ViTE: Virtual Graph Trajectory Expert Router for Pedestrian Trajectory Prediction

Ruochen Li[1], **Zhanxing Zhu**[2], **Tanqiu Qiao**[1], **Hubert P. H. Shum**[1*]

[1]Department of Computer Science, Durham University, UK
[2]School of Electrical and Computer Science (ECS), University of Southampton, UK
{ruochen.li, hubert.shum}@durham.ac.uk

Abstract

Pedestrian trajectory prediction is critical for ensuring safety in autonomous driving, surveillance systems, and urban planning applications. While early approaches primarily focus on one-hop pairwise relationships, recent studies attempt to capture high-order interactions by stacking multiple Graph Neural Network (GNN) layers. However, these approaches face a fundamental trade-off: insufficient layers may lead to under-reaching problems that limit the model's receptive field, while excessive depth can result in prohibitive computational costs. We argue that an effective model should be capable of adaptively modeling both explicit one-hop interactions and implicit high-order dependencies, rather than relying solely on architectural depth. To this end, we propose ViTE (Virtual graph Trajectory Expert router), a novel framework for pedestrian trajectory prediction. ViTE consists of two key modules: a Virtual Graph that introduces dynamic virtual nodes to model long-range and high-order interactions without deep GNN stacks, and an Expert Router that adaptively selects interaction experts based on social context using a Mixture-of-Experts design. This combination enables flexible and scalable reasoning across varying interaction patterns. Experiments on three benchmarks (ETH/UCY, NBA, and SDD) demonstrate that our method consistently achieves state-of-the-art performance, validating both its effectiveness and practical efficiency.

Code — https://github.com/Carrotsniper/ViTE

Introduction

Human trajectory prediction aims to forecast future pedestrian paths based on observed motion histories. It plays a vital role in autonomous driving for tasks such as collision avoidance and emergency braking (Bai et al. 2015; Luo et al. 2018; Liu et al. 2021), and is also essential in video surveillance for identifying suspicious activities (Luber et al. 2010; Shi et al. 2021). This task is challenging due to the uncertainty of human behavior and the varying relevance of surrounding individuals: nearby agents may not interact, while distant ones can still be coordinated. This complexity requires models that can flexibly capture interactions across multiple scales.

Early approaches primarily focus on pairwise interactions, capturing local spatial dependencies between individuals.

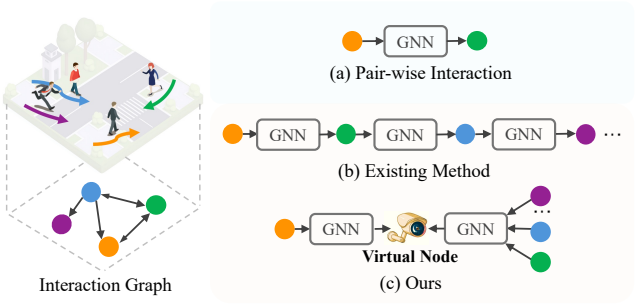

Figure 1: Comparison of interaction modeling strategies. **(a)** Traditional methods capture only one-hop interactions. **(b)** Existing methods stack multiple GNN layers to model high-order dependencies. **(c)** Our method introduces virtual nodes to capture high-order interactions efficiently.

Representative methods like Social-LSTM (Alahi et al. 2016; Gupta et al. 2018) utilize social pooling to aggregate information from neighboring agents within a fixed window, while SS-LSTM (Xue, Huynh, and Reynolds 2018) introduces occupancy maps to encode nearby spatial configurations. With the advancement of relational modeling, graph-based methods (Huang et al. 2019; Kosaraju et al. 2019; Mohamed et al. 2020; Shi et al. 2021) have become prominent, representing pedestrians as nodes and their interactions as edges in a graph, enabling data-driven learning of pairwise relationships via Graph Neural Networks (GNNs) as shown in Figure 1 (a). To capture richer interaction dynamics, more recent efforts have focused on modeling high-order interactions—influences that are not directly observable between agent pairs but emerge through multi-hop connections within a crowd. Methods such as HighGraph (Kim et al. 2024) and PCHGCN (Chen, Sang, and Zhao 2025) address this by stacking multiple GNN layers (Figure 1 (b)), allowing information to propagate across longer distances and indirectly connected agents. These models have shown improvements in capturing group behavior patterns and non-local dependencies.

Despite their effectiveness, high-order interaction modeling via deep GNNs introduces two key challenges. The first is depth-related inefficiency. Stacking too few GNN layers results in under-reaching, where the receptive field is insufficient to cover relevant agents beyond local neigh-

*Corresponding author

borhoods, leading to incomplete interaction modeling (Lu et al. 2024). Conversely, increasing depth can cause over-smoothing (Rusch, Bronstein, and Mishra 2023), where repeated message passing dilutes node-specific features, making individual agent representations less discriminative. This trade-off poses a fundamental bottleneck for graph-based trajectory models that rely solely on depth to model relational complexity. Another major limitation is the lack of contextual adaptability. In real-world crowds, the influence of other pedestrians varies across scenes and situations: in some cases, immediate neighbors dominate decision-making, while in others, long-range or indirect interactions play a critical role. However, most existing methods apply a fixed, shared relational structure or aggregation scheme to all agents, ignoring this diversity. As a result, they fail to dynamically adjust the reasoning process according to individual agents' social context and interaction scale.

To address these limitations, we propose ViTE (Virtual graph Trajectory Expert router), a novel framework for pedestrian trajectory prediction. It comprises two key components. First, the **Virtual Graph**, introduces dynamically assigned virtual nodes that act as relational mediators between pedestrians (Figure 1 (c)). These nodes serve as intermediate hubs for message exchange, enabling the model to capture long-range dependencies and high-order interactions without deep GNN stacks. This compact and learnable structure not only mitigates under-reaching but also improves efficiency and expressiveness in modeling long-range relations. Second, the **Expert Router** adopts a Mixture-of-Experts (MoE) design to enable context-aware interaction reasoning. We introduce multiple interaction experts, each specialized in a particular interaction scale (e.g., one-hop or high-order), are coordinated by a gating network that dynamically routes each agent's representation to the most relevant experts. This adaptive mechanism allows ViTE to flexibly integrate diverse relational patterns based on social context. Together, these two modules form a cohesive and scalable system that overcomes depth limitations and enables fine-grained social reasoning for trajectory prediction. Our main contributions are:

- We propose **ViTE**, a novel framework for pedestrian trajectory prediction that enables efficient and adaptive high-order interaction modeling.
- We design a **Virtual Graph** module that introduces dynamic virtual nodes as relational hubs to capture indirect social dependencies.
- We develop an **Expert Router** based on a Mixture-of-Experts mechanism to perform context-aware reasoning across multiple interaction scales.

