ART: Adaptive Relational Transformer for Pedestrian Trajectory Prediction with Temporal-Aware Relations

Ruochen Li[1], Ziyi Chang[1], Junyan Hu[1], Jiannan Li[2], Amir Atapour-Abarghouei[1], Hubert P. H. Shum[1*]

***Abstract*— Accurate prediction of real-world pedestrian trajectories is crucial for a wide range of robot-related applications. Recent approaches typically adopt graph-based or transformer-based frameworks to model interactions. Despite their effectiveness, these methods either introduce unnecessary computational overhead or struggle to represent the diverse and time-varying characteristics of human interactions. In this work, we present an Adaptive Relational Transformer (ART), which introduces a Temporal-Aware Relation Graph (TARG) to explicitly capture the evolution of pairwise interactions and an Adaptive Interaction Pruning (AIP) mechanism to reduce redundant computations efficiently. Extensive evaluations on ETH/UCY and NBA benchmarks show that ART delivers state-of-the-art accuracy with high computational efficiency.**

I. INTRODUCTION

Pedestrian trajectory prediction aims to forecast future locations of pedestrians given observed trajectories. It is a key component of human–robot interaction [1], [16], autonomous driving systems [6], [25], and smart surveillance infrastructure [21], [19]. Accurate prediction of real-world pedestrian trajectories is crucial for robot-related applications, including human-aware robot navigation [7], [10] and collision avoidance systems [38], [22]. However, pedestrian trajectory prediction remains challenging due to the inherent stochasticity of motion behaviors and social interactions.

Early approaches to pedestrian trajectory prediction modeled social interactions using pooling-based strategies that aggregate information from nearby pedestrians through fixed spatial windows [2], [14] or occupancy maps [37], [13]. To enable explicit and expressive interaction modeling, two main directions have emerged. On the one hand, graph-based methods encode pedestrians as nodes and model their pairwise interactions as edges, enabling structured message-passing via graph neural networks [20], [21], [35], [23]. Specifically, these methods propagate interaction features via a message-passing paradigm over graph structures to model social influence among pedestrians. On the other hand, transformer-based approaches [17], [30], [39], [40] leverage self-attention mechanisms [31] to model interactions by dynamically weighting the influence of surrounding agents across both spatial and temporal dimensions, providing a flexible framework for capturing long-range dependencies and complex motion patterns.

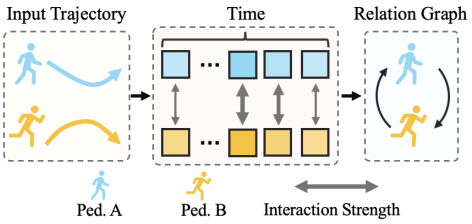

Fig. 1. Framework overview. Framework overview. The relation between pedestrians is inferred from the temporal evolution of their pairwise interactions over the observed history.

Despite these advances, previous methods either incur redundant computation or fail to capture the heterogeneous and evolving nature of pedestrian interactions. Pedestrian trajectories contain rich spatial-temporal information, yet existing methods [17], [35], [30], [29] construct pair-wise relations based on temporally aggregated representations, where trajectory sequences are compressed into static node features and relational reasoning is performed on these representations, thereby overlooking the dynamic evolution of interactions and time-varying strength. In addition, many methods construct social graphs using uniform neighbor selection, such as fully connected graphs [17], [28] or top-k based neighbor selection [21], [35], which ignore the heterogeneous nature of pedestrian interactions.

In this work, we propose an Adaptive Relational Transformer (ART), a lightweight Transformer-based framework for pedestrian trajectory prediction. First, to explicitly capture the temporal evolution of pairwise interactions, we introduce a Temporal-Aware Relation Graph (TARG). Unlike existing methods that compress multi-step trajectories into static embeddings before computing pairwise relations, TARG leverages pairwise attention [15] to model agent interactions at each time step and aggregates them via learnable weights, providing a temporally-aware relation modeling where informative time steps (e.g., moments of close proximity or directional change) receive higher weights, while less relevant moments contribute minimally to the final pairwise relations (see Figure 1). Second, to reduce redundancy in dense interaction graphs, we propose Adaptive Interaction Pruning (AIP) via top-p filtering, which adaptively selects informative neighbors based on interaction strength [24], rather than enforcing a fixed top-k neighborhood. For each agent, only neighbors whose attention weights jointly exceed a threshold p are retained, resulting in an adaptively sparsified interaction graph that retains only informative neighbors.

