Fuzzy-Logic Guided Kinematic Prior Model for 3D Human Motion Prediction

Botao Zhou[1⋆[0009−0009−9482−6260]], Jianqi Zhong[2[0000−0001−7939−1494]], Wenming Cao[1[0000−0002−8174−6167]], and Wenbin Zou[1[0000−0003−1389−9089]]

[1] Shenzhen University, Shenzhen 518060, Guangdong, China
[2] Shenzhen University of Information Technology, Shenzhen 518172, Guangdong, China
2400042053@mails.szu.edu.cn

Abstract. Accurate 3D human motion prediction from historical skeletal sequences is fundamentally critical for downstream applications ranging from human-robot interaction to autonomous driving. However, mainstream graph-based methods struggle to filter high-frequency observational noise, often over-constraining representations to local rigid topologies at the expense of non-local biomechanical synergies. To address these limitations, we propose a novel architecture that integrates kinematic priors with fuzzy logic to dynamically balance local topological constraints and global kinematic dependencies. The core of our approach lies in the synergy of two key modules: the Kinematics-Guided Fuzzy Attention (KGFA) and the Topology-free Global Mixer (TGM). Specifically, KGFA operates through two cascaded stages: the Kinematic Confidence Filter (KCF) computes physical metrics to generate fuzzy masks for kinematic filtering, while the Fuzzy Feature Refiner (FFR) subsequently purifies these representations using entropy-based channel reweighting. Finally, these fuzzy-refined features are directly fused with the unconstrained spatial representations derived from TGM. Extensive evaluation on the Human3.6M benchmark demonstrates that our approach establishes a new state-of-the-art. Compared to existing methods, it achieves average MPJPE reductions of 6.3% and 8.6% in short-term and long-term motion prediction, respectively.

Keywords: Fuzzy logic · Human motion prediction · Kinematic

1 Introduction

Human Motion Prediction (HMP) forecasts future 3D poses from historical data, serving as a fundamental technology for autonomous systems and virtual reality [6,12,13,23]. Consequently, developing precise HMP models remains a critical focus [14].

Driven by deep learning, HMP has progressed significantly in spatio-temporal modeling [24,29] via RNNs [17], ST-GCNs [9,18,30], Transformers [7,22], and

⋆ Corresponding author.

Diffusion models [2], alongside structural refinement strategies [1,3,20]. Furthermore, DCT is widely adopted to represent temporal smoothness and reduce long-term error accumulation [21,25].

Despite these advances, a critical challenge remains: dataset observation noise (e.g., MoCap jitter, missing frames) causes models to overfit and degrades generalization [4,9,16]. To address this, we propose an anti-noise framework integrating fuzzy logic and kinematic principles. By leveraging physical properties like velocity for fuzzy membership, our method dynamically evaluates input uncertainty, suppressing high-frequency noise while preserving motion semantics.

Our unified framework utilizes three novel modules: the Kinematic Confidence Filter (KCF) prior to penalize unreliable observations, Fuzzy Feature Refinement (FFR) attention for feature selection, and Topology-free Global Mixer (TGM) to synergize structural representations attenuated by fuzzy gates.

In summary, our contributions are as follows:

- We propose a fuzzy-logic guided kinematic prior framework to dynamically balance local topological constraints and global kinematic dependencies, and resolve the vulnerability to dataset observation noise in 3D human motion prediction.
- In our framework, we propose three modules: KCF, FFR, and TGM, which contain operations for kinematic-based noise suppression, fuzzy feature attention, and global spatio-temporal synergy to learn noise-resilient representation and aggregate comprehensive features for effective dynamics learning.
- We conduct experiments to quantitatively and qualitatively verify that our framework outperforms state-of-the-art works by 6.3% and 8.6% of MPJPE for short-term and long-term motion prediction on the Human3.6M dataset, respectively.

2 Related Work

2.1 Human Motion Prediction

Human motion prediction aims to forecast future human poses based on historical observations. Early works primarily rely on recurrent neural networks (RNNs) to model temporal dependencies. For instance, Martinez et al. [23] and Fragkiadaki et al. [12] employ sequence-to-sequence RNN architectures to capture motion dynamics. Subsequent studies improve long-term prediction by incorporating structured representations and probabilistic modeling [1,13].

To better model spatial dependencies among joints, graph convolutional networks (GCNs) have been widely adopted. ST-GCN [29] introduces spatial-temporal graph convolutions for skeleton modeling, while subsequent works enhance graph adaptability and multi-scale representation learning [9,18,24,30]. Other approaches further explore disentangled spatial-temporal modeling [25] and dynamic relationship learning [7]. Recently, attention-based and Transformer-based methods have demonstrated superior capability in capturing long-range

dependencies. Mao et al. [21,22] leverage motion attention mechanisms to exploit trajectory patterns. In addition, MLP-based architectures such as MotionMixer [3] provide an alternative lightweight design. More recent works explore probabilistic and generative paradigms. Diffusion-based models such as BeLFusion [2] model motion uncertainty effectively. Frequency-domain modeling has also gained attention, where MHFDN [31] captures motion dynamics in the spectral space. Furthermore, methods such as KSOF [11] and IMPRESS [10] incorporate kinematic priors and structural constraints to improve robustness under incomplete observations.

