Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(2), 672-683
Published Online March 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.152064

Hans汉斯

基于未来潜状态交叉注意力的多模态轨迹预测方法

王其程

西华大学汽车与交通学院，四川 成都

收稿日期：2026年2月3日；录用日期：2026年3月17日；发布日期：2026年3月30日

摘　要

准确预测多个交通参与者的未来运动轨迹对于实现安全可靠的自动驾驶至关重要。尽管近年来的轨迹预测方法通过建模智能体交互已展现出优异性能，但现有方法主要关注历史交互而忽视了智能体未来运动之间可能出现的复杂依赖关系，这导致预测轨迹在密集交通场景中可能缺乏全局交互一致性。为了应对这一挑战，本文设计了一种基于未来轨迹交互建模的轨迹预测算法框架，通过考虑未来潜在空间内的多模态交互关系，在解码阶段对其进行显式建模。具体而言，本方法将轨迹预测分解为基于历史的粗粒度预测和基于全局交互一致性的未来轨迹优化两个阶段。第一阶段通过矢量化场景表征，采用分层式编码机制融合局部上下文特征与全局场景交互特征，从而实现对各种交通场景中丰富的历史时空交互进行建模。为捕捉历史观测之外的交互关系，第二阶段进一步在解码器中引入未来潜在状态的空间交叉注意力模块(Future Latent Cross-Attention, FLCA)，设计了跨智能体的交互掩码机制，使每个预测模态都能关注其他智能体的未来运动，同时自身不同模态之间不出现干涉。最后，在大规模自动驾驶基准数据集Argoverse1上的实验表明，本方法能够实现更具全局交互一致性和准确性的轨迹预测。

关键词

轨迹预测，多模态轨迹，时空交互，未来潜状态

A Multimodal Trajectory Prediction Method Based on Future Latent State Cross-Attention

Qicheng Wang

School of Automotive and Transportation Engineering, Xihua University, Chengdu Sichuan

Received: February 3, 2026; accepted: March 17, 2026; published: March 30, 2026

文章引用: 王其程. 基于未来潜状态交叉注意力的多模态轨迹预测方法[J]. 人工智能与机器人研究, 2026, 15(2): 672-683. DOI: 10.12677/airr.2026.152064

Abstract

Accurate prediction of the future movement trajectories of multiple traffic participants is crucial for achieving safe and reliable autonomous driving. Although recent trajectory prediction methods have demonstrated excellent performance by modeling agent interactions, existing approaches mainly focus on historical interactions while neglecting the complex dependencies that may arise between agents' future movements. This leads to the possibility that the predicted trajectories may lack global interaction consistency in dense traffic scenarios. To address this challenge, this paper designs a trajectory prediction algorithm framework based on future trajectory interaction modeling, explicitly modeling the multimodal interaction relationships in the future potential space during the decoding stage. Specifically, this method decomposes trajectory prediction into two stages: coarse-grained prediction based on history and future trajectory optimization based on global interaction consistency. In the first stage, a vectorized scene representation is used, and a hierarchical encoding mechanism is adopted to fuse local context features and global scene interaction features, thereby enabling the modeling of rich historical spatiotemporal interactions in various traffic scenarios. To capture interaction relationships beyond historical observations, the second stage further introduces a Future Latent Cross-Attention (FLCA) module in the decoder and designs an interaction masking mechanism across agents, allowing each prediction modality to focus on the future movements of other agents while avoiding interference among different modalities of the same agent. Finally, experiments on the large-scale autonomous driving benchmark dataset Argoverse1 show that this method can generate trajectory prediction results with better global interaction consistency and accuracy.

Keywords

Trajectory Prediction, Multimodal Trajectory, Spatiotemporal Interaction, Future Latent State

1. 引言

轨迹预测对于自动驾驶系统的安全与可靠至关重要，其作用是对周边车辆在复杂交通环境中的未来位置进行动态推测[1]。精准的轨迹预测不仅能够为自动驾驶系统决策提供前瞻性的信息，使其高效、准确规划自身的行驶路径，同时能够有效规避潜在的碰撞风险。

从任务形式上看，轨迹预测通常被建模为一种流式时序任务，即系统在每一时刻接收最新的感知数据，并基于有限长度的历史观测对未来一段时间内的运动轨迹进行预测[2]。该过程需要综合考虑交通参与者自身的运动惯性特征、周围智能体的交互影响以及道路结构等因素。随着人工智能技术的发展，基于深度学习的多模态轨迹预测方法因其能够同时关注车辆运动特征，驾驶意图、交通环境等多种影响因素，及其对应的不确定性问题，并以多模态的方式输出预测结果，对轨迹预测的灵活性与准确性有显著提升，从而成为研究的主要趋势。

尽管深度学习为多模态预测提供了思路，但预测性能的进一步提升，还依赖于对交通参与者间交互关系的有效建模。基于深度学习的交互建模方法主要分为时间交互建模与空间交互建模。基于注意力机制(Transformer) [3]或图(Graph)结构[4] [5]的方法为复杂交通场景下的时空交互建模提供了全新思路，其在时间维度上提取历史轨迹数据的深层次依赖特征，并在空间维度上捕捉车辆与环境间的复杂交互关系，

从而更加全面地支持轨迹预测任务。基于注意力机制的方法通过编码历史轨迹的时间自依赖性与空间交互关系来隐式表示智能体间的交互信息，并假设这种历史交互信息足以支撑未来生成。而基于图结构的方法则显式编码智能体、环境及交互关系，支持动态权重分配与端到端多模态输出，在可扩展性、长程交互建模和复杂场景泛化性上显著优于传统方法。

尽管现有方法在多模态轨迹预测的建模能力与精度上取得了长足进步，但其交互建模范式仍存在明显局限。主流方法普遍将交互建模集中在历史阶段，模型未能将未来潜在状态视为一个具有复杂交互关系的特征空间。此外，主流方法默认在解码阶段，不同智能体、不同模态的未来轨迹之间在各个时间步上都是条件独立的。这导致模型直接基于各自的历史条件表示进行独立解码，而缺乏一个能够学习并推理预测目标间复杂空间约束关系的显式机制。