*Corresponding author

[1]Ruochen Li, Ziyi Chang, Junyan Hu, Amir Atapour-Abarghouei and Hubert P. H. Shum are with Department of Computer Science, Durham University, UK {ruochen.li, ziyi.chang, junyan.hu, amir.atapour-abarghouei, hubert.shum}@durham.ac.uk

[2]Jiannan Li is with the School of Computing and Information Systems, Singapore Management University, Singapore jiannanli@smu.edu.sg

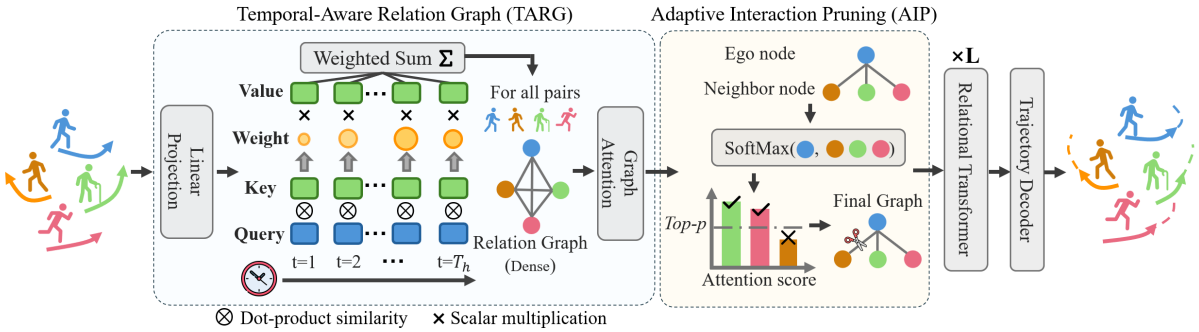

Fig. 2. Overview of ART. **Left:** Temporal-Aware Relation Graph (TARG) leverages pairwise attention to model agent interactions across time steps, assigning higher weights to informative moments. **Right:** Adaptive Interaction Pruning (AIP) uses top-p filtering to adaptively retain informative neighbors based on cumulative interaction strength, producing a sparsified graph for trajectory prediction.

The proposed ART achieves state-of-the-art (SOTA) results on both the ETH/UCY [27], [18] and NBA SportVU [26] benchmarks while maintaining superior computational efficiency over prior approaches. The main contributions are:

- We propose an Adaptive Relational Transformer (ART), a lightweight Transformer-based framework for pedestrian trajectory prediction that explicitly models social interactions while maintaining computational efficiency.
- We introduce a Temporal-Aware Relation Graph (TARG) that explicitly models pairwise pedestrian interactions by accounting for the relative importance of different time steps.
- We propose Adaptive Interaction Pruning (AIP) via top-p filtering to adaptively sparsify dense interaction graphs based on cumulative interaction strength, reducing redundancy and improving adaptivity.

II. METHOD

A. Problem Formulation

Pedestrian trajectory prediction aims to infer pedestrians' future movements from their observed trajectories. Consider a scene containing M pedestrians observed over T_h time steps. The historical trajectory of pedestrian i is represented as $\mathbf{p}_j^h = \{(x_i^t, y_i^t) \mid (x_i^t, y_i^t) \in \mathbb{R}^2, t = 1, \ldots, T_h\}$, where (x_i^t, y_i^t) denotes the 2D spatial coordinates at time step t. The corresponding ground-truth (GT) future trajectory over a prediction horizon of T_f time steps is denoted as $\mathbf{p}_j^f = \{(x_i^t, y_i^t) \mid t = T_h+1, \ldots, T_h+T_f\}$. By stacking trajectories of all pedestrians, the observed and GT trajectories can be expressed as $\mathbf{P}^h \in \mathbb{R}^{M\times T_h\times 2}$ and $\mathbf{P}^f \in \mathbb{R}^{M\times T_f\times 2}$. The training objective minimizes the discrepancy between the predicted trajectories $\hat{\mathbf{P}}^f$ and the ground-truth $\mathbf{P}^f$.

B. Temporal-Aware Relation Graph Construction

Existing graph construction methods compress $\mathbf{P}_h \in \mathbb{R}^{M\times T_h\times 2}$ into static embeddings before computing pairwise weights, losing temporal information about when interactions occur. As shown in Figure 2 (Left), we propose Temporal-Aware Relation Graph (TARG), which preserves temporal attribution by applying attention over observation sequences before spatial aggregation.

