 ASIL 策略。离线训练采用特定时间批次采样，每个批次由来自相同时间步 t 的历史行程记录组成，而不是从整个数据集中随机采样。这带来三项主要优势：（1）在相同时段内进行全局建模，从而确保路段表示的梯度更加平滑；（2）提高训练效率，因为在特定时间步共享相同道路链路的行程可以在前向和反向传播过程中重复使用已计算的路段表示；（3）与在线环境中的流式推理范式保持一致。训练完成后，模型将发布到在线服务器，以提供实时 TTE 服务。

4.2 异步在线服务

在在线服务端，为满足实时响应对高吞吐量和毫秒级延迟的要求，我们为 MixTTE 采用异步在线服务策略。以路段为中心的模型使用 TensorFlow 实现，并部署在 GPU 服务器上，通过并行专家推理实现大规模路段表示生成。由于计算成本较高，路段表示每 5 分钟更新一次，并缓存在 Redis 服务器中。以路线为中心的模型则使用 C++ 重新实现，并部署在分布式 CPU 服务器上。收到 TTE 查询后，以路线为中心的模型会获取最新的路段表示和路线特征，以便即时预测。

5 实验

本节基于滴滴当前的 TTE 系统，通过大量实验评估我们提出的框架。具体而言，我们重点关注：（1）MixTTE 相较于先进的 TTE 方法的整体性能；（2）MixTTE 的长尾性能

表1：三个数据集的统计信息。

City	北京	南京	苏州
时间跨度	2024-07-21~ 2024-09-05	2024-12-18~ 2025-02-25	2025-01-01~ 2025-04-14
行程数	22,276,096	7,954,432	16,775,680
平均时长（分钟）	18.11	13.79	13.63
平均每条行程的路线内链接	121.01	99.92	86.86
链接数	2,315,360	696,107	1,419,074
热点路段数	156,731	83,323	149,171

；（3）MixTTE 的效率；（4）MixTTE 关键模块的消融研究；（5）MixTTE 的可解释性；以及（6）MixTTE 在在线生产环境中的有效性。更多实验细节请参阅附录B。

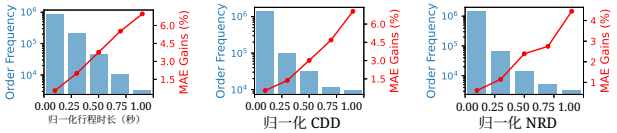
5.1 实验设置

5.1.1 训练准备。我们在北京、南京和苏州这三座大都市开展了大量实验。所有数据集均采集自滴滴网约车平台。三个数据集的详细统计信息见表1。值得注意的是，对于每座城市，我们都会识别一组具有显著交通动态的"热点路段"，用于路段嵌入生成，从而减少计算量和链路级噪声。热点路段选择的详情见附录B.1。每个数据集均取最后7天作为测试集。在**完全重新训练**设置下，我们使用每个测试日之前的110天作为训练集。在**增量学习**设置下，我们取最近一小时内完成的行程作为训练集，并取下一小时内开始的行程作为测试集。这种高频设置符合实际情况，也更适合测试增量学习策略的有效性和稳定性。对所有小时的表现取平均，以获得总体结果。

5.1.2 基线方法。 在全量重新训练设置下，我们将不采用增量学习的 MixTTE（MixTTE-*WoIL*）与以下基线方法进行比较：基于规则的方法：(1) RouteETA：通过累加路线内链接的历史平均旅行时间来预测行程时间。以路线为中心： (2) HierETA [3]：一种用于路径级行程时间估计的两级分层专家系统；(3) WDR：滴滴当前部署的、未采用增量学习的行程时间估计模型，如第 2.2.1 节所述；以链路为中心：(4) CompactETA [9]：使用图注意力网络生成链路表示，然后采用轻量级多层感知机进行路径行程时间估计；(5) ConSTGAT [5]：为链路行程时间估计设计了三维图注意力网络，并将其累加以获得路径行程时间估计；(6) BigST [15]：提出了一种线性化的全局空间建模方法，用于替代我们的以路段为中心的模型。

在 IL 设置中，我们将 MixTTE 与以路线为中心的方法 iETA [13] 进行比较。为确保公平，所有方法均输入我们整理后的特征，并且以路段为中心的模型采用第 4.1 节介绍的相同流程进行训练。更多实现细节请参见附录 B.2。

5.1.3 指标。我们使用三个指标来评估不同方法的性能：**平均绝对误差 (MAE)**：预测行程时间与实际行程时间之间绝对差值的平均值；**平均**



（a）行程时长。 （b）行程中交通状况偏差程度-程度。 （c）行程中交通状况非重复程度-次数不复现程度。

图 3：MixTTE 相较于滴滴当前 TTE 模型在头部和尾部交通场景中的相对 MAE 增益 [13]

绝对百分比误差 (MAPE)：预测行程时间与实际行程时间之间绝对百分比差值的平均值；**坏例比例 (BCR)**：预测行程时间满足 MAE ＞ 300 秒且 MAPE ＞ 20% 的行程查询所占百分比。

5.2 总体性能（研究问题1）

表2展示了整体性能。MixTTE 在所有指标上均始终优于所有基线方法。在全量重新训练设置下，MixTTE 的平均绝对误差、平均绝对百分比误差和 BCR 相较于次优结果分别最高提升 2.01%、2.41% 和 8.81%。在增量学习（IL）设置下，MixTTE 相较于 iETA 分别最高提升 2.39%、3.70% 和 10.32 %。这些增益表明，MixTTE 能有效适应并捕捉多层次、不断演变的交通模式。BCR 更显著的提升凸显了我们为长尾泛化设计的链路级方法的有效性。我们还可以进一步观察到：（1）深度学习方法优于基于规则的基线方法（RouteETA），表明它们在捕捉高阶非线性依赖关系、实现准确行程时间估计方面更具优势；（2）以路段为中心的基线方法大多不及业界以路线为中心的系统（WDR），凸显了特征工程在工业实践中的重要性；（3）在北京，将 BigST 集成后，其 MAPE 和 BCR 优于 WDR关于，证明了链路级建模的重要性。然而，集成后的模型并未在所有方面都优于 WDR，表明将链路级建模集成到高度优化的以路线为中心的行程时间系统中并非易事；（4）增量学习方法始终优于全量重新训练方法，表明增量学习对非平稳交通的有效性。然而，iETA 相较于 WDR 的提升幅度有限，这是因为 iETA 的增量学习策略缺乏模块化。