针对上述问题，本文提出了一种基于未来潜在状态交叉注意力(Future Latent Cross-Attention, FLCA)的多智能体轨迹预测方法。该方法将预测流程分为基于历史观测的粗粒度轨迹生成和基于全局交互一致性的轨迹优化两个阶段。在第一阶段，模型通过对矢量化场景进行分层编码，提取智能体的局部运动特征及其与周围环境的交互信息，并进一步整合全局空间依赖关系，形成统一的历史场景表征。基于此，解码器并行生成所有智能体的多模态粗粒度轨迹。在第二阶段，通过引入 FLCA 模块，显式建模未来潜在状态下跨智能体的交互关系。该模块利用交叉注意力机制使不同智能体的轨迹模态能够相互协调，并通过掩码机制避免模态间干扰。最终，FLCA 以残差优化方式对粗粒度轨迹进行修正，输出更准确的预测轨迹。

本文的主要贡献包括：

1) 本文提出了一种两段式的轨迹预测框架，第一阶段基于历史交互信息生成粗粒度的候选轨迹，并在第二阶段进一步考虑未来潜在状态交互，生成更具全局交互一致性的优化轨迹，实现了对预测轨迹准确性的提升。

2) 设计了一种交互掩码机制，构建了基于未来潜在状态交互的交叉注意力模型，能够在预测阶段建模跨智能体的交互，动态调整不同模态未来运动之间的相对关系，从而生成更加合理的多模态轨迹预测结果。

2. 相关工作

2.1. 轨迹预测

轨迹预测是多变量时间序列(Multi-variate Time Series, MTS)预测任务，经常使用递归类深度学习模型进行预测[6]。智能体行为与轨迹预测方法主要有动力学模型，基于运动行为的方法、深度学习方法[7]。

运动学和动力学参数通常用于早期的智能体轨迹预测任务。例如，恒横摆角速度和加速度模型常被采用，因为它假定平稳行驶期间横摆角速度和加速度不会在短时间内突然变化。卡尔曼滤波器如无迹卡尔曼滤波器和扩展卡尔曼滤波器也被用作轨迹预测的模型[8]。然而，这些模型在长期预测任务中表现不佳，因为短期假设(智能体动力学参数短期内不会突变)不再适用于长期预测。

基于智能体运动行为的模型能更好地反映智能体的长期意图，提高长视野轨迹预测的准确性。智能体行为识别主要基于智能体历史轨迹和运动状态。Xie [9]等人将物理方法和驾驶行为方法相结合，提出了一种车辆轨迹预测方法，考虑到车辆动力学与运动学参数，基于物理的轨迹预测方法能够在短期内保证精度，基于行为的预测方法通过行为估计对未来的轨迹具有长期的预测能力。将两种预测模型相结合，提出了交互式多模型轨迹预测方法。交互多模型中各模型的概率可根据各模型的预测方差递归调整。比较结果表明，交互式多模型轨迹预测可以在较长的预测周期内获得准确的预测轨迹。此外，隐马尔科夫

模型[10]、高斯过程回归(Gaussian Process, GP) [11]、支持向量机[12]和概率有限状态机[13]等机器学习方法也陆续被用于行为与轨迹预测。然而，传统方法所使用的手工特征难以表征多模态行为模式，因此其长期水平轨迹预测的精度仍然相对较低。

此后，随着深度学习技术的迅速发展，该方法也被引入到轨迹预测任务中。由于 RNN 或 LSTM 能够在上下文时间步长中提取隐藏的依赖关系，因此早期大多数相关工作都是基于 RNN 或 LSTM 等网络结构。它们的功能是对序列的每个输入项执行相同的操作，同时考虑前一个时间步的输入。由于智能体轨迹预测是一个序列对序列(Seq2Seq)的任务，因此采用 LSTM 编解码框架是非常常见的[14] [15]。Alexandre Alahi，Li Fei-Fei 等人[16]提出了一种“social-LSTM”的算法架构，引入了一个“Social”池化层，允许空间邻近序列的 LSTM 共享彼此的隐藏状态，该架构可以自动学习在时间上吻合的轨迹之间发生的典型交互。此外，注意力机制 Attention [3]、图神经网络 GNN 等也被用于轨迹预测中，以提取更丰富的交互信息。Messaoud 等人[17]使用一种注意力机制，明确强调邻近车辆对自车未来状态的重要性。该方法不仅考虑成对的车辆交互作用，并对高阶交互作用进行了建模。Multimodal Transformer [18]引入基于 Transformer 结构的神经网络预测框架，建模智能体之间的相互作用，并提取预测目标对地图路径点的注意力。Vector Net [19]通过向量的形式表示道路上的各种元素，避免采用有损渲染的栅格图输入与密集计算的卷积网络编码。各向量通过图卷积神经网络连接，以建模各向量元素之间的连接关系。它也被后续的一些工作如 Dense TNT [20]、mmTransformer [21]用作骨干网络。为了捕获更复杂的道路拓扑与长距离依赖关系，LaneGCN [22]被提出，并且通过场景与智能体之间的融合网络整合全局信息，用于场景内所有智能体的轨迹预测。HiVT [1]在此基础上提出了一种分层向量化建模框架，通过局部编码器提取个体历史运动特征，并利用全局交互模块高效建模多智能体之间的历史交互关系。同时，HiVT 通过平移不变表示和旋转不变空间学习模块，提高了模型对场景几何变换的鲁棒性，使其能够在保持高精度的同时实现高效的多智能体预测。

