Spatiotemporal-Untrammelled Mixture of Experts for Multi-Person Motion Prediction

Zheng Yin[1*], Chengjian Li[1*], Xiangbo Shu[1†], Meiqi Cao[1], Rui Yan[1], Jinhui Tang[2]

[1]Nanjing University of Science and Technology
[2]Nanjing Forestry University
alanyz@njust.edu.cn, lichengjian@njust.edu.cn, shuxb@njust.edu.cn

Abstract

Comprehensively and flexibly capturing the complex spatiotemporal dependencies of human motion is critical for multi-person motion prediction. Existing methods grapple with two primary limitations: i) Inflexible spatiotemporal representation due to reliance on positional encodings for capturing spatiotemporal information. ii) High computational costs stemming from the quadratic time complexity of conventional attention mechanisms. To overcome these limitations, we propose the Spatiotemporal-Untrammelled Mixture of Experts (ST-MoE), which flexibly explores complex spatio-temporal dependencies in human motion and significantly reduces computational cost. To adaptively mine complex spatio-temporal patterns from human motion, our model incorporates four distinct types of spatiotemporal experts, each specializing in capturing different spatial or temporal dependencies. To reduce the potential computational overhead while integrating multiple experts, we introduce bidirectional spatiotemporal Mamba as experts, each sharing bidirectional temporal and spatial Mamba in distinct combinations to achieve model efficiency and parameter economy. Extensive experiments on four multi-person benchmark datasets demonstrate that our approach not only outperforms state-of-art in accuracy but also reduces model parameter by 41.38% and achieves a 3.6× speedup in training.

Introduction

Human motion prediction aims to forecast future human movements from observed motion sequences. The field carries substantial importance for applications including human-robot interaction (Gui et al. 2018; Zhuo et al. 2019; Jiang et al. 2024), autonomous driving (Tang et al. 2023; Fang et al. 2023; Cao et al. 2025; Jiang et al. 2025), and surveillance systems (Vu et al. 2020; Qu et al. 2025a,c,b; Xing et al. 2025). By analyzing historical human motion, robotic systems can infer human intentions, enabling more effective collaboration. Although traditional approaches focus on single individuals (Butepage et al. 2017; Mao et al. 2019; Cui, Sun, and Yang 2020; Shu et al. 2021), multi-person motion prediction (MPMP) holds greater practical relevance, as real-world scenarios typically involve multiple individuals. Recent multi-person motion prediction methods leverage Transformer to learn spatial relationships between joints via self-attention (Xu et al. 2023), neglecting the importance of capturing spatiotemporal dependencies in multi-person motion.

To this end, MRT (Guo et al. 2022) employs temporal positional encoding to capture changes in human posture over time, and uses spatial positional encoding in the global encoder to enhance spatial correlations (Fig. 1(a)). However, this spatiotemporal position encoding employs a fixed pattern and lacks flexibility, leading to constrained prediction performance (Chu et al. 2021). TBIFormer (Peng, Mao, and Wu 2023) partitions human body parts and incorporates trajectory-aware relative position encoding, thereby enhancing the model's spatial perception (Fig. 1(b)). However, due to the quadratic computational complexity of the attention mechanism, the body part concatenation operation increases sequence length, significantly elevating computational costs. On the other hand, IAFormer (Xiao et al. 2024) utilizes self-attention mechanisms to explore spatiotemporal features within interactive information, significantly improving performance (Fig. 1(c)). Yet like TBIFormer, this method remains constrained by the efficiency bottleneck.

Despite the success of Transformer-based methods, these approaches continue to exhibit two key limitations: (a) Constrained model architectures fail to flexibly and adequately explore the complex spatio-temporal dependencies inherent in human motion. (b) Excessive dependence on computationally intensive attention mechanisms results in suboptimal efficiency. One question is naturally motivated to ask: *Can we devise a new efficient paradigm for multi-person motion that flexibly and comprehensively captures spatiotemporal dependencies in human movement?*

To answer this question, we propose a lightweight framework **ST-MoE** for efficient and flexible modeling of complex spatiotemporal dependencies in multi-person motion prediction (Fig. 1(d)). Specifically, inspired by the dynamic activation mechanism of subnetworks in Mixture-of-Experts (MoE), we design four heterogeneous experts that respectively model distinct spatiotemporal patterns. Unlike conventional approaches relying on attention mechanisms or explicit positional encodings, ST-MoE adaptively activates optimal experts based on varied spatiotemporal features, thereby enhancing flexible modeling for diverse

*These authors contributed equally.
†Corresponding author.