5.3 长尾性能（研究问题2）

我们还分析了长尾交通场景中的性能增益。具体而言，我们首先根据以下三个指标的取值范围，将苏州数据集中的测试行程查询均匀划分为5组：（1）行程时长（TD）、（2）行程中交通状况偏差度（CDD）和（3）行程中交通状况非重复度（NRD）。其中，行程查询 $q=(l_{ori},\ l_{des},\ \tau,\ \mathrm{R})$ 的 CDD 和 NRD 指标定义如下：

$$\mathrm{CDD}(q)=\mathbb{E}_{v\in\mathcal{R}}\|x_v^{\tau_v}-x_v^{\tau}\|_2,\quad \mathrm{NRD}(q)=\mathbb{E}_{v\in\mathcal{R}}\|x_v^{\tau_v}-\bar{x}_v^{\tau_v}\|_2,$$

其中，x_v^{τ}表示链路 v在时间步 τ.的离散交通拥堵等级特征。时间步 τ_v是车辆到达链路 v的时刻，$x_v^{\tau_v}$表示相应时刻的历史平均拥堵等级

表2：三种真实世界数据集上不同方法的整体性能比较。最佳结果以粗体显示，次佳结果加下划线，第三佳结果前带星号（*）。

训练设置	模型	北京			南京			苏州		
		平均绝对误差（秒）	平均绝对百分比误差（%）	BCR（%）	平均绝对误差(秒)	平均绝对百分比误差(%)	BCR(%)	平均绝对误差(秒)	平均绝对百分比误差(%)	BCR(%)
无训练	RouteETA	181.06	17.12	11.87	153.06	22.56	11.4	139.07	19.84	8.61
完全重新训练	HierETA	135.07	12.84	6.20	87.02	12.61	2.31	82.53	11.26	1.90
	WDR	134.02	12.06	6.06	77.29	10.59	1.59	78.60	10.64	1.53
	CompactETA	136.63	12.25	6.12	83.54	11.79	2.10	81.38	10.87	*1.46
	ConSTGAT	134.52	12.11	6.04	78.16	10.56	1.98	79.82	10.73	1.59
	BigST	134.75	11.98	6.03	77.93	*10.49	1.91	79.21	10.77	1.54
	MixTTE-*Wo*增量学习	132.43	11.86	*5.73	75.74	10.26	1.45	77.29	10.51	1.42
增量学习	iETA	*132.73	*11.93	5.71	*77.10	10.53	*1.55	*78.22	*10.59	*1.46
	MixTTE	**130.84**	**11.75**	**5.25**	**75.26**	**10.14**	**1.39**	**77.18**	**10.48**	**1.36**

步 τ_v. 如图 3 所示，网约车行程在全部三个指标上均呈现长尾分布。更重要的是，在计算不同样本组中 MixTTE 相较于滴滴当前行程时间估计模型的平均绝对误差增益后，我们观察到平均绝对误差增益从头部样本到尾部样本呈上升趋势，这验证了 MixTTE 在长尾场景中的泛化能力。

5.4 效率分析（研究问题3）

我们进一步开展实验，以验证MixTTE的效率。在全量重新训练设置下，我们将MixTTE-*WoIL* 与在路段数量方面具有线性复杂度<style id='3'>w.r.t. 的链路中心基线方法进行比较。我们采用训练吞吐量和推理延迟作为指标，其中训练吞吐量以每分钟处理的平均训练样本数进行衡量，而推理延迟仅针对每个时间步的路段表示生成进行测量。为确保公平比较，我们在每个推理步骤激活的参数数量和局部交通建模的邻域大小方面，使基线方法与MixTTE保持一致。如表3所示，MixTTE-*WoIL*在三项效率指标上均始终优于CompactETA和ConSTGAT。这是因为CompactETA需要堆叠多个计算成本高昂的图注意力网络层来捕获长程依赖关系，而ConSTGAT采用的三维图注意力网络同时执行空间注意力和时间注意力，使其成本达到单个图注意力网络层的 T 倍。此外，尽管我们的模型具有更大的模型容量，其效率仍与BigST相当，这证明了ESGMoE模块能够在不引入额外计算成本的情况下提升模型容量的有效性。

在增量学习设置下，我们将本框架与-*WoPU*进行比较，后者移除了 ASIL 策略中的参数选择性更新。我们采用所有增量学习步骤的平均训练时间和可训练参数作为指标。由于 ASIL 策略对路段数量不敏感，因此我们仅报告苏州数据集上的结果。如表4所示，与 MixTTE-*FIL*相比，MixTTE 显著减少了训练时间和可训练参数。MixTTE-*FIL* 在每个增量学习步骤中更新所有参数，而 MixTTE 仅在发生周期性偏移时选择性地激活路径级或链路级参数，从而取得了显著的效率增益。

表3：完全重新训练与推理效率对比。（北京：北京，南京：南京，苏州：苏州）。

模型	训练吞吐量（K/分钟）			推理延迟（毫秒/时间步）		
	BJ	NJ	SZ	BJ	NJ	SZ
CompactETA	14.96	23.13	16.23	541.27	286.62	476.84
ConSTGAT	10.51	17.80	12.64	704.11	358.52	650.98
BigST	**22.45**	**32.77**	23.53	**78.30**	40.05	**69.40**
MixTTE-无增量学习	21.17	32.02	**25.38**	88.97	**30.45**	75.11

