Breaking the Passive Learning Trap: An Active Perception Strategy for Human Motion Prediction

Juncheng Hu[1], Zijian Zhang[1], Zeyu Wang[2], Guoyu Wang[1], Yingji Li[1], Kedi Lyu[1]*

[1]College of Computer Science and Technology, Jilin University, Changchun, China
[2]College of Computer Science and Engineering, Dalian Minzu University, Liaoning, China
jchu@jlu.edu.cn, zhangzijian@jlu.edu.cn, 20231578@dlnu.edu.cn, wgy21@mails.jlu.edu.cn, yingjili@jlu.edu.cn, kdlyu@jlu.edu.cn

Abstract

Forecasting 3D human motion is an important embodiment of fine-grained understanding and cognition of human behavior by artificial agents. Current approaches excessively rely on implicit network modeling of spatiotemporal relationships and motion characteristics, falling into the ***passive learning trap*** that results in redundant and monotonous 3D coordinate information acquisition while lacking actively guided explicit learning mechanisms. To overcome these issues, we propose an **A**ctive **P**erceptual **S**trategy (**APS**) for human motion prediction, leveraging quotient space representations to explicitly encode motion properties while introducing auxiliary learning objectives to strengthen spatio-temporal modeling. Specifically, we first design a *data perception module* that projects poses into the quotient space, decoupling motion geometry from coordinate redundancy. By jointly encoding tangent vectors and Grassmann projections, this module simultaneously achieves geometric dimension reduction, semantic decoupling, and dynamic constraint enforcement for effective motion pose characterization. Furthermore, we introduce a *network perception module* that actively learns spatio-temporal dependencies through restorative learning. This module deliberately masks specific joints or injects noise to construct auxiliary supervision signals. A dedicated auxiliary learning network is designed to actively adapt and learn from perturbed information. Notably, APS is model agnostic and can be integrated with different prediction models to enhance *active perceptual*. The experimental results demonstrate that our method achieves the new state-of-the-art, outperforming existing methods by large margins: 16.3% on H3.6M, 13.9% on CMU Mocap, and 10.1% on 3DPW.

Introduction

Modeling 3D human motion from high-dimensional and highly stochastic historical observations to achieve accurate future human motion prediction (HMP) is extremely challenging. Given its significant implications, HMP has received diverse applications, including embodied intelligence, human-computer interaction, and autonomous driving (Salomão et al. 2022; Li et al. 2022c; Zhou and Wang 2023; Yao et al. 2023).

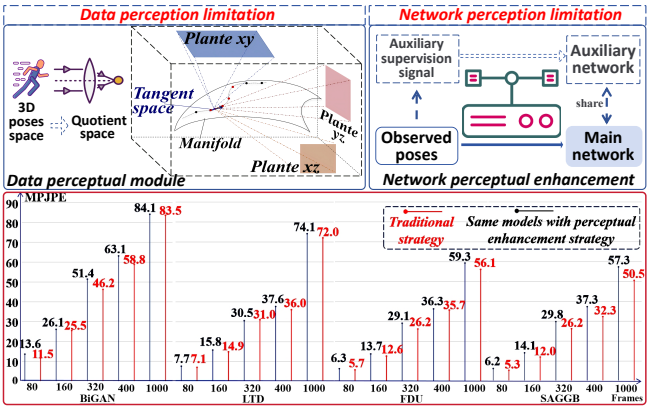

Figure 1: The upper part highlights the motivation behind APS, while the lower part demonstrates its effectiveness in mitigating premature performance bottlenecks.

In this paper, we focus on enhancing motion prediction through hierarchical perception (*i.e. spatio-temporal relationships* and *motion properties*) and capture long-range spatio-temporal dependencies to produce accurate predictions. Recent methods propose specialized network structures to model human motion. Some approaches utilize RNNs (Liu et al. 2022) and TCNs (Huang and Li 2023; Tang et al. 2023a) to model temporal dependencies. Some methods (Gu et al. 2024; Tang et al. 2023b; Wang et al. 2024a; Zhang et al. 2023) propose GCN networks with learnable weights. DMGNN (Li et al. 2020) and MST-GNN (Li et al. 2022a) further construct multi-scale body graphs to model local-global spatial features. PGBIG (Li et al. 2022b) also designs temporal graph convolution extract spatio-temporal features. SPGSN (Mao, Liu, and Salzmann 2020) proposes graph scattering networks to further model the temporal dependence from multiple graph spectral bands.

Scrutinizing the released implementations of existing methods, one observes that current methods often suffer from limitations that lead to *falling into the passive learning trap*, thus affecting prediction accuracy, as illustrated in Figure 1. We believe that this issue is mainly due to the following two reasons: ***Firstly***, *data perception limitation*. Existing methods (Martinez, Black, and Romero 2017; Li

*Corresponding Author: Kedi Lyu

et al. 2024) perform dynamic modeling in 3D pose space, inheriting the intrinsic complexity of human motion ($e.g.$, high dimensionality and stochastic nature). In such high-dimensional regimes, effective motion features often become entangled, while many implicit kinematic properties resist effective encoding within rigid 3D representations. Without active guidance, the network will be forced to learn redundant coordinate system information. ***Secondly***, *network perception limitation*. Current approaches (Zhou et al. 2021; Yu et al. 2023) predominantly rely on passive information aggregation through static architectures, lacking an explicit, actively guided learning mechanism. This inherent limitation restricts their ability to dynamically adapt to complex spatiotemporal dependencies, ultimately impairing their effectiveness in modeling intricate motion patterns.

To tackle these challenges, an active perceptual strategy (APS) for HMP is proposed to facilitate active perception of spatio-temporal dependencies and motion information to mitigate the undesired impact of ***passive learning trap***, as illustrated in Figure 2. APS consists of two major modules, namely a *data perception module* (DPM) and a *network perception module* (NPM).