Related Work

Interaction Modeling for Trajectory Prediction

Early work on pedestrian trajectory prediction focused on pairwise interactions to capture the influence of nearby agents. Social-LSTM (Alahi et al. 2016; Gupta et al. 2018) introduced a pooling mechanism to aggregate neighboring hidden states, while SS-LSTM (Xue, Huynh, and Reynolds 2018) encoded spatial layouts via occupancy maps to enhance local interaction modeling. However, these methods struggle with complex multi-agent dynamics due to their limited relational structure. To address this, graph-based approaches (Bae and Jeon 2021; Qiao et al. 2022; Bae and Jeon 2023; Qiao et al. 2024) have been proposed, representing pedestrians as nodes and their interactions as edges in a graph. ST-GAT (Huang et al. 2019) employs graph attention (Qiao et al. 2025; Shao et al. 2025) to adaptively weigh neighboring influences, Social-STGCNN (Mohamed et al. 2020) combines spatial and temporal convolutions, Social-BiGAT (Kosaraju et al. 2019) models bidirectional influence flows, and SGCN (Shi et al. 2021) applies sparse GCNs to capture spatial-temporal dependencies. These methods advance beyond pairwise designs by leveraging GNNs to jointly model spatial and temporal dependencies. To capture high-order interactions, which refer to indirect influences mediated through intermediate agents, recent works such as HighGraph (Kim et al. 2024) and PCHGCN (Chen, Sang, and Zhao 2025) stack multiple GNN layers to enable multi-hop message passing. While effective for modeling non-local dependencies, this strategy faces a trade-off: shallow networks suffer from under-reaching (Lu et al. 2024; Li et al. 2025b), while deeper ones risk over-smoothing, degrading representation quality (Rusch, Bronstein, and Mishra 2023). To this end, we introduce adaptive virtual nodes that serve as global relational hubs, effectively capturing high-order interactions by mediating indirect dependencies without relying on deep GNNs.

Mixture of Expert

Mixture of Experts (MoE) is a modular neural architecture that partitions the input space and routes each input to a subset of specialized experts, selected dynamically by a gating network (Yuksel, Wilson, and Gader 2012; Jiang et al. 2024; Mu and Lin 2025). Unlike traditional ensembles, MoE activates only a few experts per input, enabling high model capacity with reduced computational cost. This design has shown success in NLP (Jacobs et al. 1991; Shazeer et al. 2017), vision (Wang et al. 2020; Riquelme et al. 2021), and multi-modal learning (Mustafa et al. 2022). In graph learning, MoE has been applied to aggregate across neighborhoods (Abu-El-Haija et al. 2020), correct bias (Hu et al. 2022), and enhance molecular prediction (Kim et al. 2023). Recent works explore top-k input routing (Zhou et al. 2022) and multi-hop fusion (Wang et al. 2023), but often rely on fixed routing or task-specific designs, limiting adaptability. In this work, we propose a MoE-based expert router for trajectory prediction, which enables context-aware routing over one-hop and high-order graph interactions. By dynamically selecting the most relevant expert for each node, our model effectively adapts to diverse interactions in pedestrian crowds.

Methodology

Problem Formulation

Pedestrian trajectory prediction forecasts future positions from past movements. Mathematically, given a scene with N pedestrians observed over T_{obs} time steps, the trajectory of pedestrian i is denoted as $X_i = (x_t^i, y_t^i) \mid t \in [-T_{obs}+1, \ldots, 0]$ for the observed past, and $Y_i = (x_t^i, y_t^i) \mid t \in [1, \ldots, T_{pred}]$ for the future ground-truth.