2.2. 交互建模方法

1) 时序交互建模：时间交互建模关注单个智能体自身运动状态在历史时间维度上的依赖关系，其核心目标是刻画历史轨迹对未来运动的影响。这类方法通常基于序列建模框架，通过递归结构或注意力机制提取时间序列中的长期依赖关系。建模方法多采用 RNN、LSTM 或 GRU 等循环神经网络结构，对历史轨迹进行编码。随着 Transformer 架构的发展，基于时间注意力机制的方法逐渐被引入轨迹预测任务，通过自注意力机制建模不同时间步之间的依赖关系，从而提升对长期运动模式的建模能力[23]。但需要指出的是，尽管目前主流预测方法都进行了时间维度的自依赖性建模，但其核心仍然是针对历史信息建模，并未对未来潜在状态空间下的多维特征进行表示。

2) 空间交互建模：空间交互建模主要关注同一时间步下不同智能体之间的相互影响。这类方法通常基于智能体之间的相对位置、距离关系或邻接结构，对空间维度上的交互关系进行显式建模。典型方法包括社会池化(Social Pooling) [16]或图结构[24]的深度学习模型。这些方法通过聚合邻近智能体的信息，使模型能够感知交通环境中的交互关系，从而提升预测合理性。但这类方法的交互建模依然发生在历史阶段，对未来阶段不同预测目标与模态之间的空间约束关系缺乏显式表征。

3) 未来潜状态空间交互：在多模态轨迹预测框架中，模型通常会生成多条候选未来轨迹，用以描述交通场景中的不确定性。然而，现有方法往往将不同预测模态视为相互独立，仅通过模态概率进行筛选，而忽略了不同模态及不同智能体在未来空间中的潜在交互结构。实际上，不同智能体的未来轨迹之间往往存在显式或隐式的约束，例如空间碰撞冲突、避让关系以及协同行为等。这些约束并不完全由历史信息决定，而是在未来演化过程中逐渐显现。

基于上述分析，本文提出在解码阶段引入未来潜状态空间的交互建模机制，通过未来潜状态交叉注意力(Future Latent Cross-Attention, FLCA)，对不同智能体及其不同预测模态的未来潜在表示进行交互建模，从而在保持历史建模结构不变的前提下，引导模型生成更加合理的多模态预测结果。

3. 方法

3.1. 问题定义

在多智能体轨迹预测任务中，给定场景中所有交通参与者在历史时间窗口内的观测信息，目标是预测各智能体在未来时间范围内可能的运动轨迹。设场景中共有 N 个智能体，其历史轨迹集合表示为

$$H_N = \left\{ X_i^{1:T_h} \mid i = 1, 2, \cdots, N \right\} \tag{1}$$

其中，$X_i^t \in \mathbb{R}^2$ 表示第 i 个智能体在 t 时刻的二维位置坐标，T_h 为历史观测长度。

而预测目标是在给定历史信息 H_N 以及地图等静态环境信息 M 的条件下，为每个智能体生成未来 T_f 个时间步的多模态轨迹集合 F_N^k 及模态对应的置信度 pi_N^k。综上，基于历史轨迹的预测任务可以表示为以下公式：

$$\left\{ F_N^k, pi_N^k \right\}_{k=1}^{K} = \mathcal{P}\left(H_N, \mathcal{M} \right) \tag{2}$$

其中，k 表示轨迹模态数，$\mathcal{P}(\,)$ 表示预测模型。

然而，大多数现有方法默认不同模态之间在生成过程中相互独立，这一假设忽略了未来不同假设之间潜在的约束与相互影响，例如空间冲突、速度一致性或多智能体博弈关系。本文认为，多模态未来轨迹在潜在空间中并非独立样本，而应被视为一个具有内在关联的集合，即：

$$p\left(\mathcal{F}_N \mid H_N, \mathcal{M} \right) \neq \prod_{k=1}^{K} p\left(F_N^k \mid H_N, \mathcal{M} \right) \tag{3}$$

其中，$\mathcal{F}_N$ 表示由 F_N^k 组成的，所有交通参与者的多模态长序列预测轨迹的集合。

3.2. 整体架构

本文提出的 FLCA 多智能体轨迹预测方法将预测过程划分为两个阶段：第一阶段基于历史观测生成多模态的粗粒度轨迹，第二阶段则在全局交互一致性的约束下对这些轨迹进行优化。

