Social Motion Prediction with Cognitive Hierarchies

Wentao Zhu [1*] **Jason Qin** [1*] **Yuke Lou** [1] **Hang Ye** [1]
Xiaoxuan Ma [1] **Hai Ci** [1] **Yizhou Wang** [1, 2]

[1] Center on Frontiers of Computing Studies, School of Computer Science, Peking University
[2] Institute for Artificial Intelligence, Peking University

Abstract

Humans exhibit a remarkable capacity for anticipating the actions of others and planning their own actions accordingly. In this study, we strive to replicate this ability by addressing the social motion prediction problem. We introduce a new benchmark, a novel formulation, and a cognition-inspired framework. We present **`Wusi`**, a 3D multi-person motion dataset under the context of team sports, which features intense and strategic human interactions and diverse pose distributions. By reformulating the problem from a multi-agent reinforcement learning perspective, we incorporate behavioral cloning and generative adversarial imitation learning to boost learning efficiency and generalization. Furthermore, we take into account the cognitive aspects of the human social action planning process and develop a cognitive hierarchy framework to predict strategic human social interactions. We conduct comprehensive experiments to validate the effectiveness of our proposed dataset and approach.

1 Introduction

Human beings are inherently social creatures. Concretely, individuals unconsciously anticipate the actions of others and make informed decisions about their own behaviors in social contexts [13, 19, 54], which enables individuals to cooperate and compete with others in a variety of scenarios, from pedestrian traffic [25, 44] to team sports [64, 50]. Notably, task experts demonstrate exceptional skill in predicting others' movements in advance [37, 70].

To better understand and replicate this ability, the research community has proposed the task of future prediction for multiple interacting agents given their historical observations. The majority of prior work in this area focuses on modeling and predicting agent interactions at the trajectory level [42, 39, 52, 5, 34, 58, 75, 59, 43, 72] and has demonstrated promising results in applications such as autonomous driving [38, 49, 17, 36, 16, 27]. However, trajectory-based approaches can only reflect coarse-grained interactions (*e.g.*, collision avoidance, social distancing) and fail to capture the rich and fine-grained human actions. To this end, some studies have investigated the *multi-person motion prediction* problem, which aims to forecast both trajectories and poses for a group of people [2, 3, 67, 68, 22, 63, 48, 73]. Despite recent advancements, there remain several critical challenges in this field. Firstly, existing multi-person motion datasets are primarily designed for human pose estimation tasks [65, 1, 31], and consequently, do not place particular emphasis on human interactions. Individuals in these datasets tend to move casually and interact with others at random, making future predictions both difficult and less meaningful. Secondly, the majority of prior methods concentrate on developing neural network architectures for end-to-end supervised training while overlooking the cognitive aspects of human social action planning. These two challenges are closely related and necessitate a comprehensive solution.

[*] Equal Contribution.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

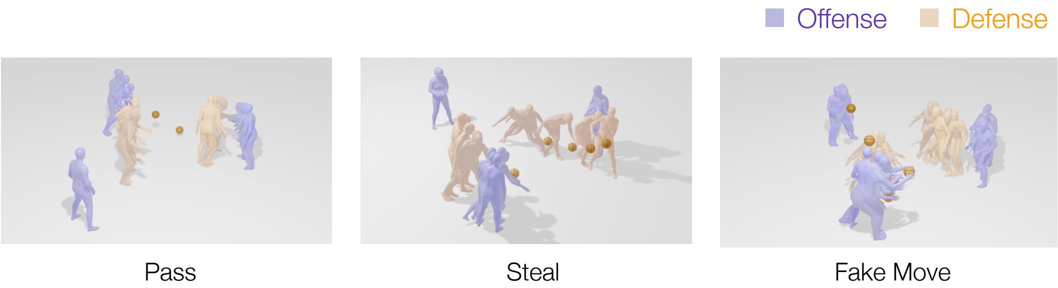


Figure 1: Example sequences from **Wusi** dataset. Three players in purple form the offensive team, and two players in orange form the defensive team. Left: a successful pass for the offensive team. Middle: a successful steal for the defensive team. Right: a successful pass following a fake pass.

In this work, we propose novel perspectives on this problem to address the aforementioned limitations. We begin by constructing a large-scale multi-person 3D motion dataset featuring intense and strategic interactions among participants. To achieve this, we turn to team sports, which offer several inherent advantages: 1) Well-defined game rules and global rewards implicitly constrain and guide individual actions. 2) Participants develop intricate interaction strategies based on their roles, such as planning cooperative actions with teammates while acting adversarially towards opponents. Skilled players even employ sophisticated techniques like deception, *etc*. 3) Human motions exhibit greater dynamism in terms of pose diversity and motion intensity, making motion prediction more challenging than in previous datasets.

Additionally, we present a new formulation of the multi-person motion prediction task as a multi-agent reinforcement learning (MARL) problem. Specifically, we model the task using imitation learning, where the objective is to learn a policy from expert demonstrations. We employ behavioral cloning (BC) [7, 51] to imitate expert behaviors from the dataset. To improve learning efficiency and generalization, we utilize generative adversarial imitation learning (GAIL) [26], aiming to render agents' policy indistinguishable from experts' policy. Furthermore, we propose a framework for human social interactions based on the cognitive hierarchy theory [10]. In particular, we assume that people base their decisions on their predictions regarding the likely actions of others, while others engage in similar decision-making processes from their perspectives. By considering the reasoning steps recursively, we posit that a level k agent takes actions based on the level $k-1$ agent actions, where k represents the depth of strategic thought. Building on this insight, we develop a computation model that can be elegantly integrated with the MARL formulation.

We summarize our contributions as follows: 1) We present **Wusi**, the first large-scale multi-human 3D motion dataset featuring intense and strategic interactions. We demonstrate that our dataset has greater motion diversity than existing datasets and poses a more significant challenge for the social motion prediction problem. 2) We propose a novel MARL formulation for the problem and develop an imitation learning baseline that combines behavioral cloning and generative adversarial imitation learning. 3) We introduce a cognitive hierarchy framework to model the strategic and game-theoretic human social interactions. Our approach outperforms the state-of-the-art methods in challenging long-term social motion predictions.

2 Related work

2.1 Multi-agent trajectory prediction

The multi-agent trajectory prediction problem has been extensively studied, especially for interacting traffic participants, *e.g*., pedestrians and vehicles. There exist multiple industrial benchmarks including Argoverse [12], nuScenes [8], Waymo Open Motion Dataset [18]. Meanwhile, various generation methods have been explored to model the interaction among different agents and predict their future trajectories [42, 39, 52, 5, 34, 58, 75, 59, 43, 72]. In this work, we aim to predict intense and strategic interactions for a group of humans with fine-grained body motions.