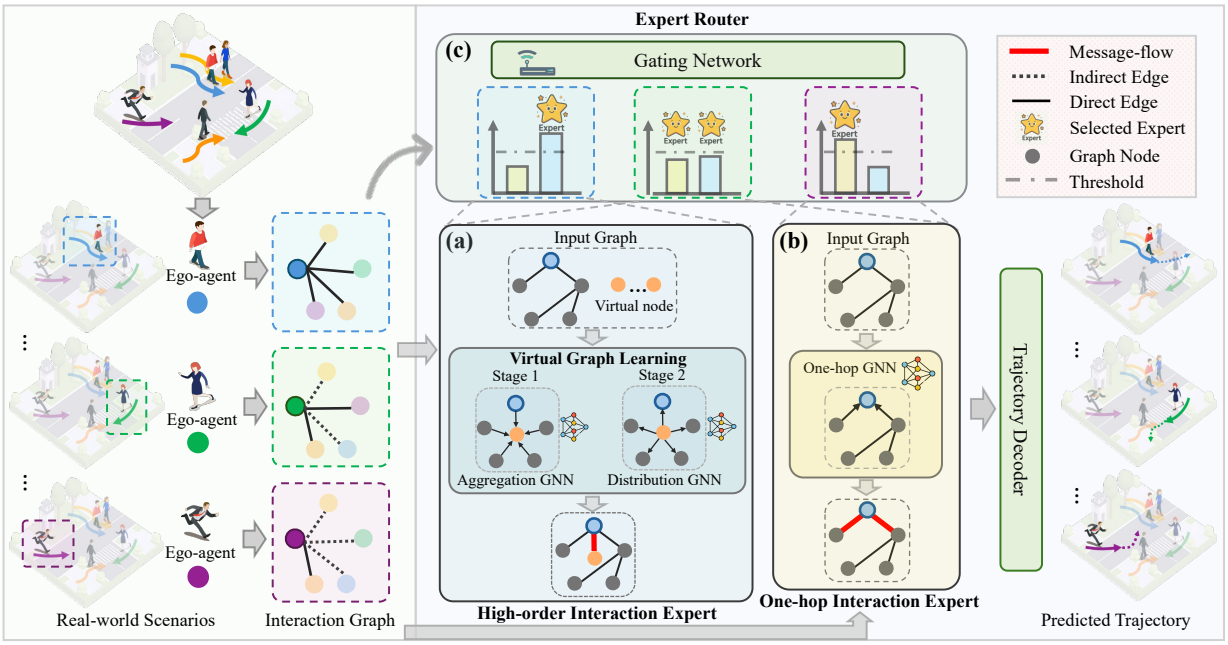

Figure 2: Overview of ViTE. Given pedestrian trajectories, we first construct interaction graphs. In (a), the high-order interaction expert captures indirect, long-range dependencies via Virtual Graph Learning, while (b) illustrates the one-hop expert modeling direct interactions. These expert outputs are then dynamically fused by a MoE-based Expert Router, as depicted in (c), enabling context-aware routing of graph information. Finally, an MLP-based decoder outputs future trajectories for each pedestrian.

Stacking all pedestrians yields the tensors $\mathbf{X} \in \mathbb{R}^{N \times T_{obs} \times 2}$ and $\mathbf{Y} \in \mathbb{R}^{N \times T_{pred} \times 2}$, representing the observed and future trajectories respectively, with each position in 2D coordinates. The goal is to minimize the error between the predicted trajectories $\hat{\mathbf{Y}}$ and ground-truth future trajectories $\mathbf{Y}$.

Feature Initialization

For each pedestrian i, we construct the input representation from two complementary feature types: absolute spatial coordinates $\mathbf{p}_i = (x, y)$ and temporal displacement vectors $\mathbf{r}_i = (r_x, r_y)$, where $\mathbf{r}_i^{(t)} = \mathbf{p}_i^{(t)} - \mathbf{p}_i^{(t-1)}$ captures the motion dynamics. The final input feature is obtained through concatenation: $\mathbf{X}_i^{\text{in}} = [\mathbf{p}_i; \mathbf{r}_i] \in \mathbb{R}^{T_{obs} \times 4}$. We represent pedestrian interactions using a spatial interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $n_i \in \mathcal{V}$ corresponds to pedestrian i, and each edge $e_{ij} \in \mathcal{E}$ models one-hop relationships. Node features are initialized as $\mathbf{n}_i^{(0)} = \mathcal{F}_{node}(\mathbf{X}_i^{\text{in}})$, where $\mathcal{F}_{node}(\cdot)$ denotes a Multi-Layer Perceptron (MLP). To capture social dependencies, we dynamically determine the graph connectivity via a k-nearest neighbor strategy based on feature similarity. For each connected pair (i, j), we compute the initial embedding feature using a relational transformer (RT) layer (Diao and Loynd 2023; Lee et al. 2024) equipped with sparse attention to ensure computational efficiency, denoted as $\mathbf{n}_{i,pair}^0 = RT(\mathbf{n}_i^{(0)}, M)$, where M is connection mask.

High-order Interaction Modeling via Virtual Graph

Capturing long-range dependencies is essential in trajectory prediction, as agents often influence one another beyond their immediate neighbors through high-order interactions. However, traditional GNNs based on first-order aggregation inherently suffer from *under-reaching*—the inability to propagate information across distant nodes within shallow architectures (Qian et al. 2024). Although stacking GNN layers helps capture high-order interactions (Kim et al. 2024; Chen, Sang, and Zhao 2025), it significantly increases computational cost.

To quantify this limitation, we calculate the *effective resistance* as a structural indicator of communication efficiency in graphs. Formally, the effective resistance between two nodes i and j is defined as:

$$R_{ij} = (e_i - e_j)^T \mathcal{L}^+ (e_i - e_j), \tag{1}$$

where $\mathcal{L}^+$ is the Moore–Penrose pseudoinverse of the graph Laplacian, and e_i, e_j are standard basis vectors. Lower R_{ij} indicates more efficient message propagation (Lu et al. 2024). For example, in the chain-structured graph shown in Figure 3 (left), node a requires four hops to reach node e, resulting in a high resistance of $R_{ae} = 4.0$, which hinders the efficiency of long-range communication.

To address this issue, we propose the **Virtual Graph**, which introduces a small set of virtual nodes as mediators to facilitate high-order long-range communication. As shown in Figure 3 (right), adding a virtual node dramatically reduces the resistance between distant nodes (e.g., $R_{ae} = 1.2$), enabling more efficient global information flow.

Specifically, we construct a high-order interaction graph $\mathcal{G}_{\text{high}} = (\mathcal{N} \cup \mathcal{V}_{\text{virtual}}, \mathcal{E}_{\text{high}})$, where $\mathcal{N}$ denotes the set of real pedestrian nodes, and $\mathcal{V}_{\text{virtual}} = \{v_1, v_2, \ldots, v_V\}$ is a set of virtual nodes acting as communication hubs. These virtual

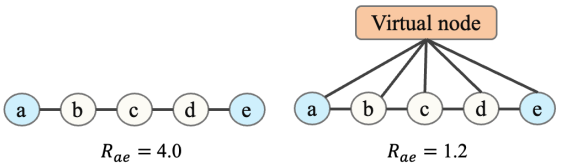

Figure 3: Comparison of effective resistance (R_{ae}) between a standard chain graph (left, $R_{ae} = 4.0$) and a virtual-node-enhanced graph structure (right, $R_{ae} = 1.2$). Lower effective resistance indicates more efficient message propagation and improved global connectivity.

nodes are instantiated per training batch and initialized with diverse embeddings sampled from a learnable distribution:

$$\mathbf{v}_k^{(0)} \sim \mathcal{D}_{\text{vn}}, \quad k = 1, \ldots, V, \tag{2}$$

where $\mathcal{D}_{\text{vn}}$ is designed to promote representational diversity. To explicitly capture high-order dependencies, we propose a structured two-stage message passing scheme.