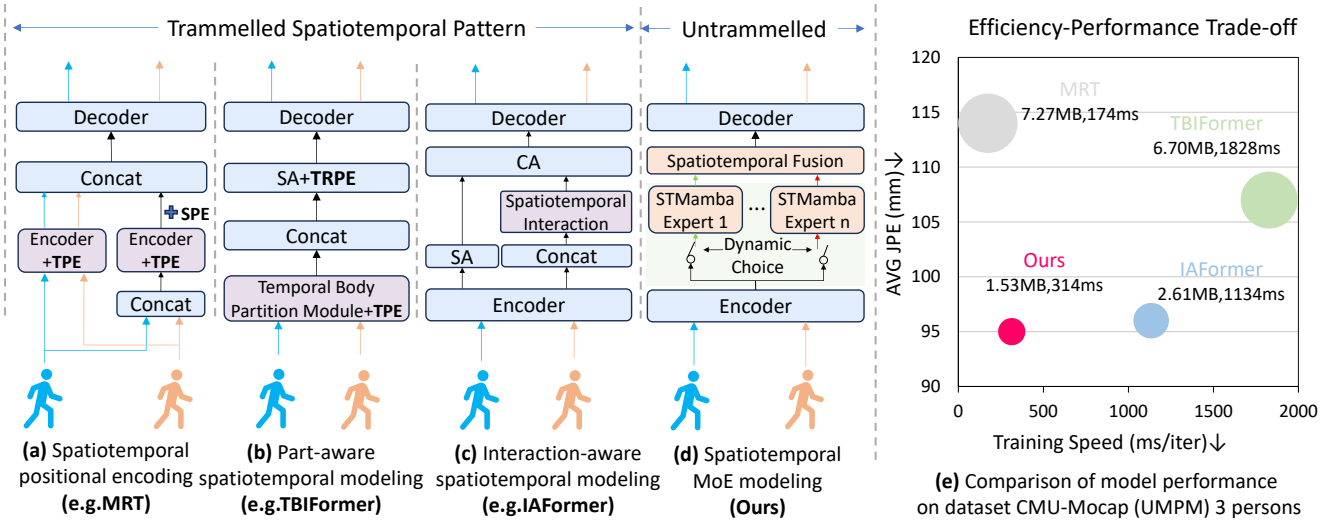

Figure 1: Insight of our work and performance of models on dataset CMU-Mocap(UMPM). (a)-(c): Limitations of prior methods using attention mechanism/spatio-temporal positional encodings, which solely capture trammelled spatiotemporal patterns. **TPE/SPE** denote temporal/spatial positional encoding. **TRPE** denotes trajectory-aware relative position encoding. **SA/CA** denote Self-Attention and Cross-Attention, respectively. (d): Our MoE-based framework with dynamic expert selection for adaptive spatiotemporal modeling. (e): Our model achieves an optimal efficiency-performance trade-off.

motion. To improve modeling efficiency, we replace the traditional high-cost spatiotemporal attention mechanism with the Mamba, which has linear complexity. Specifically, ST-MoE flexibly configures combinations of bidirectional spatial and temporal Mamba for each expert to capture diverse spatiotemporal interaction patterns, and facilitate cross-dimensional feature fusion and parameter compression by sharing spatiotemporal Mamba. Under the synergy of MoE's dynamic routing mechanism and Mamba's efficient modeling, ST-MoE achieves a significant reduction in computational cost while maintaining prediction accuracy. Results on CMU-Mocap (UMPM) (Van der Aa et al. 2011) confirm our approach's superiority over the SOTA IAFormer (Xiao et al. 2024) in prediction accuracy, while simultaneously accelerating training by 3.6× and reducing model parameters by 41.38% (Fig. 1(e)).

In summary, our main contributions are threefold:

- **First lightweight MoE for MPMP**. We propose Spatiotemporal-Untrammelled Mixture of Experts (ST-MoE), the first framework that integrates spatiotemporal Mamba with dynamic expert routing for multi-person motion prediction. This dual mechanism fundamentally resolves the efficiency-accuracy trade-off.
- **Distinct Spatiotemporal Experts.** We introduce four different spatiotemporal bidirectional mamba experts to flexibly model human motion dynamics, resolving the trammelled spatiotemporal dependency capture in existing multi-person motion prediction methods.
- **Efficient and Low-Parameter.** Extensive experiments on multi-person motion prediction benchmarks demonstrate that our framework attains state-of-the-art performance, while compressing the model size by 41.38% and achieving 3.6× faster training speed.

Related Work

Human Motion Prediction. Early research on human motion prediction focused on predicting single-person motion, utilizing RNNs to solve this problem (Martinez, Black, and Romero 2017). However, due to the inherent recursive nature of RNNs leading to problems such as error accumulation, several methods employ Graph Convolutional Networks (GCNs) to explore spatiotemporal dependencies (Ma et al. 2022; Zhong et al. 2022). Recent research has extended to the more challenging multi-person motion prediction (Wang et al. 2021; Xiao et al. 2024; Peng, Mao, and Wu 2023). MRT (Wang et al. 2021) employs spatiotemporal positional encoding to capture constrained spatiotemporal dependencies and IAFormer (Xiao et al. 2024) learns spatiotemporal interactions via attention mechanisms. These methods exhibit high computational complexity and trammelled modeling of complex spatiotemporal dependencies in motion. Conversely, our model extracts diverse spatiotemporal features more efficiently and flexibly.

Mixture of Experts. Mixture of Experts (MoE) was first proposed in (Jacobs et al. 1991) and was later integrated into the Transformer (Lepikhin et al. 2020; Fedus, Zoph, and Shazeer 2022; Du et al. 2022). This paradigm replaces Feed Forward Network (FFN) layers with MoE layers, selectively activating only relevant experts per input to achieve adaptive processing, significantly reduce computational cost, while still benefiting from a vast pool of specialized knowledge. More recently, MoE has also seen widespread applications in industrial-scale large language models (LLMs) (Liu et al. 2024; Team 2024; Yang et al. 2025). Unlike previous approaches where each expert employs an identical FFN structure, we utilize distinct spatiotemporal Mamba blocks for