The **DPM** dresses high-dimensional human pose perception challenges via a geometric framework that projects pose sequences onto the quotient space, which includes two parts. The *tangent space* capturing local motion dynamics via tangent vectors v representing instantaneous pose variations, and the *Grassmann manifold* $Gr(k, n)$ embeddings that model low-dimensional subspace states ($e.g.$, joint movement constraints). Specifically, the motion of each joint can be regarded as a trajectory on the manifold $\mathcal{M}$, and the displacements between frames are the tangent vectors v of this estimate. *First*, we project poses onto the tangent space , and utilize the tangent space operator to calculate the trajectory of joint motion, representing it as the tangent vector v. *Then*, each vector is projected onto the Grassmann manifold, *i.e.,* decomposed into three different fixed subspaces $Gr(2, 3)$. The angle between the tangent vector and the coordinate plane reflects the relative orientation of v with respect to these subspaces. The angle characterizes the orientation of the tangent vector on the Grassmann manifold, describing how v is distributed in the different subspaces. *Finally*, these subspace-projected motion data are processed by NPM to generate future motion data.

The **NPM** aims to address network perception limitation by constructing auxiliary supervision signals to force the network to actively repair the spatiotemporal relationships that are missing or disturbed by noise. NPM comprises two components. A *spatio-temporal enhancement component* (SEC) induces active perception by selectively corrupting observed pose to formulate auxiliary reconstruction tasks. Through reconstructing these masked data into their original configurations, it establishes auxiliary learning objectives that compel downstream networks to develop intrinsic motion pattern comprehension. These intentionally corrupted data are then fed into a *spatial-temporal learning component* (SLC). To adapt to this dynamically changing human motion data format, we design SLC as a data learning component. SLC employs spatio-temporal graph attention mechanisms to dynamically integrate local relationships with global motion semantics. By iteratively refining the reconstruction of corrupted poses through adaptive feature aggregation across temporal and spatial dimensions, it achieves active perception of latent motion dependencies without predefined structural priors. The attention units automatically emphasize salient spatio-temporal correlations while suppressing noise propagation, enabling the model to capture comprehensive motion dynamics through self-taught feature interactions.

Contributions. To summarize, our key contributions are as follows: i) We propose a novel active perception strategy that is model-agnostic for human motion prediction tasks. ii) We present a data perception module and a network perception module to separately enhance pose representation and dynamic context modeling, jointly enhancing the framewok's active perception capability, and elevating prediction accuracy. iii) Our method achieves SOTA results on three benchmark datasets, 16.3% on H3.6M, 13.9% on CMU Mocap, and 10.1% on 3DPW.

Our Approach

Problem formulation. Given poses sequence $P^n = \langle p_0, p_1, \ldots, p_n \rangle$, the network captures implicit spatio-temporal dependencies and predicts the future pose sequence $P^N = \langle \hat{p}_{n+1}, \hat{p}_{n+2}, \ldots, \hat{p}_{n+N} \rangle$. $p_n \in R^{J*3}$ represents the human pose at the n_{th} frame, and J denotes the total number of joints. N refer to the lengths of future movements. The objective is to learn a prediction model $F_{pred}(\cdot)$ such that the predicted future motion $\hat{P}^N = F_{pred}(P^n)$ is closed to the ground truth P^N. Inspired by the motion in the quotient space, we propose decomposing and embedding the human pose sequence into different subspace constraints, whereby the original motion can be represented as a dynamic projection on the fiber tuft η. Formally, let $T_p\mathcal{M}$ describe the instantaneous motion ($e.g.$, joint velocity or displacement) of a human pose manifold $\mathcal{M}$ at a point p (a certain pose). $Gr(k, d)$ denotes the space of all possible k-dimensional projections in a d-dimensional tangent space. Thus, the fiber bundle constructed by the human pose sequence can be represented as:

$$\eta = (T_p\mathcal{M}, Gr(k, d)). \tag{1}$$

At this point, we can describe the motion by learning the tangent vectors and their projections on subspaces.

Method Overview. In this paper, we propose an **A**ctive **P**erception **S**trategy (APS) consisting of two key modules: a *data perceptual module* (DPM) and a *network perceptual module* (NPM), as illustrated in Figure 2. Specifically, the DPM is *first* designed to mitigate data perception limitations by transforming motion sequences from 3D pose space into quotient space. The processed data is *then* fed into the NPM, where active noise injection or joint masking is applied to the raw skeletal data, thereby stimulating the network's active perception capability. *Finally*, the perturbed yet structured data is input into the learning network for predicting future motion trajectories in subspace. The following sections elucidate our proposed DPM and NPM in detail.

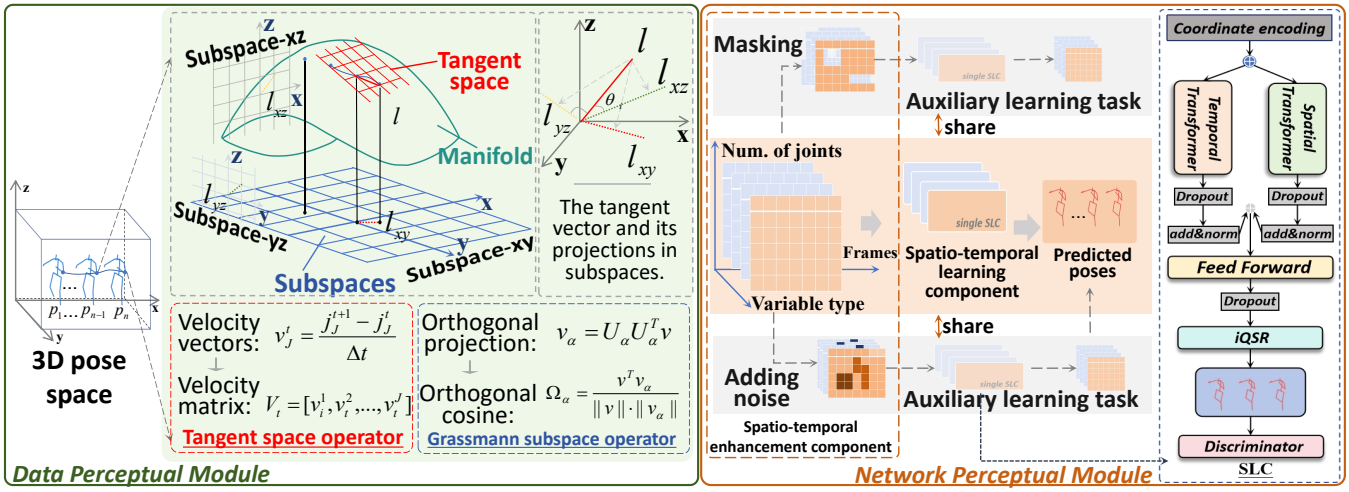

Figure 2: The architecture of the proposed method.