Despite these advances, most existing methods assume that historical observations are accurate and noise-free, which limits their robustness in real-world scenarios.

2.2 Fuzzy Logic

Fuzzy logic has been widely adopted for modeling uncertainty and handling imprecise information in complex non-linear systems. The Takagi-Sugeno (T-S) fuzzy model [26] significantly advanced the field by providing a systematic framework for system identification and predictive control. In recent years, to better address higher-order uncertainties in dynamic environments, interval type-2 fuzzy logic systems have gained prominent attention [5,15]. These systems extend traditional fuzzy logic by incorporating a footprint of uncertainty, making them highly robust against environmental noise. Furthermore, various enhancements in fuzzy rule generation and optimization [27] have expanded their applicability in time-series forecasting and spatial-temporal modeling. Integrating fuzzy logic principles offers a complementary perspective to deterministic neural network approaches, providing enhanced interpretability and noise-resilience for sequential prediction tasks. Despite its effectiveness in handling uncertainty, fuzzy logic remains underexplored in human motion prediction, particularly in modeling noisy observations and uncertain motion dynamics.

3 Methodology

3.1 Problem Formulation

Given an observed past sequence spanning T_{in} frames, denoted as $\mathbf{X}_{1:T_{in}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_{in}}\}$, human motion prediction aims to forecast a future sequence of T_{out} poses, represented by $\hat{\mathbf{X}}_{T_{in}+1:T_{in}+T_{out}} = \{\hat{\mathbf{x}}_{T_{in}+1}, \hat{\mathbf{x}}_{T_{in}+2}, \ldots, \hat{\mathbf{x}}_{T_{in}+T_{out}}\}$. Each pose $\mathbf{x}_t$ comprises J skeletal joints, where every joint is described by a D-dimensional feature vector representing its spatial configuration. Consequently, the complete human pose at frame t is defined as $\mathbf{x}_t \in \mathbb{R}^{J \times D}$, which is flattened into a $(J \cdot D)$-dimensional vector to serve as the input for our network architectures. To perform this forecasting, we learn a mapping function f parameterized by weights θ, which maps the historical sequence $\mathbf{X}_{1:T_{in}}$ to the predicted future sequence:

$$\hat{\mathbf{X}}_{T_{in}+1:T_{in}+T_{out}} = f(\mathbf{X}_{1:T_{in}}; \theta). \tag{1}$$

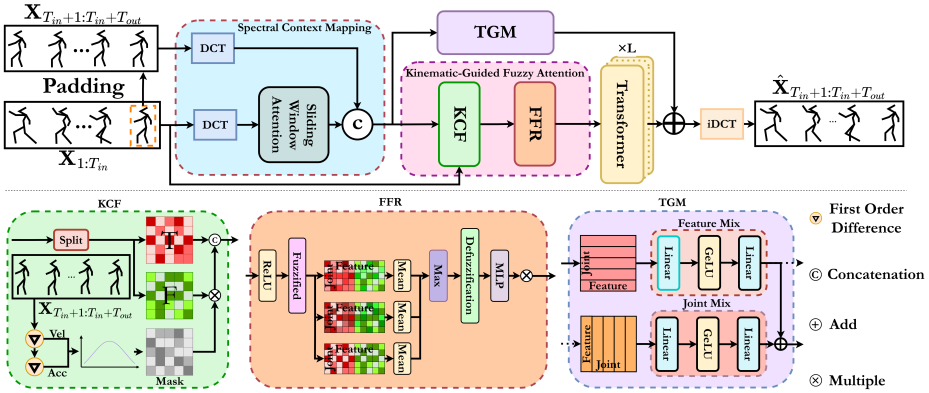

Fig. 1. Our architecture first applies discrete cosine transform (DCT) to convert the temporal sequences into the frequency domain. Then, a two-stream network is built for deep feature extraction: the primary stream employs Kinematic-Guided Fuzzy Attention (KGFA), including a Kinematic Confidence Filter (KCF) to reduce noise and a Fuzzy Feature Refiner (FFR) for spatial attention, followed by a Transformer for spatiotemporal dependencies, while the auxiliary stream utilizes a Topology-free Global Mixer (TGM) to provide a global receptive field. Finally, we build residual connections to fuse these dual-stream features and use inverse DCT (IDCT) to recover the temporal information.

3.2 Model Architecture

Our proposed architecture employs a two-stream prediction methodology driven by time-domain-assisted frequency-domain analysis. Temporal sequences are mapped to the frequency domain via DCT [21] for enhanced smoothness and efficiency, while a sliding window attention model leverages repetitive motion patterns [22]. In the primary stream, the Kinematic-Guided Fuzzy Attention (KGFA) module combines a Kinematic Confidence Filter (KCF) to attenuate noise using fuzzy kinematic priors, and a Fuzzy Feature Refiner (FFR) to compute precise spatial attention via local-global comparisons. A Transformer subsequently captures complex spatiotemporal dependencies. Concurrently, the auxiliary stream utilizes a Topology-free Global Mixer (TGM) to provide a global receptive field through sequential feature and node mixing. Finally, residual connections fuse these global synergistic features with local representations, yielding a comprehensive output (Fig. 1).

