

# Hyper-STTN: Hypergraph Augmented Spatial-Temporal Transformer Network for Trajectory Prediction

Weizheng Wang<sup>1,†</sup>, Baijian Yang<sup>1</sup>, Sungeun Hong<sup>3</sup>, Wenhai Sun<sup>1</sup>, and Byung-Cheol Min<sup>1,2</sup>

**Abstract**—Predicting crowd intentions and trajectories is critical for a range of real-world applications, involving social robotics and autonomous driving. Accurately modeling such behavior remains challenging due to the complexity of pairwise spatial-temporal interactions and the heterogeneous influence of groupwise dynamics. To address these challenges, we propose Hyper-STTN, a Hypergraph-augmented Spatial-Temporal Transformer Network for crowd trajectory prediction. Hyper-STTN constructs crowd hypergraphs with multi-scale group sizes to model groupwise correlations, captured through spectral hypergraph convolution based on hypergraph random walk. In parallel, a spatial-temporal transformer is employed to learn pedestrians’ pairwise latent interactions across multimodal dimensions. Eventually, above heterogeneous groupwise and pairwise features are subsequently incorporated and aligned via a multimodal transformer. Extensive experiments on public pedestrian motion datasets demonstrate that Hyper-STTN consistently outperforms state-of-the-art baselines and ablation models. The project website is available at <https://sites.google.com/view/hypersttn>.

## I. INTRODUCTION

Human trajectory prediction is a pivotal research topic in computer vision and robotics, aiming to anticipate agents’ future movements based on their past behaviors. The forecasting capability is critical for many real-world applications such as smart city systems and social robot navigation [1]–[4]. However, accurately forecasting human trajectories in social environments remains highly challenging due to the inherent stochasticity and unpredictability, and the diversity of crowd dynamics. These difficulties are compounded by the need to model intricate interactions among individuals and groups, involving both cooperative and competitive social behaviors, as illustrated in Fig. 1.

Human trajectory forecasting is governed by three interrelated factors: individual intrinsic states, pairwise and groupwise social interactions, and instantaneous intentions. However, lots of existing approaches have struggled in modeling higher-order interactions and in reasoning across heterogeneous feature domains. Specifically, individual intrinsic states encode correlations between an agent’s current conditions and temporal patterns, implying future velocities or locations to be inferred from historical observations through sequence

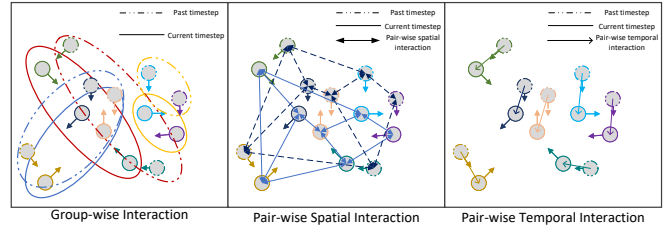


Fig. 1. HHI feature illustration: groupwise HHI captures latent correlations among high-level perspectives on group behaviors, while pairwise spatial-temporal HHI represents individual influences.

reasoning processes, as demonstrated in prior studies [5], [6]. Moreover, contemporary forecasting frameworks are still unable to accurately infer subjective intentions from sparse or ambiguous observations, even when employing advanced machine learning or neural network architectures [7]. Thus, addressing these challenges necessitates enhanced representations of latent social dynamics, achieved through the joint integration of group-level embeddings and pairwise features across spatial-temporal dimensions.

Recently, recent research has increasingly focused on estimating complex social influences among humans to better capture the uncertainty of human movements. For instance, [6], [8]–[10] employ neural networks and attention mechanisms to model pairwise social interactions among agents across spatial-temporal dimensions. Nevertheless, the absence of an explicit graph structure constrains the expressiveness of the extracted features. Alternatively, some works [9], [11], [12] incorporate spatial-temporal graphs, in which spatial edges encode interactions between agents and temporal edges represent individual temporal dependencies reflecting intrinsic states. More recently, [7], [13] extended this paradigm by constructing pedestrian group dynamics as a hypergraph to capture high-order group-level interactions for trajectory forecasting, where each hyperedge connects multiple vertices to represent collective group dynamics.

Despite state-of-the-art (SOTA) algorithms have achieved remarkable performance, inadequate modeling of human-human interaction (HHI) continues to constrain further advances on forecasting accuracy, particularly in highly dynamic environments. None of the above approaches fully model HHI across both group-level and pairwise feature dimensions [6], nor do they effectively align heterogeneous multimodal dependencies [9]. For instance, the lack of group-level interaction reasoning may obscure intergroup coordination or conflict [12]. In team sports such as basketball, different defensive strategies exemplify this challenge: zone defenses typically involve one offensive player interacting with multiple defenders, whereas man-to-man schemes entail

<sup>1</sup>School of Applied and Creative Computing, Purdue University, West Lafayette, IN, USA. wang5716@purdue.edu, byang@purdue.edu, sun841@purdue.edu. <sup>2</sup>Department of Computer Science and Department of Intelligent Systems Engineering, Indiana University Bloomington, Bloomington, IN, USA. minb@iu.edu. <sup>3</sup>Department of Applied Artificial Intelligence, Sungkyunkwan University, Suwon, South Korea. csehong@skku.edu.

<sup>†</sup>Corresponding author.

predominantly pairwise matchup. Furthermore, insufficient spatial-temporal feature inference can also exacerbate ambiguities in HHI modeling [7].

To address the aforementioned challenges, we propose Hyper-STTN, a hypergraph-augmented spatial-temporal transformer network explicitly designed to model both pairwise and groupwise social interactions across spatial-temporal dimensions, as illustrated in Fig. 1. The central idea of Hyper-STTN is to provide an effective framework for inferring HHI and social dependencies by leveraging multiscale hypergraphs. The main contributions of this work are as follows: (1) We introduce Hyper-STTN, which constructs crowd dynamics using a set of multiscale hypergraphs to jointly capture groupwise and pairwise social interactions in both spatial and temporal domains; (2) Hyper-STTN incorporates a multimodal transformer module to align heterogeneous features and interactions, thereby mitigating interpretational ambiguities; (3) Extensive experiments on publicly available pedestrian trajectory datasets demonstrate that Hyper-STTN outperforms existing SOTA approaches.