2.2 Multi-person motion dataset

The existing multi-human motion datasets can be categorized into 2D and 3D. The PoseTrack dataset [6, 29] offers video sequences with manually annotated 2D keypoints. However, 2D keypoints fail to accurately represent the proximity of individuals in the real world, and manual annotation is difficult to scale. Meanwhile, 3D multi-person motion datasets can be effectively constructed with motion capture (mocap) systems, including marker-based solutions and markerless ones. 3DPW [65] uses the inertial measurement units (IMUs) to obtain high-quality motion reconstructions. UMPM [61] employs a set of reflective markers to identify the joint positions. However, wearable sensors may struggle to capture complex movements and become expensive as the number of individuals increases. Other datasets, *e.g.*, CMU-Mocap [1], Panoptic [31], MuPoTs-3D [41], ExPI [22], employ markerless mocap systems with multi-view cameras to address the occlusion problem and obtain 3D human motion through triangulation. In this work, we construct a multi-person motion dataset featuring high motion diversity and intense social interactions using multi-view cameras. We provide a comparison with previous datasets in Section 3.3.

2.3 Multi-person motion prediction

Joo *et al.* [32] propose to predict human motion conditioned on other individuals' kinesic signals in a triadic haggling scenario. Adeli *et al.* [2] introduce the 2D social motion forecasting (SoMoF) benchmark, which aims to predict multi-person trajectory and pose on the PoseTrack dataset [6]. They also propose a baseline model using a shared GRU encoder and a pooling layer to incorporate social clues. TRiPOD [3] employs attention graphs to characterize the spatiotemporal social interactions. It also considers the joint occlusion and body invisibility issues arising from 2D observations. Futuremotion [67] provides a simple baseline using no social context, yet achieving competitive performance on SoMoF. As 2D multi-person motion prediction faces challenges such as data scarcity, depth ambiguity, and occlusions, more recent research has shifted towards exploring the multi-person motion prediction problem in 3D. Wang *et al.* [68] propose a multi-range Transformer to separately encode the local and global motion history. Guo *et al.* [22] design a cross-interaction-attention (XIA) module to model close interactions between pairs in duo dance scenarios. Vendrow *et al.* [63] introduce joint-aware attention and joint-wise query to predict the entire future sequence without recurrence. Peng *et al.* [48] present a social-aware motion attention mechanism that models both inter- and intra-individual motion relations. DuMMF [73] frames the problem as a dual-level generative task and instantiates it with various generative models. In contrast to previous work, we adopt a MARL perspective and model the strategic social interaction process using cognitive hierarchies.

2.4 Cognitive hierarchy theory

Cognitive hierarchy theory (CHT) is a model in behavioral economics and game theory that aims to describe human decision processes in strategic games. Several foundational works introduce the concept of CHT [10], and provide experimental evidence [45, 14, 23, 57, 56] supporting the existence of cognitive hierarchies in human decision-making processes. Taking it a step further, Wright *et al.* [71] develop a Poisson cognitive hierarchy model to predict human behavior in normal-form games. Li *et al.* [35] summarize how the information structure guided by the cognitive hierarchy supports belief generation and policy generation in game-theoretic multi-agent learning. In this work, we propose a cognitive hierarchy framework in conjunction with MARL, and demonstrate its effectiveness in a real-world social motion prediction scenario.

3 Dataset

3.1 Overview

We present a multi-person 3D motion dataset with a special focus on strategic interactions called **`Wusi`** (Wusi Basketball Training Dataset). In the following, we first introduce the contents of the dataset, then outline the data collection pipeline, and finally provide statistical analysis and comparison with regard to existing datasets.

Our dataset captures the no-dribble-3-on-2 basketball drills performed by a team of professional basketball athletes. In each drill, three offensive players possess the ball, while another two players

Table 1: Dataset comparison. We compare our dataset with existing multi-person motion datasets employed by previous works on the multi-person motion prediction task. † denotes multi-person subset as utilized in previous works [2, 68].

Dataset	2D/3D	Frames	Duration (min)	No. of people	Interaction
PoseTrack [6]†	2D	8K	5.5	Multiple	Weak
CMU-Mocap [1]†	3D	34K	4.9	2	Weak
3DPW [65]†	3D	5K	6.1	2	Weak
MuPoTs-3D [41]	3D	8K	4.4	2-3	Weak
ExPI [22]	3D	30K	20.0	2	Cooperative
Wusi	3D	60K	40.3	5	Strategic

are on defense. The offensive team aims to accomplish as many successful passes as possible within the given time, while the defensive team strives to get steals, deflections, and slow down the offense. Since dribbling is prohibited, the drill requires the offensive players to make better passing decisions and foresee the defense before they pass; conversely, the defensive players need to anticipate the passing directions in order to steal the ball. Figure 1 showcases the diverse and dynamic human interactions present in our dataset. On the left, the offensive player successfully passes the ball to a teammate. In the middle, the defensive player accurately anticipates the passing trajectory and successfully steals the ball. On the right, the offensive player pretends to pass the ball to the right, then swiftly executes a pass to the left.

3.2 Data collection

Our motion capture system consists of 11 synchronized and calibrated wide-baseline cameras. Our setup ensures that the players are well-surrounded by the cameras from elevated shooting angles, enabling each body joint to be covered by at least four different camera views. We utilize a markerless multi-person 3D pose estimation algorithm [74] with JDETracker [69]. Subsequently, we apply a post-processing pipeline that includes automatic failure detection and temporal filtering [11]. A final manual check is conducted to guarantee the integrity of the data sequences.

3.3 Data analysis

In this section, we examine the statistics of **Wusi** dataset and compare it with existing datasets for multi-person motion prediction. As outlined in Table 1, current datasets exhibit two primary limitations for the multi-person motion prediction task. Firstly, high-quality 3D motion data is limited in terms of scale (duration and the number of people). Consequently, prior works [68, 48, 63, 73] resort to randomly mixing the motion sequences [1, 65, 41] in order to generate more multi-person motion sequences for training. However, such practice inevitably causes unnatural human interactions. Secondly, the interaction strength in existing datasets is restricted, mainly featuring simple actions such as walking together [6, 1, 65, 41]. More recently, the ExPI dataset [22] introduces cooperative 3D motions of two Lindy-hop dancers. In contrast, our dataset offers 2 to 9 times the video duration of existing datasets and includes strategic global interactions among 5 individuals.