3.3 Spectral Context Mapping

Following [21], we introduce a trajectory memory bank mechanism. The raw sequence $\boldsymbol{X}_{1:T_{in}}$ is partitioned by a sliding window of length $K + T_{out}$ into subsequences $\boldsymbol{X}_{n:n+K+T_{out}-1}$. To retrieve relevant history, the recent observation forms the Query ($\boldsymbol{q}$), while the first K poses of each window serve as the Key ($\boldsymbol{k}_n$). Matched fragments are mapped via DCT to suppress high-frequency noise,

defining the low-frequency kinematic Value ($\boldsymbol{V}_n$). This mapping is formulated as:

$$\boldsymbol{q} = F_q(\boldsymbol{X}_{T_{in}-K+1:T_{in}}),\ \boldsymbol{k}_n = F_k(\boldsymbol{X}_{n:n+K-1}), \tag{2}$$

$$A_n = \frac{\boldsymbol{q}\boldsymbol{k}_n^T}{\sum_{n=1}^{T_{in}-K-T_{out}+1} \boldsymbol{q}\boldsymbol{k}_n^T},\ \boldsymbol{S} = \sum_{n=1}^{T_{in}-K-T_{out}+1} A_n \boldsymbol{V}_n. \tag{3}$$

where F_q and F_k are 1D convolutions. The aggregated output $\boldsymbol{S}$ is concatenated with the DCT coefficients of the padded final frame and forwarded to the next layer.

3.4 Kinematic-Guided Fuzzy Attention

To effectively mitigate observation noise within the dataset, we propose KGFA. The KGFA module adopts a cascaded "prior constraint-feature refinement" architecture. First, KCF is introduced to construct a fuzzy membership function utilizing the velocity and acceleration state of joints in physical space. This enables adaptive frequency truncation and smoothing of DCT domain features. Subsequently, FFR is designed to model the saliency of local joints from a global receptive field using fuzzy information entropy theory, thereby achieving dynamic weighting of feature channels. The integration of these two modules strikes a balance between physical constraints and data-driven learning.

Kinematic Confidence Filter (KCF) To mitigate observation noise, we exploit the distinct frequency responses of joints: vigorous motion exhibits high-frequency uncertainty, while smooth motion aligns with low-frequency priors. Thus, we compute a fuzzy "motion smoothness" degree via kinematics to adaptively weight frequency features. Given the padded sequence $\mathbf{X}_{T_{in}-K:T_{in}+T_{out}}$, we extract discrete first-order velocity $\mathbf{v}_t$ and second-order acceleration $\mathbf{a}_t$:

$$\mathbf{v}_t = \mathbf{x}_{t+1} - \mathbf{x}_t, \tag{4}$$

$$\mathbf{a}_t = \mathbf{v}_{t+1} - \mathbf{v}_t = \mathbf{x}_{t+2} - 2\mathbf{x}_{t+1} + \mathbf{x}_t. \tag{5}$$

By aggregating them into temporal matrices $\mathbf{V} = \{\mathbf{v}_t\}$ and $\mathbf{A} = \{\mathbf{a}_t\}$ and averaging their L_2 norms along the temporal dimension, we define local kinematic energy E:

$$E = \alpha\|\mathbf{V}\|_2 + \beta\|\mathbf{A}\|_2, \tag{6}$$

where α and β are learnable weights. A Gaussian function fuzzifies E into a smoothness membership $\mu \in (0, 1]$:

$$\mu = \exp\left(-\frac{E^2}{2\sigma^2 + \epsilon}\right). \tag{7}$$

A higher μ indicates a steady state. To suppress high-frequency noise while preserving primary low-frequency trajectories, we combine the scalar μ with a learned frequency penalty W_{freq} to form a fuzzy mask M_{fuzzy}:

$$M_{fuzzy} = 1 - (1 - \mu) * \text{Sigmoid}(W_{freq}). \tag{8}$$

Finally, this mask dynamically filters the first F_{dct} core components of the DCT features D via the Hadamard product ($\odot$):

$$\tilde{D}_{[:F_{dct}]} = D_{[:F_{dct}]} \odot M_{fuzzy}. \tag{9}$$

Fuzzy Feature Refiner (FFR) Standard Global Average Pooling often obscures anomalous responses of locally active joints. To address this, the FFR module introduces fuzzy T-conorm operations to explicitly localize high-uncertainty joints, sharpening attention on non-linear local dynamics. In conjunction with this dynamic filtering mechanism, given the input DCT frequency feature matrix $\mathbf{Z}$, the subsequent fuzzification utilizes two parallel paths. The first path computes a basic fuzzy membership ν_{base}, mapped via a parameterized Sigmoid:

$$\nu_{base} = \frac{1}{1 + \exp(\mathbf{W}_\nu \cdot (\mathbf{Z} + \mathbf{b}_\nu))}. \tag{10}$$

The second path evaluates the global relative saliency γ_{sal} by comparing local energy with the global average $\bar{\mathbf{Z}}$ at the same frequency component. Since saliency approaches 1 when the local amplitude significantly exceeds the global average, it is formulated as:

$$\gamma_{sal} = \frac{1}{1 + \exp(\tau \cdot (\bar{\mathbf{Z}} - \mathbf{Z}))}. \tag{11}$$