Stage 1 (Real-to-Virtual Message Aggregation). Each real node $n_i \in \mathcal{N}$ sends its node embedding to all virtual nodes, enabling them to aggregate diverse contexts into high-level representations. The embedding of each virtual node v_k is updated as:

$$\mathbf{v}_k^{(1)} = \text{GNN}_{\text{r}\to\text{v}}\left(\mathbf{v}_k^{(0)}, \text{AGG}_{\text{r}\to\text{v}}\left(\{\mathbf{n}_i^{(0)} \mid n_i \in \mathcal{N}\}\right)\right), \tag{3}$$

where $\text{AGG}_{\text{r}\to\text{v}}$ is a permutation-invariant aggregator (e.g., mean or attention), and $\text{GNN}_{\text{r}\to\text{v}}$ is a standard GNN module such as a Graph Attention Network (Veličković et al. 2018).

Stage 2 (Virtual-to-Real Message Distribution). Each updated virtual node v_k broadcasts its high-level contexts to all pedestrian nodes, enabling them to integrate indirect high-order interactions into their representations. The embedding of each pedestrian node n_i is updated as:

$$\mathbf{n}_i^{(1)} = \text{GNN}_{\text{v}\to\text{r}}\left(\mathbf{n}_i^{(0)}, \text{AGG}_{\text{v}\to\text{r}}\left(\{\mathbf{v}_k^{(1)} \mid v_k \in \mathcal{V}\}\right)\right), \tag{4}$$

where $\text{AGG}_{\text{v}\to\text{r}}$ is a permutation-invariant aggregator, and $\text{GNN}_{\text{v}\to\text{r}}$ is a standard GNN module for fusing virtual node information. The final output is then normalized and activated as: $\mathbf{n}_{i,\text{high}}^{(\text{out})} = \text{GELU}\left(\text{LayerNorm}\left(\mathbf{n}_i^{(1)}\right)\right)$.

In summary, our virtual graph interaction module integrates both localized and global context by combining one-hop and high-order message passing mechanisms. The one-hop graph captures fine-grained local interactions among neighboring agents, while the virtual-node-enhanced graph enables efficient modeling of high-order dependencies. Together, these complementary designs equip the model with a more expressive and scalable interaction representation.

One-hop Interaction Modeling

To capture one-hop interactions among pedestrians, we define a first-order directed interaction graph $\mathcal{G}_{\text{one-hop}} = (\mathcal{N}, \mathcal{E}_{\text{one-hop}})$, where edges $\mathcal{E}_{\text{one-hop}}$ encode spatial relationships between neighboring agents based on their similarities (Lee et al. 2024). This graph supports message passing operations in GNNs, enabling the modeling of one-hop dependencies among directly connected nodes (Li, Katsigiannis, and Shum 2022; Li et al. 2025a).

The expert leverages both node and edge features to update each node embedding. Specifically, each node n_i aggregates information from its neighbors as follows:

$$\mathbf{m}_{ij} = \phi_m\left(\mathbf{n}_i^{(0)}, \mathbf{n}_j^{(0)}, \mathbf{e}_{ij}^{(0)}\right), \tag{5}$$

$$\mathbf{m}_i = \text{AGG}_{\text{one-hop}}\left(\{\mathbf{m}_{ij} \mid j \in \mathcal{N}_{\text{one-hop}}(i)\}\right), \tag{6}$$

$$\mathbf{n}_{i,\text{one-hop}}^{(\text{out})} = \phi_u\left(\mathbf{n}_i^{(0)}, \mathbf{m}_i\right), \tag{7}$$

where $\phi_m(\cdot)$ denotes the message function that integrates source node, target node, and edge features; $\text{AGG}_{\text{one-hop}}(\cdot)$ is a permutation-invariant aggregation function; and $\phi_u(\cdot)$ is the update function, typically an MLP layer. This design allows the model to leverage both one-hop structural information and relation features for direct interaction modeling.

Expert Router

While one-hop and high-order graph structures capture complementary interaction patterns, existing methods typically adopt a fixed graph design, limiting adaptability to varying scene complexities. Such rigidity can result in under-reaching in complex scenarios or redundant computation in simpler ones. To address this limitation, we propose an adaptive **Expert Router** based on the MoE paradigm, which learns a context-aware routing policy to dynamically allocate interaction modeling capacity. Specifically, we treat $\mathcal{G}_{\text{one-hop}}$ and $\mathcal{G}_{\text{high}}$ as two specialized experts responsible for modeling direct and high-order interactions, respectively. By allowing each node to selectively attend to the most relevant expert based on its feature representations.

For each node feature $\mathbf{n}_i^{(0)}$, the gating network computes a soft routing distribution (Hu et al. 2025) over the two experts:

$$\mathbf{g}_i = \text{Softmax}(\phi(\mathbf{n}_i^{(0)})), \tag{8}$$

$$\phi(\mathbf{n}_i^{(0)}) = \mathbf{W}_g \mathbf{n}_i^{(0)} + \boldsymbol{\epsilon} \cdot \text{Softplus}(\mathbf{W}_n \mathbf{n}_i^{(0)}), \tag{9}$$

where $\phi(\mathbf{n}_i^{(0)}) \in \mathbb{R}^2$ are the logits over two experts, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ adds Gaussian noise for regularization. $\mathbf{W}_g$ and $\mathbf{W}_n$ are learnable projection matrices.

To enable adaptive complexity control and improve efficiency, we adopt a threshold-based Top-P mechanism that dynamically selects active experts (Huang et al. 2024; Hu et al. 2025). Unlike traditional MoE methods (Wang et al. 2023) that always fuse all experts, our approach activates experts based on confidence, enabling lightweight local processing for simple cases and triggering global reasoning in complex interaction scenarios.