苏州数据集中的表4：增量学习效率对比集。

模型	平均训练时间（秒/步）	平均可训练参数量（千/步）
MixTTE-*WoPU*	36.63	525.66
MixTTE	7.81	47.49

5.5 消融研究（研究问题4）

为验证 MixTTE 各模块的有效性，我们在南京和苏州数据集上比较以下模型变体：i）-*WoEA* 移除 STEA 模块以及 ESGMoE 层中的分层路由；ii）-*WoHR* 移除 ESGMoE 层中的分层路由；iii）-*WoMoE* 移除 ESGMoE 层；iv）-*WoZE* 移除 ESGMoE 层中的零计算专家；v）-*WoPU* 定义同效率分析。如图4所示，我们得出以下观察结果。

在全量重新训练设置中，-*WoHR* 的表现不如 MixTTE-*WoIL*，这表明分层路由对于稳定 ESGMoE 训练至关重要。比较-*WoEA* 与-*WoHR*可以看出，STEA 对增强路径上下文感知能力至关重要。此外，-*WoHR* 在苏州数据集上造成了更大的性能下降，这是因为苏州更加异质化的道路网络需要更稳定的混合专家训练。最后，-*WoMoE* 导致的性能下降揭示了 ESGMoE 层的容量；-*WoZE* 同样会导致性能下降，尤其相对于 南京的 BCR

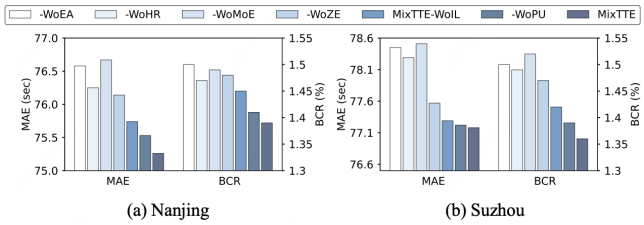


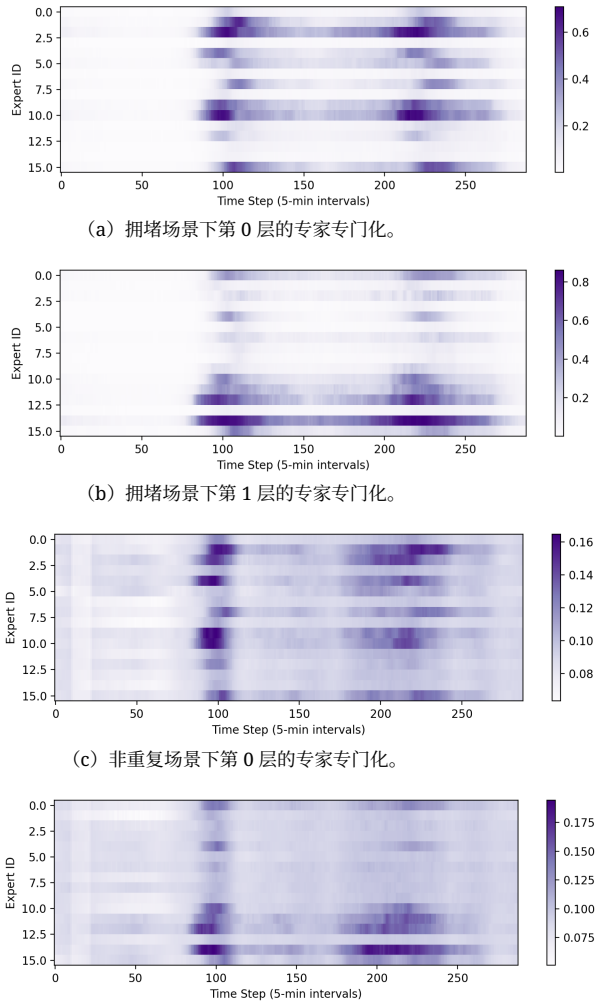
图4：南京和苏州数据集上的消融研究。

数据集而言，这证明了零计算专家对于增强长尾场景处理能力的重要性。在 IL 设置中，*-WoPU* 会导致性能下降，表明其能够有效提升既有混合TTE模型中 IL 的稳定性。

5.6 可解释性分析（RQ 5）

我们进一步分析了 ESGMoE 层在不同交通场景下的专家激活分布，以深入了解模型的决策。具体而言，在每个时间步，我们计算每条链路的 Top-2路由分布，并记录相应的专家分配。对于每一天，我们都会得到一个链路-专家分配张量 $\mathbf{A} \in \mathbb{R}^{288\times N\times 16}$，其中$N$ 表示链路数量，16 表示专家数量（第 0 至第 11 个为图专家，第 12 个为空专家，第 13 个为恒等专家，第 14 至第 15 个为常数专家），288 表示一天中 5 分钟时间步的总数。该张量使我们能够分析每个专家所专精的交通场景。

如图5所示，我们可以得出几个关键发现。(1) 专家专门化：部分专家会被激活，用于建模拥堵或非周期性交通模式，这表明我们的异构专家设计能够有效地针对长尾场景形成专门化。(2) 层间差异： 比较不同层中的激活分布可以发现，与第0层相比，第1层在长尾场景中的专家专门化更加稀疏，这与更深层通常倾向于捕捉更抽象的交通语义这一事实一致。(3) 零计算专家的行为： 首先，在第0层中，它们的激活频率远低于图专家。这是因为原始特征建模高度依赖图计算来捕捉高级模式。其次，在第1层中，它们的使用率有所提高，此时来自前一层的凝练语义能够支持更加轻量级的处理，从而避免过度建模。第三，常数专家在两层中都持续保持较高的参与度，体现了其记忆能力。空值专家和恒等专家虽然使用频率较低，但在非周期性场景和更高层中呈现出有意义的激活模式，证实了它们在摊销简单或噪声模式方面的必要性。