Synthesizing these with the inherent fuzzy deviation ξ_{dev}—which quantifies the structural variance and uncertainty of the frequency components—we apply the Algebraic Sum rule for aggregation:

$$\omega = \gamma_{sal} + \xi_{dev} - \gamma_{sal} \cdot \xi_{dev}. \tag{12}$$

Using the aggregated fuzzy weight ω, we perform weighted centroid defuzzification to yield a global context descriptor $\mathbf{z}_{pool}$:

$$\mathbf{z}_{pool} = \frac{\sum \omega \cdot \mathbf{Z}}{\sum \omega + \epsilon}. \tag{13}$$

Finally, a two-layer MLP ($\mathbf{W}_{c1}, \mathbf{W}_{c2}$) and Sigmoid activation generate channel attention weights $\mathbf{M}_{attn}$, modulating the original frequency feature:

$$\mathbf{M}_{attn} = \text{Sigmoid}(\mathbf{W}_{c2}\delta(\mathbf{W}_{c1}\mathbf{z}_{pool})), \tag{14}$$

and the ultimate output descriptor is obtained by element-wise multiplication:

$$\mathbf{Z}_{out} = \mathbf{Z} \odot \mathbf{M}_{attn}. \tag{15}$$

3.5 Topology-free Global Mixer (TGM)

The Topology-free Global Mixer processes the representation through two orthogonal mapping stages:

Feature Mixing: The first stage models intra-node representations by operating independently on the feature vector of each specific joint. A two-layer MLP projects the incoming features $\mathbf{Z}_{out}$ into a higher-dimensional hidden space to facilitate cross-channel feature fusion:

$$H = \mathbf{W}_{f2}\delta(\mathbf{W}_{f1}\mathbf{Z}_{out}), \tag{16}$$

where $\mathbf{W}_{f1}$ and $\mathbf{W}_{f2}$ are the learnable weight matrices of the feature mixer, and δ denotes the GELU activation function.

Node Mixing: To enable unrestricted global information exchange across different physical joints, the intermediate matrix H is transposed. A subsequent two-layer MLP is applied strictly along the spatial dimension, allowing every joint to dynamically interact with every other joint:

$$H_{spatial} = \left(\mathbf{W}_{n2}\delta(\mathbf{W}_{n1}H^{\top})\right)^{\top}, \tag{17}$$

where $\mathbf{W}_{n1}, \mathbf{W}_{n2}$ are the weight matrices of the node mixer. The matrix is subsequently transposed back to its original shape.

Residual Integration: To preserve the original representation and ensure stable gradient propagation during deep network training, the globally mixed features are added to the original input $\mathbf{Z}_{out}$ via a residual skip connection:

$$Y = \mathbf{Z}_{out} + H_{spatial}. \tag{18}$$

The output Y represents the globally synergized features, which are subsequently fused with the outputs of the parallel Transformer module to construct a comprehensive spatio-temporal descriptor.

3.6 Loss Function

We adopt the Mean Per Joint Position Error (MPJPE) as the base metric, defined as the average Euclidean distance over all future frames and joints. To address error accumulation in long-horizon autoregressive prediction, we introduce a Progressive Exponential Time-Weighting mechanism. The training objective is a temporally weighted MPJPE:

$$\mathcal{L} = \frac{1}{T \cdot N} \sum_{t=1}^{T} \sum_{n=1}^{N} w_t \left\| \hat{\mathbf{x}}_t^n - \mathbf{x}_t^n \right\|_2, \tag{19}$$

where the frame-wise weight is given by $w_t = \exp\left(\gamma \cdot \frac{t-1}{T-1}\right)$, controlled by a time-varying factor $\gamma \geq 0$. The factor γ follows a warm-up and linear ramp-up schedule over the training epochs: it remains zero during an initial warm-up phase (reducing to standard MPJPE for stable short-term learning), then increases linearly to a maximum value, gradually shifting emphasis toward distant frames. Global gradient clipping is applied to ensure numerical stability against the aggressive exponential weights.

4 Experiments Details

4.1 Experimental Setup

Dataset The Human3.6M [16] dataset, a definitive benchmark for 3D human pose estimation and motion tracking, provides millions of synchronized pose vectors across 15 activities. Following standard protocols, we downsample the sequences from 50 Hz to 25 Hz and utilize subjects S1, S6, S7, S8, and S9 for training, S11 for validation, and S5 for testing.

Configuration We implement our network using PyTorch 2.1.0 and train it on a single NVIDIA 4090 GPU with the Adam optimizer, utilizing a batch size of 64 and an initial learning rate of 0.0005 under a smooth exponential decay strategy. The latent feature dimension is set to 256. To forecast future frames, the model processes input sequences of length $T_{in} = 50$ and generates outputs of length $T_{out} = 10$ for short-term tasks and $T_{out} = 25$ for long-term tasks, with prediction precision evaluated using the MPJPE metric strictly following [22].