We encode the observed trajectories into high-dimensional features. Given $\mathbf{P}_h \in \mathbb{R}^{M\times T_h\times 2}$, we apply a linear projection with positional encoding to obtain temporal node features:

$$\mathbf{H} = \mathrm{PosEnc}(\mathbf{W}_{\text{in}}\mathbf{P}_h) \in \mathbb{R}^{M\times T_h\times d}, \tag{1}$$

where $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d\times 2}$ is a learnable projection matrix, d is the hidden dimension, and $\mathrm{PosEnc}(\cdot)$ denotes sinusoidal positional encoding [31] that injects temporal information.

To capture time-varying interactions between pedestrian pairs, we reshape $\mathbf{H}$ from $\mathbb{R}^{M\times T_h\times d}$ to $\mathbb{R}^{T_h\times M\times d}$ and apply multi-head attention [31] over the temporal dimension. For each attention head $h \in \{1, \ldots, H\}$, we have:

$$\mathbf{Q}^h = \mathbf{H}\mathbf{W}_Q^h, \quad \mathbf{K}^h = \mathbf{H}\mathbf{W}_K^h, \quad \mathbf{V}^h = \mathbf{H}\mathbf{W}_V^h, \tag{2}$$

where $\mathbf{W}_Q^h, \mathbf{W}_K^h, \mathbf{W}_V^h \in \mathbb{R}^{d\times d_h}$ are learnable weight matrices and $d_h = d/H$ is the dimension per head. For each pair of pedestrians (i, j), we compute time-resolved attention scores by applying attention across all time steps:

$$\alpha_{ij}^h(t) = \frac{\exp\left(\frac{\mathbf{Q}_i^h(t)^\top \mathbf{K}_j^h(t)}{\sqrt{d_h}}\right)}{\sum_{t'=1}^{T_h}\exp\left(\frac{\mathbf{Q}_i^h(t')^\top \mathbf{K}_j^h(t')}{\sqrt{d_h}}\right)}, \quad t = 1, \ldots, T_h, \tag{3}$$

where $\alpha_{ij}^h(t) \in [0, 1]$ represents the attention weight for pedestrian pair (i, j) at time step t under head h, and $\sum_{t=1}^{T_h}\alpha_{ij}^h(t) = 1$. These temporally-aware scores explicitly capture when each interaction becomes significant, preserving temporal attribution throughout the graph construction.

The attention scores are used to aggregate value representations across time:

$$\mathbf{R}_{ij}^h = \sum_{t=1}^{T_h}\alpha_{ij}^h(t)\cdot\mathbf{V}_j^h(t) \in \mathbb{R}^{d_h}, \tag{4}$$

where $\mathbf{R}_{ij}^h$ is the aggregated relation representation for pedestrian pair (i, j) under head h. By concatenating outputs from all heads and applying a linear projection, we obtain the pairwise relation features $\mathbf{R}_{ij}$.

Finally, we compute edge weights from the pairwise relation features $\mathbf{R}_{ij}$. Following the graph attention mechanism [32], we apply a learnable function to produce edge weights:

$$w_{ij} = \sigma\left(\mathbf{a}^\top[\mathbf{R}_{ij}\|\mathbf{R}_{ji}] + b\right), \tag{5}$$

where $\mathbf{a} \in \mathbb{R}^{2d}$ and $b \in \mathbb{R}$ are learnable parameters, $\sigma(\cdot)$ is the sigmoid activation function, and the concatenation of $\mathbf{R}_{ij}$ and $\mathbf{R}_{ji}$ captures bidirectional interactions. The adjacency matrix $\mathbf{W} \in \mathbb{R}^{M \times M}$ with entries w_{ij} defines the temporal-aware relation graph.