（a）拥堵场景下第 0 层的专家专门化。

（b）拥堵场景下第 1 层的专家专门化。

（c）非重复场景下第 0 层的专家专门化。

（d）非重复场景下第 1 层的专家专门化。

图5：2024-09-02 北京数据集上 ESGMoE 层的可解释性分析。每个矩阵元素的颜色表示分配给某个专家的链接中，在每个时间步经历拥堵或非重复性状况的比例。此处，若当前交通状况与其历史平均值之间的偏差超过整体历史分布的 0.9分位数，则将该状况定义为非重复性状况。

5.7 在线评估（RQ6）

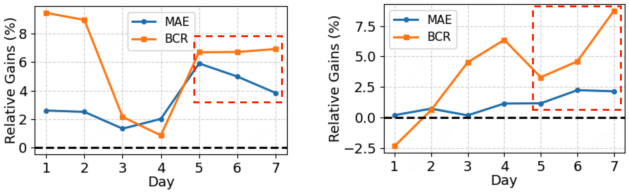
在本节中，我们在滴滴的大规模网约车平台上开展在线A/B测试，以验证 MixTTE 在实时生产环境中的效用。具体而言，我们关注每次行程开始时针对下车行程查询的行程时间估计，这类查询涵盖了各种行程距离和时长。在测试期间，我们将行程查询随机分为对照组和处理组，其中对照组的查询由当前的行程时间估计模型回答

滴滴的系统，而处理组由增强了 MixTTE 的系统响应。行程完成后，我们会记录其实际旅行时间，以便进一步进行指标计算。

表5报告了两个大都市中两组数据的 TTE 性能，这两个城市分别记为 X市和 Y市，评估周期为1周，分别为2025-04-28至2025-05-04以及2025-05-27至2025-06-02。可以看出，MixTTE 在平均绝对误差和 BCR 方面始终优于滴滴当前的 TTE 系统相对于 ，这证明了 MixTTE 在提升实时 TTE 服务方面的优越性。此外，图6显示出更显著的性能增益

表5：在线行程时间估计测试中的性能增益。

	X市		Y市	
	平均绝对误差（秒）	BCR（%）	平均绝对误差（秒）	BCR（%）
对照组	116.72	5.35	87.75	2.77
处理组	113.40	5.12	86.67	2.66
增益（%）	**3.03**	**4.41**	**1.24**	**4.30**



(a) X市。 (b) Y市。

图6：相对于对照组的每日性能增益。红色方块标记了在节假日两个指标上均有更大增益的情况。

在两个城市的节假日期间。这不仅表明整合链接级建模对于刻画未充分表示的交通动态十分重要，也证明了 MixTTE 在长尾泛化方面的有效性。

6 相关工作

行程时间估计。 TTE 通常涉及基于起讫点的 [27, 30] 和基于路径的问题设定 [10, 34]。在本文中，我们聚焦于后者，即对指定的目标路径进行估计。沿这一方向的现有研究主要可分为以路线为中心的方法和以链路为中心的方法。以路线为中心的方法将行驶路线作为一个整体进行建模，并以端到端的方式完成 TTE。早期的统计方法通过对相似历史轨迹的旅行时间求平均来进行近似估计 [35]。近期的深度学习方法利用循环神经网络等序列模型 [48, 50, 54]和注意力机制 [3, 13]，捕捉路线内链接之间更复杂的时空相关性。相比之下，以链路为中心的方法侧重于捕捉每个路线内链接周围的交通动态，以实现更具上下文感知能力的 TTE。一些研究构建了复杂的时空图神经网络（STGNNs）来估计链路行程时间，然后将其求和以得到路线 TTE [4, 5, 16, 17, 53, 55]。另一些研究则侧重于利用交通上下文表示 [9, 44] 或未来交通预测 [20]来丰富路线内链接的原始特征。然而，所有这些方法仍局限于局部交通上下文，未能充分释放整个城市道路网络中丰富的交通语义，从而增强每条路径的表示。

混合专家。 MoE 最初由 Jacobs等人 [19]提出，用于在神经网络中执行条件计算，以缓解任务间干扰。Shazeer 等人 [43] 首次利用稀疏门控机制将 MoE 扩展至十亿参数级架构。此后，路由机制 [29, 38], 专家设计 [21, 31], 分布式训练 [39],和基础设施 [31]方面的进展，进一步推动了 MoE 在包括计算机视觉 [40], 多模态学习在内的多个领域中的大规模应用，但自然语言处理除外学习 [37], 图学习 [12, 49] 和时间序列分析 [45]。MoE 技术也开始在交通领域崭露头角。ST-MoE [26] 和 DutyTTE [36] 采用稀疏门控混合专家框架，分别用于交通预测去偏和不确定性量化。TESTAM [24], CP-MoE [20] 和 MH-MoE [18] 则设计了异构专家，以捕捉多样的交通模式。然而，这些方法仅初步验证了 MoE 在交通预测中条件计算的优势，未能高效、稳定地处理具有长尾交通分布的大规模道路网络。