Table 2: Comparison of pose diversity based on different thresholds.

Threshold	50mm	100mm
Human3.6M [28]	24%	12%
ExPI [22]	52%	23%
CMU-Mocap [1]	20%	9%
MuPoTs-3D [41]	37%	19%
Wusi	**53**%	**27**%

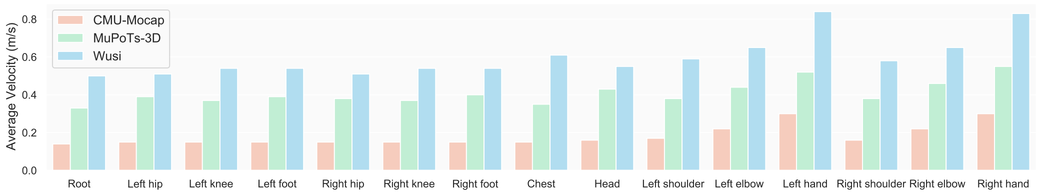


Figure 2: Comparison of motion intensity.

Additionally, we provide a quantitative comparison with existing datasets, emphasizing two critical aspects: *pose diversity* and *motion intensity*. Firstly, we examine the pose diversity within the datasets by calculating the ratio of unique poses to the total number of poses following [28, 22]. Table 2 shows that our dataset surpasses the previous datasets in terms of overall pose diversity. We further analyze the motion intensity by computing the average velocity for all the body joints. As illustrated in Table 2, our dataset exhibits more dynamic movements across all body joints. We suppose that these characteristics make **`Wusi`** not only challenging for motion prediction, but also potentially useful for other related tasks, *e.g.* motion generation [76].

4 Method

4.1 Formulation

The multi-person motion prediction problem can be formulated as follows: given the motion histories $\{x_t^p\}_{1\leq t\leq T, 1\leq p\leq P}$ of length T from P human subjects, predict their future motion $\{x_t^p\}_{T\leq t\leq T+T', 1\leq p\leq P}$ of length T', where each x_t^p represents a 3D human pose.

We model the problem using a Markov Decision Process (MDP) [60] by dividing the prediction sequence into L steps with even step length $m = \frac{T'}{L}$ following [66], and further extend the formulation to multiple agents. We use bold symbols to represent the collective variables of all agents following the convention in MARL. At each step i, The state $\boldsymbol{s}_i$ is defined as the aggregated motion history $\{x_t^i\}_{1\leq t\leq T+(i-1)\times m, 1\leq p\leq P}$ for all the agents. The action for the p-th agent a^p is defined as a sequence of velocities $\{v_t^p\}_{T+(i-1)\times m\leq t\leq T+i\times m-1}$, where $v_t^p = x_{t+1}^p - x_t^p$. Therefore, the joint action of all P agents $\boldsymbol{a}_i$ deterministically transition the MDP into the new state $\boldsymbol{s}_{i+1}$ according to:

$$x_{t^*}^p = x_{T+(i-1)\times M}^p + \sum_{t=T+(i-1)\times M}^{t^*} v_t^p, \quad T+(i-1)\times m+ \leq t^* \leq T+i\times m, 1\leq p\leq P. \tag{1}$$

The goal of this objective is to learn a policy π^p for each agent p that maps each possible state $\boldsymbol{s}$ to agent action a^p by $a^p = \pi^p(\boldsymbol{s})$. In the real world, people optimize their π^p to maximize an implicit team reward r. Since we have no access to r or the environment, we aim to learn π^p from expert demonstrations $\mathcal{D}$ via imitation learning.

We parameterize π^p with a set of parameters θ. In practice, we follow [68] to employ a global-range Transformer E_g to encode global state feature $\boldsymbol{s}_g = E_g(\boldsymbol{s})$, and a local-range Transformer E_l to encode local state feature $s_l^p = E_l(\boldsymbol{s})$ for each agent p. See Figure 3 for an overview. This approach helps to extract multi-level state features and enables all agents to share policy network parameters.

4.2 Behavioral Cloning

A straightforward method in imitation learning is Behavioral Cloning (BC) [7, 51], *i.e.*, using expert demonstrations to minimize the difference between the actions produced by the policy and those taken by the experts:

$$\theta^* = \text{argmin}_\theta \sum_{\mathcal{D}} \sum_{p=1}^{P} \ell\left(\pi_\theta^p\left(\boldsymbol{s}\right), \tilde{\pi}^p\left(\boldsymbol{s}\right)\right), \tag{2}$$

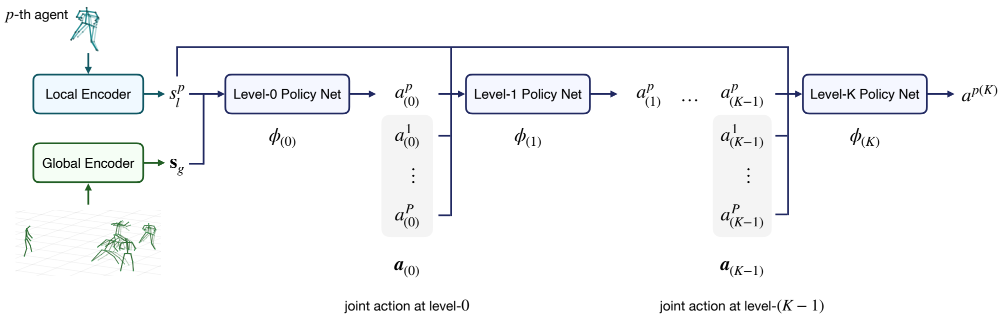


Figure 3: Framework overview. For the p-th agent, two state encoders first extract its local and global state features s_l^p and $\boldsymbol{s}_g$, from which the level-0 policy network produces an initial action $a^p_{(0)}$. The level-k policy network produces action $a^p_{(k)}$ based on s_l^p and the joint actions of the previous level $\boldsymbol{a}_{(k-1)}$ ($k \geq 1$).

where s is sampled from expert demonstrations $\mathcal{D}$, π^p_θ and $\tilde{\pi}^p$ are policies of the model and the expert for agent p, ℓ is a distance function defined in the action space. While BC enjoys the benefit of being both computation-efficient and sample-efficient [53], it also faces several drawbacks. For instance, policies tend to be overfitted with respect to the state distribution encountered by the experts, leading to suboptimal generalization ability [66, 30]. To cope with this challenge, we further introduce GAIL and cognitive hierachy reasoning.