Data Perceptual Module

As mentioned above, human motion is inherently characterized by high-dimensionality (*e.g.,* multi-joint coordination) and stochastic nature (*e.g.,* variable movement patterns). Existing methods typically model motions directly in high-dimensional pose spaces, forcing the network to learn redundant coordinate system information, and passively falling into the coupled representation of geometry and semantics. Based on this, we propose the **DPM** to utilize the quotient space representation to actively enhance the implicit dynamic information while reducing the high-dimensional poses. Specifically, by mapping to the quotient space, pose sequences are decomposed and embedded into subspace-constrained representations, where primitive motions are dynamically projected onto a fiber bundle. Formally, this reduces motion modeling to tracking tangent space variations v on poses manifold $\mathcal{M}$ at point p relative to its Grassmann structure $Gr(\cdot)$. For engineering implementation, we design the *tangent space operator* (TSO) and *Grassmann manifold operator* (GMO) to complete the mapping.

Tangent space operator. The objective of TSO is to describe the motion change of human poses. Let the human pose have J joints, the time series length be n, and the original data be $P = \{p_t\}_{t=1}^{n}$ and $p_t = [j_t^1, ..., j_t^J]$. For each joint j, the neighboring frame velocity vectors are computed:

$$v_J^t = \frac{j_J^{t+1} - j_J^t}{\Delta t}. \tag{2}$$

Then, considering the velocity field as a tangent vector on the manifold $\mathcal{M}$, the velocity matrix $V_t = [v_t^1, v_t^2, ..., v_t^J]$ can be constructed, and the tangent vector is expressed as:

$$\mathbf{v}_t = vec(V_t). \tag{3}$$

Thus, in the tangent space $T_p\mathcal{M}$, the human pose p is represented as a set of tangent vectors $vec(\cdot)$.

Grassmann manifold operator. Different from the Lie-group based methods, which calculate the rotation of poses through complex systems such as exponential mapping, logarithmic mapping and Lie algebra, we design GMO to describe the orientation change of tangent vector in the subspaces. The tangent vector v from TSO is projected onto the Grassmann manifold $Gr(k, n)$ (k-dimensional projections in a d-dimensional tangent space), which is decomposed into three different fixed subspaces $Gr(2, 3)$, each subspace is a 2D space $\mathcal{S}$, which is represented by an orthonormal basis matrix $\mathbf{U}_\alpha = [\mathbf{u}_1, \ldots, \mathbf{u}_n], \alpha \in \{xy, yz, zx\}$. For a subspace $\mathcal{S}_\alpha = \text{span}(\mathbf{U}_\alpha)$, orthogonal projection of $\mathbf{v}$ onto $\mathcal{S}$:

$$\mathbf{v}_\alpha = \mathbf{U}_\alpha \mathbf{U}_\alpha^\top \mathbf{v} = \sum_{i=1}^{k} (\mathbf{v}^\top \mathbf{u}^i)\mathbf{u}^i. \tag{4}$$

The orthogonal cosine $\mathbf{\Omega}_\alpha$ between $\mathbf{v}$ and its projection $\mathbf{v}_\alpha$ is computed as:

$$\mathbf{\Omega}_\alpha = \frac{\mathbf{v}^\top \mathbf{v}_\alpha}{\|\mathbf{v}\| \cdot \|\mathbf{v}_\alpha\|}. \tag{5}$$

The information from GMO is expressed through an orthogonal decomposition of the fine-grained dynamics rather than skeletal joint coordinates. The quotient space representation $\mathcal{Q}$ is denoted as $\mathcal{Q}^i = \{p_1, \mathbf{v}_t, \mathbf{\Omega}^i\}$, where p_1 denotes the last frame of the observations. Since the variables in GMO represent the rotation of the human pose, we only need to obtain the modulus of the tangent vector in TSO.

Network Perceptual Module

Currently, our quotient space representation mitigates the challenges of high-dimensional stochasticity through three techniques: geometric dimensionality reduction, semantic decoupling, and dynamic constraints, thereby reducing the learning burden while improving data perception. However, existing models rely excessively on networks to implicitly capture spatio-temporal dependencies, but lack an actively guided explicit learning mechanism. Therefore, we design a network perceptual module (**NPM**) to construct an auxiliary supervision signal through adversarial perturbation (mask/noise injection), compelling the network to

actively repair spatiotemporal relationships. Specifically, we *firstly* design a *spatio-temporal enhancement component* (**SEC**) that induces active perception by selectively destroying the observed attitude coordinates to formulate ***the auxiliary learning task***. This design imposes two requirements on the modeling network: i) the ability to reconstruct masked coordinates by learning spatio-temporal relationships between corrupted and intact data, and ii) robustness to incomplete motion sequences induced by the masking process. For these, we introduce a *spatio-temporal learning component* (**SLC**), implemented as a generative adversarial transformer, which dynamically infers and adapts to complex motion patterns while recovering the perturbed structure.

Spatio-temporal enhancement component. In the auxiliary learning tasks, each coordinate x_t^i in the historical motion sequence is randomly masked with probability p_m. The objective is to reconstruct these masked coordinates from the observed unmasked values. In another way, Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma)$ is added to each coordinate x_t^i with probability p_n, where σ controls the noise intensity. The goal is to recover clean motion from these corrupted inputs. Let P, P_M, and P_D denote the original motion sequence, masked sequence, and noisy sequence. While all tasks share the same backbone network, they employ distinct prediction heads to specialize in their respective objectives.

Spatio-temporal learning component. To learn incomplete motion data, we process each coordinate as an independent feature while modeling spatio-temporal dependencies through a coordinate-level attention mechanism. We introduce a learnable masking token to explicitly mark masked coordinates, seamlessly integrating this masking awareness into the attention mechanism to handle arbitrary missing values. Given the inherent complexity of disentangling and fusing such spatio-temporal information, we leverage a WGAN for training, capitalizing on its demonstrated effectiveness for prediction tasks. SLC maps DPM data to an embedding space via a linear encoder.