C. Adaptive Interaction Pruning

As discussed in the previous section, TARG produces an adjacency matrix $\mathbf{W}$ capturing pairwise interactions. However, modeling all $M(M-1)$ relations incurs $\mathcal{O}(M^2)$ computational complexity and introduces noise from spurious interactions. To reduce computational redundancy while enabling adaptive neighbor selection, we propose Adaptive Interaction Pruning (AIP) (Figure 2 (Right)) via top-p filtering [24], which allows each pedestrian to discover a personalized neighbor set based on attention distributions. For each target pedestrian i, we rank potential neighbors $\{j \mid j \neq i\}$ by their edge weights in descending order. Let π_i denote the permutation that sorts edge weights:

$$w_{i,\pi_i(1)} \geq w_{i,\pi_i(2)} \geq \cdots \geq w_{i,\pi_i(M-1)}. \tag{6}$$

We then compute the cumulative sum of sorted weights:

$$C_i(k) = \sum_{j=1}^{k} w_{i,\pi_i(j)}, \quad k = 1, \ldots, M-1. \tag{7}$$

Given threshold $p \in (0, 1]$, we retain the minimal set of neighbors whose cumulative weight exceeds p:

$$k_i^* = \min \left\{ k \mid \frac{C_i(k)}{C_i(M-1)} \geq p \right\}, \tag{8}$$

where k_i^* is the effective neighborhood size for pedestrian i. The pruned adjacency matrix $\tilde{\mathbf{W}} \in \mathbb{R}^{M \times M}$ is obtained by:

$$\tilde{w}_{ij} = \begin{cases} w_{ij} & \text{if } j \in \{\pi_i(1), \ldots, \pi_i(k_i^*)\}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Finally, we renormalize the weights for each pedestrian as:

$$\tilde{w}_{ij} = \frac{\tilde{w}_{ij}}{\sum_{j'=1}^{M} \tilde{w}_{ij'}}. \tag{10}$$

This design adaptively determines neighborhood size based on learned interaction strength, resulting in a sparsified yet informative interaction graph that improves both efficiency and robustness.

D. Relational Transformer and Trajectory Decoding

Given the pruned adjacency matrix $\tilde{\mathbf{W}} \in \mathbb{R}^{M \times M}$ from AIP, we employ L layers of Relational Transformer (RT) [17], [11] to refine node representations through edge-aware attention. Let $\mathbf{h}_i^{(0)} \in \mathbb{R}^d$ denote the initial node feature for pedestrian i obtained from TARG.

At each layer $l \in \{1, \ldots, L\}$, RT updates both node and edge features. For node updates, it applies multi-head attention modulated by edge features $\mathbf{e}_{ij}^{(l-1)}$:

$$\alpha_{ij}^{(l)} = \text{softmax}_j \left(\frac{(\mathbf{Q}_i^{(l)})^\top (\mathbf{K}_j^{(l)} + \mathbf{K}_{\mathbf{e}_{ij}}^{(l)})}{\sqrt{d_h}} \right), \tag{11}$$

where $\mathbf{Q}_i^{(l)}, \mathbf{K}_j^{(l)}$ are node-based query and key projections, $\mathbf{K}_{\mathbf{e}_{ij}}^{(l)}$ is the edge-based key, and the attention is computed only over pruned neighbors (where $\tilde{w}_{ij} > 0$). Node features are then updated via:

$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \text{FFN} \left(\sum_{j:\tilde{w}_{ij}>0} \alpha_{ij}^{(l)} (\mathbf{V}_j^{(l)} + \mathbf{V}_{\mathbf{e}_{ij}}^{(l)}) \right), \tag{12}$$

where $\mathbf{V}_j^{(l)}$ and $\mathbf{V}_{\mathbf{e}_{ij}}^{(l)}$ are value projections, and FFN denotes a feed-forward network with residual connection. Edge features are updated by aggregating source and target node information:

$$\mathbf{e}_{ij}^{(l)} = \mathbf{e}_{ij}^{(l-1)} + \text{MLP} \left([\mathbf{h}_i^{(l)} \| \mathbf{h}_j^{(l)} \| \mathbf{e}_{ij}^{(l-1)} \| \mathbf{e}_{ji}^{(l-1)}] \right), \tag{13}$$

where $\|$ denotes concatenation and the bidirectional edge features capture symmetric interactions. This edge-aware mechanism allows RT to leverage the pruned graph structure from AIP while refining pairwise relations. We use $L = 1$ layer in our experiments for efficiency.

For trajectory prediction, we concatenate the initial node features $\mathbf{h}_i^{(0)}$ and the refined features $\mathbf{h}_i^{(L)}$ after RT encoding, then feed them into K parallel prediction heads to obtain trajectories following [35], [17]:

$$\hat{\mathbf{p}}_{i,k}^f = \text{MLP}_k([\mathbf{h}_i^{(0)} \| \mathbf{h}_i^{(L)}]) \in \mathbb{R}^{T_f \times 2}, \quad k = 1, \ldots, K, \tag{14}$$

where $\|$ denotes concatenation and $\hat{\mathbf{p}}_{i,k}^f$ represents the k-th predicted future trajectory for pedestrian i.