Let $\tilde{\mathbf{g}}_i = \text{Sort}(\mathbf{g}_i, \text{descending})$ be the sorted expert probabilities. The active expert set $\mathcal{S}_i$ is defined as the minimal set whose cumulative probability exceeds a threshold p:

$$\mathcal{S}_i = \left\{ k : \sum_{j=1}^{k} \tilde{g}_{i,j} \leq p \right\} \cup \left\{ \arg\min_k \left\{ \sum_{j=1}^{k} \tilde{g}_{i,j} > p \right\} \right\}, \tag{10}$$

ETH/UCY Dataset											
Subset	GroupNet	MemoNet	MID	NPSN	EqMotion	EigenTraj	LED	SingularTraj	MART	PCHGCN	Ours
ETH	0.46/0.73	0.40/0.61	0.39/0.66	0.36/0.59	0.40/0.61	0.36/0.53	0.39/0.58	**0.35/0.42**	**0.35**/0.47	0.42/0.65	**0.35**/0.49
HOTEL	0.15/0.25	**0.11/0.17**	0.13/0.22	0.16/0.25	0.12/0.18	0.12/0.19	**0.11/0.17**	0.13/0.19	0.14/0.22	0.17/0.28	**0.11/0.17**
UNIV	0.26/0.49	0.24/0.43	0.22/0.45	0.23/0.39	0.23/0.43	0.24/0.43	0.26/0.43	0.25/0.44	0.25/0.45	**0.21/0.38**	0.23/0.42
ZARA1	0.21/0.39	0.18/0.32	**0.17**/0.30	0.18/0.32	0.18/0.32	0.19/0.33	0.18/**0.26**	0.19/0.32	**0.17**/0.29	**0.17**/0.31	0.18/0.30
ZARA2	0.17/0.33	0.14/0.24	**0.13**/0.27	0.14/0.25	**0.13**/0.23	0.14/0.24	**0.13/0.22**	0.15/0.25	**0.13/0.22**	**0.13**/0.23	**0.13/0.22**
AVG	0.25/0.44	0.21/0.35	0.21/0.38	0.21/0.36	0.21/0.35	0.21/0.34	0.21/0.33	0.21/**0.32**	0.21/0.33	0.22/0.37	**0.20/0.32**

Table 1: Performance comparison on the ETH/UCY dataset. Metrics are $\min \mathrm{ADE}_{20}/\min \mathrm{FDE}_{20}$. **Bold** and underline indicate the best and second-best results.

NBA Dataset										
Time	STAR	GroupNet	MemoNet	MID	NPSN	DynGroupNet	LED	SingularTraj	MART	Ours
1.0s	0.43/0.66	0.26/0.34	0.38/0.56	0.28/0.37	0.35/0.58	0.19/0.28	0.18/0.27	0.28/0.44	0.18/0.26	**0.17/0.25**
2.0s	0.75/1.24	0.49/0.70	0.71/1.14	0.51/0.72	0.68/1.23	0.40/0.61	0.37/0.56	0.61/1.00	**0.35/0.50**	**0.35/0.50**
3.0s	1.03/1.51	0.73/1.02	1.00/1.57	0.71/0.98	1.01/1.76	0.65/0.90	0.58/0.84	0.96/1.47	0.54/0.71	**0.53/0.70**
4.0s	1.13/2.01	0.96/1.30	1.25/1.47	0.96/1.27	1.31/1.79	0.89/1.13	0.81/1.10	1.31/1.98	0.73/**0.90**	**0.72**/0.91

Table 2: Performance comparison on the NBA dataset. Metrics are $\min \mathrm{ADE}_{20}/\min \mathrm{FDE}_{20}$. **Bold** and underline indicate the best and second-best results.

where $p \in (0, 1)$ controls the sparsity level (e.g., $p = 0.7$). The selected expert weights are then renormalized as:

$$\hat{g}_{i,k} = \begin{cases} \frac{g_{i,k}}{\sum_{j \in \mathcal{S}_i} g_{i,j}} & \text{if } k \in \mathcal{S}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Given the selected expert set $\mathcal{S}_i$ and renormalized weights $\hat{g}_{i,k}$ from Eq. 11, the final node representation is computed as a weighted sum:

$$\mathbf{n}^{(\text{out})}_{i,router} = \hat{g}_{i,1} \cdot \mathbf{n}^{(\text{out})}_{i,\text{one-hop}} + \hat{g}_{i,2} \cdot \mathbf{n}^{(\text{out})}_{i,\text{high}}. \tag{12}$$

By leveraging soft gating and selective expert activation, the proposed **Expert Router** enables context-aware routing of one-hop and high-order interactions, allowing the model to dynamically adjust its reasoning complexity based on pedestrian behavior patterns.

To prevent routing collapse and promote expert diversity, we introduce an importance-based auxiliary loss (Hu et al. 2025):

$$\mathcal{L}_{\text{imp}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{Std}(\mathbf{g}_i)}{\text{Mean}(\mathbf{g}_i) + \epsilon}, \tag{13}$$

which encourages balanced expert utilization across nodes and improves the robustness of the routing mechanism.

Trajectory Decoding

For trajectory prediction, we concatenate three feature types $[\mathbf{n}^{(0)}_i; \mathbf{n}^{(0)}_{i,pair}; \mathbf{n}^{(\text{out})}_{i,router}]$ and feed them into K parallel MLP-based prediction heads to generate diverse trajectories following (Xu et al. 2023; Lee et al. 2024). Given K predicted trajectories $\hat{\mathbf{Y}}^{(k)}_i$ from K prediction heads for pedestrian i, we compute the prediction loss by selecting the best trajectory in terms of minimum ℓ_2 distance to the ground truth:

$$\mathcal{L}_{\text{pred}} = \frac{1}{NT_{pred}} \sum_{i=1}^{N} \sum_{t=1}^{T_{pred}} \min_{k} ||\mathbf{p}^{(t)}_i - \hat{\mathbf{p}}^{(t,k)}_i||_2, \tag{14}$$

where $\mathbf{p}^{(t)}_i = (x^i_t, y^i_t)$ is the ground-truth position of pedestrian i at time t, and $\hat{\mathbf{p}}^{(t,k)}_i$ is the corresponding prediction from the k-th head. The total loss incorporates MoE regularization: $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{imp}}$.