Spatio-temporal attention. To capture the distinct patterns in joint spatial correlations and temporal dynamics, our SLC decouples these dependencies via dedicated attention mechanisms. The spatial attention module processes features within each timestamp: for features $H^t \in \mathbb{R}^{J\times D}$ at time t, we project them into queries Q^t, keys K^t, and values V^t via linear transformations $\delta(\cdot)$, then compute a masked spatial attention matrix $A_s \in \mathbb{R}^{J\times J}$. Following recent efficient attention designs, we adopt a low-rank approximation:

$$head^t = \varphi\left(Q^t\right)\left(\varphi(K)^T V^t \cdot A_s\right), \tag{6}$$

where $\varphi(\cdot)$ denotes a dimensionality-reducing nonlinear. The outputs of H are concatenated and processed by a feed-forward network $\phi(\|_{i=1}^{H}\text{head}_i^t)$. Conversely, the temporal attention module operates on joint-specific sequences: for features $H_j \in \mathbb{R}^{T\times D}$ of joint j, we derive Q_j, K_j, V_j analogously and compute weights $A_t \in \mathbb{R}^{T\times T}$:

$$head_j = \varphi\left(Q_j\right)\left(\varphi\left(K_j\right)^T V_j \cdot A_t\right). \tag{7}$$

Motion discriminator. We employs a WGAN-GP architecture to improve training stability. The adversarial system consists of: a generator based on a spatio-temporal deformable Transformer, and two specialized discriminators. The fidelity discriminator ensures pose realism, while the continuity discriminator preserves temporal coherence.

Inference Objective

To accommodate diverse task objectives (prediction, masking, and denoising), our framework applies targeted supervision to distinct sequence segments. Let M denote masked joints, and $\hat{p}$, $\hat{p}_M$, $\hat{p}_D$ represent outputs for prediction, masking, and denoising tasks respectively. The composite loss integrates task-specific terms:

$$\begin{aligned} L = \underbrace{\frac{1}{T_f J}\sum_{t,j}\|\hat{p}_j^t - p_j^t\|^2}_{L_{pred}} + \alpha_1 \underbrace{\frac{1}{|M|}\sum_{M}\|\hat{p}_{M,j}^t - p_j^t\|^2}_{L_{mask}} \\ + \alpha_2 \underbrace{\frac{1}{T_p J}\sum_{t,j}\|\hat{p}_{D,j}^t - p_j^t\|^2}_{L_{denoise}}. \end{aligned} \tag{8}$$

where α_1, α_2 balance task contributions. For adversarial training, we adopt WGAN-GP to minimize the Wasserstein distance between generated (P_g) and real (P_r) sequence distributions:

$$\begin{aligned} L_{adv} = E_{X'\sim P_g}\left[D\left(x'\right)\right] - E_{X\sim P_r}[D(x)] \\ + \lambda E_{x'\sim P_{x'}}\left[\left(\left\|\nabla_{x'} D\left(x'\right)\right\|_2 - 1\right)^2\right], \end{aligned} \tag{9}$$

where $\hat{X}$ interpolates real and generated data, and λ controls gradient penalty strength. This penalty enforces Lipschitz continuity, stabilizing GAN training. The total loss combines all terms:

$$L_g = \beta_1 L + \beta_2 L_{adv}. \tag{10}$$

Experiments

In this section, we evaluate the proposed method on large benchmark datasets. We seek to answer the following research questions. **Q1:** How is the proposed method comparing to state-of-the-art (SOTA) motion prediction approaches? **Q2:** How do the visual results of the proposed method compare to SOTA motion prediction methods? **Q3:** How much do different components of EPS contribute to its performance? Then, we first present the experimental settings, followed by answering the above research questions.

Datasets and Experimental Settings

Human 3.6 Million (**H3.6M**) dataset (Ionescu et al. 2014) contains 3.6 million human images recorded by a Vicon motion capture system. 7 subjects perform 15 different classes of actions. Following the evaluation protocol of previous work (Jain et al. 2016), duplicate points in poses are removed and downsampled to 25 FPS. S5 is utilized as the test set. CMU Motion Capture (**CMU Mocap**) dataset is released by researchers from Carnegie Mellon University. 12 infrared cameras are utilized to capture human poses.

