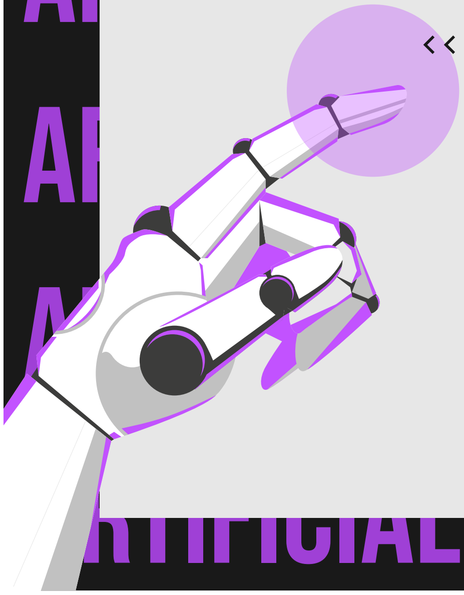
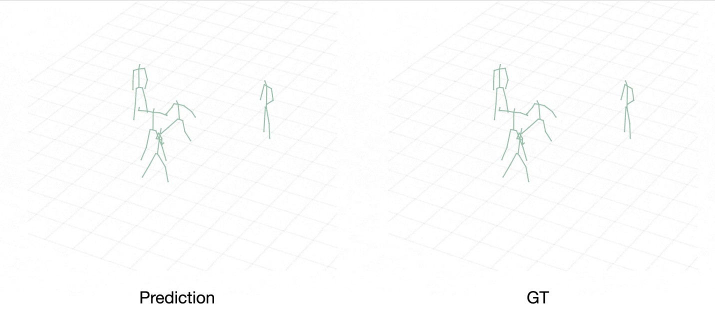


SOCIAL MOTION PREDICTION WITH COGNITIVE HIERARCHIES

By: Kelvin Zhou and Rishi Ramesh

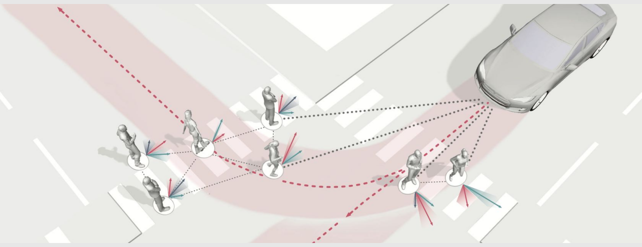
INTRODUCTION

- Humans naturally plan their actions around what they believe others around them will do
- Planning these actions are an integral part of what allows us to cooperate and compete with each other
- Our study looks at predicting the actions and movements of multiple agents (players) in the context of team sports, accounting for social context
- Introduces a novel 3D dataset and cognitive hierarchies to existing frameworks



PREVIOUS WORKS

- Dataset
  - 2D datasets don't capture the spatial nuances included in the proximity of individuals
  - 3D datasets were too expensive and didn't fully capture complex movements
  - Existing datasets were meant for pose estimation and doesn't have relevant data for making meaningful predictions
- 2D/3D Motion Prediction
  - Multi-Agent Reinforcement Learning and Cognitive Hierarchies
  - Mostly done for trajectory prediction without accounting for other factors
  - Insistent on trying to fit neural networks to this problem using supervised learning but fails to capture the strategic social aspects of human actions



WUSI

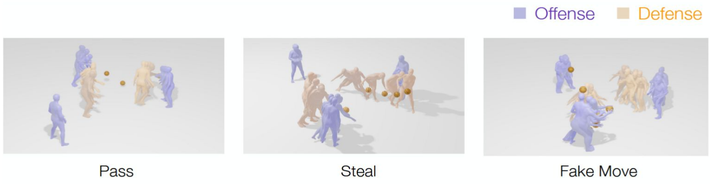
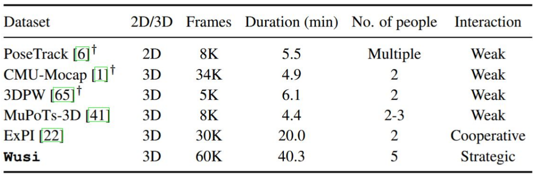


Figure 1: Example sequences from **Wusi** dataset. Three players in purple form the offensive team, and two players in orange form the defensive team. Left: a successful pass for the offensive team. Middle: a successful steal for the defensive team. Right: a successful pass following a fake pass.