## II. BACKGROUND

### A. Related Works

Early efforts to human trajectory prediction relied on handcrafted models, such as the social force model [14] and Gaussian processes [15], which leverage fixed physical or mathematical rules to abstract environmental dynamics. However, the inherent inflexibility of these traditional methods leads to overlooking latent social interactions and producing inaccurate forecasts under complex conditions. Motivated by advances in machine learning, numerous deep learning-based algorithms have been developed to further express social interactions among pedestrians [6], [8], [16], [17]. However, previous learning-based methods either fail to adaptively incorporate the extracted features or cannot effectively leverage long-term dependencies. Inspired by the success of attention mechanisms and transformers in sequence learning and pairwise feature representation [9], [18], [19], recent work has explicitly modeled social interactions across both spatial and temporal dimensions through spatial-temporal graph representations of crowd movements. While pairwise interactions have been effectively addressed through the development of transformers, aforementioned approaches still neglect to capture groupwise latent influences on individual trajectory.

Despite prior studies [7], [13] have adapted hypergraph-based networks to learn groupwise features, they still struggle to disentangle spatial interactions along the temporal dimension. Consequently, aligning groupwise and pairwise social interactions with long-term dependencies remains a critical challenge in human trajectory prediction. In this work, we address this gap by not only integrating spatial and temporal social interactions from the mask attention mechanism into the hypergraph-based interaction representation, but also introducing the cross-modal attention mechanism [20] to effectively exploit groupwise and pairwise heterogeneous multimodal dependencies.

### B. Social Interaction Reasoning

Conventional sequence learning models, such as convolutional networks [21], [22] and recurrent networks [6], [12], [19], typically adopt hierarchical architectures that process information sequentially through network depth orderly. The lack of parallelization can potentially cause information to gradually vanish or explode across many layers of computation, particularly when capturing long-term dependencies. Conversely, transformer networks directly model element-to-element correlations in a fully parallel manner, leading to substantial improvements in sequence learning and pairwise interaction representation, as exemplified by [18].

More recently, some existing works [9], [10], [16], [17] have employed transformer architectures to encode pairwise interactions across both spatial and temporal dimensions in trajectory prediction tasks. Given that agents' future trajectories are strongly correlated with their self-spatial and temporal features, those methods have achieved improvements over the SOTA performance via the better latent HHI feature representation. Nevertheless, although transformers can emphasize the importance of each pair of elements through attention mechanisms, they still tend to overlook groupwise features that capture higher-order crowd correlations.

Whereas standard graphs restrict edges to pairwise connections, hypergraphs provide a natural and expressive framework for modeling groupwise interactions, with hyperedges capable of simultaneously linking multiple vertices. Hypergraph models have been applied across diverse domains including biology, physics, recommender systems, and trajectory prediction [23], [24]. For instance, in ecological systems the correlations among predator-competitor groups are rarely linear. Especially, multiple predator-competitor interactions are difficult to approximate with pairwise edges alone. Accordingly, [23] introduced higher-order interactions among biotic communities via hypergraphs to better capture and stabilize natural ecosystems. In a similar vein, [7] employed hypergraphs to capture groupwise social interactions for forecasting human trajectories. However, most existing approaches either rely on pre-defined topologies or construct hypergraphs solely from direct distances, without accounting for the underlying data distribution. In contrast, Hyper-STTN explicitly incorporates both pedestrian distribution effects and group scale to generate multiscale hypergraphs, while simultaneously embedding crowd dependencies across spatial-temporal dynamics within the hypergraph structure.

## III. METHODOLOGY

### A. Preliminary

We construct a set of multiscale hypergraphs  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathbf{W})$  for the representation of crowd movement dynamics. Where  $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$  denotes the set of vertices presenting the  $N$  agents in the scenario, and  $\mathcal{E} = \{e_1, e_2, \dots, e_M\}$  is the set of  $M$  hyperedges among such vertices. Each hyperedge can link more than two vertices to form the groupwise interaction. The set  $\mathcal{T} = \{\tau_1, \tau_2, \dots, \tau_H\}$  specifies total  $H$  multiscale hypergraphs.

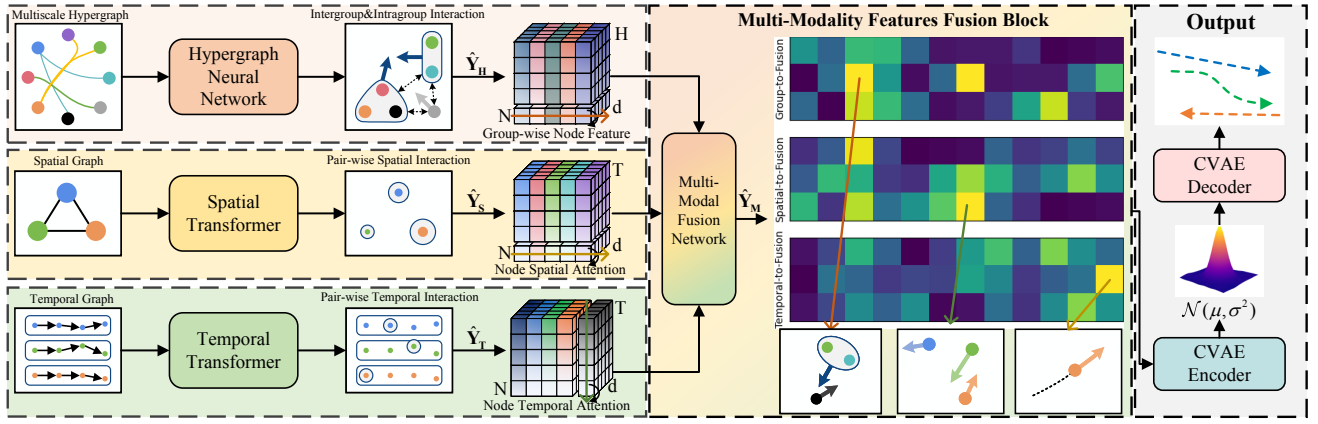


Fig. 2. Hyper-STTN neural network framework: (a) Spatial Transformer leverages a masked multi-head attention layer network along the time-dimension to represent spatial attention features; (b) Temporal Transformer is deployed to capture individual agent’s long-term temporal attention dependencies; and (c) Multi-Modal Transformer fuses heterogeneous spatial and temporal features via a multi-head cross-modal transformer block and a self-transformer block to abstract the uncertainty of multimodality crowd movements.

The weight matrix of each hyperedge is defined as  $\mathbf{W} = \text{diag}(w_{e_1}, w_{e_2}, \dots, w_{e_M}) \in \mathbb{R}^{M \times M}$ . Moreover, the incident matrix  $\mathbf{H} \in \mathbb{R}^{N \times M}$  of hypergraph  $\mathcal{G}$  is defined as follows:  $\mathbf{H}(v, e) = 1$ , if  $v \in e$ ;  $\mathbf{H}(v, e) = 0$ , if  $v \notin e$ . The diagonal vertex matrix  $\mathbf{M}_v \in \mathbb{R}^{N \times N}$  and diagonal edge matrix  $\mathbf{M}_e \in \mathbb{R}^{M \times M}$  are composed by the degree of vertex  $d(v) = \sum_{e \in \mathcal{E}} w_{(e)} \mathbf{H}(v, e)$  and the degree of edge  $d(e) = \sum_{v \in \mathcal{V}} \mathbf{H}(v, e)$ , separately.

