EqMotion: Equivariant Multi-agent Motion Prediction with Invariant Interaction Reasoning – Supplementary Material –

A. Theoretical Proofs

In this section, we prove Theorem 1 in our paper which shows EqMotion's equivariance property and the interaction reasoning module's invariance property. Note that, here we treat all the vectors to be row vectors since we multiply the rotation matrix by right.

1. For the initialization layer $\mathcal{F}_{\rm IL}(\cdot)$, the initial geometric feature is equivariant and the initial pattern feature is invariant:

$$\mathbb{G}^{(0)}\mathbf{R}+\mathbf{t},\, \mathbf{H}^{(0)} = \mathcal{F}_{\rm IL}(\mathbb{X}\mathbf{R}+\mathbf{t}).$$

Proof: For the ith agent, we show its initial geometric feature is equivariant to the input motion under Euclidean transformation. When transforming the past motion, we have

$$\begin{aligned}
&\phi_{\rm init_g}\big((\mathbf{X}_i\mathbf{R}+\mathbf{t})-\overline{\mathbb{X}}\mathbf{R}+\mathbf{t})\big)+\overline{\mathbb{X}}\mathbf{R}+\mathbf{t}\\
&=\mathbf{W}_{\rm init_g}\big((\mathbf{X}_i\mathbf{R}-\overline{\mathbb{X}}\mathbf{R})\big)+\overline{\mathbb{X}}\mathbf{R}+\mathbf{t}\\
&=(\mathbf{W}_{\rm init_g}(\mathbf{X}_i-\overline{\mathbb{X}})+\overline{\mathbb{X}})\mathbf{R}+\mathbf{t}\\
&=\mathbb{G}^{(0)}\mathbf{R}+\mathbf{t}
\end{aligned} \tag{1}$$

Thus we show the initial geometric feature is equivariant to the input motion under Euclidean transformation. We also show its initial pattern feature is invariant to the input motion under Euclidean transformation. When transforming the past motion, we have,

$$\begin{aligned}
&\triangle\,(\mathbf{X}_i\mathbf{R}+\mathbf{t}) = \triangle(\mathbf{X}_i)\mathbf{R}=\mathbf{V}_i\mathbf{R},\\
&||\mathbf{V}_i^t\mathbf{R}||_2^2=\mathbf{V}_i^t\mathbf{R}\mathbf{R}^\top{\mathbf{V}_i^t}^\top={\mathbf{V}_i^t}{\mathbf{V}_i^t}^\top=||\mathbf{V}_i^t||_2^2=\rho_i^t,\\
&\text{angle}(\mathbf{V}_i^t\mathbf{R},\mathbf{V}_i^{t-1}\mathbf{R})=\frac{\mathbf{V}_i^t\mathbf{R}(\mathbf{V}_i^{t-1}\mathbf{R})^\top}{||\mathbf{V}_i^t\mathbf{R}||_2||\mathbf{V}_i^{t-1}\mathbf{R}||_2}\\
&=\frac{\mathbf{V}_i^t\mathbf{R}\mathbf{R}^\top{\mathbf{V}_i^{t-1}}^\top}{||\mathbf{V}_i^t||_2||\mathbf{V}_i^{t-1}||_2}=\frac{\mathbf{V}_i^t{\mathbf{V}_i^{t-1}}^\top}{||\mathbf{V}_i^t||_2||\mathbf{V}_i^{t-1}||_2}\\
&=\text{angle}(\mathbf{V}_i^t,\mathbf{V}_i^{t-1})=\theta_i^t,\\
&\phi_{\rm init_h}([\,\rho_i;\theta_i\,])=\mathbf{h}_i^{(0)}.
\end{aligned} \tag{2}$$

Thus we show the initial pattern feature is invariant to the input motion under Euclidean transformation.

2. The reasoning module $\mathcal{F}_{\rm IRM}(\cdot)$ along with reasoned interaction categorical vectors $\{\mathbf{c}_{ij}\}$ is invariant:

$$\{\mathbf{c}_{ij}\}=\mathcal{F}_{\rm IRM}(\mathbb{G}^{(0)}\mathbf{R}+\mathbf{t},\mathbf{H}^{(0)}).$$