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

DATA COLLECTION AND ANALYSIS

- 11 synchronized and calibrated wide-baseline cameras
- Markerless multi-person 3D pose estimation algorithm
- Limitations of previous work data analysis:
  - Motion Sequence Mixing
  - Limited Interaction Strength → less poses
- Wusi offers 2-9x times video duration → more strategic interactions amongst agents

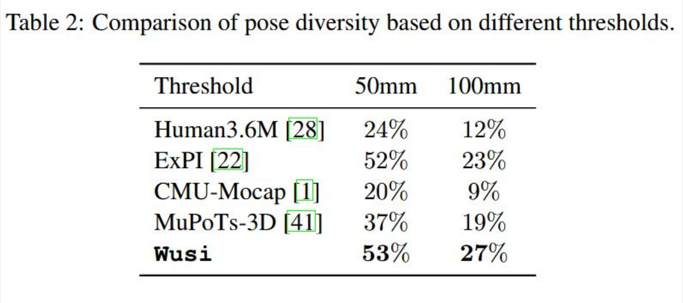


Table 2: Comparison of pose diversity based on different thresholds.

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%**

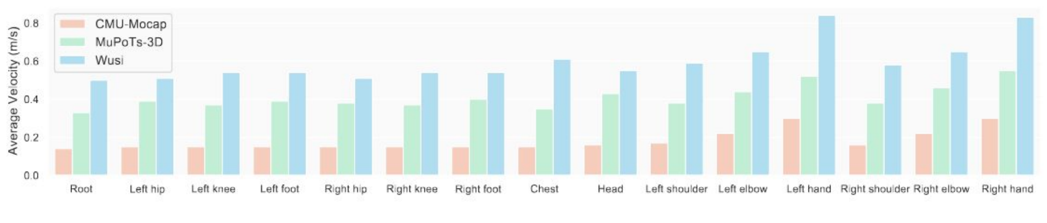
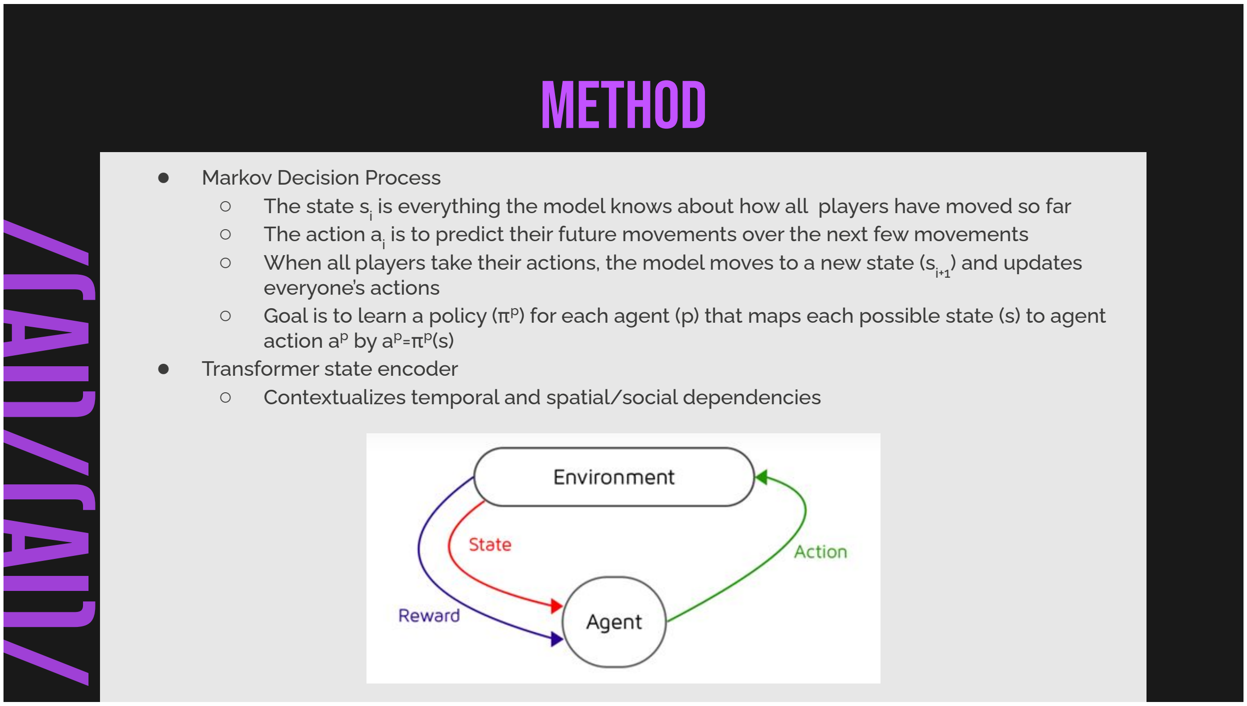