Additionally, the adjacency matrix of hypergraph  $\mathcal{G}$  is given by  $\mathbf{A} \in \mathbb{R}^{N \times N}$ , where  $\mathbf{A} = \mathbf{H}\mathbf{W}\mathbf{H}^\top - \mathbf{M}_v$ . Notably, Hyper-STTN not only enables the modeling of high-order HHI features via multiscale hypergraphs, but also represents the pairwise dependencies with respect to the lowest-scale hypergraph  $\tau_1$  that degenerates into a standard graph in the scale-1. Let  $\mathbf{X} \in \mathbb{R}^{N \times T_i \times d}$  and  $\hat{\mathbf{X}} \in \mathbb{R}^{N \times T_o \times d}$  denote, separately, the past and future trajectory data of crowd, where  $T_i = 8$  and  $T_o = 12$  are the input and output sequence lengths with the dimension  $d = 2$ . The position of  $i$ -th agent at time  $t$  is expressed as  $\mathbf{x}_i^t = (x, y)$ . Eventually, Hyper-STTN describes the pedestrian trajectory distribution  $\mathcal{P}$  as:

$$\{\hat{\mathbf{X}}_1, \dots, \hat{\mathbf{X}}_N\} \sim \mathcal{P}(\{\mathbf{X}_1, \dots, \mathbf{X}_N\} | \mathcal{G}; \text{Hyper-STTN}) \quad (1)$$

where  $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$  and  $\{\hat{\mathbf{X}}_1, \dots, \hat{\mathbf{X}}_N\}$  are the complete sets of input and output sequences for all agents in the scenario. For the  $i$ -th agent, the input sequence and corresponding output sequence are  $\mathbf{X}_i = \{\mathbf{x}_i^{-T_i+1}, \dots, \mathbf{x}_i^0\}$ ,  $\hat{\mathbf{X}}_i = \{\mathbf{x}_i^1, \dots, \mathbf{x}_i^{T_o}\}$ .

### B. Hyper-STTN Architecture

Hyper-STTN captures both groupwise and pairwise interactions to reason out HHI for trajectory prediction, as illustrated in Fig. 2. Specifically, the spatial-temporal transformer network and the hypergraph convolution neural network are leveraged in parallel to encode pairwise HHI and groupwise HHI features, with respect to crowd hypergraph. Subsequently, the heterogeneous interaction embeddings are fused via the cross-attention from the multimodal transformer. Eventually, forecasting trajectories are then decoded by a CVAE decoder [25].

### C. Spatial-Temporal Transformer

Drawing inspiration from [1], [9], [16], [26], we designed a spatial-temporal transformer network to abstract the correlation of pairwise agents across spatial-temporal dimensions. Whereas spatial and temporal dependencies are represented by the relative importance in the attention maps (Fig. 6). For the spatial transformer, the input sequence  $\mathbf{X}$  is first processed by a positional encoding layer to incorporate sequence information [18]. The pre-processed data then passes sequentially through a layer normalization block (LN), a multihead mask attention layer, and a feed-forward network (FFN) (Fig. 4), with residual connections employed to stabilize training. At the core of the spatial-temporal transformer is the attention mechanism, which the vanilla multihead attention mechanism is defined as follows:

$$\begin{aligned} \text{Atten}(\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i) &= \text{softmax}\left(\frac{\mathbf{Q}^i(\mathbf{K}^i)^\top}{\sqrt{d_h}}\right)\mathbf{V}^i \\ \text{Multi}(\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i) &= f_{\text{fc}}(\text{head}_1, \dots, \text{head}_h) \\ \text{head}_{(i)} &= \text{Atten}_{(i)}(\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i) \end{aligned} \quad (2)$$

where  $f_{\text{fc}}$  is a fully connected layer,  $\text{head}_{(i)}$  denotes  $i$ -th head in the multihead attention with  $i \in [1, \dots, h]$ ,  $\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i$  are  $i$ -th query matrix, key matrix, and value matrix with the dimension  $d_h$ . Additionally, due to limitations in data collection, certain timesteps of individual trajectories may be unrecorded in the temporal transformer, while some pedestrians may appear intermittently in the spatial transformer. To address this variability, a masked attention mechanism [27] is employed, enabling the spatial-temporal transformer to robustly handle length-varying sequence data.

$$\text{MAtten}(\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i) = \text{softmax}\left(\mathcal{M}^i + \frac{\mathbf{Q}^i(\mathbf{K}^i)^\top}{\sqrt{d_h}}\right)\mathbf{V}^i \quad (3)$$

where attention mask matrix  $\mathcal{M}$  is defined to handle the issue of varying length data and to encode relative distance feature  $\omega(n, t) = f_{\text{fc}}(\text{dis}(\text{agent-pair}))$  into the attention as follows:

$$\mathcal{M}(n, t) = \begin{cases} -\infty & \mathbf{X}(n, t) = \text{none} \\ \omega(n, t) & \text{otherwise} \end{cases} \quad (4)$$