4.3 Generative Adversarial Imitation Learning

Generative Adversarial Imitation Learning (GAIL) [26] is a model-free imitation learning approach. It leverages the adversarial training framework of Generative Adversarial Networks (GAN) [21] and the concept of inferring reward functions from Inverse Reinforcement Learning (IRL) [46]. Direct estimation of reward signals from expert demonstrations can be very difficult; therefore, GAIL transforms the IRL problem into its equivalent dual problem of occupancy measure matching. Specifically, the policy network is regularized to match its state-action pair distribution to that of the experts' policy through adversarial training. We implement GAIL using a global discriminator D parameterized by ω, which differentiates the state-action pairs produced by agents' joint policy $\boldsymbol{\pi}$ and experts' joint policy $\tilde{\boldsymbol{\pi}}$. The optimization objective can be formulated as:

$$\min_\theta \max_\omega \mathbb{E}_{\boldsymbol{\pi}_\theta} \left[\log D_\omega \left(\boldsymbol{s}, \boldsymbol{\pi}_\theta \left(\boldsymbol{s}\right)\right)\right] + \mathbb{E}_{\tilde{\boldsymbol{\pi}}} \left[\log \left(1 - D_\omega \left(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{\pi}} \left(\tilde{\boldsymbol{s}}\right)\right)\right)\right], \tag{3}$$

where $\boldsymbol{s}$ and $\tilde{\boldsymbol{s}}$ are sampled from expert demonstrations $\mathcal{D}$ independently.

4.4 Cognitive hierarchies

Furthermore, we notice that in real-world situations, individuals would predict others' behaviors by forming beliefs about others' policies and then acting accordingly to maximize their own payoffs [20, 9]. This can be formulated as a hierarchical, game-theoretic decision-making process, in which lower-level agents adopt straightforward strategies, while higher-level players anticipate the strategies of lower-level agents and respond accordingly. We explicitly model this recursive reasoning process to learn more interpretable and robust agent policies. We adopt a specific type of cognitive hierarchy model [10] called *level-k thinking* [57, 45, 15] to represent this process. Specifically, agents at each level (except for the lowest one) assume that others are reasoning at the previous level. As shown in Figure 3, a straightforward policy is to take actions based on local and global state features s_l^p and $\boldsymbol{s}_g$, which we denote as level-0 policy $\pi^p_{(0)}$. Level-0 actions are thus defined as:

$$a^p_{(0)} = \pi^p_{(0)}(\boldsymbol{s}) = \phi_{(0)}(s_l^p, \boldsymbol{s}_g), \tag{4}$$

where $\phi_{(0)}$ is the level-0 policy network. Then for a level-k agent ($k \geq 1$), it takes actions based on the joint agent actions $\boldsymbol{a}_{(k-1)}$ at the previous level and its local state features s_l^p:

$$a^p_{(k)} = \pi^p_{(k)}(\boldsymbol{s}) = \phi_{(k)}(s^p_l, \boldsymbol{a}_{(k-1)}), \quad 1 \leq k \leq K \tag{5}$$

where $\phi_{(k)}$ is the level-k policy network, K is the maximum strategic depth.

4.5 Training objectives

Finally, we introduce our training objectives which effectively integrate the aforementioned components. For level-k ($1 \leq k \leq K$) agents, we apply GAIL to regularize the distance between their joint policy and the experts' policy:

$$\mathcal{L}_{\text{GAIL}} = \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\pi}_{(k)}} \left[\log D\left(\boldsymbol{s}, \boldsymbol{\pi}_{(k)}\left(\boldsymbol{s}\right)\right)\right] + \mathbb{E}_{\tilde{\boldsymbol{\pi}}}\left[\log\left(1 - D\left(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{\pi}}\left(\tilde{\boldsymbol{s}}\right)\right)\right)\right]. \tag{6}$$

In addition, we employ BC on level-K agent policies:

$$\mathcal{L}_{\text{BC}} = \sum_{\mathcal{D}} \sum_{p=1}^{P} \ell\left(\pi^p_{(K)}\left(\boldsymbol{s}\right), \tilde{\pi}^p\left(\boldsymbol{s}\right)\right). \tag{7}$$

We train the discriminator D and the policy networks $\boldsymbol{\phi}$ alternatively, where D aims to maximize $\mathcal{L}_{\text{GAIL}}$, and $\boldsymbol{\phi}$ is optimized to minimize a linear combination of $\mathcal{L}_{\text{BC}}$ and $\mathcal{L}_{\text{GAIL}}$ using a policy gradient algorithm following [30]:

$$\mathcal{L}_{\boldsymbol{\phi}} = \mathcal{L}_{\text{BC}} + \lambda\mathcal{L}_{\text{GAIL}}, \tag{8}$$

where λ is a constant for balancing the loss terms.

5 Experiments

5.1 Setup

Implementation details. We implement the presented framework to train and test on the proposed **Wusi** dataset. We employ Transformer encoder [62] for both the local and global state encoders, as well as Transformer decoders for the policy networks. Each Transformer consists of 3 layers with 8 attention heads. We share parameters for policy networks $\phi_{(1)} \ldots \phi_{(K)}$. We set the strategic reasoning depth $K = 3$ unless otherwise stated. For all the methods, we provide 1s motion history and predict future 1s motion. For additional experimental details, please refer to the appendix.

Evaluation metrics. We compute the Mean Per Joint Position Error (MPJPE) between the predicted future motion and ground-truth. In order to disentangle the results of trajectory prediction and pose prediction, we further calculate the mean root position error and mean local pose error (MPJPE after root alignment). We report all the errors in millimeters.

5.2 Evaluation and comparison

In this section, we evaluate the proposed dataset and method, comparing with the prior works. We conduct experiments with the following baseline methods including a naive baseline and multiple state-of-the-art approaches: 1) *Frozen* is a naive baseline that simply replicates the last frame of the input motion. 2) *Social-STGCNN* [42] is a multi-agent trajectory prediction method using spatio-temporal graph convolutional networks. For trajectory prediction methods, we train and test using the root (global) trajectories and replicate the last frame for the root-relative (local) pose. 3) *SocialVAE* [72] is a multi-agent trajectory prediction method using timewise variational autoencoder. 4) *History Repeats Itself (HRI)* [40] is a single-person motion prediction method that allows absolute coordinate inputs. 5) *Multi-Range Transformers (MRT)* [68] predicts multi-person motion utilizing a local-range encoder and a global-range encoder. 6) *Social Motion Transformer (SoMoFormer)* [63] employs the joints of all people as queries and predicts their future motion in parallel.