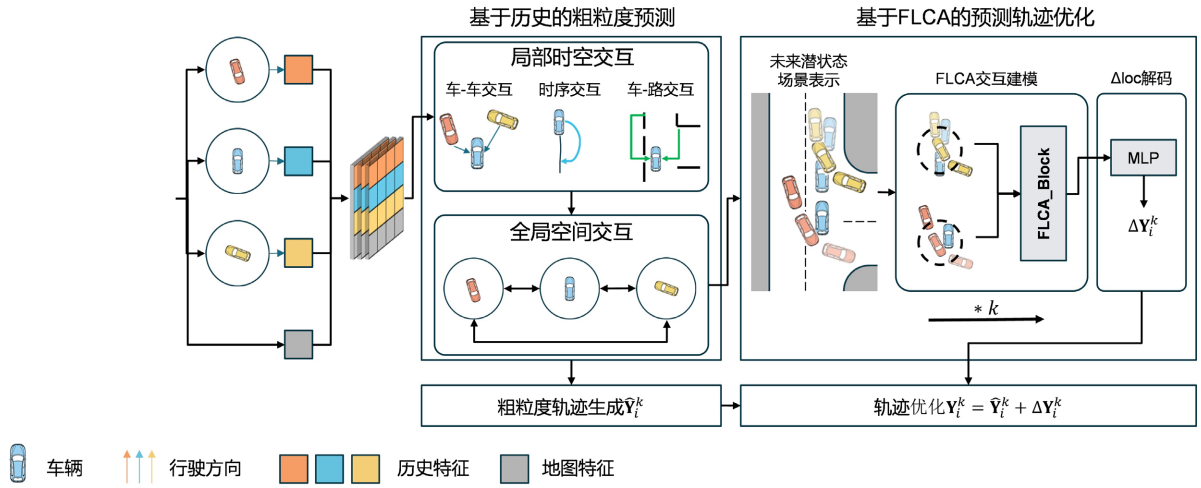

Figure 1. FLCA prediction algorithm framework
图 1. FLCA 预测算法框架

在基于历史的粗粒度预测阶段，模型首先对场景进行分层编码。通过矢量化表征提取智能体的局部上下文特征，由于周边智能体的位置与行为是对中心智能体影响最直接的因素，因此先在每个独立时间步上建模车–车交互，然后在时间维度上融合这些社交化特征，最后引入相对稳定的环境约束，即车–路交互。为进一步整合场景的全局信息，避免局部特征提取带来的感知野受限，信息遗失等问题，模型将这些局部特征输入一个全局交互模块，以捕捉长距离的空间依赖关系，最终形成对场景历史的统一表征。基于该表征，解码器为所有智能体并行生成一组多模态的未来粗轨迹。算法框架如图 1 所示。

在随后的轨迹优化阶段，模型专注于提升这些粗粒度轨迹的未来交互。为此，引入了 FLCA 模块，在解码过程中对未来潜在状态进行交互建模。该模块通过交叉注意力机制，使每个智能体的每一个轨迹模态都能感知其他智能体所有可能的未来运动状态，并利用交互掩码防止同一智能体内不同模态间的相互干扰。FLCA 模块以残差学习的方式输出轨迹偏移量，对第一阶段生成的初始轨迹进行细粒度调整，最终输出更准确的多模态预测结果。

3.3. 基于历史的粗粒度预测

为了使模型能够在各种交通环境中准确预测周边智能体的未来轨迹，本文采用分层式编码器结构对各智能体之间的时空交互关系进行充分的建模。该编码器由局部时空编码和全局交互编码两阶段组成。

1) 局部时空编码器首先将整个交通场景划分，对于整个场景中的第 i 个智能体，构建一个以其为中心的局部块 $\mathbb{B}_i$。由于在轨迹预测任务中，交通参与者之间的交互行为以及道路拓扑结构对未来运动行为具有决定性影响。因此在局部时空编码器中进一步对该局部块内的车车交互与车路交互进行空间维度上的交叉注意力建模。

对于车–车交互建模，以目标智能体 i 的历史轨迹序列作为空间交叉注意力的查询，邻居智能体 j 的历史轨迹序列作为键和值，即

$$Emb_i^{\mathrm{A-A}} = \mathrm{Attn}_{\mathrm{A-A}}\left(H_i, H_j\right) \tag{4}$$

其中，$Emb_i^{\mathrm{A-A}}$ 表示模型对第 i 个智能体的车车交互建模后的空间层嵌入。

$$\mathrm{Attn}_{\mathrm{A-A}}\left(Q_{\mathrm{A-A}}, K_{\mathrm{A-A}}\right) = \mathrm{softmax}\left(\frac{Q_{\mathrm{A-A}} K_{\mathrm{A-A}}^{\top}}{\sqrt{d}}\right) V_{\mathrm{A-A}} \tag{5}$$

其中，$Q_{\mathrm{A-A}} = H_i W_Q^{\mathrm{A-A}}$，$K_{\mathrm{A-A}} = H_j W_K^{\mathrm{A-A}}$，$V_{\mathrm{A-A}} = H_j W_V^{\mathrm{A-A}}$。

这种建模方式使得目标智能体可以自适应地关注对其未来运动影响更为显著的周边智能体。

其次，在时间维度上对目标智能体历史轨迹序列进行建模的具体实现为：

$$Emb_i^{\mathrm{Temp}} = \mathrm{Attn}_{\mathrm{temp}}\left(Emb_i^{\mathrm{A-A}}\right) \tag{6}$$

其中，Emb_i^{Temp} 表示模型对第 i 个智能体的时间维度特征进行建模计算后得到的时间层嵌入变量。

$$\mathrm{Attn}_{\mathrm{temp}}\left(H_i\right) = \mathrm{softmax}\left(\frac{Q_t K_t^{\top}}{\sqrt{d}}\right) V_t \tag{7}$$

其中，$Q_t = H_i W_Q^t$，$K_t = H_i W_K^t$，$V_t = H_i W_V^t$。

该结构能够充分考虑不同预测目标在不同历史时间步之间的依赖关系，自适应的分配注意力权重。

在完成了动态目标的时空交互建模以后，模型进一步考虑同一局部块 $\mathbb{B}_i$ 内地图信息 $\mathcal{M}_i$ 对预测目标未来行为的交互影响，即为车–路交互建模。该模块以整合了车车交互与自身时序依赖信息的 Emb_i^{Temp} 作为空间交叉注意力的查询，以地图信息 $\mathcal{M}_i$ 作为键和值，具体实现为：

$$Emb_i^{\mathrm{L}} = \mathrm{Attn}_{\mathrm{A-M}}\left(Emb_i^{\mathrm{temp}}, \mathcal{M}_i\right) \tag{8}$$

$$\mathrm{Attn}_{\mathrm{A-M}}\left(Q_{\mathrm{A-M}},K_{\mathrm{A-M}}\right)=\mathrm{softmax}\left(\frac{Q_{\mathrm{A-M}}K_{\mathrm{A-M}}^{\top}}{\sqrt{d}}\right)V_{\mathrm{A-M}} \tag{9}$$

其中，$Q_{\mathrm{A-M}}=Emb_i^{\mathrm{temp}}W_Q^{\mathrm{A-M}}$，$K_{\mathrm{A-M}}=\mathcal{M}_tW_K^{\mathrm{A-M}}$，$V_{\mathrm{A-M}}=\mathcal{M}_tW_V^{\mathrm{A-M}}$。