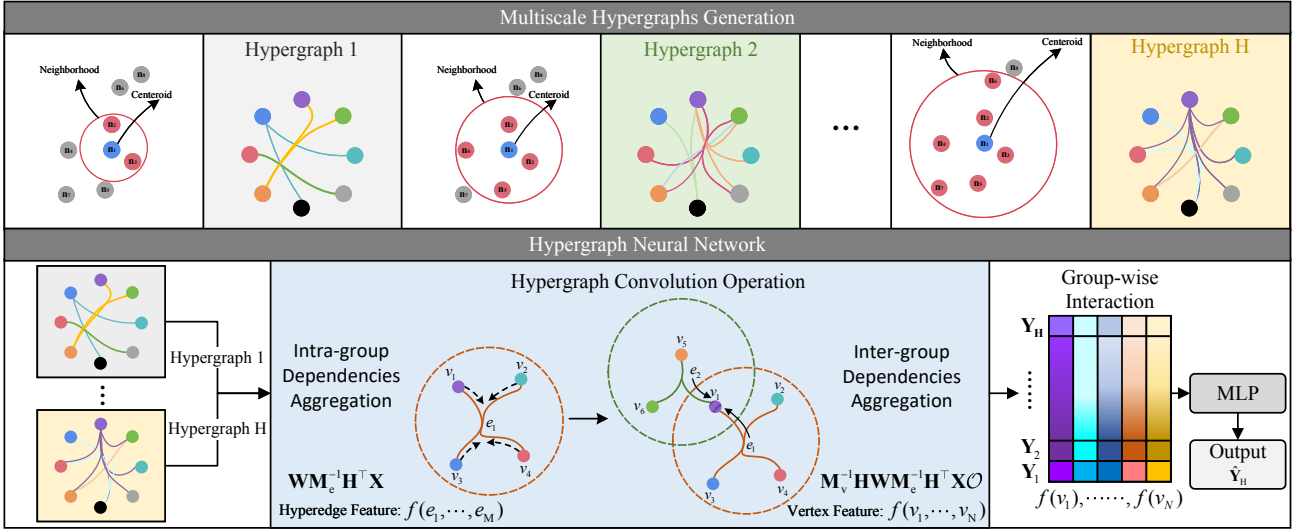


Fig. 3. Groupwise HHI Representation: i) We construct groupwise HHI with a set of multiscale hypergraphs, where each agent is queried in the feature space with varying 'k' in KNN to link multiscale hyperedges. ii) After constructing HHI hypergraphs, groupwise dependencies are captured by point-to-edge and edge-to-point phases with hypergraph spectral convolution operations.

The framework of temporal transformer and spatial transformer are illustrated as Fig. 4. The spatial transformer captures pairwise spatial interactions  $\hat{\mathbf{Y}}_S \in \mathbb{R}^{N \times T_i \times d}$  among all agents at each timestep, while the temporal transformer summarizes individual temporal interactions  $\hat{\mathbf{Y}}_T \in \mathbb{R}^{N \times T_i \times d}$  for agents, as  $\hat{\mathbf{Y}}_S; \hat{\mathbf{Y}}_T = \text{Trans}_{\text{spatial}}(\mathbf{X}); \text{Trans}_{\text{temporal}}(\mathbf{X})$ .

#### D. Hypergraph Neural Network

The crowd groups of Hyper-STTN are constructed with respect to neighborhoods' feature dimensional Mahalanobis distance [28], considering both interactive correlation distribution and physical spatial distance. In particular, the pedestrians' motion similarity is captured by the covariance matrix of Mahalanobis distance to cover the situations that a long-distance person presents a high potentiality to join the group. We construct crowd hypergraphs by vertex classifications of social group which are formulated as a spectral hypergraph  $k$ -way partitioning problem [29], [30]. We define a group at time  $t$  as a set of agents whose latent motion embeddings are mutually similar under a Mahalanobis metric computed over short-horizon kinematics (position, velocity, heading) and context crowd distribution features; proximity alone is not sufficient. Hyperedges connect an agent to its top- $k$  neighbors under this metric; varying  $k_i$  yields multiscale groups with respect to  $\tau_i$ . The normalized one cut spectral hypergraph partitioning task is defined as follows:

$$\begin{aligned} \arg \min_f \frac{1}{2} \sum_{e \in \mathcal{E}} \sum_{v_i, v_j \in \mathcal{V}} \frac{w(e)}{d(e)} \left[ \frac{f(v_i)}{\sqrt{d(v_i)}} - \frac{f(v_j)}{\sqrt{d(v_j)}} \right]^2 \\ = \arg \min_f f^T \Delta f \end{aligned} \quad (5)$$

where  $f(\cdot)$  is a classification function, and  $\Delta$  is the positive semi-definite hypergraph Laplacian.

Assuming the  $k$ -way hypergraph partition by a set of vertices subsets  $\{\mathcal{V}_1, \dots, \mathcal{V}_p\}$  from  $\mathbf{F} = [f_1 \dots f_p]$ . Subsequently, the spectral hypergraph  $k$ -way partitioning problem as a combinatorial optimization problem can be relaxed with

respect to minimizing  $\beta(\mathcal{V}_1, \dots, \mathcal{V}_p)$  as follows:

$$\arg \min_{\beta} \{\beta(\mathcal{V}_1, \dots, \mathcal{V}_p)\} = \left\{ \sum_{i=1}^p f_i^T \Delta f_i \right\} = \{ \text{trace}(\mathbf{F}^T \Delta \mathbf{F}) \} \quad (6)$$

To address above  $k$ -way hypergraph partition task, the  $K$ -nearest neighbor (KNN) [31] method is leveraged for multiscale hypergraphs generation. Hyper-STTN iteratively applies KNN to each vertex to identify its interactive neighborhoods based on the Mahalanobis feature distance. Particularly, the hypergraph at scale  $\tau_1 = 2$  degenerates to a standard graph, representing only pairwise interaction. In details, the low-level trajectory embeddings of each agent  $\{\mathbf{q}(x_i), \dots, \mathbf{q}(x_N)\}$ , where  $\mathbf{q}(x_i) \in \mathbb{R}^d$  are obtained via a fully connected layer from the input sequences  $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ . The similarity matrix  $\mathcal{S}$  is then filled by feature dimensional agent pairs' Mahalanobis distance [28], which incorporate interaction attributes and motion distributions to mitigate Euclidean distance based distortions arising from feature correlations and heterogeneous distributions. The Mahalanobis distance of  $(i, j)$ -th vertex pair  $Dis(i, j)$  is defined as follows:

$$Dis(i, j) = \sqrt{[\mathbf{q}(x_i) - \mathbf{q}(x_j)]^T \sum^{-1} [\mathbf{q}(x_i) - \mathbf{q}(x_j)]} \quad (7)$$

where  $\sum^{-1}$  is the covariance matrix of sample distribution.

The  $(i, j)$  element of similarity matrix  $\mathcal{S} \in \mathbb{R}^{N \times N}$  is defined as follows:

$$\mathcal{S}(i, j) = \exp\left[-\frac{Dis(i, j)^2}{\varrho^2}\right] \quad (8)$$

where  $\varrho$  is the mean of all vertex pairs' feature distances.