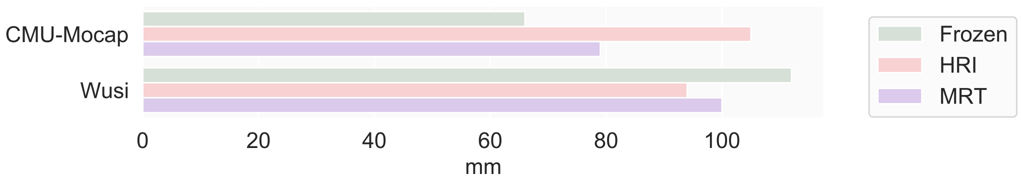


Figure 4: Comparison of baseline methods on CMU-Mocap [1] and our dataset. We compute the 1s mean local pose prediction errors. The baseline methods are trained and tested on two datasets separately.

Table 3: Performance comparison with the baseline methods. We compute the mean prediction errors of global pose, local pose, and root, respectively.

milliseconds	Global				Local				Root			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
Frozen	119.2	165.7	207.3	244.5	62.8	82.8	98.9	112.1	99.8	139.3	175.0	207.6
Social-STGCNN [42]	188.2	246.4	298.8	346.5	62.8	82.8	98.9	112.1	173.1	225.6	273.7	317.9
SocialVAE [72]	84.0	127.2	171.7	215.1	62.8	82.8	98.9	112.1	52.6	89.7	130.3	170.8
HRI [40]	**50.0**	106.8	145.5	189.2	**39.5**	62.9	78.4	94.1	**40.1**	88.0	119.6	157.7
MRT [68]	66.9	103.2	140.2	176.4	49.4	68.8	85.4	99.5	50.6	79.0	109.3	140.1
SoMoFormer [63]	53.2	88.8	124.9	160.0	42.3	62.6	79.7	93.9	42.5	70.7	100.8	131.2
Ours	54.6	**86.2**	**119.3**	**152.5**	43.7	**60.8**	**74.6**	**86.6**	41.7	**66.9**	**94.8**	**124.0**

5.2.1 Quantitative results

We first benchmark three baseline methods to assess the motion prediction difficulty of our dataset in comparison to the commonly-used existing dataset [1]. Figure 4 reveals that for the CMU-Mocap dataset [1], merely repeating the last-frame pose (*Frozen*) leads to significantly lower errors than the model predictions. This finding indicates that the motions in this dataset are relatively "static", making it less challenging and informative for the social motion prediction task. In contrast, the dynamic nature of our dataset gives rise to considerably higher prediction errors of the naive *Frozen* baseline. The results demonstrate that the proposed dataset serves as a more convincing benchmark for social motion prediction and introduces more challenges to the field.

Furthermore, we conduct a quantitative evaluation of the proposed framework's performance and compare it with state-of-the-art approaches. As illustrated in Table 3, our method achieves competitive results with SOTA approaches. It is worth noting that the single-person method HRI [40] demonstrates respectable short-term motion prediction accuracy; however, its performance over longer horizons is constrained by the absence of global context, particularly in root trajectory prediction. In contrast, our approach significantly surpasses SOTA methods in long-term social motion prediction, highlighting its effectiveness in modeling long-range human interactions. In the subsequent section, we will demonstrate the critical role that social context plays in predicting long-term human motion via examples.

5.2.2 Qualitative results

In order to better understand the performance improvement of our method, we visualize its motion prediction results and compare with the baseline methods qualitatively. Figure 5 illustrates an representative result in which the input motion shows that a defensive player (highlighted) is switching the target of defense. Our approach correctly predicts its future direction and orientations as the offensive player at the other side will likely be covered by the teammate. In contrast, other approaches fail to reason this, predicting that the player would continue to move along the input trajectory or simply stop at the middle.

In Figure 6, three players to be noticed are highlighted in rose, purple and blue (we also call them this way below). Additionally, we visualize the actions produced by the intermediate policy networks to

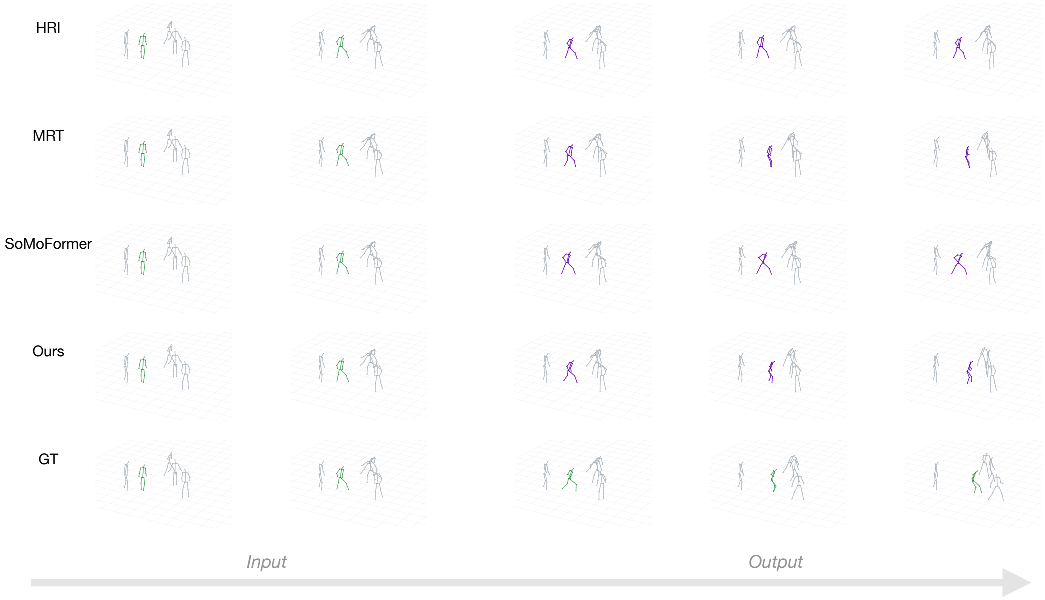


Figure 5: Qualitative comparison with existing methods. Left two columns are past motion and right three columns are future motion. We highlight the prominent player with bright colors. Real motion are shown in green, and model predictions are shown in purple.