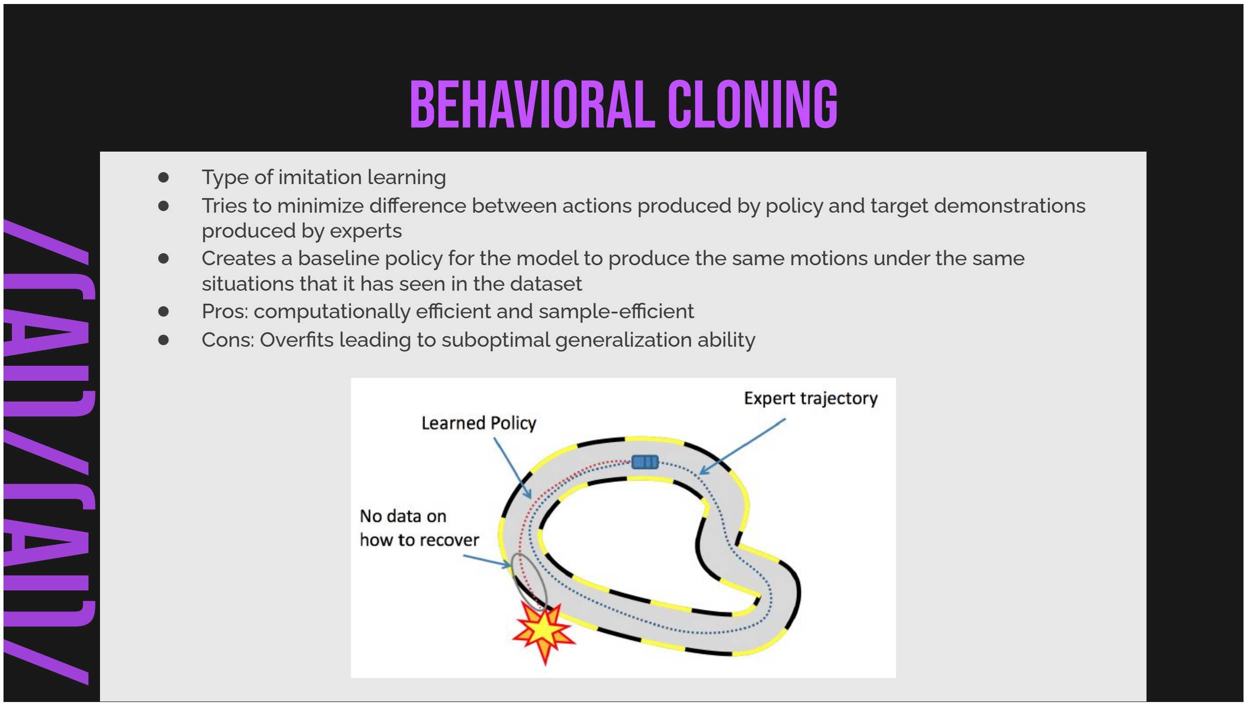
Figure 2: Comparison of motion intensity.