Time (ms)	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
	Phoning						Eating						Purchases					
HMR	12.5	21.3	39.3	58.6	71.3	112.8	9.2	13.9	34.6	47.1	61.3	84.8	15.3	30.6	64.7	73.9	97.5	122.7
GA-MIN	8.3	17.8	37.9	44.8	63.0	101.5	**5.8**	12.5	25.3	33.8	47.3	65.2	12.4	28.5	60.0	72.9	89.9	135.2
FDU	7.8	17.2	37.5	47.3	65.1	96.7	6.3	13.7	29.1	36.3	49.0	71.1	11.8	27.2	**41.3**	52.1	94.8	130.7
AMHGCN	8.1	18.2	38.9	48.3	64.9	99.9	6.1	13.7	28.7	35.6	47.7	72.2	11.6	27.3	58.4	71.4	93.0	134.6
SAGGB	7.4	17.1	37.8	47.9	65.7	101.9	6.2	14.1	29.8	37.3	51.1	75.1	**10.9**	26.8	59.8	73.9	96.9	137.5
Ours	**6.7**	**12.8**	**30.9**	**39.5**	**58.1**	**78.7**	6.0	**12.1**	**21.2**	**27.1**	**40.3**	**57.3**	11.0	**23.8**	50.2	**50.5**	**72.1**	**105.8**
Time (ms)	80	160	320	400	560		80	160	320	400	560	1000	80	160	320	400	560	1000
	Directions						Sitting Down						Taking Photo					
HMR	23.3	25.0	47.2	61.5	80.9	116.9	9.6	18.6	41.1	57.7	101.7	148.3	7.9	19.0	31.5	57.3	83.5	108.5
GA-MIN	6.8	15.3	42.1	50.2	68.1	100.0	14.5	25.5	56.3	70.3	90.8	135.2	8.3	16.6	38.2	49.0	59.7	115.3
FDU	6.6	16.4	39.6	50.1	68.1	97.2	13.9	25.6	54.2	67.2	94.3	145.3	8.1	18.0	39.2	50.6	72.2	116.1
AMHGCN	6.4	16.3	39.4	49.9	67.5	100.7	13.3	26.8	55.6	69.2	93.8	146.0	8.1	18.3	41.1	52.1	71.8	114.2
SAGGB	**6.2**	16.0	39.0	50.0	70.6	101.8	12.8	26.3	55.9	70.3	96.9	150.2	7.8	17.9	41.3	52.9	77.3	118.6
Ours	6.3	**10.7**	**17.1**	**29.8**	**45.0**	**59.8**	**8.9**	**17.8**	**38.2**	**55.6**	**63.5**	**99.8**	**7.8**	**13.5**	**27.2**	**43.1**	**55.5**	**94.1**
Time (ms)	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
	Sitting						Posing						Greeting					
HMR	12.6	25.6	44.7	60.7	76.4	118.4	13.6	23.5	62.5	114.1	126.3	143.6	12.9	31.9	55.6	82.5	104.3	123.2
GA-MIN	8.1	18.5	41.9	53.2	70.2	109.1	7.8	19.3	43.4	56.0	83.2	150.2	12.8	26.3	61.8	75.8	95.3	121.5
FDU	8.7	18.9	42.1	53.2	72.3	114.5	**7.5**	19.3	47.1	62.0	93.3	149.5	13.0	30.7	63.1	78.2	109.4	141.8
AMHGCN	8.5	18.7	42.3	53.7	73.8	115.8	9.7	24.7	60.6	77.8	108.1	169.2	13.4	32.1	70.3	85.8	109.1	146.3
SAGGB	**8.1**	18.4	42.3	54.1	74.7	116.6	9.1	23.3	57.4	74.6	109.4	165.8	12.5	30.4	68.6	85.4	110.0	141.7
Ours	8.2	**16.1**	**37.2**	**48.9**	**78.5**	**108.4**	9.6	**18.8**	**45.2**	**50.0**	**65.1**	**108.2**	**9.2**	**21.3**	**40.1**	**63.2**	**73.3**	**102.3**
Time (ms)	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
	Waiting						Walking Dog						Average					
HMR	12.8	24.5	45.2	85.1	87.5	121.9	30.1	41.4	78.4	100.1	134.7	157.4	14.5	25.0	49.5	72.6	93.2	123.5
GA-MIN	7.5	17.2	41.1	52.3	70.7	102.2	18.9	38.5	70.9	84.0	104.1	138.2	10.1	21.5	47.2	58.4	76.6	115.8
FDU	8.2	18.4	41.3	52.1	70.0	116.1	14.5	32.7	63.8	76.0	94.6	123.1	9.7	21.6	45.3	56.8	80.3	118.4
AMHGCN	8.3	19.3	43.4	54.4	73.1	105.3	18.0	39.3	76.0	89.3	108.4	142.4	10.1	23.2	50.4	62.5	82.8	122.4
SAGGB	7.6	17.9	41.1	52.3	73.3	104.1	**16.0**	36.0	72.0	85.2	103.8	137.3	9.5	22.2	49.5	62.2	84.5	122.8
Ours	**7.3**	**15.8**	**38.1**	**49.5**	**65.4**	**92.6**	16.1	**32.1**	**63.0**	**70.4**	**98.4**	**109.2**	**8.8**	**17.7**	**37.1**	**48.0**	**65.0**	**92.4**

Table 1: Short-term and long-term prediction results on the H3.6M.

Following previous works (Mao et al. 2019), we adopt the same training/test splits. 3D Poses in the Wild (**3DPW**) dataset (von Marcard et al. 2018) is proposed primarily for wild scenes that are recorded by a handheld smartphone camera or IMU. It contains 60 video sequences with more than 51,000 indoor or outdoor poses. We build our model on the PyTorch with a NVIDIA 3090Ti GPU. The Adam Optimizer is utilized with a learning rate of 0.001. The model is trained for 15 epochs with a batch size of 16. The loss weight $\alpha_1 = \alpha_2 = 1$, $\beta_1 = 0.9$, and $\beta_2 = 0.1$. We evaluate our proposed approach by measuring the mean per joint position error (MPJPE) after the alignment of the root joint. Experimental results at 80 ms, 160 ms, 320 ms, 400 ms, and 1,000 ms in the future are shown for comparisons.

Comparisons with Existing Approaches (RQ1)

Results on H3.6M. We compare our method with SOTA approaches: AMHGCN (Li et al. 2024) and HMR (Liu et al. 2019) (RNN-based); FDU (Gao et al. 2023) (graph-based); GA-MIN (Wang et al. 2024b) (DCT-Transformer hybrid); and SAGGB (Wang et al. 2024c) (temporal-graph convolutional). These methods are retrained using the authors' official implementations to ensure fair evaluation. For a detailed introduction of the relevant works, please refer to the supplementary documents. As shown in Table 1, our method demonstrates substantial improvements across both prediction horizons. ***For short-term prediction*** (0-400 ms), APS achieves an average improvement of 7.3-46.2% over baselines, with particularly significant gains of 12.8-39.5% against SAGGB. The effectiveness in capturing spatio-temporal patterns is especially evident in complex motions, where APS reduce errors by 29.8-73.8% compared to SAGGB. ***For long-term prediction*** (400-1,000ms), our method maintains superior performance with 17.1-59.8% lower errors on average. Notably, we outperform SAGGB by 19.9-44.7% across all time horizons, with particular results (31.6-59.8%) for challenging sequences (*e.g.,* Directions). The consistent accuracy across different prediction windows confirms the robustness of APS in modeling long-term dependencies. Quantitative results demonstrate our method superior performance in both prediction horizons, with strong results for non-periodic motions (e.g., 59.8 mm vs 101.8 mm error for 'Directions' at 1,000 ms) compared to periodic ones (e.g., 109.2 mm vs 137.3 mm for 'Walking Dog'). This performance gap suggests our quotient space modeling is effective for complex motions.