通过上述建模，最后输出结果 Emb_i^{L} 即为整个局部时空编码器的编码输出。该变量既包含了丰富的场景智能体空间交互关系，道路拓扑结构等地图信息对预测目标未来运动的约束、导向等影响关系，又在时间维度上对目标运动的内在约束进行了充分的考虑。

2) 全局交互编码器，用于建模不同局部区域之间的长距离依赖关系，从而补偿局部建模带来的感受野限制。该编码器将局部时空编码阶段得到的局部块 $\mathbb{B}_i$ 的丰富交互表示 Emb_i^{L} 视为建模计算单位，并在 N 个局部块之间显式建模跨区域交互，即

$$Emb_i^{\mathrm{G}}=\sum_{j=1}^{N}\alpha_{ij}\cdot\left(Emb_j^{\mathrm{L}}W_V^{G}\right) \tag{10}$$

其中，α_{ij} 表示注意力权重，定义为：

$$\alpha_{ij}=\frac{\exp\left(\frac{\left(Emb_i^{\mathrm{L}}W_Q^{G}\right)\left(Emb_j^{\mathrm{L}}W_K^{G}\right)^{\top}}{\sqrt{d}}\right)}{\sum_{k=1}^{N}\exp\left(\frac{\left(Emb_i^{\mathrm{L}}W_Q^{G}\right)\left(Emb_k^{\mathrm{L}}W_K^{G}\right)^{\top}}{\sqrt{d}}\right)} \tag{11}$$

该过程允许每一个局部上下文根据其与其他区域的相关性，自适应地聚合全局信息，从而补偿局部建模在空间尺度上的不足。最终得到的全局表示 Emb_i^{G} 作为条件信息，为后续的未来潜状态空间提供了丰富的场景表征。

在分层式历史编码器完成对历史时序依赖与空间交互的建模后，解码器接收到的潜在特征表示已包含了智能体的未来运动趋势信息。对其直接解码，可视为一种粗粒度预测结果，主要反映智能体在历史条件约束下的运动趋势，即

$$\hat{\mathbf{Y}}_i^{k}=\mathrm{MLP}_{\mathrm{loc}}\left(Emb_i^{\mathrm{G}}\right)\in\mathbb{R}^{F\times 2} \tag{12}$$

其中 F 表示预测时间步数。

3.4. 基于 FLCA 的预测轨迹优化

尽管分层式历史编码器的输出 Emb_i^{G} 尚未被解码为显式的空间轨迹，但其语义上已经包含了智能体未来运动行为的高层表征。这是因为该变量是基于历史信息建模得到的，编码过程已隐式学习了历史交互模式与未来演化趋势之间的映射关系。因此，本方法将 $\mathbf{Z}$ 其视为一个定义在未来潜在空间中的多模态状态集合，即

$$\mathbf{Z}=\left\{\mathrm{z}_i^{(k)}=f_{\mathrm{latent}}\left(Emb_i^{\mathrm{G}},k\right)\mid i=1,2,\cdots,N;k=1,\cdots,K\right\} \tag{13}$$

其中，$\mathrm{z}_i^{(k)}\in\mathbb{R}^D$ 表示第 i 个智能体在第 k 个预测模态下的未来潜在表示。$\mathbf{Z}\in\mathbb{R}^{K\times N\times D}$，第一维对应模态维度，第二维对应智能体维度。

为对未来潜在状态集合进行统一建模，将所有未来潜在状态展平为一个 token 集合

$$\mathbf{T}=\left\{\mathrm{t}_m\right\}_{m=1}^{KN},\mathrm{t}_m\in\mathbb{R}^D \tag{14}$$

其中每一个 token 对应一个二元索引(i, k)即某一智能体的某一预测模态。

在 FLCA 模块中，所有 token 之间通过多头注意力机制进行交互建模，如图 2 所示。

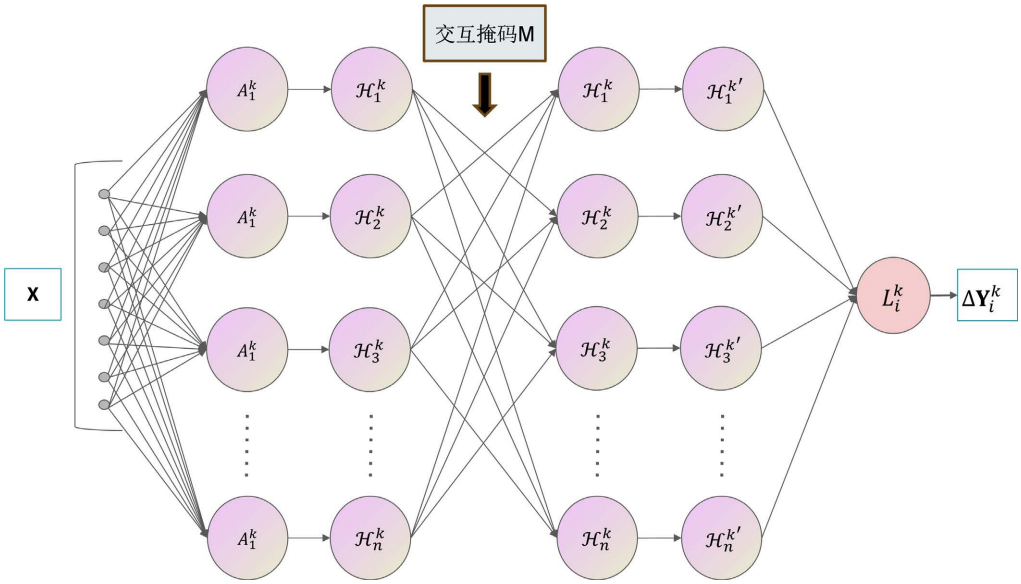

Figure 2. Schematic diagram of FLCA interaction mechanism
图 2. FLCA 交互机制示意图

具体而言，对第 h 个注意力头有

$$Q^{(h)}=\mathbf{T}W_Q^{(h)},K^{(h)}=\mathbf{T}W_K^{(h)},V^{(h)}=\mathbf{T}W_V^{(h)} \tag{15}$$