增量学习。 一般而言，增量学习（IL）有三种设置，包括类增量学习（CIL）、任务增量学习（TIL）和域增量学习（DIL）[46]。由于旅行时间不存在明确的类别边界，因此本文聚焦于 DIL，其目标是在持续从新的数据分布中学习的同时，避免灾难性遗忘问题。现有的 DIL 方法可分为三类：（1）基于正则化的方法 [23], ；（2）基于样本回放的方法 [41], ；以及（3）参数隔离与扩展技术 [42]。具体而言，在交通领域，这三类 DIL 方法均已用于应对交通分布偏移。例如，iETA [13] 采用自蒸馏正则化和周期性行程数据回放，以减轻增量学习过程中的灾难性遗忘。PECPM [47] 和 TEAM [22] 维护代表性模式库，以稳定地适应演化中的道路网络。TFMoE [25] 识别道路网络中变化最剧烈和最稳定的节点，以进行选择性更新。EAC [2] 引入轻量级提示，扩展模型的适应能力。本文开发了一种专为多层级 TTE 框架设计的基于参数隔离的增量学习策略，以支持高频适应，同时维持可持续的计算成本。

7 结论

本文提出了 MixTTE，这是一种可扩展且自适应的框架，通过协同整合链路级建模，克服滴滴当前以路线为中心的 TTE 系统感受野有限和长尾瓶颈的问题。借助时空外部注意力模块和外部稳定化图 MoE，MixTTE 能够高效捕获百万规模道路网络中的全局依赖关系和异质交通模式。量身定制的异步增量学习策略进一步使系统能够以稳定、高效的方式模块化适应频繁的交通变化。大量实验以及在滴滴网约车平台上的成功部署，验证了该方法的实用价值。未来，我们计划将 MixTTE 扩展为基础框架，充分发挥链路级建模的潜力，从而提升 TTE 以外的各类预测和决策服务。

8 致谢

本研究得到国家自然科学基金（项目编号：62572417、92370204）、国家重点研发计划（项目编号：2023YFF0725004）、广州市基础与应用基础研究计划（项目编号：2024A04J3279）、广州市教育局以及CCF-DiDi GAIA合作研究基金的资助。