Given K predicted trajectories $\{\hat{\mathbf{p}}_f^{i,k}\}_{k=1}^K$ for each pedestrian i, we adopt the best-of-K training strategy by selecting the trajectory with minimum ℓ_2 distance to the ground truth:

$$\mathcal{L} = \frac{1}{MT_f} \sum_{i=1}^{M} \sum_{t=1}^{T_f} \min_{k \in \{1,\ldots,K\}} \left\| \mathbf{p}_i^{f,t} - \hat{\mathbf{p}}_{i,k}^{f,t} \right\|_2, \tag{15}$$

where $\mathbf{p}_i^{f,t} = (x_i^t, y_i^t)$ is the ground-truth position of pedestrian i at time t, and $\hat{\mathbf{p}}_{i,k}^{f,t}$ is prediction from the k-th head.

III. EXPERIMENTS

A. Datasets and Evaluation Metrics

We evaluate ART on the ETH/UCY [27], [18] and NBA SportVU [26] benchmarks. The ETH/UCY dataset consists of five subsets, including ETH, HOTEL, UNIV, ZARA1, and ZARA2 captured in diverse social scenarios. Following the standard protocol in [17], we use 3.2 seconds (8 frames) of historical trajectories to predict the subsequent 4.8 seconds (12 frames). The NBA SportVU dataset contains trajectories of 10 players from basketball games. Consistent with [33], [17], we use 2.0 seconds (10 frames) to predict the next 4.0 seconds (20 frames). We evaluate performance using $\min \text{ADE}_k$ and $\min \text{FDE}_k$, following [14], [17]. Average Displacement Error (ADE) computes the mean Euclidean distance across all predicted time steps, whereas Final Displacement Error (FDE) evaluates the distance at the last time step.

TABLE I

QUANTITATIVE RESULT ON THE ETH/UCY DATASET. METRICS ARE $\min \mathrm{ADE}_{20}/\min \mathrm{FDE}_{20}$. BEST AND SECOND-BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE.

ETH/UCY Dataset										
Subset	GroupNet [33]	MemoNet [34]	MID [12]	NPSN [4]	EqMotion [35]	ET [3]	LED [26]	SingularTraj [5]	MART [17]	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	**0.35**/0.47
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	0.13/0.21
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	0.24/0.43
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**/0.28
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	**0.12/0.21**
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	**0.20/0.32**

TABLE II

QUANTITATIVE RESULT ON THE NBA DATASET. METRICS ARE $\min \mathrm{ADE}_{20}/\min \mathrm{FDE}_{20}$. BEST AND SECOND-BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE.

NBA Dataset										
Time	STAR [39]	GroupNet [33]	MemoNet [34]	MID [12]	NPSN [4]	DynGroupNet [36]	LED [26]	SingularTraj [5]	MART [17]	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**	**0.53/0.71**
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**	**0.72/0.90**

B. Quantitative results

We evaluate our model against a comprehensive set of SOTA baselines on both ETH/UCY and NBA datasets, including STAR [39], GroupNet [33], MemoNet [34], MID [12], NPSN [4], EqMotion [35], ET [3], DynGroupNet [36], LED [26], SingularTraj [5], and MART [17].

The quantitative results are summarized in Tables I and II. Our method achieves the best overall results on the ETH/UCY datasets with the lowest average $\min \mathrm{ADE}_{20}/\min \mathrm{FDE}_{20}$ of 0.20/0.32, improving over strong baselines with 0.21 average ADE (e.g., MART/SingularTraj/LED) by 4.8% while matching the best average FDE. It also attains the best result on ZARA2 (0.12/0.21), demonstrating strong generalization across diverse scenes. On the NBA dataset, our model achieves consistently superior performance across different time scales in terms of both ADE and FDE, highlighting effective long-term interaction modeling and robust prediction accuracy.

C. Ablation Study and Model Analysis

1) Ablation Study on Relation Weighting Strategies: In this section, we evaluate different relation weighting strategies in the proposed Temporal-Aware Relation Graph, including cosine similarity, which assigns static similarity-based weights, random weighting, which ignores temporal structure by randomly sampling weights, and uniform weighting, which applies equal weights to all pairwise relations. As shown in Table III, our temporally-aware weighting achieves the best performance on the ETH/UCY dataset, demonstrating that explicitly modeling temporal attribution in pairwise relations leads to more effective interaction representations.