其中 $W_Q^{(h)}$，$W_K^{(h)}$，$W_V^{(h)}\in\mathbb{R}^{D\times d_h}$，$d_h=D/H$，$H$ 表示注意力头数。

第 h 个注意力头的输出定义为

$$\mathrm{Attn}^{(h)}(\mathbf{T})=\mathrm{softmax}\left(\frac{Q^{(h)}\left(K^{(h)}\right)^{\top}}{\sqrt{d_h}}+M\right)V^{(h)} \tag{16}$$

其中，M 表示跨智能体掩码矩阵。

最终，FLCA 模块的输出通过多头拼接得到，表示为

$$L_i^k=\mathrm{FLCA}(\mathbf{T})=\bigoplus_{h=1}^{H}\mathrm{Attn}^{(h)}(\mathbf{T}) \tag{17}$$

其中，$\oplus$ 表示在特征维度上的拼接操作。

需要指出的是，与传统自注意力不同，FLCA 的注意力计算发生在未来潜在状态空间中，其建模对象不再是历史轨迹或时序特征，而是对未来运动之间关系的显式建模。

对于跨智能体掩码矩阵 M，定义为

$$M_{(i,k),(j,k')}=\begin{cases}-\infty, & i=j\\ 0, & i\neq j\end{cases} \tag{18}$$

其中，$(i,k),(j,k')$ 分别表示 token 对应的智能体–模态索引。

如此定义该掩码矩阵，使得当两个 token 来自于同一智能体，即 $i=j$ 时，其间的注意力连接关系被完全屏蔽。而当两个 token 来自不同智能体，即 $i\neq j$ 时，其注意力连接不受限制。

该掩码机制隐式引入了同一智能体的不同模态保持条件独立的约束。不同模态本质上对应同一智能

体未来行为，但不同模态之间本就是互斥的，若允许其在潜在空间中相互注意，会导致模态间的信息泄漏与塌缩。因此，本机制强制每一个未来潜在状态仅能关注其他智能体的所有模态，从而显式建模跨智能体、跨模态的潜在协同与制约关系。

在 FLCA 模块中，跨智能体的未来潜在交互已被显式建模，其输出特征编码了不同预测之间的空间冲突与协同关系。采用残差优化策略，即令 FLCA 模块输出未来轨迹的偏移量，且对其施加幅值约束，表示为

$$\Delta \mathbf{Y}_i^f = \tanh\left(\mathrm{MLP}_\Delta\left(L_i^k\right)\right)\cdot \lambda \tag{19}$$

其中，$\Delta \mathbf{Y}_i^f \in \mathbb{R}^{F\times 2}$，$\lambda$ 为缩放系数，用于控制最大偏移幅度。

因此，最终优化后的轨迹表示为

$$\mathbf{Y}_i^k = \hat{\mathbf{Y}}_i^k + \Delta \mathbf{Y}_i^k \tag{20}$$

初始轨迹主要由历史运动模式决定，而偏移量则显式刻画未来阶段的交互影响。二者在功能上实现了解耦，使模型能够保持对历史运动规律的稳定建模，同时将 FLCA 的表达能力集中用于未来交互引起的局部调整。

4. 实验

本节在大规模自动驾驶基准数据集 Argoverse1 [25]上进行了实验，以综合评估本方法在不同交通场景上对未来轨迹的预测能力。第 4.1 节介绍了实施细节和评估指标。在第 4.2 节中，在 Argoverse1 数据集上将本研究的方法与基线方法进行了比较。在第 4.3 节中，通过对 FLCA 模块进行消融研究，旨在验证其设计对网络性能的必要性与贡献。结果证明提出的方法可以(i) 建模未来潜在空间内的轨迹交互；(ii) 提升多模态轨迹预测的精度。

4.1. 实验设置

1) 数据集：实验基于大规模 Argoverse1 运动预测数据集开展，该数据集提供智能体轨迹与高精地图数据，共包含 323,557 个真实驾驶场景，划分为训练集(205,942 样本)、验证集(39,472 样本)及测试集(78,143 样本)。训练与验证场景均为 10 Hz 采样的 5 秒序列，测试集仅公开前 2 秒轨迹。本方法利用 Argoverse1 运动预测数据集提供的历史 2 秒观测，预测智能体未来 3 秒的运动轨迹。

2) 评估指标：采用运动预测领域标准评估指标，包括最小平均位移误差(minADE)、最小最终位移误差(minFDE)与漏检率(MR)。评估允许模型为每个智能体生成至多 6 条轨迹。minADE 计算最佳预测轨迹与真实轨迹之间所有未来时间步的 L2 距离平均值，minFDE 仅计算最终时间步的误差，其中最佳预测轨迹指终点误差最小的轨迹。MR 定义为真实轨迹终点与最佳预测轨迹终点距离超过 2.0 米的场景比例。

3) 实施细节：模型在 NVIDIA RTX3090 GPU 上训练 64 周期，使用 AdamW 优化器，批次大小、初始学习率、权重衰减与丢弃率分别设置为 32、3×10^{-4}、1×10^{-4} 与 0.1。遵循基线设定，预测模态数取值为 6。

4.2. 实验结果

在 Argoverse1-val 数据集上的定量实验结果如表 1 所示。所提出的 FLCA 方法在关键评估指标上均取得最优性能。具体而言，FLCA 在 minADE 与 minFDE 两项指标上显著优于现有方法，分别达到 0.6866 米与 1.0214 米，相较于表现最佳的基线方法 HiVT-128，分别降低了约 11.2%与 12.6%。在反映预测可靠性的漏检率(MR)指标上，FLCA 取得了 0.1020 的成绩，仅略差于表现最好的 HOME 方法(0.0846)，但仍