参考文献

[1] Lei Bai、Lina Yao、Can Li、Xianzhi Wang 和 Can Wang。2020。用于交通预测的自适应图卷积循环网络。载于《神经信息处理系统进展》第33卷。[2] Wei Chen 和 Yuxuan Liang。 [n. d.]。扩展与压缩：探索持续时空图预测的调优原则。载于《第十三届国际学习表征会议》。[3] Zebin Chen、Xiaolin Xiao、Yue-Jiao Gong、Jun Fang、Nan Ma、Hua Chai 和 Zhiguang Cao。2022。从多视角解释轨迹：用于到达时间估计的层次化自注意力网络。载于《第28届ACM SIGKDD知识发现与数据挖掘会议》。2771–2779。[4] Austin Derrow-Pinion、Jennifer She、David Wong、Oliver Lange、Todd Hester、Luis Perez、Marc Nunkesser、Seongjae Lee、Xueying Guo、Brett Wiltshire 等。2021。利用Google 地图中的图神经网络预测ETA。载于《第30届ACM信息与知识管理国际会议论文集》。3767–3776。[5] Xiaomin Fang、Jizhou Huang、Fan Wang、Lingke Zeng、Haijin Liang 和 Haifeng Wang。2020。ConSTGAT：用于百度地图行程时间估计的上下文时空图注意力网络。载于《第26届ACM SIGKDD知识发现与数据挖掘会议》。2697–2705。[6] William Fedus、Jeff Dean 和 Barret Zoph。2022。深度学习中的稀疏专家模型综述。CoRR abs/2209.01667（2022）。[7] William Fedus、Barret Zoph 和 Noam Shazeer。2022。Switch Transformer：通过简单高效的稀疏性扩展至万亿参数模型。《机器学习研究杂志》23（2022），120:1–120:39。[8] Xiushi Feng、Shuncheng Liu、Haitian Chen 和 Kai Zheng。2023。利用不确定性感知生成式记忆回放进行持续轨迹预测。载于《IEEE国际数据挖掘会议（ICDM 2023）》，中国上海，2023年12月1日至4日。IEEE，1013–1018。[9] Kun Fu、Fanlin Meng、Jieping Ye 和 Zheng Wang。2020。CompactETA：用于旅行时间预测的快速推理系统。载于《第26届ACM SIGKDD知识发现与数据挖掘会议》。3337–3345。[10] Kun Fu、Fanlin Meng、Jieping Ye 和 Zheng Wang。2020。CompactETA：用于旅行时间预测的快速推理系统。载于《第26届ACM SIGKDD知识发现与数据挖掘会议》。3337–3345。[11] Meng-Hao Guo、Zheng-Ning Liu、Tai-Jiang Mu 和 Shi-Min Hu。2023。超越自注意力：用于视觉任务的双线性层外部注意力。《IEEE模式分析与机器智能汇刊》45卷5期（2023），5436–5447。[12] Zihao Guo、Qingyun Sun、Haonan Yuan、Xingcheng Fu、Min Zhou、Yisen Gao 和 Jianxin Li。2025。GraphMoRE：通过黎曼专家混合缓解拓扑异质性。载于《第三十九届AAAI人工智能会议论文集》。11754–11762。[13] Jindong Han、Hao Liu、Shui Liu、Xi Chen、Naiqiang Tan、Hua Chai 和 Hui Xiong。2023。iETA：用于到达时间估计的稳健且可扩展的增量学习框架。载于《第29届ACM SIGKDD知识发现与数据挖掘会议论文集》。4100–4111。[14] Jindong Han、Hao Liu、Haoyi Xiong 和 Jing Yang。2023。通过自监督分层图神经网络进行半监督空气质量预测。《IEEE知识与数据工程汇刊》35卷5期（2023），5230–5243。[15] Jindong Han、Weijia Zhang、Hao Liu、Tao Tao、Naiqiang Tan 和 Hui Xiong。2024。BigST：用于大规模道路网络交通预测的线性复杂度时空图神经网络。《VLDB基金会会刊》17卷5期（2024），1081–1090。[16] Huiting Hong、Yucheng Lin、Xiaoqing Yang、Zang Li、Kun Fu、Zheng Wang、Xiaohu Qie 和 Jieping Ye。2020。HetETA：用于到达时间估计的异构信息网络嵌入。载于《第26届ACM SIGKDD知识发现与数据挖掘会议》。2444–2454。[17] Jizhou Huang、Zhengjie Huang、Xiaomin Fang、Shikun Feng、Xuyi Chen、Jiaxiang Liu、Haitao Yuan 和 Haifeng Wang。2022。DuETA：利用高效图学习对交通拥堵传播模式建模，以预测百度地图上的ETA。载于《第31届ACM信息与知识管理国际会议论文集》。3172–3181。[18] Shaohan Huang、Xun Wu、Shuming Ma 和 Furu Wei。2024。MH-MoE：多头混合专家。CoRR abs/2411.16205（2024）。[19] Robert A. Jacobs、Michael I. Jordan、Steven J. Nowlan 和 Geoffrey E. Hinton。1991。自适应局部专家混合。《神经计算》3卷1期（1991），79–87。[20] Wenzhao Jiang、Jindong Han、Hao Liu、Tao Tao、Naiqiang Tan 和 Hui Xiong。2024。用于城市交通拥堵预测的可解释级联专家混合。载于《第30届ACM SIGKDD知识发现与数据挖掘会议论文集》。5206–5217。[21] Peng Jin、Bo Zhu、Li Yuan 和 Shuicheng Yan。2025。混合专家++：利用零计算专家加速混合专家方法。载于《第十三届国际学习表征会议》。[22] Duc Kieu、Tung Kieu、Peng Han、Bin Yang、Christian S. Jensen 和 Bac Le。2024。TEAM：用于交通预测的拓扑演化感知框架。《VLDB基金会会刊》18卷2期（2024），265–278。[23] James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、Joel Veness、Guillaume Desjardins、Andrei A Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska 等，2017。克服神经网络中的灾难性遗忘。《美国国家科学院院刊》114，13（2017），3521–3526。[24] Hyunwook Lee 和 Sungahn Ko。2024。TESTAM：一种融合专家混合的时间增强型时空注意力模型。载于第十二届国际学习表征会议。[25] Sanghyun Lee 和 Chanyoung Park。2024。通过专家混合实现持续交通预测。CoRR abs/2406.03140（2024）。[26] Shuhao Li、Yue Cui、Yan Zhao、Weidong Yang、Ruiyuan Zhang 和 Xiaofang Zhou。2023。ST-MoE：用于交通预测去偏的时空专家混合。载于第32届ACM信息与知识管理国际会议论文集。1208–1217。[27] Yaguang Li、Kun Fu、Zheng Wang、Cyrus Shahabi、Jieping Ye 和 Yan Liu。2018。用于行程时间估计的多任务表征学习。载于第24届ACM SIGKDD知识发现与数据挖掘国际会议论文集。1695–1704。[28] Jianqing Liang、Min Chen 和 Jiye Liang。2024。图外部注意力增强型Transformer。载于第41届国际机器学习会议论文集，第235卷。29560–29574。[29] Mengqi Liao、Wei Chen、Junfeng Shen、Shengnan Guo 和 Huaiyu Wan。2025。HMoRA：借助分层LoRA专家混合提升大语言模型的效能。载于第十三届国际学习表征会议。[30] Yan Lin、Huaiyu Wan、Jilin Hu、Shengnan Guo、Bin Yang、Youfang Lin 和 Christian S. Jensen。2023。面向基于地图的服务的起讫点旅行时间预言器。《ACM数据管理会刊》1，3（2023），217:1–217:27。[31] Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 等。2024。Deepseek-v3技术报告。arXiv预印本 arXiv:2412.19437（2024）。[32] Hangchen Liu、Zheng Dong、Renhe Jiang、Jiewen Deng、Jinliang Deng、Quanjun Chen 和 Xuan Song。2023。时空自适应嵌入使原始Transformer成为交通预测领域的最先进模型。载于第32届ACM信息与知识管理国际会议论文集。4125–4129。[33] Hao Liu、Wenzhao Jiang、Shui Liu 和 Xi Chen。2023。面向滴滴按需网约车服务的具备不确定性感知能力的概率旅行时间预测。载于第29届ACM SIGKDD知识发现与数据挖掘大会论文集。4516–4526。[34] Hao Liu、Wenzhao Jiang、Shui Liu 和 Xi Chen。2023。面向滴滴按需网约车服务的具备不确定性感知能力的概率旅行时间预测。载于第29届ACM SIGKDD知识发现与数据挖掘大会论文集。4516–4526。[35] Wuman Luo、Haoyu Tan、Lei Chen 和 Lionel M. Ni。2013。在大规模轨迹数据中查找基于时间段的最频繁路径。载于2013年ACM SIGMOD国际数据管理会议论文集。713–724。[36] Xiaowei Mao、Yan Lin、Shengnan Guo、Yubin Chen、Xingyu Xian、Haomin Wen、Qisen Xu、Youfang Lin 和 Huaiyu Wan。2025。DutyTTE：解析起讫点旅行时间估计中的不确定性。载于第三十九届AAAI人工智能会议论文集。12390–12398。[37] Basil Mustafa、Carlos Riquelme、Joan Puigcerver、Rodolphe Jenatton 和 Neil Houlsby。2022。使用LIMoE进行多模态对比学习：语言图像专家混合。载于《神经信息处理系统进展》第35卷。[38] Joan Puigcerver、Carlos Riquelme Ruiz、Basil Mustafa 和 Neil Houlsby。2024。从稀疏专家混合到软专家混合。载于第十二届国际学习表征会议。[39] Samyam Rajbhandari、Conglong Li、Zhewei Yao、Minjia Zhang、Reza Yazdani Aminabadi、Ammar Ahmad Awan、Jeff Rasley 和 Yuxiong He。2022。DeepSpeed-MoE：推进专家混合推理与训练，赋能下一代AI规模化发展。载于第39届国际机器学习会议论文集，第162卷。18332–18346。[40] Carlos Riquelme、Joan Puigcerver、Basil Mustafa、Maxim Neumann、Rodolphe Jenatton、André Susano Pinto、Daniel Keysers 和 Neil Houlsby。2021。利用稀疏专家混合扩展视觉模型。载于《神经信息处理系统进展》第34卷。8583–8595。[41] David Rolnick、Arun Ahuja、Jonathan Schwarz、Timothy Lillicrap 和 Gregory Wayne。2019。用于持续学习的经验回放。《神经信息处理系统进展》第32卷（2019）。[42] Amir Rosenfeld 和 John K. Tsotsos。2020。通过深度适应实现增量学习。《IEEE模式分析与机器智能汇刊》42，3（2020），651–663。[43] Noam Shazeer、Azalia Mirhoseini、Krzysztof Maziarz、Andy Davis、Quoc V. Le、Geoffrey E. Hinton 和 Jeff Dean。2017。超大规模神经网络：稀疏门控专家混合层。载于第五届国际学习表征会议。[44] Yibin Shen、Cheqing Jin、Jiaxun Hua 和 Dingjiang Huang。2022。TTPNet：一种基于张量分解和图嵌入的旅行时间预测神经网络。《IEEE知识与数据工程汇刊》34，9（2022），4514–4526。[45] Xiaoming Shi、Shiyu Wang、Yuqi Nie、Dianqi Li、Zhou Ye、Qingsong Wen 和 Ming Jin。 [无日期。]Time-MoE：十亿规模时间序列基础模型与