Inspired by [30], [32], we initialize multiscale hypergraphs by performing KNN searches for each vertex and its  $k$  nearest neighbors with respect to the similarity matrix  $\mathcal{S}$ , as shown in Fig. 3. In this representation, each vertex corresponds to a single agent, while hyperedges link multiple

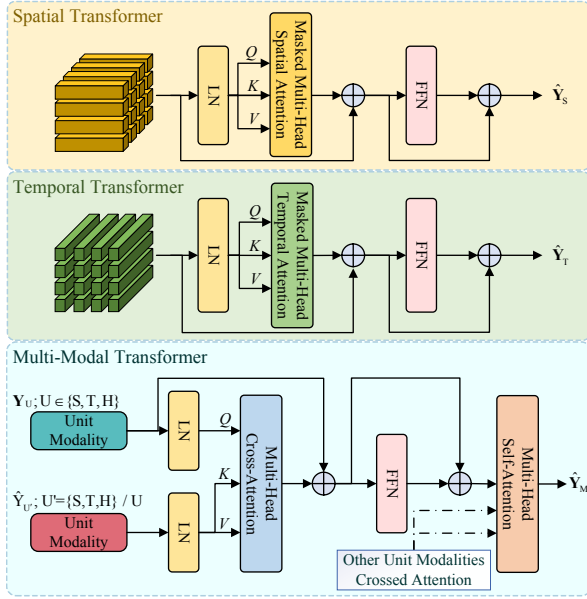


Fig. 4. Hybrid Spatial-Temporal Transformer Framework: Pedestrians' motion intents and dependencies are abstracted as spatial and temporal attention maps by multi-head attention mechanism of spatial-temporal transformer. Additionally, a multi-head cross attention mechanism is employed to align heterogeneous groupwise and pairwise features.

agents for the representation of groupwise interaction. The  $k$ -nearest neighbor selection is kept fixed per training episode. We refresh neighbor sets every  $E$  episodes to account for updated embedding structures during training.

After constructing multiscale crowd hypergraphs, a hypergraph convolution neural network [30] is employed to estimate the groupwise interactions. The random walk probability [29] of hypergraph aggregates the weighted dependencies of all sub-vertices into hyperedges features, which are then combined into vertex features, as shown in Fig. 3. Let a vertex  $v_i$  be stochastically selected with one of its hyperedges  $e_o$ , and the probability of hypergraph random walk  $\mathcal{O}_{(v_i, v_j)}$  from  $v_i$  to  $v_j$  on hyperedge  $e_o$  can be defined as follows:

$$\mathcal{O}_{(v_i, v_j)} = \sum_{e \in \mathcal{E}} \mathbf{W}_{e_o} \frac{\mathbf{H}(v_i, e_o) \mathbf{H}(v_j, e_o)}{\mathbf{M}_v(i, i) \mathbf{M}_e(o, o)} \quad (9)$$

The matrix normalized form of hypergraph random walk can be expressed as follows:

$$\mathcal{O} = \mathbf{M}_v^{-\frac{1}{2}} \mathbf{H} \mathbf{W} \mathbf{M}_e^{-1} \mathbf{H}^T \mathbf{M}_v^{-\frac{1}{2}} \quad (10)$$

Based on the definition of random walk probability, the hypergraph interaction features, representing groupwise interaction information are calculated by a hypergraph spectral convolution operation grounded, which aggregates the global dependencies between vertices and hyperedges. Following [30], [33], the spectral hypergraph convolution of an input signal  $\mathbf{x} \in \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  with a filter  $\mathbf{g}$  is defined as follows:

$$\mathbf{g} \otimes \mathbf{x} := \sum_{u=0}^U \theta_u \mathbf{T}_u(\tilde{\Delta}) \mathbf{x} \quad (11)$$

where  $\Delta = \mathbf{I} - \mathcal{O}$  is the regularized hypergraph Laplacian matrix. And  $\mathbf{T}_u(\tilde{\Delta})$  is the Chebyshev polynomial of order  $u$  with scaled Laplacian  $\tilde{\Delta} = 2\Delta/\lambda_{\max} - \mathbf{I}$ , where  $\lambda_{\max}$  is

the largest eigenvalue of  $\Delta$ . The parameter  $\theta$  is a weighted parameter, and  $U$  is the kernel size of the graph convolution.

Decomposing the hypergraph Laplacian matrix and approximating it by the first-order Chebyshev polynomial as a hypergraph convolution operation with single scale groupwise interaction dependence as output  $\mathbf{Y} \in \mathbb{R}^{N \times d}$ .

$$\mathbf{Y} = \mathbf{M}_v^{-1} \mathbf{H} \mathbf{W} \mathbf{M}_e^{-1} \mathbf{H}^T \mathbf{X} \mathcal{O} \quad (12)$$

Finally, Hyper-STTN aggregates all the multiscale hypergraphs features together to present crowded miscellaneous groupwise interactions as follows:

$$\hat{\mathbf{Y}}_H = f_{MLP}[\text{Concat}_{h=1}^H(\mathbf{Y})] \quad (13)$$

where  $f_{MLP}$  is a MLP neural network, and  $\hat{\mathbf{Y}}_H \in \mathbb{R}^{N \times H \times d}$  is the groupwise interaction feature.

#### E. Multi-Modal Fusion Network and CVAE Decoder

As shown in Fig. 4, the multi-modal transformer [20] is developed to fuse heterogeneous spatial-temporal features, using a cross-attention layer and self-attention layer. Wherein the multihead cross-attention mechanism  $\text{CMA}(\cdot)$  captures cross-modality features as follows:

$$\text{CMA}(\hat{\mathbf{Y}}_U) = \text{Multi}(\mathbf{Q}_U^{\text{head}_j}, \mathbf{K}_{U'}^{\text{head}_j}, \mathbf{V}_{U'}^{\text{head}_j}) \quad (14)$$

where  $U \in \{S, T, H\}$ , and  $\mathbf{Y}_U^{\text{head}_j}$  present the  $j$ -th head cross-modal attention of arbitrary unit modal with  $j \in \{1, \dots, h\}$ .

Subsequently, the groupwise HHI features  $\hat{\mathbf{Y}}_H$  and pairwise HHI dependencies  $\hat{\mathbf{Y}}_S, \hat{\mathbf{Y}}_T$  are aligned by a multi-head self-attention network as final crowd dynamics representation  $\hat{\mathbf{Y}}_M = \text{Trans}_{\text{multimodal}}(\hat{\mathbf{Y}}_H, \hat{\mathbf{Y}}_S, \hat{\mathbf{Y}}_T)$ , where neural network  $\text{Trans}_{\text{multimodal}}$  refers to [20].