明显优于其他多数对比模型。此外，如图 3 所示的定性结果，绿色实线轨迹代表预测目标的未来轨迹真值，红色虚线代表未来轨迹预测值，取多模态轨迹中的概率最大值进行可视化分析对比。结果表明 FLCA 方法可以在复杂交通场景中同时对多个智能体做出准确的，多模态的预测。

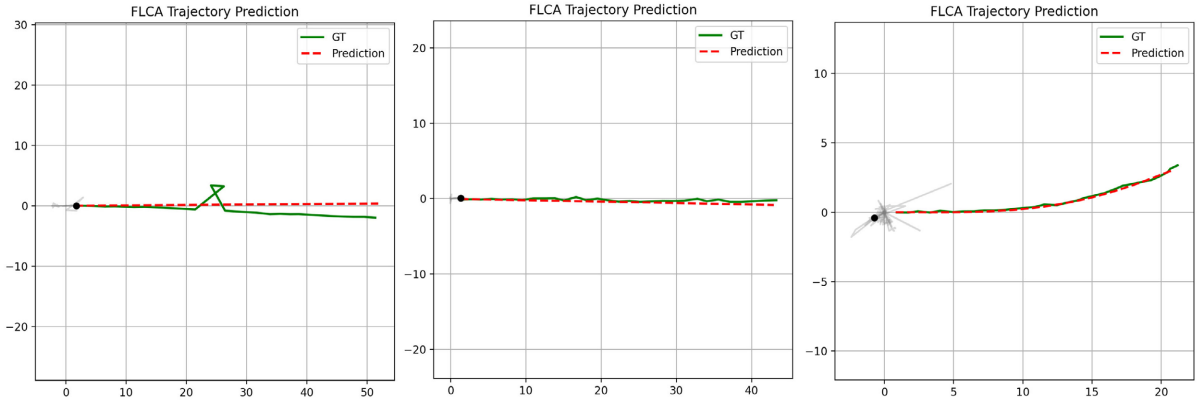

Figure 3. Prediction effect
图 3. 预测效果

Table 1. The performance of the proposed FLCA method was systematically compared with that of the baseline model on the Argoverse1 trajectory prediction dataset
表 1. 在 Argoverse1 轨迹预测数据集上系统对比了所提 FLCA 方法与基线模型的性能

方法	minADE	minFDE	MR
LaneGCN	0.8679	1.3640	0.1634
Scene Transformer	0.8026	1.2321	0.1255
DenseTNT	0.8817	1.2815	0.1258
mmTransformer	0.8436	1.3383	0.1540
HOME	0.8904	1.2919	**0.0846**
HiVT-128	0.7735	1.1693	0.1267
FLCA	**0.6866**	**1.0214**	0.1021

综合分析表明，FLCA 在轨迹预测的精度上实现了有效的提升，其性能提升主要得益于模型中引入了对未来潜在状态交互的建模分析，从而生成更贴近真实轨迹的预测结果。

4.3. 消融实验

为评估 FLCA 中不同的交互建模方法对网络的影响，本文在 Argoverse1 数据集上设计并进行了消融实验，通过交替移除其中的一个交互层来展示每个模块对预测性能的影响。此外，对于是否引入轨迹残差优化机制(Δloc)也进行了对比。消融实验结果如表 2 所示。

消融实验表明，完整模型(0.69 minADE, 1.02 minFDE, 0.10 MR)在所有配置中性能最佳。移除时序交互模块导致性能显著下降，表明其对建模运动连续性至关重要；而去掉空间交互模块，仅对时序建模则使 minFDE 升至 1.47，说明场景内车－车交互与车－路交互对轨迹精度影响重大。引入对未来交互的建模对于仅基于历史时空交互建模具有更优的性能，而轨迹残差优化机制 Δloc 相较于完整模型则使局部轨迹平滑性与贴合度下降。实验表明，各模块共同支撑了模型在复杂交通场景下的准确、可靠的轨迹预测。

Table 2. Ablation experiment on the Argoverse1 dataset
表 2. 在 Argoverse1 数据集上的消融实验

Temporal	Space	Future	Δloc	minADE	minFDE	MR
√				0.92	1.47	0.17
	√			1.00	1.56	0.21
√	√			0.77	1.17	0.13
√	√	√		0.73	1.12	0.12
√	√	√	√	**0.69**	**1.02**	**0.10**

5. 结论

本文针对多智能体多模态轨迹预测任务中交互建模主要局限于历史阶段、未来预测结果之间缺乏显式协调的问题，提出了一种基于未来潜在状态交叉注意力(FLCA)的多模态轨迹预测方法。该方法将轨迹预测明确划分为基于历史的粗粒度预测与基于未来潜在状态交互的预测优化两个阶段。在分层式历史编码器的基础上，引入未来潜在状态空间的交互建模机制，在预测阶段显式刻画跨智能体的空间交互关系，并采用轨迹残差优化策略，从而提升多模态预测结果的准确性。在 Argoverse1 数据集上的实验结果表明，所提出的方法在多项主流评估指标上均优于基线模型，验证了未来潜在状态交互建模在提升多模态轨迹预测性能方面的有效性。未来的研究将进一步探索将在未来潜在状态交互建模过程中引入物理约束或行为先验，以进一步提升预测结果的可解释性与准确性。另一方面，将研究 FLCA 模块在更大规模场景及端到端自动驾驶系统中的应用潜力，并考虑模型结构的轻量化与推理效率优化，以满足实际自动驾驶系统对实时性的需求。

参考文献

[1] Zhou, Z., Ye, L., Wang, J., Wu, K. and Lu, K. (2022) HiVT: Hierarchical Vector Transformer for Multi-Agent Motion Prediction. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 8813-8823. https://doi.org/10.1109/cvpr52688.2022.00862

[2] Zhou, Z., Wang, J., Li, Y. and Huang, Y. (2023) Query-Centric Trajectory Prediction. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 17863-17873. https://doi.org/10.1109/cvpr52729.2023.01713

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems* (*NIPS*), Long Beach, 4 December 2017, 30.