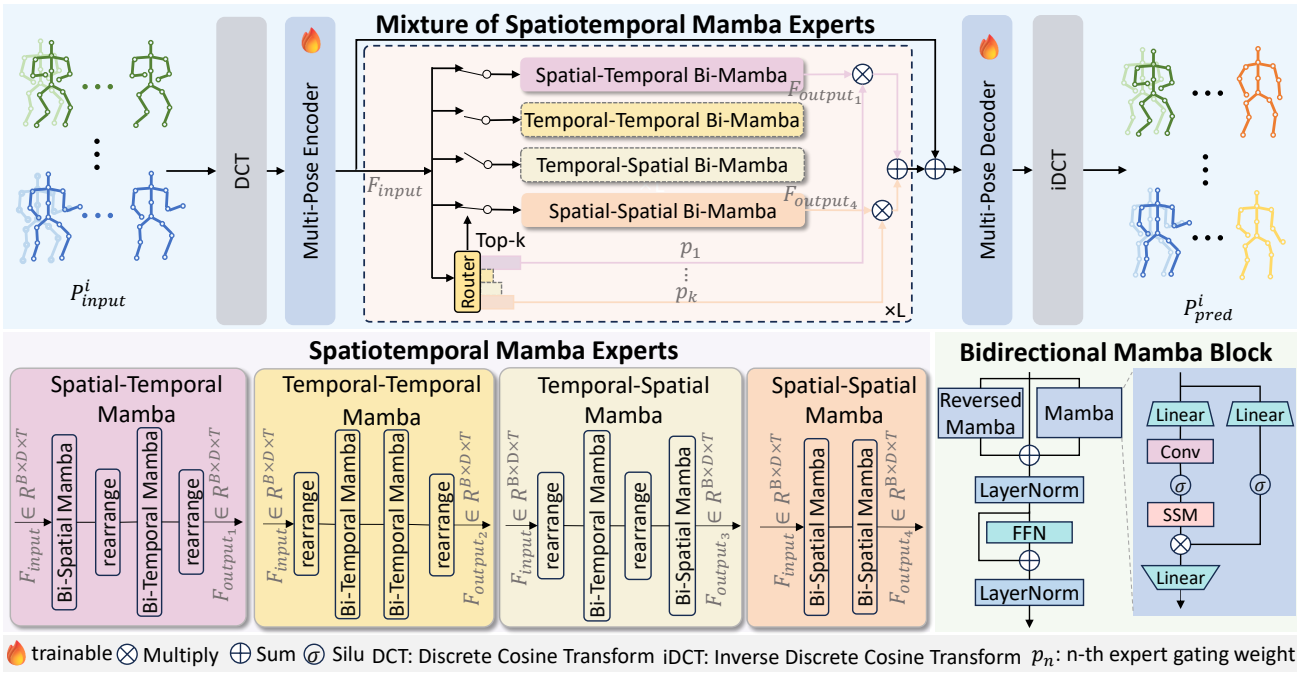

Figure 2: Overview of Spatiotemporal-Untrammelled Mixture of Experts (ST-MoE). The padded input motion sequence is first encoded and then adaptively routed by a gating router to distinct spatiotemporal experts. ST-MoE aggregates the outputs of the selected experts and decodes the merged features to predict the future motion sequences. Each expert in the expert pool consists of pairwise combinations of bidirectional temporal Mamba and bidirectional spatial Mamba, with shared parameters across all experts for both temporal and spatial Mamba components. This design enables comprehensive learning of spatiotemporal dependencies in human motion while maintaining lightweight.

individual experts to model human motion more flexibly.

State Space Model. State Space Models (SSMs) have achieved significant success in numerous sequence modeling tasks due to their ability to capture long-range dependencies (Gu and Dao 2023; Fu et al. 2022; Gu, Goel, and Ré 2021; Li et al. 2024). Mamba (Gu and Dao 2023) proposes a selective scanning mechanism that dynamically filters irrelevant inputs while maintaining linear inference time and efficient training, establishing itself as a strong Transformer alternative. MoE-Mamba (Pióro et al. 2024) combines MoE and Mamba to scale SSMs, matching Transformer performance while retaining computational gains. However, the exploration of Mamba's potential in multi-person motion prediction remains untapped. In this paper, we employ a MoE framework integrating Mamba to capture spatiotemporal dependencies in human motion, significantly boosting model efficiency. Unlike MoE-Mamba (Pióro et al. 2024), our lighter framework integrates Mamba blocks within each expert, while MoE-Mamba alternates Mamba blocks with FFN-based MoE layers.

Method

Overview

The overview of the proposed ST-MoE is illustrated in Fig. 2. First, the motion sequence is mapped from pose space to feature space via Discrete Cosine Transform (DCT) and a Multi-Pose Encoder, resulting in motion feature $\boldsymbol{F}_{\text{input}}$. Subsequently, $\boldsymbol{F}_{\text{input}}$ is routed to the Mixture of Spatiotemporal Mamba Experts (MoSTME), where each expert consists of a paired combination of bidirectional spatial and temporal Mamba modules in four possible configurations. The router activates specific spatiotemporal experts to flexibly explore the spatiotemporal dependencies in human motion. The outputs $\boldsymbol{F}_{\text{output}_i}$ from all activated experts are then aggregated, where i denotes the i-th activated expert. Finally, the aggregated feature is decoded back from the feature space to the pose space via a Multi-Pose Decoder and Inverse DCT (iDCT), yielding the predicted future motion.

Problem Statement

Multi-person motion prediction aims to forecast the future joint positions of multiple individuals based on their historical motion. Mathematically, we define the historical motion sequences of the i-th person $\boldsymbol{P}^i_{1:t} = \{\boldsymbol{P}^i_1, \boldsymbol{P}^i_2, \dots, \boldsymbol{P}^i_t\} \in \mathbb{R}^{D\times t}$ and the future motion $\boldsymbol{P}^i_{t+1:T} \in \mathbb{R}^{D\times(T-t)}$. The timestamp t indicates the last observed frame, while T corresponds to the final predicted frame. The pose dimension D equals to $J \times 3$, where J represents the number of joints. To facilitate the model's learning of future poses, we follow prior work (Mao et al. 2019; Dang et al. 2021; Xiao et al. 2024) by replicating the last observed frame $\boldsymbol{P}^i_t$ for $T-t$ times and appending to $\boldsymbol{P}^i_{1:t}$, yielding the input sequence $\boldsymbol{P}^i_{\text{input}} = \{\boldsymbol{P}^i_1, \dots, \boldsymbol{P}^i_t, \dots, \boldsymbol{P}^i_t\} \in \mathbb{R}^{D\times T}$. Our aim is to pre-

dict future motion $\boldsymbol{P}^i_{\text{pred}} = \{\boldsymbol{P}^i_1, \ldots, \boldsymbol{P}^i_t, \hat{\boldsymbol{P}}^i_{t+1}, \ldots, \hat{\boldsymbol{P}}^i_T\} \in \mathbb{R}^{D\times T}$ from padded motion sequences $\boldsymbol{P}^i_{\text{input}}$.

Multi-Pose Encoder and Decoder

GCNs have demonstrated superior capability in capturing spatiotemporal dependencies among joints in human motion, and are therefore widely adopted in motion prediction tasks (Mao et al. 2019). For example, IAFormer (Xiao et al. 2024) employs GCN-based encoder and decoder modules to extract features from multi-person motion. Inspired by this design, we adopt the multi-pose encoder/decoder architecture from IAFormer to enhance the model's capacity in representing complex pose dynamics. In addition, to further improve representation compactness and effectively capture the smooth dynamics of human motion, we apply the DCT before the encoder and iDCT after the decoder (Mao et al. 2019; Wang et al. 2021). The encoding process is as follows:

$$\boldsymbol{F}^i_{\text{input}} = \texttt{ME}(\texttt{DCT}(\boldsymbol{P}^i_{\text{input}})), \tag{1}$$

where $\texttt{ME}(\cdot)$ represents Multi-Pose Encoder and $\texttt{DCT}(\cdot)$ represents Discrete Cosine Transform. The $\boldsymbol{F}^i_{\text{input}}$ denotes the motion feature in the feature space of the i-th person.