METHOD

- Markov Decision Process
  - The state s_i is everything the model knows about how all players have moved so far
  - The action a_i is to predict their future movements over the next few movements
  - When all players take their actions, the model moves to a new state (s_{i+1}) and updates everyone's actions
  - Goal is to learn a policy (π^p) for each agent (p) that maps each possible state (s) to agent action a^p by $a^p=\pi^p(s)$
- Transformer state encoder
  - Contextualizes temporal and spatial/social dependencies

Environment
State
Action
Reward
Agent

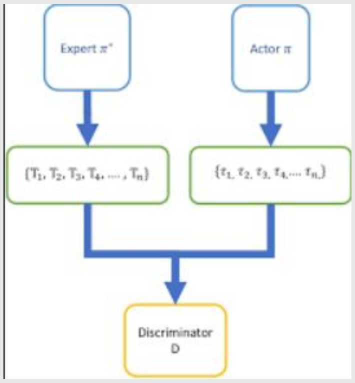


BEHAVIORAL CLONING

- Type of imitation learning
- Tries to minimize difference between actions produced by policy and target demonstrations produced by experts
- Creates a baseline policy for the model to produce the same motions under the same situations that it has seen in the dataset
- Pros: computationally efficient and sample-efficient
- Cons: Overfits leading to suboptimal generalization ability

GENERATIVE ADVERSARIAL IMITATION LEARNING

- Improves generalization by using 2 main components
  - Generator → agent you want to train; generates actions in states, trying to imitate the expert
  - Discriminator → A neural network that tries to tell whether a (state, action) pair came from the expert or from the agent


Expert π^*
Actor π
$\{T_1, T_2, T_3, T_4, \ldots, T_n\}$
$\{\tau_1, \tau_2, \tau_3, \tau_4, \ldots \tau_n\}$
Discriminator
D

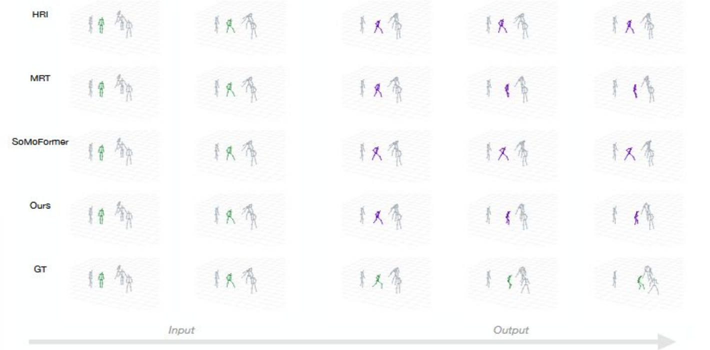
COGNITIVE HIERARCHIES

- Comes from behavioral game theory
- Models how people make strategic decisions when interacting with others
- Core idea: people reason in levels
- The higher levels base their actions from predicting what the lower levels will do
- BC and GAIL used to train a network at each level of cognitive hierarchy where the input to that network comes from the predictions of the previous level's network

$$a^p_{(0)} = \pi^p_{(0)}(\boldsymbol{s}) = \phi_{(0)}(s^p_l, \boldsymbol{s}_g),$$

RESULTS

- The model works better with long term predictions for a group of people
- Struggles with short term predictions for a single person
- Also works well in social context like team sports
- For a switch situation where a defensive player has to cover a new opponent, their model predicted correctly that the player will make the switch whereas other related models predicted the player would continue to move in the same direction or stop in the middle



CONTRIBUTIONS AND FUTURE DIRECTIONS

- Largest-ever 3D motion dataset
- Reframes motion prediction as MARL problem
- Proposes new Cognitive Hierarchy framework
- Limitation: May not be true for all social interactions

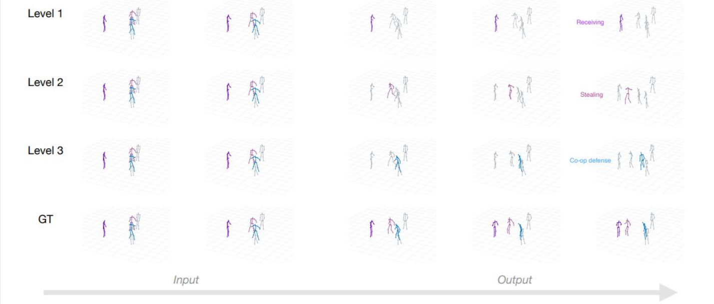


Figure 6: Actions produced by policy networks at different levels. Left two columns are past motions and right three columns are future motions.