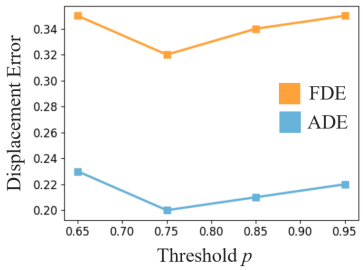

Fig. 3. Ablation study of Top-p threshold on the ETH/UCY dataset.

TABLE III

ABLATION STUDY OF RELATION WEIGHTING STRATEGIES ON THE ETH/UCY DATASET.

Weighting Strategies	ETH/UCY Dataset	
	$\min \mathrm{ADE}_{20}$	$\min \mathrm{FDE}_{20}$
Cosine Similarity	0.22	0.36
Random Weighting	0.23	0.37
Uniform Weighting	0.23	0.36
Ours	**0.20**	**0.32**

2) Ablation Study on Top-p Threshold: Figure 3 illustrates the impact of different Top-p thresholds on prediction performance. As p decreases from 0.95 to 0.75, both ADE and FDE consistently improve, suggesting that moderate sparsification effectively suppresses weak or noisy interactions while preserving informative relations. When $p = 0.95$, the behavior closely resembles the non-sparsified setting, leading

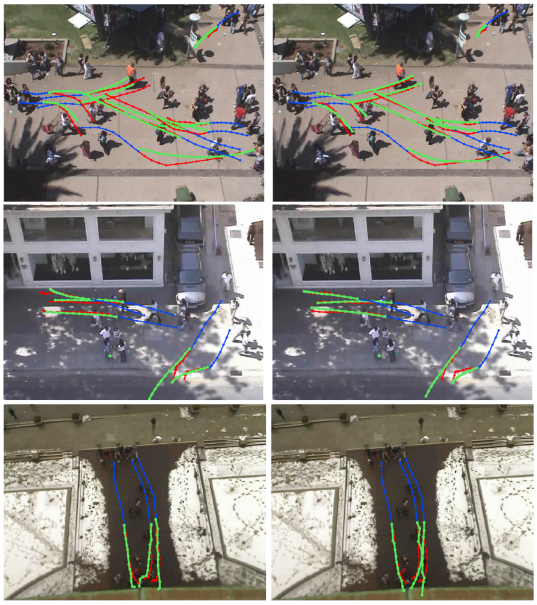

Fig. 4. Qualitative comparisons with MART [17] on the ETH/UCY dataset. Past trajectories are shown in blue, ground truth in red, and model predictions in green.

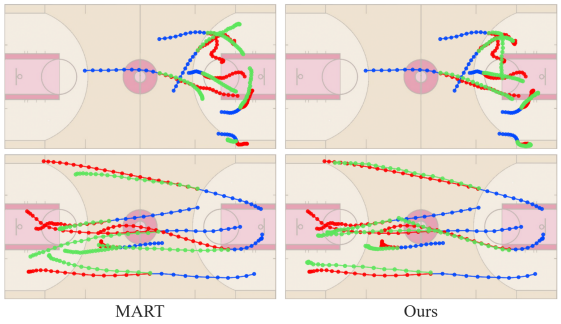

Fig. 5. Qualitative comparisons with MART [17] on the NBA dataset. Past trajectories are shown in blue, ground truth in red, and predictions in green.

to degraded performance due to redundant interaction modeling. Conversely, overly aggressive sparsification ($p = 0.65$) also harms performance by discarding useful interactions. Overall, these results demonstrate that a moderate Top-p threshold ($p = 0.75$) achieves the best balance between interaction selectivity and information preservation.

3) Complexity and Efficiency Analysis: As shown in Table IV, our model achieves competitive parameter efficiency with the lowest MACs among all methods, highlighting its efficiency for trajectory prediction. Compared with existing approaches that rely on either larger model sizes or substantially higher computational costs, our method offers a more favorable trade-off between accuracy and efficiency.

D. Qualitative Results

1) Trajectory Prediction Visualizations: In this section, we present qualitative visualizations to illustrate the superiority of our method. We choose MART [17] as the comparison baseline since it is a prior SOTA approach that adopts the same trajectory decoder and relational Transformer architecture, enabling a fair comparison. Figure 4 illustrates the qualitative results on the ETH/UCY dataset across three scenes of increasing interaction complexity, ranging from relatively sparse environments to densely crowded settings. Compared with MART, our predictions more closely follow the ground-truth trajectories and exhibit fewer unrealistic overlaps and collisions, with the advantages becoming more pronounced as scene complexity increases.