专家混合。载于第十三届国际学习表征会议。[46] Gido M. van de Ven、Tinne Tuytelaars 和 Andreas S. Tolias。2022。增量学习的三种类型。《自然·机器智能》4，12（2022），1185–1197。[47] Binwu Wang、Yudong Zhang、Xu Wang、Pengkun Wang、Zhengyang Zhou、Lei Bai 和 Yang Wang。2023。演化图上的模式扩展与巩固：用于持续交通预测。载于第*29*届*ACM SIGKDD*知识发现与数据挖掘大会论文集。2223–2232。[48] Dong Wang、Junbo Zhang、Wei Cao、Jian Li 和 Yu Zheng。2018。你何时抵达？基于深度神经网络的行程时间估计。载于第三十二届*AAAI*人工智能会议论文集。2500–2507。[49] Haotao Wang、Ziyu Jiang、Yuning You、Yan Han、Gaowen Liu、Jayanth Srinivasa、Ramana Kompella 和 Zhangyang Wang。2023。图专家混合：通过显式多样性建模在大规模图上学习。载于《神经信息处理系统进展》第*36*卷。[50] Zheng Wang、Kun Fu 和 Jieping Ye。2018。学习估计旅行时间。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。858–866。[51] Felix Wu、Amauri H. Souza Jr.、Tianyi Zhang、Christopher Fifty、Tao Yu 和 Kilian Q. Weinberger。2019。简化图卷积网络。载于第*36*届国际机器学习会议论文集，第97卷。6861–6871。[52] Haixu Wu、Huakun Luo、Haowen Wang、Jianmin Wang 和 Mingsheng Long。2024。Transolver：一种适用于一般几何形状偏微分方程的快速Transformer求解器。载于第*41*届国际机器学习会议论文集，第235卷。29560–29574。[53] Haitao Yuan、Guoliang Li 和 Zhifeng Bao。2022。重新审视道路网络中的路径旅行时间估计：异质性、邻近性、周期性与动态性。《*VLDB*基金会会刊》16，3（2022），393–405。[54] Hanyuan Zhang、Hao Wu、Weiwei Sun 和 Baihua Zheng。2018。DeepTravel：一种具备辅助监督的基于神经网络的行程时间估计模型。载于第二十七届国际人工智能联合会议论文集。3655–3661。[55] Hanyuan Zhang、Xinyu Zhang、Qize Jiang、Liang Li、Baihua Zheng 和 Weiwei Sun。2024。通过跨视图位置对齐增强的空间拓扑感知双变换器实现行程时间估计。《*IEEE*智能交通系统汇刊》25，12（2024），20508–20522。

A 方法详情

A.1 STEA 中的时间内空间层次建模

尽管外部记忆单元能够连接交通模式在空间和时间上的交互，但在实际应用中，强化时间步内的全局空间相关性建模仍然十分重要。如第 3.1 节和图 2 所述，我们采用空间层次建模方法 [14, 52] 来实现这一目标。通过将细粒度的链路级交通上下文投影为语义令牌，我们能够更高效地建模全局空间相关性，同时与 EKR 步骤无缝衔接。下面将进一步介绍实现层次建模所必需的软聚类和广播步骤。给定路段表示$\mathbf{H} \in \mathbb{R}^{N\times d}$，软聚类步骤利用一个可学习的权重矩阵$\mathbf{W}_{sc} \in \mathbb{R}^{d\times U_{in}}$，并通过 Softmax 激活为每条链路生成软聚类权重，即，

$$\alpha_i = \text{Softmax}\left(\mathbf{h}_i \cdot \mathbf{W}_{sc}\right) \in \mathbb{R}^{U_{in}}, \quad \forall i = 1, \ldots, N. \tag{10}$$

然后，我们根据软聚类权重聚合路段表示，得到交通语义令牌，即，

$$\mathbf{h}_i^{se} = \sum_{j=1}^{N} \alpha_{ij} \cdot \mathbf{h}_j \mathbf{W}_{proj} \in \mathbb{R}^d, \quad \forall i = 1, \ldots, U_{in}, \tag{11}$$