Proof: We first show the column-wise ℓ_2-distance of geometric feature $||\mathbf{G}_i^{(0)}-\mathbf{G}_j^{(0)}||_{2,\rm col}$ is invariant since for the cth column ($c=1,\cdots,C$), we have $||\mathbf{g}_{i,c}^{(0)}\mathbf{R}+\mathbf{t}-(\mathbf{g}_{j,c}^{(0)}\mathbf{R}+\mathbf{t})||_2=||\mathbf{g}_{i,c}^{(0)}\mathbf{R}-\mathbf{g}_{j,c}^{(0)}\mathbf{R}||_2=||\mathbf{g}_{i,c}^{(0)}-\mathbf{g}_{j,c}^{(0)}||_2$. Since the initial pattern feature is invariant, thus we have the edge feature $\mathbf{m}'_{ij}$, the aggregated edge feature $\mathbf{p}'_i$ and the updated node feature $\mathbf{h}'_i$ all to be invariant. Finally, we have the interaction categorical $\mathbf{c}_{ij}$ vector being invariant since $\mathbf{h}'_i$ and $||\mathbf{G}_i^{(0)}-\mathbf{G}_j^{(0)}||_{2,\rm col}$ are invariant,

$$\text{sm}\Big(\phi_{\rm rc}\big([\mathbf{h}'_i;\mathbf{h}'_j;||\mathbf{G}_i^{(0)}-\mathbf{G}_j^{(0)}||_{2,\rm col}]\big)/\tau\Big)=\mathbf{c}_{ij}.$$

3. The ℓth geometric feature learning layer $\mathcal{F}_{\rm EGFL}^{(\ell)}(\cdot)$ is equivariant:

$$\mathbb{G}^{(l+1)}\mathbf{R}+\mathbf{t}=\mathcal{F}_{\rm EGFL}^{(\ell)}(\mathbb{G}^{(\ell)}\mathbf{R}+\mathbf{t},\mathbf{H}^{(\ell)},\{\mathbf{c}_{ij}\}).$$

Proof: We show the result by indicating the inner-agent attention, inter-agent aggregation and non-linear function all to be equivariant. We first show the inner-agent attention is equivariant. When transforming the input geometric feature, for every ith agent ($i=1,2,\cdots,M$),

$$\begin{aligned}
&\phi_{\rm att}^{(\ell)}(\mathbf{h}_i^{(\ell)})\cdot(\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}-(\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t}))+\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t}\\
&=\phi_{\rm att}^{(\ell)}(\mathbf{h}_i^{(\ell)})\cdot(\mathbf{G}_i^{(\ell)}-\overline{\mathbb{G}}^{(\ell)})\mathbf{R}+\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t}\\
&=(\phi_{\rm att}^{(\ell)}(\mathbf{h}_i^{(\ell)})\cdot(\mathbf{G}_i^{(\ell)}-\overline{\mathbb{G}}^{(\ell)})+\overline{\mathbb{G}}^{(\ell)})\mathbf{R}+\mathbf{t}\\
&\to\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}
\end{aligned} \tag{3}$$

Thus the inner-agent attention is equivariant. We then show the inter-agent aggregation is equivariant. When transforming the input geometric feature, we show the column-wise ℓ_2-distance of geometric feature $||\mathbf{G}_i^{(\ell)}-\mathbf{G}_j^{(\ell)}||_{2,\rm col}$ is invariant since for the cth column ($c=1,\cdots,C$), we have $||\mathbf{g}_{i,c}^{(\ell)}\mathbf{R}+\mathbf{t}-(\mathbf{g}_{j,c}^{(\ell)}\mathbf{R}+\mathbf{t})||_2=||\mathbf{g}_{i,c}^{(\ell)}\mathbf{R}-\mathbf{g}_{j,c}^{(\ell)}\mathbf{R}||_2=||\mathbf{g}_{i,c}^{(\ell)}-\mathbf{g}_{j,c}^{(\ell)}||_2$. Thus the learned aggregation weights $\mathbf{e}_{ij}^{(\ell)}$ is invariant. We have the inter-agent aggregation's equivariance,

$$
\begin{aligned}
&\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}+\sum_{j\in\mathcal{N}_i}\mathbf{e}_{ij}^{(\ell)}\cdot(\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}-(\mathbf{G}_j^{(\ell)}\mathbf{R}+\mathbf{t}))\\
&=\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}+\sum_{j\in\mathcal{N}_i}\mathbf{e}_{ij}^{(\ell)}\cdot(\mathbf{G}_i^{(\ell)}-\mathbf{G}_j^{(\ell)})\mathbf{R}\\
&=\big(\mathbf{G}_i^{(\ell)}+\sum_{j\in\mathcal{N}_i}\mathbf{e}_{ij}^{(\ell)}\cdot(\mathbf{G}_i^{(\ell)}-\mathbf{G}_j^{(\ell)})\big)\mathbf{R}+\mathbf{t}\\
&\to\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}
\end{aligned}
\tag{4}
$$