Eventually, to estimate the stochasticity of human movements, a conditional variational auto encoder (CVAE)-based decoder [25] is employed to approximate maximum likelihood in potential distributions of motion uncertainty. The environmental dynamics feature  $\hat{\mathbf{Y}}_M$  and observed data  $\mathbf{X}$  are encoded to present Gaussian distribution  $\mathcal{N}(\mu, \sigma^2)$ . And the latent variable  $z$  is sampled by above Gaussian distribution  $z \sim \mathcal{N}(0, \sigma_T^2 \mathbf{I})$  in the testing process. Lastly, the maximum likelihood forecasting result is calculated by the decoder based on the concatenation of latent variable and observation embedding in the CVAE block as follows:

$$\begin{aligned} z &\sim \{\mathcal{N}(\mu, \sigma^2) = \text{Encoder}(\mathbf{X}, \hat{\mathbf{Y}}_M)\} \\ \hat{\mathbf{X}} &\sim \text{Decoder}(\hat{\mathbf{X}} | z, \text{Encoder}(\mathbf{X})) \end{aligned} \quad (15)$$

where  $\mu, \sigma$  are the mean and variance of the approximate distribution, and  $\text{Encoder}(\cdot)$  is Hyper-STTN backbone framework, and  $\text{Decoder}(\cdot)$  is a spectral temporal graph network from [34].

## IV. EXPERIMENTS AND RESULTS

### A. Experiments

1) *Dataset*: We conduct experiments on several widely used public pedestrian trajectory datasets collected from real-world scenarios. The training and validation data comprise

the ETH-UCY and NBA [7] datasets, which record pedestrians’ temporal location information in world coordinates. Consistent with prior works, our experiments adopt a leave-one-out cross-validation pre-process method, reserving only the last subsection of each scene for testing.

2) *Evaluation Metrics*: In the experiments, we evaluate prediction accuracy using the average displacement error (ADE) and final displacement error (FDE) metrics [6]. The ADE<sub>20</sub>/FDE<sub>20</sub> quantify performance with the Euclidean distance between predicted and ground-truth trajectories, either averaged over all timesteps or at the final timestep, which evaluate performance on the top 20 sampled predictions.

### B. Comparison Configuration

1) *Baselines*: We compared our model with several existing SOTA algorithms. The algorithms included in the comparison are Social-Attention [19], Social-GAN [8], Social-STGCNN [21], Trajectron++ [12], STAR [9], PECNet [35], Social-Implicit [22], GroupNet [7], EqMotion [36], LED [16], and IMLE [17], as shown in Table I and Table II.

2) *Ablation Models*: We design two ablation models to evaluate the contributions of distinct components for overall performance. The ablation model STTN, infers pedestrian interactions using only the hybrid spatial-temporal transformer, excluding the hypergraph network block from Hyper-STTN. The second mode, HGNN, estimates crowd motion dependencies solely through hypergraph convolution operations without transformers. In both ablation models, the CVAE structure is retained as the decoder. All training configurations, including datasets and initialization parameters, are kept identical to those used for Hyper-STTN.

3) *Training Details*: We use two NVIDIA RTX-4090 GPUs to train the model via the Adam optimizer. To improve robustness, Gaussian noise is injected during training.

In our training procedure, we utilize the distance loss function  $\mathcal{L}_{dis}$ , angle loss function  $\mathcal{L}_{ang}$ , and encoder loss function  $\mathcal{L}_{enc}$  to update the network as follows:

$$\begin{aligned} \mathcal{L}_{dis} &= \kappa_1 \|\hat{\mathbf{X}} - \hat{\tilde{\mathbf{X}}}\|_2 + \kappa_2 f_{\text{KL}}(\mathcal{N}(\mu, \sigma^2) \|\mathcal{N}(0, \sigma_{\mathbf{T}_1}^2 \mathbf{I}) \\ \mathcal{L}_{ang} &= \kappa_3 \sum_{t \in T} \sum_{j \in T, j > t} \|\angle(\hat{\mathbf{X}}_t, \hat{\mathbf{X}}_j) - \angle(\mathbf{X}_t, \mathbf{X}_j)\|_2 \quad (16) \\ \mathcal{L}_{enc} &= \kappa_4 \|\text{Encoder}(\mathbf{X}) - \mathbf{X}\|_2 \end{aligned}$$

where  $\kappa$  is the weight of each loss term, and the Kullback-Leibler (KL) divergence term  $f_{\text{KL}}(\cdot \|\cdot)$  is used to update the encoder and decoder in the CVAE block, and  $\hat{\tilde{\mathbf{X}}}$  denotes the groundtruth of dataset. The  $\angle(\cdot)$  function computes the angle between two vectors representing points.

Finally, the overall loss function for Hyper-STTN  $\mathcal{L}$  is constructed as the sum of the above equations to minimize the total loss as follows:  $\mathcal{L} = \mathcal{L}_{dis} + \mathcal{L}_{ang} + \mathcal{L}_{enc}$ .

### C. Quantitative Results

We conducted the trajectory prediction task with respect to stochastic conditions, performing Hyper-STTN’s SOTA quantitative results in ETH-UCY and NBA datasets, as shown in Table I and Table II. Notably, due to the datasets’ size and the rapid advancement of expressive models, the performance on the ETH-UCY datasets has plateaued in

recent years that is approaching the inherent errors of machine learning approaches on such datasets. However, Hyper-STTN still exhibits comparable performance against existing advanced models both on ADE<sub>20</sub> and FDE<sub>20</sub> metrics. For instance, in the ETH-UCY dataset, Hyper-STTN not only enhances the 12.5% ADE<sub>20</sub> and 1.7% FDE<sub>20</sub> from EqMotion’s [36] 0.40/0.58 to 0.35/0.57 on ETH dataset, but also improves the average 14.7%/6.5%/3.4%/10.6% ADE<sub>20</sub> and 22.9%/14.7%/7.6%/10.1% FDE<sub>20</sub> performance from Groupnet’s [7] on NBA dataset. Highlighting the effectiveness of Hyper-STTN in crowd latent interaction inference and trajectory forecasting.

Deep learning-based trajectory prediction frameworks have catalyzed a shift toward modeling latent social interactions across spatial and temporal dimensions for pedestrian motion forecasting. Methods built upon recurrent architectures [6], attention mechanisms [19], and, more recently, transformer models [9], [10] have demonstrated progressively stronger predictive performance by enabling richer high-level feature representations. Following this trend, Hyper-STTN adopts transformer backbones to explicitly model fundamental social interactions and temporal dependencies in dense crowds. Its ablation variant, STTN, which is purely transformer-driven (akin to STAR [9], [10]), isolates the contribution of these components. Experimental results indicate that the observed performance gains arise from the introduction of masked-attention and cross-modal attention modules. These modules mitigate agent-flicker effects relative to vanilla self-attention under fixed, predefined transformer data scales, while simultaneously improving the adaptability of spatial-temporal fusion.