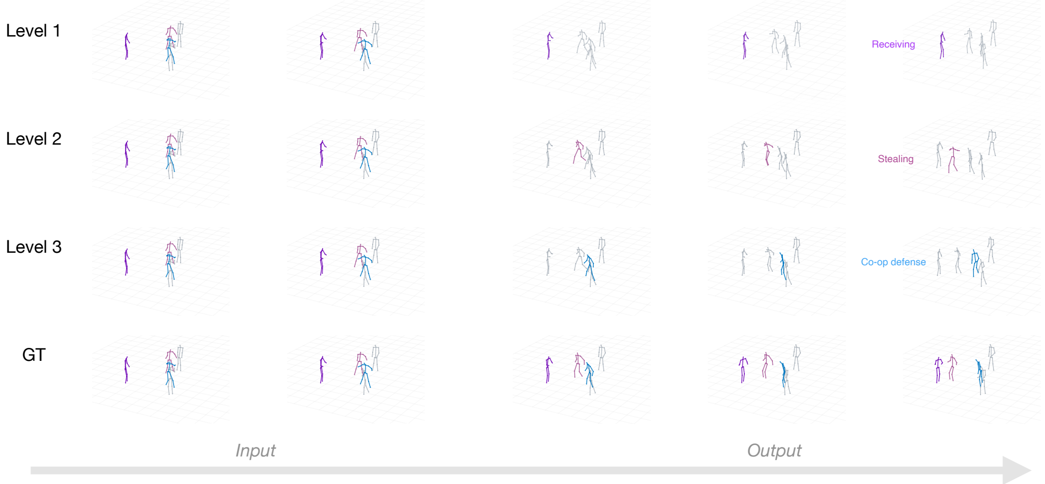


Figure 6: Actions produced by policy networks at different levels. Left two columns are past motions and right three columns are future motions.

interpret the decision-making process. Shown on the top row, player *purple* at level 1 policy network simply follows the past trend to receive the ball. As displayed in the second row, defender *rose* at level 2 gains knowledge from the first level that player *purple* is receiving the ball, and implements a steal attempt. Nevertheless, such aggressive actions could create defensive gaps. Defender *blue* at level 3 senses such gaps caused by defender *rose* from level 2, therefore executes cooperative defense upwards to fill the gap.

5.3 Ablation study

Finally, we conduct ablation studies to investigate the influence of each component. In Table 4 (a)-(e), we change the policy network depth K. $K = 0$ corresponds to producing the final action directly without any recursive reasoning. We find that introducing the recursive reasoning process ($K > 0$) clearly lowers the prediction errors, proving the effectiveness of our framework design based on cognitive hierarchy. From (f) to (h), we keep $K = 3$ and study the influence of training regularizations. (f) removes the GAIL regularization for intermediate levels $1 \ldots K-1$, and (g)

Table 4: Ablation study of cognition levels and architecture designs.

milliseconds	Global				Local				Root			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
(a) $K = 0$	60.1	94.3	129.2	163.7	47.4	66.2	82.0	95.3	45.7	72.7	101.8	131.8
(b) $K = 1$	55.2	88.0	121.9	155.5	44.3	62.1	77.0	89.7	42.3	68.0	96.3	125.4
(c) $K = 2$	54.8	87.2	120.7	154.2	43.5	60.6	75.0	87.3	**41.7**	67.6	95.9	125.3
(d) $K = 4$	**54.5**	86.3	119.7	153.5	43.6	60.5	74.7	86.6	42.0	67.3	95.5	125.2
(e) $K = 3$, full	54.6	**86.2**	**119.3**	**152.5**	43.7	60.8	74.6	86.6	41.7	**66.9**	**94.8**	**124.0**
(f) *w/o* mid GAIL	54.9	87.5	121.4	155.4	43.7	**60.4**	**74.5**	**86.4**	41.8	67.8	96.6	126.9
(g) *w/o* GAIL	54.6	87.1	121.1	155.1	**43.4**	60.4	74.7	86.8	41.9	68.1	97.0	127.0
(h) *w/o* weight sharing	57.1	91.0	125.8	160.2	44.4	62.2	77.1	89.7	43.6	70.3	99.5	129.5

Table 5: Quantitative comparison of actions produced by policy networks at different levels.

milliseconds	Global				Local				Root			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
Level-1	76.8	122.4	170.3	217.7	50.1	69.0	85.2	99.1	61.4	100.4	143.2	186.9
Level-2	61.0	99.9	144.6	192.5	47.1	66.0	82.5	96.6	46.5	78.0	116.2	158.8
Level-3	**54.6**	**86.2**	**119.3**	**152.5**	**43.7**	**60.8**	**74.6**	**86.6**	**41.7**	**66.9**	**94.8**	**124.0**

completely removes GAIL. Training without GAIL increases the long-term trajectory prediction errors, while slightly lowers the local pose prediction error. We also observe that models without GAIL tend to produce motions that appear less natural (e.g. limb twisting), leading to inferior visual effects. In (h), we use different policy network instances for levels $1 \dots K$, and it turns out that sharing the policy networks leads to better generalization.

In addition to the qualitative visualization of actions generated by policy networks at different reasoning levels, we also provide a quantitative comparison of their distance from ground-truth. Table 5 demonstrates that policy networks at different levels learn different actions, and policy networks at higher reasoning levels produce actions with less prediction errors. We hypothesize that, through the proposed recursive reasoning process, policy networks at different levels display a range of cognitive types, with behavior spanning from relatively naive to substantially rational. Higher-level policies more closely approximate an equilibrium strategy, which is usually in alignment with expert demonstrations. We provide several example motion sequences in the video demo to further illustrate this observation.

6 Conclusion

In this work, we propose the first large-scale multi-person 3D motion dataset featuring strategic human social interactions. This dataset surpasses the existing ones in scale, diversity, dynamics, and interaction, thus posing new challenges to the social motion prediction problem. We reformulate the problem using a MARL perspective. Building on this, we propose a framework that effectively combines behavioral cloning (BC), generative adversarial imitation learning (GAIL), and cognitive hierarchy. Our approach demonstrates strong generalization capabilities and improved interpretability for modeling strategic human social interactions.

Limitations: Our approach is based on the cognitive hierachy assumption and may not hold true for all type of social interactions. This work do not explicitly model social roles, which can be explored as future work.

Acknowledgment This work was partially supported by National Key R&D Program of China (2022ZD0114900). We thank Xuesong Fan and Rujie Wu for their generous support during the data collection process.

References

[1] Cmu graphics lab motion capture database. `http://mocap.cs.cmu.edu/`.

[2] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020.