Thus the inner-agent attention is equivariant. We then show the non-linear function is equivariant. When transforming the input geometric feature, the inner product of the query coordinate and the key coordinate $\langle\mathbf{q}_{i,c}^{(\ell)},\mathbf{k}_{i,c}^{(\ell)}\rangle=\mathbf{q}_{i,c}^{(\ell)}\mathbf{k}_{i,c}^{(\ell)\top}$ is invariant for every channel $c=1,2,\cdots,C$ since

$$
\begin{aligned}
&\mathbf{W}_{\mathrm{Q}}^{(\ell)}\big(\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}-(\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t})\big)=\mathbf{Q}_i^{(\ell)}\mathbf{R},\\
&\mathbf{W}_{\mathrm{K}}^{(\ell)}\big(\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}-(\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t})\big)=\mathbf{K}_i^{(\ell)}\mathbf{R},\\
&\mathbf{Q}_i^{(\ell)}\mathbf{R}(\mathbf{K}_i^{(\ell)}\mathbf{R})^\top=\mathbf{Q}_i^{(\ell)}\mathbf{R}\mathbf{R}^\top\mathbf{K}_i^{(\ell)\top}=\mathbf{Q}_i^{(\ell)}\mathbf{K}_i^{(\ell)\top}.
\end{aligned}
\tag{5}
$$

We also can have the equivariance of the two equations under different conditions,

$$
\mathbf{q}_{i,c}^{(\ell)}\mathbf{R}+\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t}=(\mathbf{q}_{i,c}^{(\ell)}+\overline{\mathbb{G}}^{(\ell)})\mathbf{R}+\mathbf{t}=\mathbf{g}_{i,c}^{(\ell+1)}\mathbf{R}+\mathbf{t}
\tag{6}
$$

and

$$
\begin{aligned}
&\mathbf{q}_{i,c}^{(\ell)}\mathbf{R}-\left\langle\mathbf{q}_{i,c}^{(\ell)}\mathbf{R},\frac{\mathbf{k}_{i,c}^{(\ell)}\mathbf{R}}{||\mathbf{k}_{i,c}^{(\ell)}\mathbf{R}||_2}\right\rangle\frac{\mathbf{k}_{i,c}^{(\ell)}\mathbf{R}}{||\mathbf{k}_{i,c}^{(\ell)}\mathbf{R}||_2}+\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t}\\
&=\mathbf{q}_{i,c}^{(\ell)}\mathbf{R}-\left\langle\mathbf{q}_{i,c}^{(\ell)}\mathbf{R},\frac{\mathbf{k}_{i,c}^{(\ell)}\mathbf{R}}{||\mathbf{k}_{i,c}^{(\ell)}||_2}\right\rangle\frac{\mathbf{k}_{i,c}^{(\ell)}\mathbf{R}}{||\mathbf{k}_{i,c}^{(\ell)}||_2}+\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t}\\
&=\mathbf{q}_{i,c}^{(\ell)}\mathbf{R}-\left\langle\mathbf{q}_{i,c}^{(\ell)},\frac{\mathbf{k}_{i,c}^{(\ell)}}{||\mathbf{k}_{i,c}^{(\ell)}||_2}\right\rangle\frac{\mathbf{k}_{i,c}^{(\ell)}\mathbf{R}}{||\mathbf{k}_{i,c}^{(\ell)}||_2}+\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t}\\
&=\left(\mathbf{q}_{i,c}^{(\ell)}-\left\langle\mathbf{q}_{i,c}^{(\ell)},\frac{\mathbf{k}_{i,c}^{(\ell)}}{||\mathbf{k}_{i,c}^{(\ell)}||_2}\right\rangle\frac{\mathbf{k}_{i,c}^{(\ell)}}{||\mathbf{k}_{i,c}^{(\ell)}||_2}+\overline{\mathbb{G}}^{(\ell)}\right)\mathbf{R}+\mathbf{t}\\
&=\mathbf{g}_{i,c}^{(\ell+1)}\mathbf{R}+\mathbf{t}
\end{aligned}
\tag{7}
$$

Since the criterion is invariant and two equations under two conditions are both equivariant, the non-linear function is equivariant. Finally, combining the equivariance of inner-agent attention, inter-agent aggregation and nonlinear function, we show the equivariance of the geometric feature learning layer.