Figure 5 presents qualitative comparisons on the NBA dataset. Although long-term trajectory prediction remains challenging for both methods due to highly dynamic and coordinated player movements, our predictions are consistently more aligned with the ground truth than those of MART, demonstrating improved robustness in complex multi-agent sports environments.

2) Temporal-Aware Relation Visualization: We qualitatively verify the effectiveness of the proposed TARG in capturing temporally salient interactions. Figure 6 provides an intuitive visualization of the proposed TARG on the NBA dataset. By tracking the attention scores between selected player pairs over time, we observe that the model assigns low weights when players are spatially distant and significantly increases the attention as they approach or interact. This behavior indicates that the proposed method can effectively identify and emphasize critical time steps where interactions become salient, thereby preserving temporal attribution rather than uniformly aggregating interactions across time.

TABLE IV

MODEL COMPLEXITY COMPARISON. BEST AND SECOND-BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE.

Method	#Param.	MACs
STAR [39]	**1.0**M	12.0G
MemoNet [34]	10.7M	6.0G
GroupNet [33]	2.2M	411.5M
MID [12]	9.0M	40.3G
EqMotion [35]	3.0M	147.1M
LED [26]	10.9M	15.0G
MART [17]	1.5M	43.3M
Ours	**1.0**M	**40.0**M

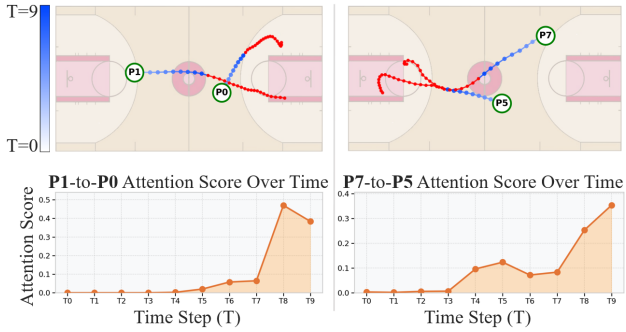

Fig. 6. Visualization of temporal-aware relation weights on the NBA dataset. Blue denotes historical trajectories, ground-truth trajectories are in red, and orange indicates the temporal evolution of attention scores.

IV. CONCLUSION

We propose ART to model pedestrian interactions through a temporal-aware relation graph and adaptive interaction pruning. By explicitly preserving temporal attribution in pairwise relations and dynamically sparsifying interaction graphs, ART achieves a favorable balance between expressive interaction modeling and computational efficiency. Our method achieves superior performance on ETH/UCY and NBA, with strong generalization in multi-agent scenarios. Future work will explore more complex and heterogeneous environments using probabilistic models like diffusion models [8] in real-world robotic systems for interactive decision-making. Other informative features such as whole-body motions [9] can also be considered in the future.

REFERENCES

[1] Rina Akabane and Yuka Kato. Pedestrian trajectory prediction based on transfer learning for human-following mobile robots. *IEEE Access*, 9:126172–126185, 2021.

[2] Alexandre Alahi, Krataith Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, pages 961–971, 2016.

[3] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *ICCV*, pages 10017–10029, 2023.

[4] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Non-probability sampling network for stochastic human trajectory prediction. In *CVPR*, pages 6477–6487, 2022.

[5] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *CVPR*, pages 17890–17901, 2024.

[6] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online pomdp planning for autonomous driving in a crowd. In *ICRA*, pages 454–460, 2015.

[7] Rashmi Bhaskara, Hrishikesh Viswanath, and Aniket Bera. Trajectory prediction for robot navigation using flow-guided markov neural operator. In *ICRA*, pages 15209–15216. IEEE, 2024.

[8] Ziyi Chang, George A Koulieris, Hyung Jin Chang, and Hubert PH Shum. On the design fundamentals of diffusion models: A survey. *Pattern Recognition*, 169:111934, 2026.

[9] Ziyi Chang, He Wang, George Koulieris, and Hubert PH Shum. Large-scale multi-character interaction synthesis. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025.

[10] Zhixian Chen, Chao Song, Yuanyuan Yang, Baoliang Zhao, Ying Hu, Shoubin Liu, and Jianwei Zhang. Robot navigation based on human trajectory prediction and multiple travel modes. *Applied Sciences*, 8(11):2205, 2018.