其中，$\mathbf{W}_{proj} \in \mathbb{R}^{d\times d}$ 是一个可学习的权重矩阵，用于将交通切片编码投影到统一的语义空间中。对这些语义令牌执行自注意力和 EKR 步骤后，我们将得到外部增强的语义令牌 $\mathbf{H}^{se,ex} \in \mathbb{R}^{U_{in}\times d}$ 这些令牌同时吸收了时间内和时间间的全局交通相关性。最后，我们根据软聚类权重，将这些语义令牌广播（BC）回链路，从而生成外部增强的路段表示，即，

$$\mathbf{h}_i^{ex} = \sum_{j=1}^{U_{in}} \alpha_{ij} \cdot \mathbf{h}_j^{se,ex} \in \mathbb{R}^d, \quad \forall i = 1, \ldots, N. \tag{12}$$

B 实验细节

B.1 热门链接选择

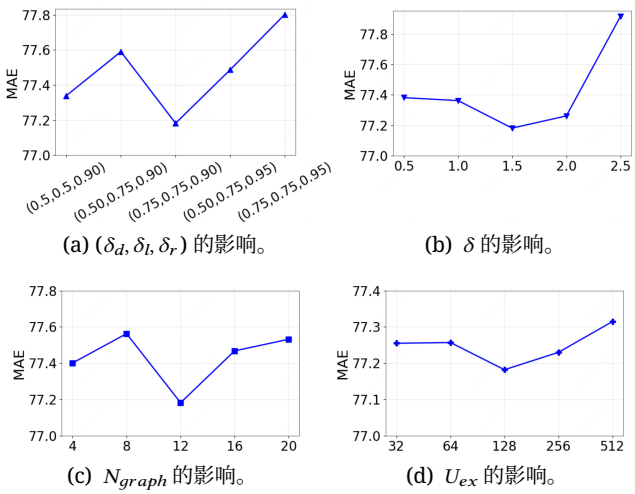
每天，我们综合考虑每条链接的拥堵程度以及经过该链接的行程数量，生成其热门度特征。超过预定义热门度阈值的链接被视为“热点路段”。随后，我们每周汇总过去一个月内的热点路段，形成下一周链接级模型训练所用的目标热点路段集合。在实际应用中，我们还会纳入所选热点路段的相邻路段，以避免它们在交通网络中孤立。对于百万规模的城市道路网络，这一过程不仅高效，而且有助于识别呈现显著交通模式的路段。这些路段应是链路级建模的重点，以便为以路线为中心的 TTE 系统提供切实的交通上下文信息。

B.2 实现细节

本节介绍 MixTTE 的实现细节。MixTTE 的隐藏维度在链路级模型中设为 64，在以路线为中心的模型中设为 256。在 STEA 模块中，外部记忆单元数量为 128，每个记忆单元的维度为 64。我们还在 EKR 和自注意力步骤中均采用了 8 头多头注意力。在 ESGMoE 模块中，ESGMoE 共包含 2 层，每层由 16 个专家组成，包括 12 个图专家、 1 个零专家、 1 个恒等专家和 2 个常量专家。每个图专家的感受野为 8 跳。每次激活排名前 2 的专家。北京的专家负载均衡损失系数设为 1000，南京和苏州设为 100。北京的零计算专家权重设为 0.5，南京和苏州设为 1.5。在 ASIL 模块中，检测异常链接的百分位数设为 0.75。用于激活链路级模型更新以及路径级和链路级模型更新的百分位数分别设为 0.75 和 0.9。我们采用 Adam 优化，学习率为 10^{-3}，权重衰减率为 5×10^{-7}，dropout 率为 0.3，并训练 2 个 epoch。

B.3 超参数敏感性分析

我们在苏州数据集上测试 MixTTE 对以下 四个超参数的敏感性：(1) ASIL 中的阈值（δ_d，δ_l，δ_r），(2) 零计算专家的负载均衡权重（δ），(3) 图专家的数量（N_{graph}），以及 (4) 外部单元的数量（U_{ex}）。MAE指标的性能变化如图 7所示。总体而言，*MixTTE* 的性能波动处于可接受范围内，表明其对这些超参数具有良好的鲁棒性。具体来说，随着每个超参数变化，性能会上下波动。对于（，δ_l，δ_r），阈值过低会导致模型受到噪声影响而不稳定，而阈值过高则会延迟对



(a) $(\delta_d, \delta_l, \delta_r)$ 的影响。 (b) δ 的影响。

(c) N_{graph} 的影响。 (d) U_{ex} 的影响。

图 7：苏州数据集上的参数敏感性分析。

真实交通变化的适应。对于 δ，取值过小可能会将复杂模式错误地路由至零计算专家，而取值过大则可能在频繁模式上浪费图专家。对于 N_{graph}，数量过少会导致模型容量不足，而数量过多则会造成训练不足和专业化程度低。对于 U_{ex}，数量不足会限制模型对全局上下文的感知能力，而数量过多则会增加过拟合风险。

C 未来工作

展望未来，我们计划从三个关键维度拓展 MixTTE 的能力，将其发展为更全面的基础框架。首先，我们计划纳入天气状况和社会事件等多模态数据，以进一步缓解城市场景中的部分可观测性问题，并提升模型的鲁棒性和泛化能力。STEA 模块的模态无关特性可自然支持将这些多模态因素整合到统一嵌入空间中，从而构建更全面的记忆库。其次，我们将研究跨城市迁移学习技术，以在不同城市网络间泛化交通模式，大幅降低将系统部署到新城市所需的校准成本。外部记忆和异构混合专家编排已为抽象更多跨城市知识奠定了架构基础。最后，我们计划探索轻量级模型蒸馏技术，以便部署我们计划构建的大规模跨城市、跨模态基础模型。