4.2 Baseline

We benchmark our approach against state-of-the-art methods using MPJPE to demonstrate its effectiveness: PGBIG [20], Eqmotion [28], MHFDN [31], KSOF [11], IMPRESS [10], MT-GCN [8], SPGSN [19].

4.3 Quantitative Comparison

Short-term Prediction Short-term motion prediction aims to predict the poses within 400 milliseconds. First, Table 1 presents the prediction errors of our method and previous methods across all short-term horizons. We see that, our method achieves the best or near-best performance with the lowest overall average error. Notably, at the challenging 400 ms timestamp, our model attains 51.6, clearly outperforming MHFDN (57.0) [31] and IMPRESS (57.1) [10]. Besides, learning the complex spatio-temporal dynamics, the advantage of our method is more pronounced in complex motion categories. Compared to the baselines, our method substantially reduces the errors on posing (from 72.5 to 41.9), sitting down (48.9), and discussion (40.0). These results demonstrate its effectiveness in mitigating error accumulation and generalizing to non-periodic human motion.

Long-term Prediction Long-term motion prediction aims to predict the poses over 400 milliseconds, which is challenging due to the pose variation and elusive human intention. To evaluate the stability, Table 2 presents the prediction errors of various methods at the 560 ms and 1000 ms on the Human3.6M dataset [16]. We see that, our method significantly outperforms all competing models in average long-term horizons, achieving the lowest average error at 560 ms. Crucially,

Table 1. Detailed short-term prediction results on Human3.6M at 80, 160, 320, and 400 ms. Best and second-best performances are indicated in **bold** and underlined, respectively.

Action	**Walking**				**Eating**				**Smoking**				**Discussion**			
Time (ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
MT-GCN [8]	11.5	18.8	34.1	41.7	9.1	17.2	35.2	42.3	9.0	15.7	30.2	39.7	10.8	22.7	53.3	64.6
PGBIG [20]	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7
SPGSN [19]	10.1	19.4	34.8	41.5	7.1	14.9	30.5	37.9	6.7	13.8	28.0	34.6	10.4	23.8	53.6	67.1
Eqmotion [28]	9.2	18.7	34.3	41.1	6.2	14.1	30.0	37.3	5.8	13.1	27.0	33.7	8.9	22.6	52.8	65.8
MHFDN [31]	**8.5**	18.6	35.7	42.6	6.7	15.2	28.6	35.9	6.4	14.5	28.3	35.7	8.8	19.7	48.5	60.9
KSOF [11]	9.4	19.1	35.3	38.2	6.6	14.7	30.6	37.8	6.2	13.7	25.3	31.2	9.3	23.0	48.1	60.2
IMPRESS [10]	9.0	**17.5**	**30.7**	**35.6**	6.0	13.5	**28.0**	**34.8**	6.1	13.4	27.5	33.9	9.1	22.9	53.3	66.7
Ours	9.1	18.1	32.6	38.2	**5.7**	**13.2**	29.0	36.8	**5.7**	**11.8**	**23.0**	**28.6**	**8.0**	**16.9**	**30.8**	**40.0**
Action	**Direction**				**Greeting**				**Phoning**				**Posing**			
Time (ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
MT-GCN [8]	8.4	23.2	42.7	56.7	13.9	**28.9**	**65.2**	**78.6**	8.8	17.4	37.3	**47.1**	9.6	20.1	54.0	77.5
PGBIG [20]	7.2	17.6	40.9	51.5	15.2	34.1	71.6	87.1	8.3	18.3	38.7	48.4	10.7	25.7	60.0	76.6
SPGSN [19]	7.4	17.1	39.8	50.3	14.6	32.6	70.6	86.4	8.7	18.3	38.7	48.5	10.7	25.3	59.9	76.5
Eqmotion [28]	6.6	**15.8**	39.0	50.0	12.7	30.0	69.1	86.3	**7.7**	**17.3**	38.2	48.4	9.3	23.7	59.6	77.5
MHFDN [31]	7.4	16.4	**38.8**	50.7	13.8	32.2	70.4	85.9	8.5	17.6	36.8	47.2	9.8	23.8	56.8	72.5
KSOF [11]	6.6	17.0	40.4	49.8	13.4	31.4	67.7	85.9	7.8	18.0	37.9	47.7	9.2	23.1	57.4	74.1
IMPRESS [10]	**6.2**	16.1	38.9	**49.5**	13.2	31.1	68.8	84.8	**7.7**	17.6	37.8	47.3	9.4	24.1	58.9	75.7
Ours	6.6	18.2	47.0	60.1	**12.1**	29.3	66.0	81.3	7.8	17.6	38.7	49.3	**7.2**	**16.6**	**32.6**	**41.9**
Action	**Purchases**				**Sitting**				**Sitting Down**				**Taking Photo**			
Time (ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
MT-GCN [8]	13.4	28.9	67.9	75.1	9.8	21.1	48.5	54.4	13.7	29.6	57.3	70.8	8.9	18.0	**39.5**	**48.1**
PGBIG [20]	12.5	28.7	60.1	73.3	8.8	19.2	42.4	53.8	13.9	27.9	57.4	71.5	8.4	18.9	42.0	53.3
SPGSN [19]	12.8	28.6	61.0	74.4	9.3	19.4	42.3	53.6	14.2	27.7	56.8	70.7	8.7	18.9	41.5	52.7
Eqmotion [28]	**11.3**	**26.8**	**59.2**	73.0	**8.0**	**18.1**	41.4	53.0	13.2	26.3	56.0	70.4	8.1	**17.8**	41.0	52.7
MHFDN [31]	11.7	27.1	59.9	73.5	8.3	**18.1**	**40.9**	52.3	13.7	27.9	55.2	69.3	**7.4**	18.4	40.6	50.5
KSOF [11]	11.7	28.1	61.5	75.1	8.4	19.0	42.2	**51.4**	13.8	28.4	58.1	72.3	9.2	18.7	42.0	55.3
IMPRESS [10]	11.4	27.2	59.4	**72.7**	8.5	19.0	42.9	54.5	13.4	27.0	56.0	69.8	8.0	18.2	41.5	52.9
Ours	11.9	29.3	62.6	76.4	8.5	19.3	44.7	57.1	**12.1**	**23.1**	**39.6**	**48.9**	8.0	18.8	44.5	57.1
Action	**Waiting**				**Walking Dog**				**Walking Together**				**Average**			
Time (ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
MT-GCN [8]	8.4	18.7	42.9	**52.7**	21.1	41.1	77.0	90.6	8.4	21.3	33.7	44.2	11.0	22.9	47.9	58.9
PGBIG [20]	8.9	20.1	43.6	54.3	18.8	39.3	73.7	86.4	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
SPGSN [19]	9.2	19.8	43.1	54.1	18.2	37.3	71.3	84.2	8.9	18.2	33.8	40.9	10.4	22.3	47.1	58.3
Eqmotion [28]	7.9	**17.9**	**41.5**	53.1	16.6	**35.9**	**72.0**	85.7	8.1	17.1	32.4	39.2	9.4	20.9	46.2	57.9
MHFDN [31]	8.2	18.4	42.0	53.3	–	–	–	–	8.9	17.9	35.7	41.2	9.6	21.4	45.9	57.0
KSOF [11]	8.2	19.3	43.0	53.6	**15.9**	36.9	72.2	**84.7**	7.8	18.0	34.0	40.5	9.5	21.3	46.1	57.4
IMPRESS [10]	8.0	18.9	42.1	52.9	17.3	37.3	73.0	87.5	8.0	17.1	**31.7**	**37.7**	9.4	21.4	46.0	57.1
Ours	**7.6**	18.5	43.7	55.8	18.3	40.2	75.8	89.0	**7.6**	**16.4**	33.1	40.1	**9.1**	**20.5**	**42.9**	**53.4**