Experiments

Experimental Setup

Datasets. We evaluate our model on three widely-used trajectory prediction benchmarks: ETH/UCY (Pellegrini et al. 2009; Lerner, Chrysanthou, and Lischinski 2007), Stanford Drone Dataset (SDD) (Robicquet et al. 2016), and NBA SportVU (Mao et al. 2023). The ETH/UCY dataset comprises five subsets (ETH, HOTEL, UNIV, ZARA1, and ZARA2), capturing pedestrian movements across diverse social scenarios. SDD is a large-scale dataset collected from a university campus. For both ETH/UCY and SDD, we follow the standard setting in (Lee et al. 2024; Xu et al. 2022b), using 3.2 seconds (8 frames) of observed trajectories to predict the next 4.8 seconds (12 frames). For ETH/UCY, we adopt a leave-one-out training protocol, training on four subsets and testing on the remaining one. The NBA SportVU dataset provides trajectories of 10 players and the ball in real basketball games. We follow (Xu et al. 2022a; Mao et al. 2023; Lee et al. 2024) that use 2.0 seconds (10 frames) of past motion to predict the next 4.0 seconds (20 frames).

Evaluation Metrics. To evaluate the performance, we adopt two metrics: $\min \mathrm{ADE}_k$ and $\min \mathrm{FDE}_k$, following

SDD Dataset										
Time	PECNet	GroupNet	MemoNet	MID	NPSN	DynGroupNet	EigenTraj	LED	MART	Ours
4.8s	9.96/15.88	9.31/16.11	8.56/12.66	9.73/15.32	8.56/11.85	8.42/13.58	8.05/13.25	8.48/**11.66**	7.43/11.82	**7.42**/11.90

Table 3: Performance comparison on the SDD dataset. Metrics are $\min \mathrm{ADE}_{20}$/$\min \mathrm{FDE}_{20}$. **Bold** and underline indicate the best and second-best results.

the evaluation protocol in (Gupta et al. 2018; Lee et al. 2024). The Average Displacement Error (ADE) measures the mean Euclidean distance between the predicted and ground-truth trajectories over all time steps, while the Final Displacement Error (FDE) focuses on the distance between the predicted and actual final positions. Given k sampled predictions for each agent, we report the minimum ADE and FDE among them, denoted as $\min \mathrm{ADE}_k$ and $\min \mathrm{FDE}_k$.

Comparison with State-of-the-Art Methods

The results on the ETH/UCY dataset are presented in Table 1, our method achieves the best overall performance, with the lowest average ADE/FDE of 0.20/0.32 across all five subsets. Compared to PCHGCN, a high-order graph-based method, our approach reduces ADE by 9.1% and FDE by 13.5%. Moreover, our method consistently ranks among the top performers on each individual subset, demonstrating strong generalization across diverse crowd scenarios. For NBA datasets shown in Table 2, our model ranks first at 1.0s, 2.0s, and 3.0s, and achieves the lowest ADE and second-best FDE at 4.0s, highlighting its robustness in modeling long-term multi-agent interactions in dynamic sports environments. Table 3 presents the results on the SDD dataset. Our method achieves the best ADE and a competitive FDE compared to existing approaches, confirming its effectiveness in forecasting pedestrian trajectories in real-world scenes.

Ablation Study and Analysis

Effect of Expert. We conduct ablation studies to assess the impact of different configurations of Expert Router (Table 4). Removing both experts results in a clear performance drop, confirming the necessity of incorporating expert mechanisms into the model. When comparing single-expert settings, the one-hop expert yields better results than the high-order counterpart, indicating that certain interaction structures are more informative for prediction. Most notably, our adaptive MoE outperforms both individual experts and the fixed mixed configuration, demonstrating that dynamic, input-dependent expert selection offers clear advantages over static alternatives. These findings validate the effectiveness of context-aware expert routing in modeling diverse trajectory patterns.

Effect of Virtual Nodes. To assess the effectiveness of the Virtual Graph, we compare our model with conventional multi-layer GCNs. As shown in Table 6, our approach achieves the best performance with the fewest parameters (1.00×), outperforming both 2-layer and 4-layer GCNs. Notably, simply increasing GCN depth does not yield better results—the 4-layer model even slightly underperforms the 2-layer variant, despite having more parameters. This suggests that stacking layers is neither efficient nor sufficient for capturing high-order interactions. In contrast, our use of virtual nodes offers a more effective and parameter-efficient solution for modeling such interactions.

Expert Configuration	ETH/UCY Dataset	
	$\min \mathrm{ADE}_{20}$	$\min \mathrm{FDE}_{20}$
One-hop Expert Only	0.22	0.34
High-order Expert Only	0.23	0.36
Mixed Expert	0.25	0.36
No Expert	0.25	0.38
Ours	**0.20**	**0.32**

Table 4: Ablation study of expert selection evaluated by $\min \mathrm{ADE}_{20}$/$\min \mathrm{FDE}_{20}$ on ETH/UCY dataset. **Bold** and underline indicate the best and second-best results.

Method	#Param.	MACs
PECNet	2.1M	259.2M
STAR	**1.0**M	12.0G
MemoNet	10.7M	6.0G
GroupNet	2.2M	411.5M
MID	9.0M	40.3G
EqMotion	3.0M	147.1M
LED	10.9M	15.0G
MART	1.5M	43.3M
Ours	**1.0**M	**23.1**M

Table 5: Model complexity comparison. **Bold** and underline indicate the best and second-best results.

Efficiency Comparisons. Table 5 compares model complexity in terms of parameters and Multiply–Accumulate Operations (MACs). Following the evaluation protocol of (Lee et al. 2024), we compute MACs using scenes with 10 agents from the ETH/UCY dataset. Our method achieves the lowest computational complexity (23.1M MACs) and the smallest parameter count (1.0M) among all baselines. These results highlight the efficiency of our model and its suitability for real-world deployment.

Qualitative Results

Visualization of Predicted Trajectory. Figure 4 shows qualitative comparisons on the ETH/UCY dataset among EigenTraj, MART, and our model. Our predictions (green) align more closely with ground-truth trajectories (red), capturing pedestrians' subtle movements and social interactions more accurately. Figure 5 presents trajectory predictions on

Figure 4: Qualitative results on ETH/UCY datasets. Historical trajectories are in blue, ground-truth trajectories are in red, and predicted trajectories are in green.