[11] Cameron Diao and Ricky Loynd. Relational attention: Generalizing transformers for graph-structured tasks. In *ICLR*, 2023.

[12] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *CVPR*, pages 17113–17122, 2022.

[13] Ke Guo, Wenxi Liu, and Jia Pan. End-to-end trajectory distribution prediction based on occupancy grid maps. In *CVPR*, pages 2242–2251, 2022.

[14] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018.

[15] Seungwoong Ha and Hawoong Jeong. Learning heterogeneous interaction strengths by trajectory prediction with graph neural network. *arXiv*, 2022.

[16] Zhe Huang, Ruohua Li, Kazuki Shin, and Katherine Driggs-Campbell. Learning sparse interaction graphs of partially detected pedestrians for trajectory prediction. *IEEE RAL*, 7(2):1198–1205, 2022.

[17] Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyoobin Lee. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. In *ECCV*, pages 89–107, 2024.

[18] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664, 2007.

[19] Ruochen Li, Stamos Katsigiannis, Tae-Kyun Kim, and Hubert PH Shum. Bp-sgcn: Behavioral pseudo-label informed sparse graph convolution network for pedestrian and heterogeneous trajectory prediction. *TNNLS*, 2025.

[20] Ruochen Li, Stamos Katsigiannis, and Hubert PH Shum. Multiclass-sgcn: Sparse graph-based trajectory prediction with agent class embedding. In *ICIP*, pages 2346–2350. IEEE, 2022.

[21] Ruochen Li, Tanqiu Qiao, Stamos Katsigiannis, Zhanxing Zhu, and Hubert PH Shum. Unified spatial-temporal edge-enhanced graph networks for pedestrian trajectory prediction. *TCSVT*, 2025.

[22] Ruochen Li, Zhanxing Zhu, Tanqiu Qiao, and Hubert PH Shum. Vite: Virtual graph trajectory expert router for pedestrian trajectory prediction. *arXiv*, 2025.

[23] Ruochen Li, Zhanxing Zhu, Tanqiu Qiao, and Hubert PH Shum. Vite: Virtual graph trajectory expert router for pedestrian trajectory prediction. In *AAAI*, volume 40, pages 17535–17543, 2026.

[24] Chaofan Lin, Jiaming Tang, Shuo Yang, Hanshuo Wang, Tian Tang, Boyu Tian, Ion Stoica, Song Han, and Mingyu Gao. Twilight: Adaptive attention sparsity with hierarchical top-p pruning. In *Neurips*, 2025.

[25] Yuanfu Luo, Panpan Cai, Aniket Bera, David Hsu, Wee Sun Lee, and Dinesh Manocha. Porca: Modeling and planning for autonomous driving among many pedestrians. *IEEE RAL*, 3(4):3418–3425, 2018.

[26] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *CVPR*, 2023.

[27] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, pages 261–268, 2009.

[28] Tanqiu Qiao, Ruochen Li, Frederick WB Li, Yoshiki Kubotani, Shigeo Morishima, and Hubert PH Shum. Geometric visual fusion graph neural networks for multi-person human-object interaction recognition in videos. *arXiv*, 2025.

[29] Tanqiu Qiao, Ruochen Li, Frederick WB Li, and Hubert PH Shum. From category to scenery: An end-to-end framework for multi-person human-object interaction recognition in videos. In *ICPR*, pages 262–277, 2024.

[30] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *ICCV*, pages 9675–9684, 2023.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neurips*, 30, 2017.

[32] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[33] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. *CVPR*, pages 6488–6497, 2022.

[34] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *CVPR*, pages 6488–6497, 2022.

[35] Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. EqMotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *CVPR*, 2023.

[36] Chenxin Xu, Yuxi Wei, Bohan Tang, Sheng Yin, Ya Zhang, Siheng Chen, and Yanfeng Wang. Dynamic-group-aware networks for multi-agent trajectory prediction with relational reasoning. *Neural Networks*, 170:564–577, 2024.

[37] Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *WACV*, pages 1186–1194, 2018.

[38] Jing Yang, Yuehai Chen, Shaoyi Du, Badong Chen, and Jose C Principe. Ia-lstm: Interaction-aware lstm for pedestrian trajectory prediction. *IEEE transactions on cybernetics*, 54(7):3904–3917, 2024.

[39] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *ECCV*, pages 507–523. Springer, 2020.

[40] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *ICCV*, pages 9813–9823, 2021.