Results on CMU Mocap. To verify the effectiveness and generalization of APS, experiments are carried out on the CMU Mocap, which has a more complex action performance than the H3.6M, and the results are detailed in Table 2. Existing advanced methods are compared with APS, including Res-Gru (Martinez, Black, and Romero 2017), FC-GCN (Cui, Sun, and Yang 2020), GA-MIN (Wang et al. 2024b), and DPnet (Tang et al. 2023c). It can be observed that our method performs optimally for ***the short-term***, and 11.2% (400 ms) in comparison to the DPnet. Compared to the DPnet, APS improves 23.5% (80 ms) and 11.2% (400 ms) on average prediction accuracy. ***In***

Time (ms)	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
	Basketball						Walking						Directing Traffic					
Res-GRU	18.4	33.8	59.5	70.5	75.5	106.7	8.2	13.7	21.9	24.5	67.2	94.2	15.2	29.6	55.1	66.1	182.3	127.1
FC-GCN	14.0	25.4	49.6	61.4	77.4	106.1	7.6	12.5	23.0	27.5	27.2	40.2	7.4	15.1	31.7	42.2	70.3	152.4
GA-MIN	10.3	19.8	40.3	51.8	—	88.8	5.2	8.9	16.2	18.2	—	26.2	5.7	10.8	27.2	33.4	—	137.8
DPnet	10.7	17.8	38.4	49.5	58.1	98.4	5.8	9.0	17.2	21.4	24.9	**34.1**	5.9	11.8	26.6	33.5	66.6	143.3
Ours	10.5	**16.8**	**33.2**	42.5	**55.1**	**82.4**	**5.1**	**7.9**	15.2	18.2	**23.1**	40.2	4.3	**9.2**	**18.1**	**27.3**	**40.6**	**99.1**
Time (ms)	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
	Soccer						Basketball Signal						Jumping					
Res-GRU	20.3	39.5	71.3	84.0	79.7	129.6	12.7	23.8	40.3	46.7	65.5	77.5	36.0	68.7	125.0	145.5	132.3	195.5
FC-GCN	12.1	21.8	41.9	52.9	82.6	117.5	3.5	6.1	11.7	15.2	25.3	53.9	22.4	44.0	87.5	106.3	131.4	164.6
GA-MIN	9.8	18.3	39.0	49.4	—	93.2	2.5	4.6	10.5	15.3	—	58.3	14.2	28.2	71.8	91.1	—	162.1
DPnet	9.0	17.1	35.8	48.7	87.1	115.0	2.6	**4.4**	10.0	13.4	30.1	61.2	12.4	28.3	70.2	89.2	**100.1**	166.1
Ours	**6.5**	**12.5**	**26.4**	**40.8**	**69.9**	**89.1**	**2.2**	4.9	**9.2**	**10.1**	**18.9**	**45.2**	**10.6**	**20.4**	**51.8**	**80.2**	109.1	**140.2**
Time (ms)	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
	Wash Window						Running						Average					
Res-GRU	8.4	15.8	29.3	24.5	85.3	102.7	25.8	48.9	88.2	100.8	124.5	158.1	18.1	34.2	61.3	70.3	101.5	123.9
FC-GCN	5.9	11.9	19.4	23.1	53.0	79.3	26.0	36.6	38.8	39.5	26.1	58.2	12.4	21.7	38.0	46.0	61.7	96.5
GA-MIN	4.5	9.9	27.8	35.2	—	69.2	17.5	22.3	22.1	26.1	—	40.1	8.7	15.4	31.9	40.1	—	84.5
DPnet	4.5	9.8	27.3	36.7	55.5	72.1	16.7	18.4	19.6	25.1	**28.3**	40.1	8.5	14.6	30.6	39.7	56.3	91.3
Ours	**4.2**	**8.2**	**24.3**	**32.5**	**44.4**	**58.2**	**8.3**	**13.1**	**24.6**	**30.1**	29.1	**38.8**	**6.5**	**11.6**	**25.4**	**35.2**	**48.8**	**74.2**

Table 2: Short-term and long-term prediction on the CMU MoCap.

Time (ms)	200	400	600	800	1000
Res-GRU	37.3	67.8	94.5	109.7	123.6
DMGNN	37.3	67.8	94.5	109.7	123.6
MSR-GCN	37.8	71.3	93.9	110.8	121.5
FDU	26.1	54.2	72.3	87.2	94.5
Ours	**19.0**	**47.5**	**67.8**	**79.3**	**89.3**

Table 3: Average prediction errors on 3DPW.

terms of long-term prediction, our method shows significant accuracy improvement of 18.7%-40.0%. The experimental results demonstrate that the spatio-temporal perception enhancement model can effectively capture the spatio-temporal dependence under longer pose sequences (400-1,000 ms). In addition, our method shows optimal prediction results on two major datasets, which further proves the effectiveness and robustness.

Results on 3DPW. The 3DPW poses significant challenges due to its capture of human motions in uncontrolled environments using handheld smartphones. As shown in Table 3, APS establishes new SOTA performance across all prediction horizons. Specifically, at 200 ms, our method achieves an error of 19.0 mm, a 27.2% improvement over FDU (26.1 mm). This advantage becomes more pronounced for longer-term predictions, reducing the error from 94.5 mm to 89.3 mm (5.5%) at 1,000 ms. Our method maintains consistent superiority, demonstrating 49.1% lower error than Res-GRU (123.6 mm) and 26.5% lower than MSR-GCN (121.5 mm) at 1,000 ms. With an average improvement of 10. 2% over FDU, these results validate the robustness in handling real-world motion variability in three datasets.

Further analysis. Our perceptual strategy achieves success through multiple synergistic advantages. By employing physically interpretable operators, it decouples implicit motion features from raw pose data, reducing the feature extraction burden on the network. The quotient space reduces high-dimensional pose space to low-dimensional equivalence classes, while its projection angle characteristics naturally align with anatomical planes to enable effective motion pattern separation. The differential modeling approach suppresses static displacement artifacts while maintaining dynamic sensitivity for a precise motion response. Through active joint information modification, the method enforces meaningful spatiotemporal relationship learning.

Visual Experiments (RQ2)

As illustrated in Figure 3, we evaluate APS on the H3.6M dataset and compare it with three SOTA approaches: DMGNN, HRI, and SAGGB. We select the "Photo" action for analysis, as it engages all four limbs, providing a comprehensive test of spatiotemporal feature capture. APS produces predictions that align more closely with ground-truth poses than competing approaches. Notably, this action involves subtle motions, which often cause models to prematurely converge to static states. Our approach maintains accurate motion dynamics over extended horizons, demonstrating superior spatiotemporal dependency modeling. This robustness comes from our method's ability to actively learn motion relations through controlled joint perturbation. Figure 3 further presents results on the CMU Mocap for the "Jump" action, characterized by large spatial displacements. For clarity, we overlay each predicted pose with a semi-transparent ground-truth reference. Prediction results from our method exhibit significantly tighter alignment with real sequences compared to alternatives. Crucially, our model accurately tracks subtle upper-body motions despite aggressive lower-limb movements, underscoring its resilience to amplitude variations. This empirically validates the effectiveness in decoupling and reconstructing motion patterns.