Mixture of Spatiotemporal Mamba Experts

As illustrated in Fig. 1(a)-(c), existing approaches often employ fixed spatio-temporal positional encodings or spatio-temporal attention mechanisms to capture spatio-temporal features, resulting in trammelled patterns. Inspired by the flexible modeling capability of MoE (Shazeer et al. 2017; Yun et al. 2024), we introduce the Mixture of Spatiotemporal Mamba Experts to address the challenges mentioned above. First, the encoded features $\boldsymbol{F}_{\text{input}} \in \mathbb{R}^{B\times D\times T}$ are input into the Mamba Expert Pool and the router, where B denotes batch size. Furthermore, the router implements sparse activation by dynamically selecting only the top-k experts through an MLP-based gating network. It then calculates weights specifically for these activated experts and aggregates their outputs via weighted summation. This entire process can be expressed by the following formula:

$$\begin{aligned} \boldsymbol{E}_{\text{output}} &= \sum_{e=1}^{N} \boldsymbol{f}_e(\boldsymbol{F}_{\text{input}})\boldsymbol{p}_e, \\ \boldsymbol{p}_e &= \texttt{softmax}(\texttt{TopK}(g(\boldsymbol{F}_{\text{input}}), k))_e, \end{aligned} \tag{2}$$

where N denotes the total number of experts, $\boldsymbol{p}_e$ represents the dynamic weight of the e-th expert, $\boldsymbol{f}_e(\cdot)$ corresponds to the e-th expert, and $g(\cdot)$ denotes gating function. The $\texttt{TopK}(\cdot, k)$ preserves only the original values of the top-k entries in the vector while setting all other entries to $-\infty$. After the softmax operation, these entries assigned $-\infty$ effectively approximate zero, achieving sparse activation. To enable holistic learning of spatiotemporal motion features, we choose to activate all experts, which is empirically validated by our later experimental results.

Bidirectional Spatiotemporal Mamba

Previous approaches incorporate temporal or spatial information as biases into self-attention mechanisms to extract spatio-temporal features (Peng, Mao, and Wu 2023; Xu et al. 2023; Xiao et al. 2024). However, due to the quadratic complexity of self-attention, these approaches suffer from high computational costs in spatiotemporal modeling. To address this issue, we introduce Mamba with linear complexity into each expert module to reduce computational overhead. Specifically, we construct four structurally distinct experts by combining bidirectional temporal Mamba and bidirectional spatial Mamba in different configurations. Unlike traditional MoE designs that employ a single feed-forward network (FFN) architecture shared by all experts, this design allows each expert to specialize in capturing distinct types of spatiotemporal features, thereby significantly enhancing the model's flexibility and expressiveness in modeling complex spatiotemporal dynamics.

As illustrated in Fig. 2 , we devise four distinct types of spatiotemporal Mamba experts. Taking the Spatial-Temporal Mamba expert as an example, the input features $\boldsymbol{F}_{\text{input}}$ first pass through a bidirectional spatial Mamba to extract spatial features, followed by temporal processing. In this way, Spatial-Temporal Mamba can be formulated by:

$$\begin{aligned} \boldsymbol{F}'' &= \texttt{rearrange}(\texttt{Bi-SMamba}(\boldsymbol{F}_{\text{input}})), \\ \boldsymbol{F}_{\text{output}_1} &= \texttt{rearrange}(\texttt{Bi-TMamba}(\boldsymbol{F}'')), \end{aligned} \tag{3}$$

where $\boldsymbol{F}_{\text{input}} \in \mathbb{R}^{B\times D\times T}$ denotes the input feature, $\texttt{Bi-SMamba}(\cdot)$ and $\texttt{Bi-TMamba}(\cdot)$ denotes bidirectional spatial Mamba and bidirectional temporal Mamba, respectively. $\texttt{rearrange}(\cdot)$ represents the tensor transposition operation. The $\boldsymbol{F}'' \in \mathbb{R}^{B\times T\times D}$ denotes the feature after transposition and $\boldsymbol{F}_{\text{output}_1} \in \mathbb{R}^{B\times D\times T}$ denotes the final output. Other experts adopt similar operations, differing only in the order of data processing. To facilitate the learning of spatio-temporal features, each expert shares bidirectional temporal Mamba and bidirectional spatial Mamba, which can further reduce model parameters.

Moreover, the unidirectional modeling nature of the original Mamba limits its ability to capture global dependencies. To address this issue, we introduce a bidirectional scanning mechanism into both spatial and temporal Mamba modules to enhance global dependency modeling in complex motion sequences. First, the spatial Mamba performs bidirectional scanning along the spatial dimension of input features to capture spatial dependencies, then optimizes learning through residual connections. Similarly, the temporal Mamba operates along the temporal dimension. It can be formulated as follows:

$$\begin{aligned} \boldsymbol{f}^s_o &= \texttt{SMamba}(\overrightarrow{\boldsymbol{f}_s}) + \texttt{SMamba}(\overleftarrow{\boldsymbol{f}_s}) + \overrightarrow{\boldsymbol{f}_s}, \\ \boldsymbol{f}^t_o &= \texttt{TMamba}(\overrightarrow{\boldsymbol{f}_t}) + \texttt{TMamba}(\overleftarrow{\boldsymbol{f}_t}) + \overrightarrow{\boldsymbol{f}_t}, \end{aligned} \tag{4}$$

where $\overrightarrow{\boldsymbol{f}_s} \in \mathbb{R}^{B\times D\times T}$ and $\overrightarrow{\boldsymbol{f}_t} \in \mathbb{R}^{B\times T\times D}$ respectively represent the input features of the spatial Mamba and temporal Mamba. $\overleftarrow{\boldsymbol{f}_s}$ and $\overleftarrow{\boldsymbol{f}_t}$ denote spatially-reversed feature and temporally-reversed feature, respectively. $\texttt{SMamba}(\cdot)$ and $\texttt{TMamba}(\cdot)$ denote spatial Mamba and temporal Mamba, respectively. $\boldsymbol{f}^s_o$ and $\boldsymbol{f}^t_o$ denotes the output of bidirectional spatial Mamba and bidirectional temporal Mamba, respectively. Second, we employ Layer Normalization to stabilize

the training process and leverage an FFN to enhance feature representation capabilities. Subsequently, residual connections are utilized to facilitate model learning:

$$\boldsymbol{F}_o^\star = \texttt{LN}(\texttt{LN}(\boldsymbol{f}_o^\star) + \texttt{FFN}(\texttt{LN}(\boldsymbol{f}_o^\star))), \tag{5}$$

where the superscript is $\star \in \{s, t\}$, $\boldsymbol{F}_o^\star$ is the output of bidirectional spatiotemporal Mamba, $\texttt{LN}(\cdot)$ is Layer Normalization, and $\texttt{FFN}(\cdot)$ is Feed Forward Network. Then, $\boldsymbol{F}_o^s$ and $\boldsymbol{F}_o^t$ undergo distinct propagation orders across spatiotemporal experts, yielding expert-specific outputs $\{\boldsymbol{F}_{\text{output}_i}\}_{i=1}^4$. Afterwards, we aggregate the outputs of activated experts to generate the l-th layer representation $\boldsymbol{E}_{\text{output}}^l$ as input to the next MoE layer. Finally, we apply residual connections and decode the final features $\boldsymbol{E}_{\text{output}}^L$ to predict future motion:

$$\boldsymbol{P}_{\text{pred}}^i = \texttt{iDCT}(\texttt{MD}(\boldsymbol{F}_{\text{input}}^i + \boldsymbol{E}_{\text{output}}^L)), \tag{6}$$

where $\boldsymbol{P}_{\text{pred}}^i$ denotes the final predicted motion of the i-th person, $\texttt{iDCT}(\cdot)$ denotes Inverse Discrete Cosine Transform and $\texttt{MD}(\cdot)$ represents Multi-Pose Decoder.

Loss Function

To constrain history and future joint positions in human motion, we employ spatial loss L_s, which is calculated as:

$$\begin{aligned} L_s = &\frac{\lambda}{J \cdot M \cdot t} \sum_{m=1}^{M} \sum_{j=1}^{J} \sum_{i=1}^{t} \left\| \hat{\boldsymbol{P}}_{i,j}^m - \boldsymbol{P}_{i,j}^m \right\|^2 \\ &+ \frac{1}{J \cdot M \cdot (T-t)} \sum_{m=1}^{M} \sum_{j=1}^{J} \sum_{i=t+1}^{T} \left\| \hat{\boldsymbol{P}}_{i,j}^m - \boldsymbol{P}_{i,j}^m \right\|^2, \end{aligned} \tag{7}$$

where $\hat{\boldsymbol{P}}_{i,j}^m$ and $\boldsymbol{P}_{i,j}^m$ represent the predicted and ground-truth position of the j-th joint for the m-th person at timestamp i. M is the number of humans and λ is the weight coefficient.

To mitigate temporal jitter in predicted movements, following IAFormer (Xiao et al. 2024), we employ temporal consistency loss L_t, which is formulated as:

$$L_t = \texttt{MSE}(\texttt{Conv}(\boldsymbol{P}_{\text{pred}}), \texttt{Conv}(\boldsymbol{P}_{\text{gt}})), \tag{8}$$

where $\texttt{Conv}(\cdot)$ is Convolutional Neural Networks for feature mapping. $\texttt{MSE}(\cdot)$ represents Mean Squared Error.

Finally, we perform end-to-end training by optimizing the aggregated final loss function, defined as:

$$L = \alpha L_s + \beta L_t, \tag{9}$$

where α and β are the weight coefficient.

Experiments

Datasets

Following IAFormer (Xiao et al. 2024), to verify the effectiveness of ST-MoE, we conduct experiments on four multi-person motion datasets: CMU-Mocap (CMU-Graphics-Lab 2003), UMPM (Van der Aa et al. 2011), Mix1 and Mix2 (Peng, Mao, and Wu 2023), CHI3D (Fieraru et al. 2020).

CMU-Mocap (UMPM). CMU-Mocap (UMPM) is a synthetic three-person motion dataset created by integrating UMPM into CMU-Mocap. The training set contains 13,000 sequences, the test set contains 3,000 sequences, and each sequence comprises 75 frames.

Mix1 and Mix2. To validate the model's generalization capability, we train on the CMU-Mocap (UMPM) dataset and evaluate on the Mix1 and Mix2 datasets. Mix1 contains 6 individuals while Mix2 comprises 10 individuals. Both datasets are constructed by MuPoTS-3D (Mehta et al. 2018), 3DPW (Von Marcard et al. 2018), and test data from CMU-Mocap(UMPM). Each dataset consists of 1,000 motion sequences, with each sequence containing 75 frames.

CHI3D. Unlike artificially mixed datasets, CHI3D is a lab-based accurate 3d motion capture dataset containing two individuals. It better reflects real-world scenarios, making it particularly suitable for capturing complex spatiotemporal dependencies in multi-person motions.

Metrics

Following IAFormer (Xiao et al. 2024), we adopt the mean per Joint Position Error (JPE) and Aligned Position Error (APE) as evaluation metrics. JPE measures global joint position errors, including overall body displacement. APE aligns the root joint to remove global movement, focusing on evaluating pose-specific errors. *More details in appendix.*

Implementation Details

Network Architecture. The Multi-Pose Encoder/Decoder employs a 3-layer GCN structure, while the MoE module uses a single-layer design with a one-layer MLP gating function. Observed sequences are set to 50 frames (2s) to predict future 25 frames (1s). The pose dimension D is set to 45. **Reproducibility.** We train our model with a batch size of 96 and adopt the Adam (Kingma and Ba 2014) optimizer with an initial learning rate of 0.01. The learning rate is decayed exponentially by a factor of $0.1^{1/50}$ per epoch. The loss weighting coefficients are configured with $\alpha = 1$, $\beta = 1$, $\lambda = 0.1$. Training is performed on one RTX 3090 GPU.

Comparison with SOTA Methods

Results on CMU-Mocap (UMPM). As shown in Table 1, our method ST-MoE achieves SOTA results on CMU-Mocap (UMPM), outperforming JRFormer (Xu et al. 2023) by 4 mm JPE and 6 mm APE in average metrics. This improvement stems from ST-MoE's ability to adaptively capture complex spatiotemporal dependencies through four distinct experts, whereas existing methods exhibit suboptimal performance due to their constrained capacity in modeling diverse spatiotemporal patterns.