at 1000 ms, our method is the only one to successfully keep the mean error below the 100-threshold (97.7). In contrast, all other baselines degrade rapidly, with the second-best method reaching 106.9, and the most recent baseline diverging to 110.7. This substantial performance gap explicitly validates the superiority of

Table 2. Comparisons of long-term motion prediction. The best results are highlighted in **bold**, and the second best results are underlined.

Method	Walking		Eating		Smoking		Greeting		Phoning		Posing		Average	
Time (ms)	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
MT-GCN [8]	–	60.4	–	74.9	–	70.8	–	138.6	–	103.4	–	154.3	–	110.7
PGBIG [20]	48.1	56.4	51.1	76.0	46.5	69.5	110.2	143.5	65.9	102.7	106.1	164.8	76.9	110.3
SPGSN [19]	46.9	53.6	49.8	73.4	46.7	68.6	111.0	143.2	66.7	102.5	110.3	165.4	77.4	109.6
Eqmotion [28]	43.4	52.8	48.4	73.0	41.0	63.4	108.7	142.0	64.7	101.0	84.9	139.4	73.4	106.9
MHFDN [31]	47.1	54.7	49.9	74.6	46.2	68.4	**74.0**	**115.2**	73.9	117.7	71.7	102.9	76.3	109.4
KSOF [11]	**42.3**	**49.2**	52.4	75.7	48.4	70.2	113.2	145.3	67.6	102.9	109.4	173.6	77.7	109.7
IMPRESS [10]	44.6	53.7	**47.6**	**71.6**	45.6	67.7	109.1	145.3	**64.6**	**99.9**	106.8	165.1	75.7	109.3
Ours	48.7	58.8	57.2	81.8	**22.9**	**31.4**	118.8	155.0	74.9	114.8	**31.0**	**45.3**	**69.5**	**97.7**

our architecture in modeling complex, long-range spatio-temporal dependencies while preserving structural fidelity over extended durations.

4.4 Qualitative Comparison

To further evaluate the performance of our proposed model, we conduct a qualitative analysis by visualizing the predicted human motion sequences. We compare the predictions of our model against the baseline and the ground truth across a variety of complex actions, including *walking*, *sitting down*, *discussion*, and *greeting*. Fig. 1 shows a sequence of 3D skeletal poses from -400 ms to 1000 ms in *posing*, *smoking*, and *waiting* movements with an interval of 200 ms.