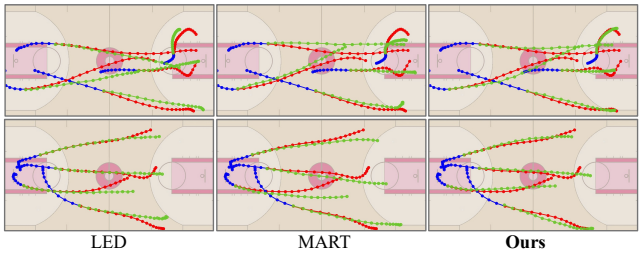

Figure 5: Qualitative results on NBA datasets. Historical trajectories are in blue, ground-truth trajectories are in red, and predicted trajectories are in green.

Variant	ETH/UCY Dataset		
	$\min \mathrm{ADE}_{20}$	$\min \mathrm{FDE}_{20}$	Param. Ratio
GCN (2-layer)	<u>0.22</u>	<u>0.36</u>	0.97x
GCN (4-layer)	0.23	<u>0.36</u>	1.11x
Ours	**0.20**	**0.32**	1.00x

Table 6: Ablation study on virtual graph effectiveness evaluated by $\min \mathrm{ADE}_{20}$/$\min \mathrm{FDE}_{20}$ on ETH/UCY dataset. **Bold** and <u>underline</u> indicate the best and second-best results.

the NBA dataset comparing LED, MART, and our approach. Our method consistently generates smoother and more precise trajectories, effectively capturing complex dynamics and long-term interactions among multiple agents. These visual results further confirm our model's superior predictive capability across diverse scenarios.

Expert Weight Analysis. Figure 6 illustrates the learned expert weights for different trajectory scenarios. The results indicate that the one-hop expert generally receives higher weights in simpler interactions. Conversely, the high-order expert dominates in more complex, globally interactive scenarios. This adaptive expert weighting confirms the effectiveness of our model in dynamically balancing the usage of one-hop expert and high-order expert.

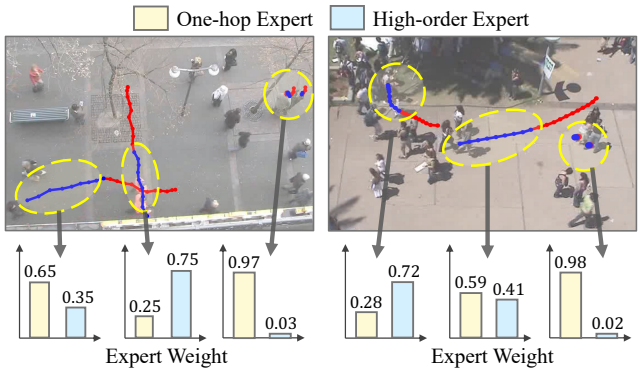

Figure 6: Visualization of learned expert weights across different scenarios. The model dynamically adjusts one-hop and high-order expert contributions by interaction complexity.

Conclusion

We propose ViTE, a novel trajectory prediction framework combining Virtual Graph for high-order interactions and Expert Router for adaptive expert routing. This design balances multi-scale reasoning and computational efficiency. Experiments on ETH/UCY, NBA, and SDD confirm state-of-the-art performance. Future work will incorporate contextual image information to further enhance scene understanding. Integrating visual semantics such as obstacles, road structure, or group behavior cues could provide stronger priors for trajectory prediction. We also aim to explore more flexible expert architectures that dynamically adjust their granularity or number based on scene complexity.

Acknowledgment

This project is supported in part by the EPSRC NorthFutures project (ref: EP/X031012/1).

References

Abu-El-Haija, S.; Kapoor, A.; Perozzi, B.; and Lee, J. 2020. N-gcn: Multi-scale graph convolution for semi-supervised node classification. In *UAI*, 841–851.

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 961–971.

Bae, I.; and Jeon, H.-G. 2021. Disentangled multi-relational graph convolutional network for pedestrian trajectory prediction. In *AAAI*, volume 35, 911–919.

Bae, I.; and Jeon, H.-G. 2023. A set of control points conditioned pedestrian trajectory prediction. In *AAAI*, volume 37, 6155–6165.

Bai, H.; Cai, S.; Ye, N.; Hsu, D.; and Lee, W. S. 2015. Intention-aware online POMDP planning for autonomous driving in a crowd. In *ICRA*, 454–460.

Chen, W.; Sang, H.; and Zhao, Z. 2025. PCHGCN: Physically Constrained Higher-Order Graph Convolutional Network for Pedestrian Trajectory Prediction. *IEEE Internet of Things Journal*, 12(13): 25033–25045.

Diao, C.; and Loynd, R. 2023. Relational Attention: Generalizing Transformers for Graph-Structured Tasks. In *ICLR*.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2255–2264.

Hu, F.; Wang, L.; Wu, S.; Wang, L.; and Tan, T. 2022. Graph classification by mixture of diverse experts. *IJCAI*.

Hu, Y.; Zhang, G.; Liu, P.; Lan, D.; Li, N.; Cheng, D.; Dai, T.; Xia, S.-T.; and Pan, S. 2025. TimeFilter: Patch-Specific Spatial-Temporal Graph Filtration for Time Series Forecasting. In *ICML*.

Huang, Q.; An, Z.; Zhuang, N.; Tao, M.; Zhang, C.; Jin, Y.; Xu, K.; Xu, K.; Chen, L.; Huang, S.; and Feng, Y. 2024. Harder Task Needs More Experts: Dynamic Routing in MoE Models. In *ACL*, 12883–12895.

Huang, Y.; Bi, H.; Li, Z.; Mao, T.; and Wang, Z. 2019. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *ICCV*, 6272–6281.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1): 79–87.

Jiang, W.; Han, J.; Liu, H.; Tao, T.; Tan, N.; and Xiong, H. 2024. Interpretable cascading mixture-of-experts for urban traffic congestion prediction. In *SIGKDD*, 5206–5217.