4. The ℓth pattern feature learning layer $\mathcal{F}_{\mathrm{IPFL}}^{(\ell)}(\cdot)$ is invariant:

$$\mathbf{H}^{(l+1)}=\mathcal{F}_{\mathrm{IPFL}}^{(\ell)}(\mathbb{G}^{(\ell)}\mathbf{R}+\mathbf{t},\mathbf{H}^{(\ell)}).$$

Proof: Similar with the invariance of reasoning module, we first have the column-wise ℓ_2-distance of geometric feature $||\mathbf{G}_i^{(\ell)}-\mathbf{G}_j^{(\ell)}||_{2,\mathrm{col}}$ is invariant Thus we have the variable in the message passing $\mathbf{m}_{ij}^{(\ell)}$, $\mathbf{p}_i^{(\ell)}$ all invariant. Finally the next layer's pattern feature $\mathbf{h}^{(l+1)}$ is invariant.

5. The output layer $\mathcal{F}_{\mathrm{EOL}}(\cdot)$ is equivariant:

$$\widehat{\mathbb{Y}}\mathbf{R}+\mathbf{t}=\mathcal{F}_{\mathrm{EOL}}(\mathbb{G}^{(L)}\mathbf{R}+\mathbf{t}).$$

Proof: When transforming the input geometric feature,

$$
\begin{aligned}
&\mathcal{F}_{\mathrm{EOL}}(\mathbb{G}^{(L)}\mathbf{R}+\mathbf{t})\\
&=\big(\mathbf{W}_{\mathrm{out}}(\mathbf{G}_i^{(\ell)}\mathbf{R}+\mathbf{t}-(\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t})\big)+\overline{\mathbb{G}}^{(\ell)}\mathbf{R}+\mathbf{t}\\
&=\big(\mathbf{W}_{\mathrm{out}}(\mathbf{G}_i^{(\ell)}-\overline{\mathbb{G}}^{(\ell)})+\overline{\mathbb{G}}^{(\ell)}\big)\mathbf{R}+\mathbf{t}\\
&=\widehat{\mathbb{Y}}\mathbf{R}+\mathbf{t}
\end{aligned}
\tag{8}
$$

Thus the output layer is equivariant.

B. Optional Operations

B.1. DCT Processing

To have a compact representation of input motion data, here we apply an optional discrete cosine transform (DCT) along the time axis to convert the input motion into the frequency domain. Mathematically, for the input motion $\mathbf{X}_i$ of agent i, we transform it by $\mathbf{X}_i\leftarrow\mathbf{W}_{\mathrm{DCT}}(\mathbf{X}_i-\overline{\mathbb{X}})$ where $\mathbf{W}_{\mathrm{DCT}}\in\mathbb{R}^{T_{\mathrm{p}}\times T_{\mathrm{p}}}$ is the DCT coefficients matrix. Correspondingly, we transform the predicted motion by an inverse DCT (iDCT) operation: $\widehat{\mathbf{Y}}_i\leftarrow\mathbf{W}_{\mathrm{iDCT}}\widehat{\mathbf{Y}}_i+\overline{\mathbb{X}}$ and $\mathbf{W}_{\mathrm{iDCT}}\in\mathbb{R}^{T_{\mathrm{f}}\times T_{\mathrm{f}}}$ is the iDCT coefficients matrix. The remove-and-add operation about the mean location $\overline{\mathbb{X}}$ is to ensure the translation equivariance. Since the DCT process is equivariant, adding this process will maintain whole network's equivariance.

B.2. Adding Velocity Information

We also introduce an optional operation to directly add the velocity information into the geometric feature by

$$\mathbf{G}_i^{(\ell)}\leftarrow\phi_\rho(\rho_i)+\mathbf{G}_i^{(\ell)},\tag{9}$$

where ρ_i is the velocity magnitude sequence and function $\phi_\rho(\cdot)$ is implemented by MLP. Since the velocity magnitude sequence is invariant, thus the operation is equivariant. This operation is placed before the nonlinear function.

C. Modification for Multi-prediction

To make EqMotion perform multiple predictions in pedestrian trajectory prediction, we slightly modify the network by using multiple prediction heads in parallel. Each prediction head consists of a feature learning layer and an output layer. Assuming the ith output produced by the ith prediction head is $\widehat{\mathbb{Y}_i}$, we use a minimum ℓ_2 prediction loss formulated by,

$$\mathcal{L}=\min_i||\mathbb{Y}-\widehat{\mathbb{Y}_i}||_2^2.\tag{10}$$

Through the loss, the optimal prediction will be optimized.