Qualitative Analysis Visually, our model outperforms the PGBIG [20] baseline in preserving motion expressiveness across challenging sequences (e.g., discussion, posing, smoking). While PGBIG suffers motion attenuation and collapses into a static "mean pose" during long-term prediction, our approach maintains dynamic, temporally coherent trajectories. Specifically, our method prevents skeletal drift in posing and captures subtle, high-frequency kinematics (like nuanced hand movements) in smoking, which PGBIG completely smooths out. Ultimately, our approach successfully circumvents the skeletal collapse and motion homogenization typical of traditional baselines, yielding predictions with superior realism, physical plausibility, and fine-grained details.

4.5 Ablation Study

To evaluate the contribution of each core component to the final prediction performance, we conducted four sets of ablation experiments under consistent dataset and experimental settings. These experiments analyze the **KCF**, the **FFR**, the **TGM**, and the impact of **Transformer Layer Count**. The quantitative results are summarized in Table 3 and Table 4.

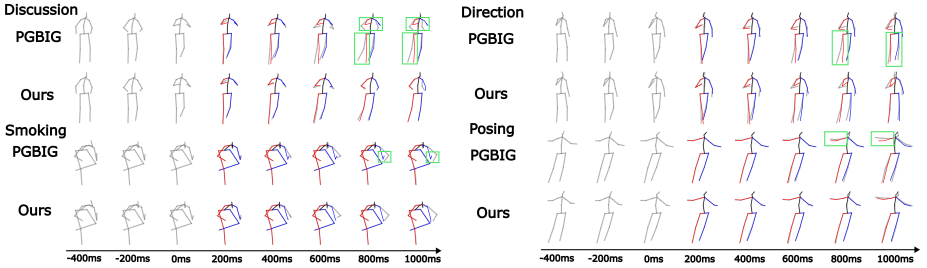

Fig. 2. Comparison visualization across different long-term forecasting scenarios on H3.6M dataset.

Effectiveness of KCF KCF leverages physical kinematics as a prior to generate fuzzy masks on DCT coefficients, constraining predictions that violate human motion laws. We removed KCF, allowing the input features to bypass the fuzzy mask filtering. The prediction error increases significantly for long-term horizons without KCF (see Table 3). This suggests that in the absence of kinematic constraints, the model tends to accumulate errors, leading to high-frequency jitter. KCF effectively filters noise through adaptive fuzzy logic, proving the necessity of physical priors.

Effectiveness of FFR FFR adaptively re-weights the feature stream through fuzzification and de-fuzzification mapping, enhancing the representation of critical joint nodes. We skip the FuzzyFeatureAttention module in the main feature flow, and results indicate that the model without FFR suffers in capturing local joint details (see Table 3). FFR dynamically regulates feature representations via probabilistic gating, successfully preserving accurate motion semantics even for highly non-linear joints experiencing sudden, unpredictable transitions.

Effectiveness of TGM The TGM serves as a parallel branch to the backbone, utilizing an MLP-based architecture to alternate between feature-wise and node-wise mixing. We disconnected the Topology-free Global Mixer branch, deriving the final fused DCT features solely from the backbone output (see Table 3). Removing the TGM branch results in poorer structural consistency across the skeleton. This confirms that the TGM plays a vital role in capturing global spatial dependencies, acting as a powerful complement to the transformer-based backbone.

Effect of Transformer Layer Count The depth of the Transformer layers directly influences the receptive field and parameter capacity. We varied the number of stages $L \in \{12, 13, 14, 15, 16\}$ while keeping other modules constant (see Table 4). Performance improves significantly as L increases from 12 to 14. However, increasing L to 15 or 16 yields diminishing returns and slight performance

Table 3. Quantitative results of module ablation studies at different time steps.

Model Variants	KCF	FFR	TGM	160ms	400ms	560ms	1000ms
Variant A	×	✓	✓	26.21	61.93	78.64	109.48
Variant B	✓	×	✓	25.42	60.80	74.05	108.69
Variant C	✓	✓	×	25.12	55.17	76.95	107.65
Ours (Full)	✓	✓	✓	**20.48**	**53.36**	**69.52**	**97.69**

degradation due to over-fitting or the over-smoothing effect. Consequently, we set $L = 14$ as the optimal balance.

Table 4. Impact of the number of Transformer layers (L) across different time horizons.

Layers (L)	Params (M)	160ms	400ms	560ms	1000ms
12	9.63	23.61	46.96	71.30	100.50
13	10.36	23.76	47.12	72.14	101.80
14 (Ours)	11.08	20.48	53.36	**69.52**	**97.69**
15	11.81	23.74	**46.71**	69.95	99.84
16	12.53	23.86	47.19	72.24	101.74

5 Conclusion

In conclusion, we propose a novel two-stream architecture for 3D human motion prediction, effectively addressing the inherent challenges of observation noise and complex spatiotemporal dependencies. The core contribution of this work lies in KGFA, which innovatively integrates kinematic priors with fuzzy logic theory to adaptively filter high-frequency sensor noise while precisely localizing active joint dynamics. Complementing this local feature refinement, the auxiliary TGM facilitates unrestricted global information exchange, ensuring holistic spatial synergy independent of predefined skeletal graphs. Through the seamless integration of physical constraints and data-driven learning, the proposed framework extracts comprehensive representations that yield highly accurate and stable predictions on Human3.6M.