Ablation Experiments (RQ3)

To rigorously evaluate our components of APS, we conduct an ablation study on H3.6M (Table 4), analyzing three modules: the ***D***ata Perceptual Module (D), which processes motion in quotient space; the Spatio-Temporal ***E***nhancement

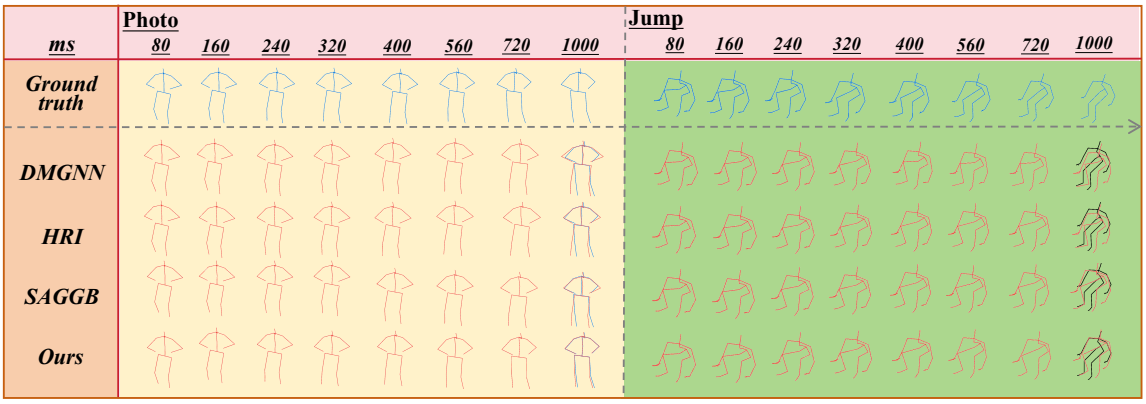

Figure 3: Visual results of our model on H3.6M and CMU datasets.

D	E	L	80	160	320	400	1000
	✓	✓	10.5	25.6	45.4	58.3	110.5
✓		✓	8.5	20.1	41.0	55.3	98.3
✓	✓		8.6	19.7	39.3	50.7	93.5
✓	✓	✓	8.8	17.7	37.1	48.0	92.4

Table 4: Ablation study results.

	Current methods					Current methods + APS				
Time (ms)	80	160	320	400	1000	80	160	320	400	1000
BiGAN	13.6	26.1	51.4	63.1	84.1	11.5	25.5	46.2	58.8	83.5
LTD	7.7	15.8	30.5	37.6	74.1	7.1	14.9	31.0	36.0	72.0
FDU	6.3	13.7	29.1	36.3	59.3	5.7	12.6	26.2	35.7	56.1
SAGGB	6.2	14.1	29.8	37.3	57.3	5.3	12.0	26.2	32.3	50.5

Table 5: Studies on model-agnostic learning.

Component (E), refining joint correlations; and the Spatio-Temporal ***L***earning Component (L), featuring our novel attention mechanism. Key findings emerge: Replacing quotient space with conventional 3D poses (removing D) degrades performance significantly ($e.g.$, 110.5 mm vs. 92.4 mm at 1,000 ms), confirming its efficacy in preserving dynamics while reducing dimensionality. Removing D (retaining E+L) causes notable drops, especially in medium-term predictions ($e.g.$, 20.1 mm vs. 17.7 mm at 160 ms), validating our joint perturbation-recovery strategy for capturing spatio-temporal relationships. Substituting L with a standard Transformer (keeping D+E) leads to consistent degradation, underscoring the superiority of our decoupled spatio-temporal attention over joint modeling approaches. The whole model (D+E+L) achieves optimal performance ($e.g.$, 92.4 mm at 1,000 ms), demonstrating the complementary roles of the modules. These results systematically validate our design choices, revealing how each component addresses distinct challenges in motion prediction.

Studies on Model-agnostic Learning

Our proposed APS is a model-agnostic framework, which can be easily integrated with existing methods. To demonstrate its effectiveness, we conduct experiments by utilizing various well-established methods such as LTD (Mao et al. 2019), DMGNN (Li et al. 2020), MSR-GCN (Li et al. 2021), and FDU (Gao et al. 2023). These methods represent different types of networks including RNNs, GCNs, and GANs, thereby covering a wide range of network classes. The results, as shown in Table 5, clearly indicate that APS significantly improves the prediction accuracy of these existing methods. Notably, APS mitigates the performance degradation typically observed in long-term predictions ($e.g.$, 1,000 ms). For instance, SAGGB+APS achieves a 12% lower error (50.5 mm vs. 57.3 mm) at 1,000ms, while FDU+APS reduces errors by 5.4% (56.1 mm vs. 59.3 mm). Crucially, the framework delays performance saturation by enhancing temporal coherence and error correction, which is especially vital for complex motion dynamics. These results validate APS as a universally applicable tool that not only boosts accuracy but also extends the effective prediction range of diverse architectures, addressing key limitations in HMP.

Conclusion

We introduce an Active Perceptual Strategy for 3D human motion prediction, addressing the issue through two key innovations: i) APS actively separates motion dynamics from coordinate redundancy via quotient space projection (Grassmann manifold and tangent vectors), compelling the network to focus on fundamental kinematic constraints and semantic motion. ii) We design an auxiliary training signal through adversarial perturbation, forcing the network to actively recover corrupted spatiotemporal relationships, thereby overcoming the limitations of passive data fitting. Our framework shifts the conventional paradigm to an active approach combining "disentangled representation" with "adversarial refinement". Extensive results on human motion datasets demonstrate the competency of our approach.

Acknowledgments

This work is supported by the National Key Research and Development Program, Unified Integrated Development technology of scientific computing language and engineering physics modeling language (Grant No.2024YFB3310200), and the Key R&D plan of Jilin Province (Grant No. 20250201076GX).