[4] Bruna, J., Zaremba, W., Szlam, A. and LeCun, Y. (2014) Spectral Networks and Locally Connected Networks on Graphs. *Proceedings of the International Conference on Learning Representations* (*ICLR*), Banf, 14-16 April 2014. arXiv:1312.6203, 2013

[5] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. 2017 *Proceedings of the International Conference on Learning Representations* (*ICLR*), Toulon, 24-26 April 2017. arXiv:1609.02907, 2016

[6] Hu, J. and Zheng, W. (2020) Multistage Attention Network for Multivariate Time Series Prediction. *Neurocomputing*, **383**, 122-137. https://doi.org/10.1016/j.neucom.2019.11.060

[7] 杨超. 自动驾驶汽车行为预测综述[J]. 汽车文摘, 2022(10): 11-18.

[8] Toledo-Moreo, R. and Zamora-Izquierdo, M.A. (2009) Imm-Based Lane-Change Prediction in Highways with Low-Cost GPS/INS. *IEEE Transactions on Intelligent Transportation Systems*, **10**, 180-185. https://doi.org/10.1109/tits.2008.2011691

[9] Xie, G., Gao, H., Qian, L., Huang, B., Li, K. and Wang, J. (2018) Vehicle Trajectory Prediction by Integrating Physics- and Maneuver-Based Approaches Using Interactive Multiple Models. *IEEE Transactions on Industrial Electronics*, **65**, 5999-6008. https://doi.org/10.1109/tie.2017.2782236

[10] Firl, J., Stubing, H., Huss, S.A. and Stiller, C. (2012) Predictive Maneuver Evaluation for Enhancement of Car-to-X Mobility Data. 2012 *IEEE Intelligent Vehicles Symposium*, Madrid, 3-7 June 2012, 558-564. https://doi.org/10.1109/ivs.2012.6232217

[11] Laugier, C., Paromtchik, I.E., Perrollaz, M., Yong, M.Y., Yoder, J., Tay, C., *et al*. (2011) Probabilistic Analysis of Dynamic Scenes and Collision Risks Assessment to Improve Driving Safety. *IEEE Intelligent Transportation Systems Magazine*, **3**, 4-19. https://doi.org/10.1109/mits.2011.942779

[12] Aoude, G.S., Luders, B.D., Lee, K.K.H., Levine, D.S. and How, J.P. (2010) Threat Assessment Design for Driver Assistance System at Intersections. 13*th International IEEE Conference on Intelligent Transportation Systems*, Funchal, 19-22 September 2010, 1855-1862. https://doi.org/10.1109/itsc.2010.5625287

[13] Hulnhagen, T., Dengler, I., Tamke, A., Dang, T. and Breuel, G. (2010) Maneuver Recognition Using Probabilistic Finite-State Machines and Fuzzy Logic. 2010 *IEEE Intelligent Vehicles Symposium*, La Jolla, 21-24 June 2010, 65-70. https://doi.org/10.1109/ivs.2010.5548066

[14] 郭景华, 何智飞, 罗禹贡, 等. 人机混驾环境下基于深度学习的车辆切入轨迹预测[J]. 汽车工程, 2022, 44(2): 153-160.

[15] Chandra, R., Guan, T., Panuganti, S., Mittal, T., Bhattacharya, U., Bera, A., *et al*. (2020) Forecasting Trajectory and Behavior of Road-Agents Using Spectral Clustering in Graph-LSTMS. *IEEE Robotics and Automation Letters*, **5**, 4882-4890. https://doi.org/10.1109/lra.2020.3004794

[16] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., *et al*. (2016) Social LSTM: Human Trajectory Prediction in Crowded Spaces. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 961-971. https://doi.org/10.1109/cvpr.2016.110

[17] Messaoud, K., Yahiaoui, I., Verroust-Blondet, A. and Nashashibi, F. (2021) Attention Based Vehicle Trajectory Prediction. *IEEE Transactions on Intelligent Vehicles*, **6**, 175-185. https://doi.org/10.1109/tiv.2020.2991952

[18] Huang, Z.Y., Mo, X.Y. and Lv, C. (2021) Multi-Modal Motion Prediction with Transformer-Based Neural Network for Autonomous Driving.

[19] Gao, J.Y., Sun, C., Zhao, H., Shen, Y., *et al*. (2020) VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-18 June 2020, 11525-11533.

[20] Gu, J.R., Sun, C. and Zhao, H. (2021) Densetnt: End-to-End Trajectory Prediction from Dense Goal Sets. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 15303-15312.

[21] Liu, Y.C., Zhang, J.H., Fang, L.J., *et al*. (2021) Multimodal Motion Prediction with Stacked Transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 7577-7586.

[22] Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., *et al*. (2020) Learning Lane Graph Representations for Motion Forecasting. In *European Conference on Computer Vision*, Springer International Publishing, 541-556. https://doi.org/10.1007/978-3-030-58536-5_32

[23] Giuliari, F., Hasan, I., Cristani, M. and Galasso, F. (2020) Transformer Networks for Trajectory Forecasting. 25*th International Conference on Pattern Recognition* (*ICPR*), Milan, 10-15 January 2020, 10335-10342.

[24] Casas, S., Gulino, C., Liao, R. and Urtasun, R. (2020) SpAGNN: Spatially-Aware Graph Neural Networks for Relational Behavior Forecasting from Sensor Data. 2020 *IEEE International Conference on Robotics and Automation* (*ICRA*), Paris, 31 May-31 August 2020, 9491-9497. https://doi.org/10.1109/icra40945.2020.9196697

[25] Chang, M., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., *et al*. (2019) Argoverse: 3D Tracking and Forecasting with Rich Maps. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 8748-8757. https://doi.org/10.1109/cvpr.2019.00895