Kim, S.; Chi, H.-g.; Lim, H.; Ramani, K.; Kim, J.; and Kim, S. 2024. Higher-order Relational Reasoning for Pedestrian Trajectory Prediction. In *CVPR*, 15251–15260.

Kim, S.; Lee, D.; Kang, S.; Lee, S.; and Yu, H. 2023. Learning topology-specific experts for molecular property prediction. In *AAAI*, volume 37, 8291–8299.

Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; and Savarese, S. 2019. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Neurips*, 32.

Lee, S.; Lee, J.; Yu, Y.; Kim, T.; and Lee, K. 2024. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. In *ECCV*, 89–107.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer graphics forum*, volume 26, 655–664.

Li, R.; Katsigiannis, S.; Kim, T.-K.; and Shum, H. P. 2025a. BP-SGCN: Behavioral Pseudo-Label Informed Sparse Graph Convolution Network for Pedestrian and Heterogeneous Trajectory Prediction. *TNNLS*.

Li, R.; Katsigiannis, S.; and Shum, H. P. 2022. Multiclass-SGCN: Sparse Graph-Based Trajectory Prediction with Agent Class Embedding. In *ICIP*, 2346–2350. IEEE.

Li, R.; Qiao, T.; Katsigiannis, S.; Zhu, Z.; and Shum, H. P. 2025b. Unified Spatial-Temporal Edge-Enhanced Graph Networks for Pedestrian Trajectory Prediction. *TCSVT*.

Liu, J.; Mao, X.; Fang, Y.; Zhu, D.; and Meng, M. Q.-H. 2021. A survey on deep-learning approaches for vehicle trajectory prediction in autonomous driving. In *ICRB*, 978–985.

Lu, W.; Guan, Z.; Zhao, W.; Yang, Y.; and Jin, L. 2024. Nodemixup: Tackling under-reaching for graph neural networks. In *AAAI*, volume 38, 14175–14183.

Luber, M.; Stork, J. A.; Tipaldi, G. D.; and Arras, K. O. 2010. People tracking with human motion predictions from social forces. In *ICRA*, 464–469.

Luo, Y.; Cai, P.; Bera, A.; Hsu, D.; Lee, W. S.; and Manocha, D. 2018. Porca: Modeling and planning for autonomous driving among many pedestrians. *RAL*, 3(4): 3418–3425.

Mao, W.; Xu, C.; Zhu, Q.; Chen, S.; and Wang, Y. 2023. Leapfrog Diffusion Model for Stochastic Trajectory Prediction. In *CVPR*.

Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *CVPR*, 14424–14432.

Mu, S.; and Lin, S. 2025. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *Arxiv*.

Mustafa, B.; Riquelme, C.; Puigcerver, J.; Jenatton, R.; and Houlsby, N. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *NeurIPS*, 35: 9564–9576.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 261–268.

Qian, C.; Manolache, A.; Morris, C.; and Niepert, M. 2024. Probabilistic Graph Rewiring via Virtual Nodes. In *Neurips*.

Qiao, T.; Li, R.; Li, F. W.; Kubotani, Y.; Morishima, S.; and Shum, H. P. 2025. Geometric visual fusion graph neural networks for multi-person human-object interaction recognition in videos. *ESWA*.

Qiao, T.; Li, R.; Li, F. W.; and Shum, H. P. 2024. From category to scenery: An end-to-end framework for multi-person human-object interaction recognition in videos. In *ICPR*, 262–277.

Qiao, T.; Men, Q.; Li, F. W. B.; Kubotani, Y.; Morishima, S.; and Shum, H. P. H. 2022. Geometric Features Informed Multi-person Human-object Interaction Recognition in Videos. In *ECCV*.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *NeurIPS*, 34: 8583–8595.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 549–565.

Rusch, T. K.; Bronstein, M. M.; and Mishra, S. 2023. A survey on oversmoothing in graph neural networks. *Arxiv*.

Shao, M.; Wang, Z.; Duan, H.; Huang, Y.; Zhai, B.; Wang, S.; Long, Y.; and Zheng, Y. 2025. Rethinking Brain Tumor Segmentation From the Frequency Domain Perspective. *TMI*, 44(11): 4536–4553.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*.

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Zhou, M.; Niu, Z.; and Hua, G. 2021. SGCN: Sparse graph convolution network for pedestrian trajectory prediction. In *CVPR*, 8994–9003.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph attention networks. In *ICLR*.

Wang, H.; Jiang, Z.; You, Y.; Han, Y.; Liu, G.; Srinivasa, J.; Kompella, R.; Wang, Z.; et al. 2023. Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling. *Neurips*, 36: 50825–50837.

Wang, X.; Yu, F.; Dunlap, L.; Ma, Y.-A.; Wang, R.; Mirhoseini, A.; Darrell, T.; and Gonzalez, J. E. 2020. Deep mixture of experts via shallow embedding. In *UAI*, 552–562.

Xu, C.; Li, M.; Ni, Z.; Zhang, Y.; and Chen, S. 2022a. GroupNet: Multiscale Hypergraph Neural Networks for Trajectory Prediction with Relational Reasoning. *CVPR*, 6488–6497.

Xu, C.; Mao, W.; Zhang, W.; and Chen, S. 2022b. Remember intentions: Retrospective-memory-based trajectory prediction. In *CVPR*, 6488–6497.

Xu, C.; Tan, R. T.; Tan, Y.; Chen, S.; Wang, Y. G.; Wang, X.; and Wang, Y. 2023. EqMotion: Equivariant Multi-Agent Motion Prediction With Invariant Interaction Reasoning. In *CVPR*.

Xue, H.; Huynh, D. Q.; and Reynolds, M. 2018. SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. In *WACV*, 1186–1194.

Yuksel, S. E.; Wilson, J. N.; and Gader, P. D. 2012. Twenty Years of Mixture of Experts. In *TNNLS*.

Zhou, Y.; Lei, T.; Liu, H.; Du, N.; Huang, Y.; Zhao, V. Y.; Dai, A. M.; Chen, Z.; Le, Q. V.; and Laudon, J. 2022. Mixture-of-Experts with Expert Choice Routing. In *NeurIPS*.