Results on Mix1 and Mix2. On Mix1 and Mix2 datasets, ST-MoE demonstrates strong generalization as the number of individuals increases. It consistently outperforms baselines: on Mix1, gains are 4 mm JPE and 1 mm APE over IAFormer (Xiao et al. 2024); on Mix2, gains are 17 mm JPE and 6 mm APE over JRFormer. This demonstrates superior capability of ST-MoE to capture more complex spatiotemporal dependencies in scenarios with more individuals.

Results on CHI3D. Compared to manually mixed datasets, CHI3D reflects spatio-temporal dependencies in multi-person scenarios more authentically (Xiao et al. 2024).

	Method	CMU-Mocap (UMPM) 3 persons				Mix1 6 persons				Mix2 10 persons			
		0.2s↓	0.6s↓	1.0s↓	Avg↓	0.2s↓	0.6s↓	1.0s↓	Avg↓	0.2s↓	0.6s↓	1.0s↓	Avg↓
JPE	MSR-GCN (Dang et al. 2021)	53	146	231	143	49	132	220	134	60	153	243	152
	HRI (Mao, Liu, and Salzmann 2020)	49	130	207	129	51	141	233	142	52	140	224	139
	MRT* (Wang et al. 2021)	36	115	193	114	37	122	212	124	38	126	214	126
	TBIFormer* (Peng, Mao, and Wu 2023)	**30**	109	182	107	34	121	209	121	34	118	198	117
	JRFormer* (Xu et al. 2023)	32	104	161	99	**32**	109	**184**	**108**	36	125	211	124
	T2P* (Jeong, Park, and Yoon 2024)	38	102	158	99	-	-	-	-	-	-	-	-
	IAFormer* (Xiao et al. 2024)	32	96	159	96	36	112	193	114	36	108	181	108
	ST-MoE* (Ours)	31	**95**	**158**	**95**	34	**108**	187	110	**34**	**106**	**179**	**107**
APE	MSR-GCN (Dang et al. 2021)	46	106	137	96	41	92	120	84	48	110	148	102
	HRI (Mao, Liu, and Salzmann 2020)	41	97	130	89	38	92	122	84	41	100	133	91
	MRT* (Wang et al. 2021)	36	108	159	101	36	109	166	104	38	115	178	110
	TBIFormer* (Peng, Mao, and Wu 2023)	27	84	118	76	28	81	113	74	30	89	124	81
	JRFormer* (Xu et al. 2023)	**20**	78	114	71	**21**	73	105	66	**22**	82	120	75
	T2P* (Jeong, Park, and Yoon 2024)	34	84	116	78	-	-	-	-	-	-	-	-
	IAFormer* (Xiao et al. 2024)	23	71	**103**	66	23	71	101	65	24	76	**108**	**69**
	ST-MoE* (Ours)	22	**70**	104	**65**	22	**69**	**100**	**64**	23	**75**	109	**69**

Table 1: Performance comparison (in mm) on mixed multi-person datasets. * means multi-person motion prediction method. The best results are in **bold** and the second-best ones are underlined.

	Method	0.2s↓	0.4s↓	0.6s↓	0.8s↓	1.0s↓	Avg↓
JPE	PGBIG (Ma et al. 2022)	69	130	181	223	258	172
	TBIFormer*	45	95	145	192	233	142
	IAFormer*	**39**	83	129	176	218	129
	ST-MoE* (Ours)	44	**79**	**123**	**161**	**200**	**121**

Table 2: JPE results (in mm) on CHI3D dataset.

Method	JPE				APE			
	0.2s↓	0.6s↓	1.0s↓	Avg↓	0.2s↓	0.6s↓	1.0s↓	Avg↓
Baseline	35.7	113.2	184.3	111.1	24.7	81.1	114.1	73.3
+ST	37.1	105.8	170.7	104.5	26.0	77.3	108.9	70.7
+TT	33.2	99.1	162.1	98.1	23.0	72.1	104.1	66.4
+TS	34.7	100.8	164.8	100.1	24.6	74.7	106.9	68.7
+SS	33.6	99.1	162.4	98.3	23.9	74.9	105.8	68.2
+All	**31.4**	**95.3**	**158.3**	**95.0**	**22.1**	**70.4**	**103.8**	**65.4**

Table 3: Ablation study for effectiveness of distinct experts on CMU-Mocap (UMPM) dataset.

As shown in Table 2, ST-MoE achieves state-of-the-art performance on CHI3D, exhibiting 8 mm lower average JPE than IAFormer and 21 mm lower than TBIFormer (Peng, Mao, and Wu 2023). These results demonstrate ST-MoE's powerful capability to capture spatiotemporal dependencies in real-world scenarios.

Model Efficiency and Parameter Economy. As shown in Fig. 1(e), ST-MoE surpasses IAFormer in performance while achieving 3.6x faster training and 41.38% fewer parameters, attributed to Mamba's linear time complexity, avoiding the high computational cost of attention mechanisms. For fair comparison, we train all models with a batch size of 96. *More results in Appendix.*

Ablation Studies

Effectiveness of Distinct Experts. To validate the effectiveness of collaborative expert interactions, we systematically replaced the four heterogeneous spatiotemporal experts with each individual type. The number of experts is set to 4. The baseline model incorporates solely Multi-Pose Encoder/Decoder, while ST, TT, TS, and SS correspond to the four expert models illustrated in Fig. 2. In Table 3, we can observe that: i) Integrating any single expert type enhances performance over the baseline, exemplified by TT expert reducing average JPE by 13.0mm and average APE by 6.9mm; ii) Combining heterogeneous experts significantly outperforms models using uniform expert types, demonstrating the critical value of specialized collaborative modeling.

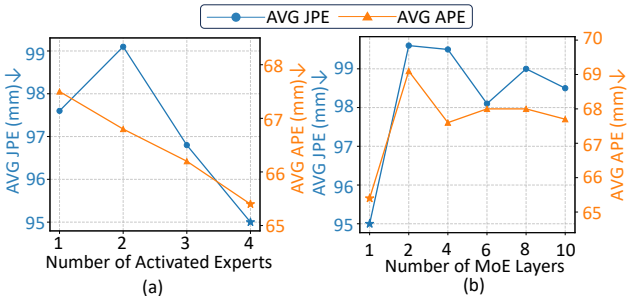

Figure 3: (a): Impact of different numbers of activated experts. (b): Impact of different numbers of MoE layers. Both are tested on the CMU-Mocap (UMPM) dataset.