References

Cui, Q.; Sun, H.; and Yang, F. 2020. Learning Dynamic Relationships for 3D Human Motion Prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 6518–6526. Computer Vision Foundation / IEEE.

Gao, X.; Du, S.; Wu, Y.; and Yang, Y. 2023. Decompose More and Aggregate Better: Two Closer Looks at Frequency Representation Learning for Human Motion Prediction. In *CVPR 2023,*, 6451–6460.

Gu, B.; Tang, J.; Ding, R.; Liu, X.; Yin, J.; and Zhang, Z. 2024. April-GCN: Adjacency Position-velocity Relationship Interaction Learning GCN for Human motion prediction. *Knowl. Based Syst.*, 292: 111613.

Huang, B.; and Li, X. 2023. Human Motion Prediction via Dual-Attention and Multi-Granularity Temporal Convolutional Networks. *Sensors*, 23(12): 5653.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE TPAMI*, 36(7): 1325–39.

Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2016. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *CVPR 2016*, 5308–5317.

Li, J.; Wang, J.; Wu, L.; Wang, X.; Luo, X.; and Xu, Y. 2024. AMHGCN: Adaptive multi-level hypergraph convolution network for human motion prediction. *Neural Networks*, 172: 106153.

Li, M.; Chen, S.; Zhang, Z.; Xie, L.; Tian, Q.; and Zhang, Y. 2022a. Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. 18–36.

Li, M.; Chen, S.; Zhang, Z.; Xie, L.; Tian, Q.; and Zhang, Y. 2022b. Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. In *ECCV 2022*, volume 13666 of *Lecture Notes in Computer Science*, 18–36.

Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; and Tian, Q. 2020. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *CVPR, 2020*, 211–220.

Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. Multiscale Spatio-Temporal Graph Neural Networks for 3D Skeleton-Based Motion Prediction. *IEEE Trans. Image Process.*, 30: 7760–7775.

Li, P.; Pei, X.; Chen, Z.; Zhou, X.; and Xu, J. 2022c. Human-like motion planning of autonomous vehicle based on probabilistic trajectory prediction. *Appl. Soft Comput.*, 118: 108499.

Liu, Z.; Wu, S.; Jin, S.; Ji, S.; Liu, Q.; Lu, S.; and Cheng, L. 2022. Investigating Pose Representations and Motion Contexts Modeling for 3D Motion Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–16.

Liu, Z.; Wu, S.; Jin, S.; Liu, Q.; Lu, S.; Zimmermann, R.; and Cheng, L. 2019. Towards Natural and Accurate Future Motion Prediction of Humans and Animals. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 10004–10012.

Mao, W.; Liu, M.; and Salzmann, M. 2020. History Repeats Itself: Human Motion Prediction via Motion Attention. In *ECCV, 2020*, 474–489.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning Trajectory Dependencies for Human Motion Prediction. In *ICCV*, 9488–9496.

Martinez, J.; Black, M. J.; and Romero, J. 2017. On Human Motion Prediction Using Recurrent Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 4674–4683.

Salomão, A.; Andaló, F.; Prim, G. S.; Vieira, M. L. H.; and Romeiro, N. C. 2022. Case Studies of Motion Capture as a Tool for Human-Computer Interaction Research in the Areas of Design and Animation. In *Human-Computer Interaction. Theoretical Approaches and Design Methods*, volume 13302, 302–311.

Tang, J.; Sun, J.; Lin, X.; Zhang, L.; Zheng, W.; and Hu, J. 2023a. Temporal Continual Learning with Prior Compensation for Human Motion Prediction. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tang, J.; Zhang, J.; Ding, R.; Gu, B.; and Yin, J. 2023b. Collaborative Multi-Dynamic Pattern Modeling for Human Motion Prediction. *IEEE Trans. Circuits Syst. Video Technol.*, 33(8): 3689–3700.

Tang, J.; Zhang, J.; Ding, R.; Gu, B.; and Yin, J. 2023c. Collaborative Multi-Dynamic Pattern Modeling for Human Motion Prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8): 3689–3700.

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 2018*, 614–631.

Wang, X.; Cui, Q.; Chen, C.; and Liu, M. 2024a. GCNext: Towards the Unity of Graph Convolutions for Human Motion Prediction. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 5642–5650. AAAI Press.

Wang, Z.; Zhou, Y.; Zhang, N.; Yang, X.; Xiao, J.; and Wang, Z. 2024b. Existence Is Chaos: Enhancing 3D Human Motion Prediction with Uncertainty Consideration. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 5841–5849. AAAI Press.

Wang, Z.; Zhou, Y.; Zhang, N.; Yang, X.; Xiao, J.; and Wang, Z. 2024c. Existence Is Chaos: Enhancing 3D Human Motion Prediction with Uncertainty Consideration. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 5841–5849. AAAI Press.

Yao, J.; Li, C.; Sun, K.; Cai, Y.; Li, H.; Ouyang, W.; and Li, H. 2023. NDC-Scene: Boost Monocular 3D Semantic Scene Completion in Normalized Device Coordinates Space. In *ICCV 2023*, 9421–9431.

Yu, H.; Fan, X.; Hou, Y.; Pei, W.; Ge, H.; Yang, X.; Zhou, D.; Zhang, Q.; and Zhang, M. 2023. Toward Realistic 3D Human Motion Prediction With a Spatio-Temporal Cross-Transformer Approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10): 5707–5720.

Zhang, W.; Zhao, S.; Meng, F.; Wu, S.; and Liu, M. 2023. Dynamic Compositional Graph Convolutional Network for Efficient Composite Human Motion Prediction. In El-Saddik, A.; Mei, T.; Cucchiara, R.; Bertini, M.; Vallejo, D. P. T.; Atrey, P. K.; and Hossain, M. S., eds., *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 2856–2864. ACM.

Zhou, H.; Guo, C.; Zhang, H.; and Wang, Y. 2021. Learning Multiscale Correlations for Human Motion Prediction. In *ICDL 2021*, 1–7.

Zhou, Z.; and Wang, B. 2023. UDE: A Unified Driving Engine for Human Motion Generation. In *CVPR 2023*, 5632–5641.