[3] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021.

[4] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[5] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[6] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018.

[7] Michael Bain and Claude Sommut. A framework for behavioural cloning. *Machine intelligence*, 15(15):103, 1999.

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[9] Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press, 2011.

[10] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

[11] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1 filter: A simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 25272530, New York, NY, USA, 2012. Association for Computing Machinery.

[12] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019.

[13] Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

[14] Miguel A Costa-Gomes and Vincent P Crawford. Cognition and behavior in two-person guessing games: An experimental study. *American economic review*, 96(5):1737–1768, 2006.

[15] Vincent P Crawford and Nagore Iriberri. Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770, 2007.

[16] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2090–2096. IEEE, 2019.

[17] Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1468–1476, 2018.

[18] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[19] Chris D Frith and Uta Frith. The neural basis of mentalizing. *Neuron*, 50(4):531–534, 2006.

[20] Drew Fudenberg, Fudenberg Drew, David K Levine, and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[22] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022.

[23] Ernan Haruvy, Dale O Stahl, and Paul W Wilson. Evidence for optimistic and pessimistic behavior in normal-form games. *Economics Letters*, 63(3):255–259, 1999.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Dirk Helbing, Péter Molnár, Illés J Farkas, and Kai Bolay. Self-organizing pedestrian movement. *Environment and planning B: planning and design*, 28(3):361–383, 2001.

[26] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

[27] Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. *arXiv preprint arXiv:2303.05760*, 2023.

[28] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, 2014.

[29] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2020, 2017.

[30] Rohit Jena, Changliu Liu, and Katia Sycara. Augmenting gail with bc for sample efficient imitation learning. In *Conference on Robot Learning*, pages 80–90. PMLR, 2021.

[31] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[32] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10873–10883, 2019.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[34] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, S Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *arXiv preprint arXiv:1907.03395*, 2019.

[35] Tao Li, Yuhan Zhao, and Quanyan Zhu. The role of information structures in game-theoretic multi-agent learning. *Annual Reviews in Control*, 2022.

[36] Xin Li, Xiaowen Ying, and Mooi Choo Chuah. Grip: Graph-based interaction-aware trajectory prediction. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3960–3966. IEEE, 2019.

[37] Florian Loffing and Rouwen Cañal-Bruland. Anticipation in sport. *Current opinion in psychology*, 16:6–11, 2017.

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

[39] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European Conference on Computer Vision*, pages 759–776. Springer, 2020.

[40] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 474–489. Springer, 2020.

[41] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[42] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020.

[43] Abduallah Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian Claudel. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 463–479. Springer, 2022.

[44] Mehdi Moussaïd, Niriaska Perozo, Simon Garnier, Dirk Helbing, and Guy Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS one*, 5(4):e10047, 2010.

[45] Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326, 1995.

[46] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[48] Xiaogang Peng, Yaodi Shen, Haoran Wang, Binling Nie, Yigang Wang, and Zizhao Wu. Somoformer: Social-aware motion transformer for multi-person motion prediction. *arXiv preprint arXiv:2208.09224*, 2022.

[49] Pongsathorn Raksincharoensak, Takahiro Hasegawa, and Masao Nagai. Motion planning and control of autonomous driving intelligence system based on risk potential optimization framework. *International Journal of Automotive Engineering*, 7(AVEC14):53–60, 2016.

[50] Angel Ric, Carlota Torrents, Bruno Gonçalves, Jaime Sampaio, and Robert Hristovski. Soft-assembled multilevel dynamics of tactical behaviors in soccer. *Frontiers in psychology*, 7:1513, 2016.

[51] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635, 2011.

[52] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[53] Caude Sammut. Behavioral cloning. *Encyclopedia of Machine Learning*, pages 93–97, 2010.

[54] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. Joint action: bodies and minds moving together. *Trends in cognitive sciences*, 10(2):70–76, 2006.

[55] Khurram Soomro, Haroon Idrees, and Mubarak Shah. Online localization and prediction of actions and interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[56] Dale O Stahl and Paul W Wilson. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

[57] Dale O Stahl II and Paul W Wilson. Experimental evidence on players' models of other players. *Journal of economic behavior & organization*, 25(3):309–327, 1994.

[58] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–669, 2020.

[59] Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. *arXiv preprint arXiv:2103.07854*, 2021.

[60] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[61] NP Van der Aa, Xinghan Luo, Geert-Jan Giezeman, Robby T Tan, and Remco C Veltkamp. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pages 1264–1269. IEEE, 2011.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[63] Edward Vendrow, Satyajit Kumar, Ehsan Adeli, and Hamid Rezatofighi. Somoformer: Multi-person pose forecasting with transformers. *arXiv preprint arXiv:2208.14023*, 2022.

[64] Luis Vilar, Duarte Araújo, Keith Davids, and Chris Button. The role of ecological dynamics in analysing performance in team sports. *Sports Medicine*, 42:1–10, 2012.

[65] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018.

[66] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7124–7133, 2019.

[67] Chenxi Wang, Yunfeng Wang, Zixuan Huang, and Zhiwen Chen. Simple baseline for single human motion forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2260–2265, 2021.

[68] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021.

[69] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *The European Conference on Computer Vision (ECCV)*, 2020.

[70] A Mark Williams, Paul Ward, John M Knowles, and Nicholas J Smeeton. Anticipation skill in a real-world task: measurement, training, and transfer in tennis. *Journal of Experimental Psychology: Applied*, 8(4):259, 2002.

[71] James Wright and Kevin Leyton-Brown. Beyond equilibrium: Predicting human behavior in normal-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 901–907, 2010.

[72] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision*, pages 511–528. Springer, 2022.

[73] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *The Eleventh International Conference on Learning Representations*, 2023.

[74] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 142–159. Springer, 2022.

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

[76] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *arXiv preprint arXiv:2307.10894*, 2023.