Acknowledgements This work was supported by the National Natural Science Foundation of China under grant 62571346, and the Fundamental Research Foundation of Shenzhen under Grant JCYJ20230808105705012.

References

1. Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3D human motion modelling. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7144–7153 (2019)

2. Baradel, F., et al.: BeLFusion: Latent diffusion for 3D human motion prediction. arXiv preprint arXiv:2210.15243 (2022)
3. Bouazizi, A., Holte, M.B., et al.: MotionMixer: MLP-based 3D human body pose forecasting. In: European Conference on Computer Vision (ECCV). Springer, Cham (2022)
4. Cai, Y., Huang, L., Wang, Y., Cham, T.-J., Cai, J., Yuan, J., Liu, J., et al.: Learning progressive joint propagation for human motion prediction. In: European Conference on Computer Vision (ECCV), pp. 226–242. Springer, Cham (2020)
5. Castillo, O., Melin, P.: A review on interval type-2 fuzzy logic applications in intelligent control. Information Sciences **524**, 1–12 (2020)
6. Chiu, H.-K., Adeli, E., Wang, B., Huang, D.-A., Niebles, J.C.: Action-agnostic human pose forecasting. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1423–1432 (2019)
7. Cui, Q., Sun, H., Yang, F.: Learning dynamic relationships for 3D human motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6519–6527 (2020)
8. Cui, Q., Sun, H.: Towards accurate 3D human motion prediction from incomplete observations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4799–4808. IEEE (2021)
9. Dang, L., Nie, Y., Long, C., Zhang, C., Li, G.: MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11467–11476 (2021)
10. Deng, H., Li, J., Li, J., Wen, J., Xu, Y.: IMPRESS: Incomplete human motion prediction via motion recovery and structural-semantic fusion. IEEE Transactions on Image Processing **35**, 2393–2406 (2026)
11. Ding, R., Qu, K., Tang, J.: KSOF: Leveraging kinematics and spatio-temporal optimal fusion for human motion prediction. Pattern Recognition **161**, 111206 (2025)
12. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: IEEE International Conference on Computer Vision (ICCV), pp. 4346–4354 (2015)
13. Ghosh, P., Song, J., Aksan, E., Hilliges, O.: Learning human motion models for long-term predictions. In: International Conference on 3D Vision (3DV), pp. 458–466. IEEE (2017)
14. Guo, X., Choi, J.: Human motion prediction via learning local structure representations and temporal dependencies. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 2580–2587 (2019)
15. Hagras, H.: Type-2 FLCs: A new generation of fuzzy controllers. IEEE Computational Intelligence Magazine **2**(1), 30–43 (2007)
16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2013)
17. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-RNN: Deep learning on spatio-temporal graphs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5308–5317 (2016)
18. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multi-scale graph neural networks for 3D skeleton based human motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 214–223 (2020)

19. Li, M., Chen, S., Zhang, Z., Xie, L., Tian, Q., Zhang, Y.: Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. In: Avidan, S., et al. (eds.) Computer Vision – ECCV 2022. LNCS, vol. 13666, pp. 18–36. Springer, Cham (2022)
20. Ma, T., Nie, Y., Long, C., Zhang, C., Li, G.: Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6437–6446 (2022)
21. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9489–9497 (2019)
22. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: European Conference on Computer Vision (ECCV), pp. 474–489. Springer, Cham (2020)
23. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2891–2900 (2017)
24. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12026–12035 (2019)
25. Sofianos, I., Sampieri, A., Franco, L., Galasso, F.: Space-time-separable graph convolutional network for pose forecasting. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11209–11218 (2021)
26. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man, and Cybernetics **15**(1), 116–132 (1985)
27. Wu, D., Mendel, J.M.: Enhancements to intersecting rule classes for interval type-2 fuzzy logic systems. IEEE Transactions on Fuzzy Systems **17**(4), 733–746 (2009)
28. Xu, C., et al.: EqMotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5250–5260 (2023)
29. Yan, S., Xiong, Y., Lin, D.: Spatial-temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
30. Zhong, C., Hu, L., Zhang, Z., Ye, Y., Xia, S.: Spatio-temporal gating-adjacency GCN for human motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6447–6456 (2022)
31. Zhou, X., Liu, R., Dong, J., Yi, P., Wang, L.: MHFDN: Multi-Branch Hybrid Frequency Domain Networks for 3D Human Motion Prediction. In: 9th International Conference on Signal and Image Processing (ICSIP), pp. 696–701. IEEE (2024)
32. Zhong, J., Cao, W.: Geometric algebra-based multiscale encoder-decoder networks for 3D motion prediction. Applied Intelligence **53**, 26967–26987 (2023)
33. Cao, W., Chen, X., Zhong, J.: AsymFormer: Asymmetric interaction-joints dynamics transformer for multi-person 3D pose forecasting. IEEE Transactions on Automation Science and Engineering **23**, 1151–1164 (2026).
34. Zhong, J., Huang, J., Cao, W.: Tacking over-smoothing: Target-guide progressive dynamic graph learning for 3D skeleton-based human motion prediction. Expert Systems with Applications **256**, 124914 (2024)