Despite the remarkable effectiveness of transformers in abstracting pairwise interactions, their limited consideration of group-level dynamics constrains further performance improvements, particularly in dense scenes where individual motions are strongly influenced by the movements of surrounding groups. For example, GroupNet [7] achieves superior performance (0.26/0.49) compared to STAR [9] (0.31/0.62) on the UNIV dataset, which contains more densely crowded scenarios. To address this gap, we propose to integrate hypergraph construction for inferring group-wise interactions with transformer architectures for pairwise feature modeling, thereby capturing richer strategies for crowded movement prediction. In addition, the ablation model HGNN exhibits performance comparable to GroupNet across several datasets, underscoring the effectiveness of the hypergraph component.

### D. Discussion

Fig. 5 illustrates the trajectory predictions of Hyper-STTN under both sparse and dense scenarios, alongside two ablation models. In all visualizations, blue dots indicate the observation inputs, while green dots denote the pedestrian ground truth trajectories, allowing clear differentiation from the predicted trajectories. As shown, Hyper-STTN consistently produces highly accurate multi step location forecasts and infers more realistic motion orientations. Fig. 6

TABLE I

THE CROWDED TRAJECTORY FORECASTING BEST-OF-20 STOCHASTIC SAMPLED RESULTS OF MINADE<sub>20</sub> / MINFDE<sub>20</sub> ON ETH-UCY DATASET.

Stochastic ADE <sub>20</sub> / FDE <sub>20</sub>	Social- Attention [19] ICRA18	Social- GAN [8] CVPR18	Trajectron ++ [12] ECCV20	STAR [9] ECCV20	Social- Implicit [22] ECCV22	GroupNet [7] CVPR22	EqMotion [36] CVPR23	LED [16] CVPR23	IMLE [17] CVPR25	STTN (ablation)	HGNN (ablation)	Hyper-STTN (ours)
ETH	1.39/2.39	0.87/1.62	0.61/1.02	0.36/0.65	0.61/1.08	0.46/0.73	0.40/0.61	<b>0.39/0.58</b>	0.40/0.58	0.52/0.75	0.42/0.73	<b>0.35/0.57</b>
HOTEL	2.51/2.91	0.67/1.37	0.19/0.28	0.17/0.36	0.33/0.63	0.15/0.25	0.12/0.18	<b>0.11/0.17</b>	0.12/0.18	0.21/0.25	0.21/0.32	0.14/0.18
UNIV	1.25/2.54	0.76/1.52	0.30/0.54	0.31/0.62	0.52/1.11	0.26/0.49	<b>0.23/0.43</b>	0.26/0.44	<b>0.23/0.39</b>	0.25/0.43	0.30/0.53	<b>0.23/0.41</b>
ZARA1	1.01/2.17	0.35/0.68	0.24/0.42	0.26/0.55	0.32/0.66	0.21/0.39	0.18/0.32	0.18/0.26	<b>0.16/0.26</b>	0.31/0.39	0.28/0.53	0.18/0.28
ZARA2	0.88/1.75	0.42/0.84	0.18/0.32	0.22/0.46	0.43/0.85	0.17/0.33	<b>0.13/0.23</b>	<b>0.13/0.22</b>	<b>0.13/0.22</b>	0.25/0.33	0.23/0.28	0.15/0.22
<b>AVG</b>	1.41/2.35	0.61/1.21	0.30/0.51	0.26/0.53	0.33/0.67	0.25/0.44	<b>0.21/0.35</b>	<b>0.21/0.33</b>	<b>0.21/0.33</b>	0.31/0.43	0.29/0.48	<b>0.21/0.33</b>

TABLE II

THE CROWDED TRAJECTORY FORECASTING BEST-OF-20 STOCHASTIC SAMPLED RESULTS OF MINADE<sub>20</sub> / MINFDE<sub>20</sub> ON NBA DATASET.

Stochastic ADE <sub>20</sub> / FDE <sub>20</sub>	Social-GAN [8] CVPR18	Social-STGCNN [21] CVPR20	Trajectron++ [12] ECCV20	PECNet [35] ECCV20	GroupNet [7] CVPR22	STTN (ablation)	HGNN (ablation)	Hyper-STTN (ours)
NBA-1s	0.41/0.62	0.34/0.48	<b>0.30/0.38</b>	0.35/0.58	0.34/0.48	0.39/0.58	0.37/0.59	<b>0.30/0.37</b>
NBA-2s	0.81/1.32	0.71/0.94	<b>0.59/0.82</b>	0.68/1.23	0.62/0.95	0.72/1.05	0.64/0.95	<b>0.58/0.81</b>
NBA-3s	1.19/1.94	1.09/1.77	<b>0.85/1.24</b>	1.01/1.76	0.87/1.31	1.07/1.51	0.98/1.44	<b>0.84/1.21</b>
NBA-4s	1.59/2.41	1.53/2.26	1.15/1.57	1.31/1.79	<b>1.13/1.69</b>	1.24/1.99	1.18/1.89	<b>1.01/1.52</b>

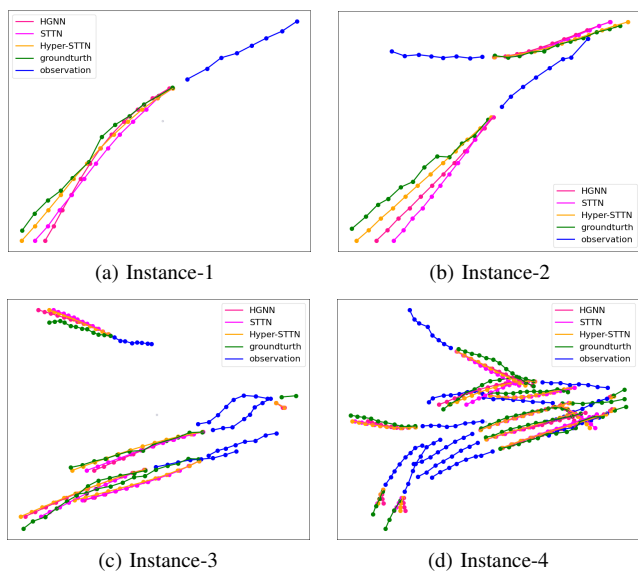


Fig. 5. Comparison of Trajectories Visualizations: The trajectories visualized for Hyper-STTN and other algorithms tested on the same test case.

further highlights how the spatial transformer module enables Hyper-STTN to capture relative spatial interactions among agent pairs, whereas the temporal transformer encodes individual motion attributes. Moreover, in dense scenarios with a growing number of pedestrian groups, inter-group and intra-group crowd correlations exert a stronger influence than in sparse conditions. Consequently, explicitly inferring group-wise interactions yields more accurate long-term forecasts, as evidenced by the trajectories and ADE/FDE metrics achieved by Hyper-STTN and HGNN.