A Appendix

A.1 Dataset

A.1.1 Data Collection

We record basketball drill videos from 11 camera views at a frame rate of 25 FPS. Multi-person 3D poses are estimated using Faster-VoxelPose [74]. To obtain 2D heatmaps, we first resize the image to 960×512 pixels and employ the ResNet [24] backbone. The 3D space size is set to $16m \times 16m \times 3m$, with the granularity of human detection networks at $80 \times 80 \times 20$. To estimate the joint locations of each subject, we set the human bounding box size to $2m \times 2m \times 2.4m$ and use a voxel size of $64 \times 64 \times 64$. We further apply JDETracker [69] with a tracking frame rate of 25FPS and a threshold of $0.2m$. Next, we conduct an automatic failure detection process. We first discard motion clips with incorrect person counts and then filter outliers with root position changes exceeding a $0.4m$ threshold compared to neighboring frames. Subsequently, we adopt the one-euro filter [11] for temporal smoothing. We set different parameters for different body joints because they exhibit different moving patterns. For upper limb joints such as wrists and elbows, we set $f_{\text{cutoff}} = 0.0287$ and $\beta = 0.5$; for lower body joints including ankles and knees, we set $f_{\text{cutoff}} = 0.0274$, $\beta = 1$; and for the rest of the body, we set $f_{\text{cutoff}} = 1$, $\beta = 1$. Finally, we perform a manual check to ensure the correctness of the data sequences.

A.1.2 Data Analysis

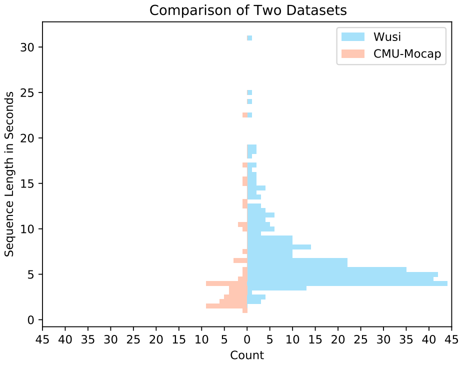
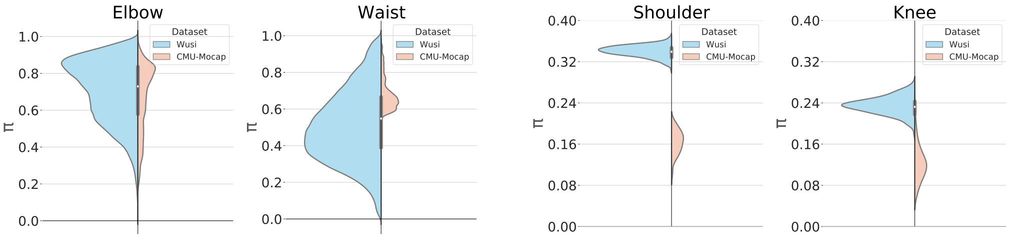


Figure 7: Clip length distribution comparison between our dataset and CMU-Mocap [1].

We visualize the clip lengths of our dataset and compare them with CMU-Mocap [1]. Figure 7 shows that our dataset features a longer overall duration and a larger number of long sequences.



(a) Overall joint angle distribution

(b) Intra-motion joint angle *std* distribution

Figure 8: Distribution comparison between our dataset and CMU-Mocap [1]. (a): We plot the joint angle distribution over the entire dataset. (b): We calculate the joint angle standard deviation for each motion sequence, then plot the *std* distribution.

Furthermore, we assess pose diversity and motion intensity utilizing joint angle distributions. As Figure 8(a) shows, our dataset encompasses a broader range of plausible body joint angles, indicating a superior pose diversity. We then investigate motion intensity using intra-motion variation. Figure 8(b) illustrates that motion sequences in our dataset have higher joint angle variances, indicating more dynamic motions.

A.2 Implementation details

We represent human motion with 3D skeleton sequences following previous works [68, 48, 55]. Specifically, each $x_t^p \in \mathbb{R}^{J \times 3}$ represents a 3D human pose for person p at timestep t, containing the Cartesian coordinates of J skeleton joints.

We apply Discrete Cosine Transform (DCT) [4] to the input motions and Inverse Discrete Cosine Transform (IDCT) for the decoded actions following [40, 68]. We employ a local-range Transformer encoder and a global-range Transformer encoder [68] as the state encoders. We utilize Transformer decoders as the policy networks. We concatenate the local state features s_l^p with global state features $\boldsymbol{s}_g$ (for level-0) or joint agent actions at the previous level $\boldsymbol{a}_{(k-1)}$ (for level-k, $k \geq 1$). For GAIL training, we implement a discriminator D that takes a pair of state $\boldsymbol{s}$ and joint actions $\boldsymbol{a}$ as input and aims to classify whether the actions are from the policy network or expert demonstrations. The discriminator consists of a global-range Transformer encoder, a temporal average pooling layer, and an MLP with one hidden layer. We implement the proposed framework with PyTorch [47] using a Linux machine with 1 NVIDIA V100 GPU. We set $\lambda = 0.002$, batch size 32, learning rate 0.0001, and train for 60 epochs using Adam [33] optimizer.

A.3 Ablation Study

Table 6: Ablation study of λ.

milliseconds	Global				Local				Root			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
$\lambda = 0.01$	61.3	94.1	127.5	160.2	50.3	69.3	84.9	97.4	45.6	71.1	99.2	128.0
$\lambda = 0.005$	54.8	86.9	120.5	154.6	43.8	61.1	74.9	87.4	41.8	67.4	97.6	126.2
$\lambda = 0.002$	54.6	86.2	**119.3**	**152.5**	43.7	60.8	74.6	86.6	41.7	66.9	**94.8**	**124.0**
$\lambda = 0.001$	54.3	**86.1**	119.5	153.5	43.1	60.0	**74.2**	**86.1**	**41.4**	**66.8**	95.2	125.3
$\lambda = 0.0002$	**54.2**	86.5	120.2	154.1	**43.0**	**59.9**	74.2	86.2	41.6	67.6	96.2	126.1

We provide an additional ablation study on loss weight λ in Table 6.

A.4 Discussions

Broader Impacts Our approach does not rely on explicit biometric markers like unique fingerprints, retina scans, or body markers. The skeletal structure representation utilized in our dataset is based on generalized joint positions and movements. Our approach should only be employed with legally obtained human motion data. Meanshile, our approach may contribute to more interpretable decision-making processes and holds potential for applications in sports tactical analysis and education.

Assets To collect our dataset, we acquire written consent from the participating athletes and coaches. We plan to release the processed 3D human motion sequences under a license agreement that permits their use for non-commercial scientific research purposes.

Human Subjects We inform the participating athletes about the data collection process and its intended uses, ensuring that we obtain their written consent. The instructions are provided by their coach, and the collected drills are part of their regular training routines. All participants receive compensation in accordance with their standard practices. To ensure privacy, we have implemented rigorous measures. Our data release strategy does not involve disclosing raw video data that may contain identifiable information, such as facial features. Instead, we will exclusively release processed 3D human motion sequences derived from the video data, guaranteeing the anonymity of the athletes.