Impact of Different Numbers of Activated Experts. To investigate the optimal number of activated experts for enhanced performance, we progressively activate the four distinct spatiotemporal experts by increasing the parameter k in Eq. (2) from 1 to 4. As illustrated in Fig. 3(a), the average APE and JPE demonstrate an overall decreasing trend as more experts are activated. Optimal performance is achieved when all four experts are fully activated. This indicates that as the number of activated experts increases, spatio-temporal features can be more flexibly processed by different experts,

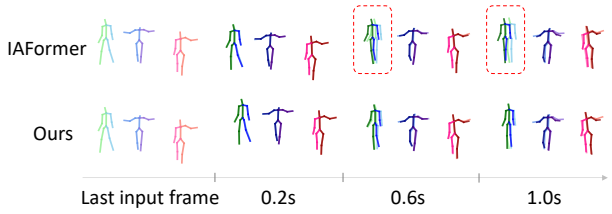

Figure 4: Visualization comparison on CMU-Mocap (UMPM) dataset. Dark-colored lines represent predicted motion, while light-colored lines indicate ground truth.

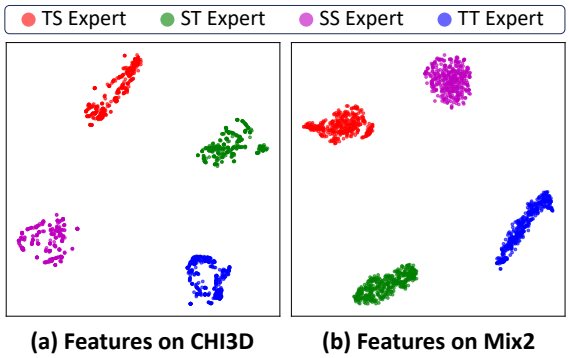

Figure 5: The t-SNE visualization of features learned by four distinct experts on CHI3D and Mix2 dataset. 300 samples are randomly selected. Best view in color.

thereby enhancing prediction performance.

Impact of Varying Numbers of MoE Layers. As shown in Fig. 3(b), we investigate the impact of varying numbers of MoE layers on predictive performance using the CMU-Mocap (UMPM) dataset. We observe that employing a single MoE layer yields the most effective results, as stacking more layers may lead to overfitting and training becomes more difficult. *More results in Appendix.*

Qualitative Analysis

Visualization of Prediction Results. Fig. 4 presents a comparative visualization of ST-MoE against IAFormer (Xiao et al. 2024) on the CMU-Mocap (UMPM) dataset. The left column depicts the last input frames, while the three right columns show predictions at different timestamps. The individual on the left begins moving leftward before abruptly stopping. Due to insufficient modeling of spatiotemporal dependencies, IAFormer fails to capture this "dynamic-to-static" motion pattern, resulting in prediction drift. In contrast, our approach captures this transition by leveraging four specialized spatiotemporal experts, achieving higher accuracy in motion forecasting. *More visualization results in Appendix.*

The t-SNE of Features Learned by Four Experts. As shown in Fig. 5, we visualize the feature spaces of the four distinct spatiotemporal experts via t-SNE on both CHI3D and Mix2 datasets. The former embodies more authentic spatiotemporal dependencies, while the latter contains a

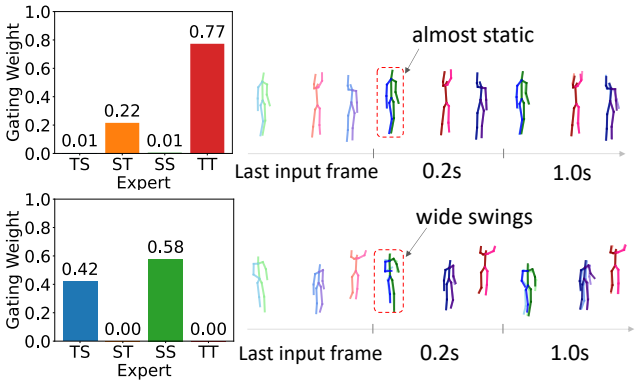

Figure 6: Visualization of adaptive gating weights of four experts and corresponding predicted motion on CMU-Mocap (UMPM). Distinct experts capture different spatiotemporal patterns with adaptive weights.

larger number of individuals exhibiting more complex spatiotemporal patterns. With 300 randomly selected samples processed by all experts, we observe that each expert forms well-separated clusters with pronounced inter-cluster distinctions. This empirically verifies that the four experts capture divergent spatiotemporal motion patterns, while our ST-MoE framework adaptively integrates these representations to achieve enhanced prediction precision.

Flexible Spatiotemporal Modeling. To investigate how the four experts collaborate, we visualize the adaptive gating weights corresponding to each expert and predicted motions on the CMU-Mocap (UMPM) dataset. As illustrated in Fig. 6, adaptive selection of TT/ST experts yields almost static motions, reflecting their focus on spatiotemporal patterns with limited spatial variation. Conversely, when SS/TS experts are activated, the corresponding motion visualization exhibits wide arm swings during running, demonstrating their strength in modeling spatially dynamic patterns. ST-MoE dynamically allocates weights to distinct spatiotemporal experts via adaptive gating, enabling more flexible capture of spatiotemporal dependencies in multi-person motion.

Conclusion

In this work, we propose the first lightweight framework **ST-MoE** for efficient multi-person motion prediction. To overcome trammelled spatiotemporal modeling, we introduce four specialized experts that capture distinct spatiotemporal patterns, enabling comprehensive and flexible learning of motion pattern through adaptive expert selection. To resolve the high computational complexity of existing methods, we design bidirectional spatiotemporal Mamba blocks as the backbone network for each expert, significantly reducing computational overhead. Exhaustive experiments validate that our approach achieves an optimal trade-off between efficiency and performance.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Grant No. U25A20442, 62222207, 62427808, 62472208).

References

Butepage, J.; Black, M. J.; Kragic, D.; and Kjellstrom, H. 2017. Deep representation learning for human motion prediction and classification. In *CVPR*, 6158–6166.

Cao, M.; Shu, X.; Jiang, X.; Yan, R.; Yao, Y.; and Tang, J. 2025. Exploiting Frequency Dynamics for Enhanced Multimodal Event-based Action Recognition. In *ICCV*, 5969–5979.

Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; and Shen, C. 2021. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*.

CMU-Graphics-Lab. 2003. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/.

Cui, Q.; Sun, H.; and Yang, F. 2020. Learning dynamic relationships for 3d human motion prediction. In *CVPR*, 6519–6527.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *CVPR*, 11467–11476.

Du, N.; Huang, Y.; Dai, A. M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A. W.; Firat, O.; et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*, 5547–5569.

Fang, S.; Wang, Z.; Zhong, Y.; Ge, J.; and Chen, S. 2023. Tbp-former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving. In *CVPR*, 1368–1378.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120): 1–39.

Fieraru, M.; Zanfir, M.; Oneata, E.; Popa, A.-I.; Olaru, V.; and Sminchisescu, C. 2020. Three-dimensional reconstruction of human interactions. In *CVPR*, 7214–7223.

Fu, D. Y.; Dao, T.; Saab, K. K.; Thomas, A. W.; Rudra, A.; and Ré, C. 2022. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Gui, L.-Y.; Wang, Y.-X.; Liang, X.; and Moura, J. M. 2018. Adversarial geometry-aware human motion prediction. In *ECCV*, 786–803.

Guo, W.; Bie, X.; Alameda-Pineda, X.; and Moreno-Noguer, F. 2022. Multi-person extreme motion prediction. In *CVPR*, 13053–13064.

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

Jeong, J.; Park, D.; and Yoon, K.-J. 2024. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *CVPR*, 1617–1628.

Jiang, X.; Cao, M.; Tang, H.; Shen, F.; and Li, Z. 2025. Fine-grained Image Retrieval via Dual-Vision Adaptation. *arXiv preprint arXiv:2506.16273*.

Jiang, X.; Tang, H.; Gao, J.; Du, X.; He, S.; and Li, Z. 2024. Delving into multimodal prompting for fine-grained visual classification. In *AAAI*, 2570–2578.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Li, C.; Shu, X.; Cui, Q.; Yao, Y.; and Tang, J. 2024. FT-MoMamba: Motion generation with frequency and text state space models. *arXiv preprint arXiv:2411.17532*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ma, T.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2022. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *CVPR*, 6437–6446.

Mao, W.; Liu, M.; and Salzmann, M. 2020. History repeats itself: Human motion prediction via motion attention. In *ECCV*, 474–489.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning trajectory dependencies for human motion prediction. In *ICCV*, 9489–9497.

Martinez, J.; Black, M. J.; and Romero, J. 2017. On human motion prediction using recurrent neural networks. In *CVPR*, 2891–2900.

Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Sridhar, S.; Pons-Moll, G.; and Theobalt, C. 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 120–130.

Peng, X.; Mao, S.; and Wu, Z. 2023. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *CVPR*, 17121–17130.

Pióro, M.; Ciebiera, K.; Król, K.; Ludziejewski, J.; Krutul, M.; Krajewski, J.; Antoniak, S.; Miłoś, P.; Cygan, M.; and Jaszczur, S. 2024. Moe-mamba: Efficient selective state space models with mixture of experts. *arXiv preprint arXiv:2401.04081*.

Qu, H.; Wei, J.; Shu, X.; and Wang, W. 2025a. Learning clustering-based prototypes for compositional zero-shot learning. *arXiv preprint arXiv:2502.06501*.

Qu, H.; Wei, J.; Shu, X.; Yao, Y.; Wang, W.; and Tang, J. 2025b. OmniGaze: Reward-inspired Generalizable Gaze Estimation In The Wild. In *NeurIPS*.

Qu, H.; Yan, R.; Shu, X.; Gao, H.; Huang, P.; and Xie, G.-S. 2025c. MVP-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. *IEEE TMM*.

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. *arXiv preprint arXiv:1701.06538*.

Shu, X.; Zhang, L.; Qi, G.-J.; Liu, W.; and Tang, J. 2021. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *IEEE TPAMI*, 44(6): 3300–3315.

Tang, B.; Zhong, Y.; Xu, C.; Wu, W.-T.; Neumann, U.; Zhang, Y.; Chen, S.; and Wang, Y. 2023. Collaborative uncertainty benefits multi-agent multi-modal trajectory forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13297–13313.

Team, Q. 2024. Qwen1. 5-moe: Matching 7b model performance with 1/3 activated parameters. *https://qwenlm. github. io/blog/qwen-moe*.

Van der Aa, N.; Luo, X.; Giezeman, G.-J.; Tan, R. T.; and Veltkamp, R. C. 2011. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *ICCV Workshops*, 1264–1269.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 601–617.

Vu, T.-H.; Ambellouis, S.; Boonaert, J.; and Taleb-Ahmed, A. 2020. Anomaly detection in surveillance videos by future appearance-motion prediction. In *ICCV Theory and Applications*, 484–490.

Wang, J.; Xu, H.; Narasimhan, M.; and Wang, X. 2021. Multi-person 3d motion prediction with multi-range transformers. *NeurIPS*, 34: 6036–6049.

Xiao, P.; Xie, Y.; Xu, X.; Chen, W.; and Zhang, H. 2024. Multi-person Pose Forecasting with Individual Interaction Perceptron and Prior Learning. In *ECCV*, 402–419.

Xing, L.; Wang, A. J.; Yan, R.; Shu, X.; and Tang, J. 2025. Vision-centric Token Compression in Large Language Model. In *NeurIPS*.

Xu, Q.; Mao, W.; Gong, J.; Xu, C.; Chen, S.; Xie, W.; Zhang, Y.; and Wang, Y. 2023. Joint-relation transformer for multi-person motion prediction. In *CVPR*, 9816–9826.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yun, S.; Choi, I.; Peng, J.; Wu, Y.; Bao, J.; Zhang, Q.; Xin, J.; Long, Q.; and Chen, T. 2024. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *NeurIPS*, 37: 98782–98805.

Zhong, C.; Hu, L.; Zhang, Z.; Ye, Y.; and Xia, S. 2022. Spatio-temporal gating-adjacency gcn for human motion prediction. In *CVPR*, 6447–6456.

Zhuo, T.; Cheng, Z.; Zhang, P.; Wong, Y.; and Kankanhalli, M. 2019. Unsupervised online video object segmentation with motion property understanding. *IEEE TIP*, 29: 237–249.