1) *Effect of Hypergraph Neural Network:* Hyper-STTN consists of two main components: a hypergraph neural network and a spatial-temporal transformer with a fusion module. As shown in Table I, the HGNN branch achieves performance comparable to GroupNet on several datasets, indicating that hypergraph-based interaction modeling is effective, particularly in crowded scenes. This advantage is

further reflected in Fig. 5, where the STTN ablation exhibits noticeably larger forecasting errors than HGNN in predicted agent trajectories. These results suggest that the performance gain comes from explicitly modeling higher order groupwise interactions, which are not captured by STTN alone but are effectively introduced by HGNN.

2) *Effect of Spatial-Temporal Transformer:* Compared with previous transformer-based methods such as STAR, Hyper-STTN improves the average performance from 0.26/0.53 to 0.21/0.33. It also consistently outperforms its STTN and HGNN ablations, reducing the average error from 0.31/0.43 and 0.29/0.48 to 0.21/0.33, respectively. These gains demonstrate the effectiveness of combining transformer-based multi-modal fusion with hypergraph-based group interaction modeling. Moreover, as shown in Fig. 5, Hyper-STTN generates trajectories that are closer to the ground truth than both STTN and HGNN across distinct interaction scenarios.

## V. CONCLUSION

In this paper, we proposed Hyper-STTN, a hypergraph-based hybrid spatial-temporal transformer for trajectory prediction tasks. Hyper-STTN jointly captures groupwise and pairwise interactions in crowd dynamics through hypergraph and transformer networks. Extensive experiments on the ETH-UCY and NBA datasets demonstrate that Hyper-STTN outperforms existing state-of-the-art prediction algorithms. Future research will focus on enhancing its capability to handle more complex scenarios and on scaling the model for real-time deployments.

## REFERENCES

- [1] W. Wang, R. Wang, L. Mao, and B.-C. Min, "Navistar: Socially aware robot navigation with hybrid spatio-temporal graph transformer and preference learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 11 348–11 355.
- [2] W. Wang, L. Mao, R. Wang, and B.-C. Min, "Multi-robot cooperative socially-aware navigation using multi-agent reinforcement learning," *2024 IEEE International Conference on Robotics and Automation*.

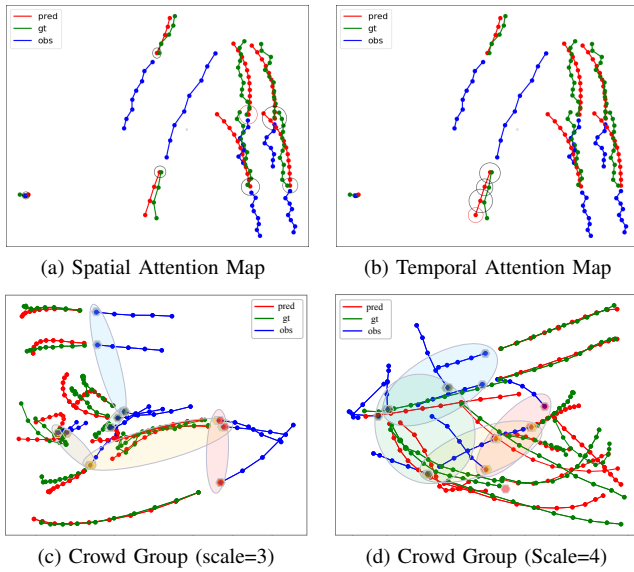


Fig. 6. Hyper-STTN pairwise and groupwise interactions visualization: Subfigures (a) and (b) show pairwise attention maps. Subfigures (c) and (d) illustrate groupwise multi-scale hypergraphs.

[3] W. Wang, A. Bera, and B.-C. Min, "Hypergraph-based coordinated task allocation and socially-aware navigation for multi-robot systems," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 16 225–16 232.

[4] W. Wang, C. Yu, Y. Wang, and B.-C. Min, "Human-robot cooperative distribution coupling for hamiltonian-constrained social navigation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 10 808–10 815.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[6] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[7] C. Xu, M. Li, Z. Ni, Y. Zhang, and S. Chen, "Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6498–6507.

[8] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2255–2264.

[9] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020.

[10] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9813–9823.

[11] J. Li, F. Yang, M. Tomizuka, and C. Choi, "Evolvegraph: multi-agent trajectory prediction with dynamic relational reasoning," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[12] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 683–700.

[13] S. Lee, J. Lee, Y. Yu, T. Kim, and K. Lee, "Mart: Multiscale relational transformer networks for multi-agent trajectory prediction," in *European Conference on Computer Vision*. Springer, 2024, pp. 89–107.

[14] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *CVPR 2011*. IEEE, 2011.

[15] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 283–298, 2007.

[16] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, "Leapfrog diffusion model for stochastic trajectory prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5517–5526.

[17] Y. Fu, Q. Yan, L. Wang, K. Li, and R. Liao, "Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 282–17 293.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *2018 IEEE international Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4601–4607.

[20] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Association for Computational Linguistics*, vol. 2019. NIH Public Access, 2019, p. 6558.

[21] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[22] A. Mohamed, D. Zhu, W. Vu, M. Elhoseiny, and C. Claudel, "Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 463–479.

[23] J. Grilli, G. Barabás, M. J. Michalska-Smith, and S. Allesina, "Higher-order interactions stabilize dynamics in competitive network models," *Nature*, vol. 548, no. 7666, pp. 210–213, 2017.

[24] C. Ke and J. Honorio, "Exact inference in high-order structured prediction," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[25] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[26] C. Chen, Y. Liu, L. Chen, and C. Zhang, "Bidirectional spatial-temporal adaptive transformer for urban traffic flow forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[27] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[28] F. Wang and J. Sun, "Survey on distance metric learning and dimensionality reduction in data mining," *Data mining and knowledge discovery*, vol. 29, no. 2, pp. 534–564, 2015.

[29] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," *Advances in neural information processing systems*, vol. 19, 2006.

[30] Y. Gao, Y. Feng, S. Ji, and R. Ji, "HGNN+: General hypergraph neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3181–3199, 2023.

[31] P. Cunningham and S. J. Delany, "k-nearest neighbour classifiers-a tutorial," *ACM computing surveys*, vol. 54, no. 6, pp. 1–25, 2021.

[32] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3558–3565.

[33] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[34] D. Cao, Y. Wang, J. Duan, C. Zhang, X. Zhu, C. Huang, Y. Tong, B. Xu, J. Bai, J. Tong *et al.*, "Spectral temporal graph neural network for multivariate time-series forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17 766–17 778, 2020.

[35] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 759–776.

[36] C. Xu, R. T. Tan, Y. Tan, S. Chen, Y. G. Wang, X. Wang, and Y. Wang, "Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1410–1420.