D. Experiment Details

D.1. Dataset Description

D.1.1 Particle Dynamics

We use the particle N-body simulation environment [6] in a 3-dimensional space similar to [3, 11]. The system contains 5 interacted particles. In the reasoning task, in the Springs simulation, particles will be randomly connected by a spring with a probability of 0.5. The particles connected by springs interact via forces given by Hooke's law. In the Charged simulation, particles will be randomly charged or uncharged. The charged particles will repel or attract others via Coulomb forces. The probability of positive charged, uncharged and negative charged is 0.25, 0.5, and 0.25. We predicted the future motion of 20 timestamps given the historical observations of 20 timestamps. We use a downsampling rate of 100. We use 5k, 2k and 2k samples for training, validating and testing, respectively. In the prediction task, the setting is similar except the probability of positive charged, uncharged and negative charged is 0.5, 0, and 0.5.

D.1.2 Molecule Dynamics

We adopt the MD17 [2] dataset which contains the motions of different molecules generated via a molecular dynamics simulation environment. The goal is to predict the motions of every atom of the molecule. We randomly pick four kinds of molecules: Aspirin, Benzene, Ethanol and Malonaldehyde. We learn a prediction model for each molecule. We predicted the future motion of 10 timestamps given the observation of 10 timestamps. The raw data is a long sequence and we sample the trajectory with a sampling rate of 20 and a sampling gap of 400. We randomly pick 5k, 2k and 2k samples for training, validating and testing.

D.1.3 3D Human Skeleton Motion

Human 3.6M (H3.6M) dataset [5] contains 7 subjects performing 15 classes of actions, and each subject has 22 body joints. All sequences are downsampled by two along time. Following previous paradigms [8, 9], the models are trained on the segmented clips in the 6 subjects and tested on the clips in the 5th subject.

D.1.4 Pedestrian Trajectories

ETH-UCY dataset [7, 10], contains 5 subsets, ETH, HOTEL, UNIV, ZARA1, and ZARA2. In the dataset, pedestrian trajectories are captured at 2.5Hz in multi-agent social scenarios. Following the standard setting [1, 4, 12], we use 3.2 seconds (8 timestamps) to predict the 4.8 seconds (12 timestamps). We use the leave-one-out approach, training on 4 sets and testing on the remaining set.

Table 1. Effect of different numbers of learning layers on H3.6M.

Layers	80ms	160ms	320ms	400ms	Average
1	9.5	21.4	46.7	58.3	34.0
2	9.3	20.7	45.4	56.5	33.0
3	**9.1**	20.3	44.3	55.7	32.4
4	**9.1**	**20.1**	**43.7**	**55.0**	**32.0**
5	**9.1**	20.2	43.9	55.2	32.1

D.2. Implementation Details

In all the experiments, we set the number of feature learning layers L to 4. We use the Adam optimizer to train the model on a single NVIDIA RTX-3090 GPU. All the MLPs have 2 layers with a ReLU activation function.

Particle Dynamics We set the number of coordinates in the geometric feature C as 64 and the dimension of the pattern feature D as 64. The predefined category number L is 2. We set the batch size to 50 and use a learning rate of 5e-4. The model is trained for 200 epochs.

Molecule Dynamics We set the number of coordinates in the geometric feature C as 64 and the dimension of the pattern feature D as 64. The predefined category number L is 2. We set the batch size to 50 and use a learning rate of 5e-4. The model is trained for 300 epochs.

Human Skeleton Motion For short-term motion prediction, we set the number of coordinates in the geometric feature C as 72 and the dimension of the pattern feature D as 64. The predefined category number L is 4. We set the batch size to 100 and use a learning rate of 5e-4. The model is trained for 80 epochs. For long-term motion prediction, we set the number of coordinates in the geometric feature C as 96 and the dimension of the pattern feature D as 64. The predefined category number L is 4. We set the batch size to 100 and use an initial learning rate of 5e-4 with a decay rate of 0.8 for every 2 epochs. The model is trained for 100 epochs.

Pedestrian Trajectories We set the number of coordinates in the geometric feature C as 64 and the dimension of the pattern feature D as 64. The predefined category number L is 4. We set the batch size to 100 and use an initial learning rate of 8e-4/5e-4/1e-3/5e-4/1e-3 with a decay rate of 0.8/0.8/0.95/0.8/0.9 for every 2/2/2/2/2 epochs on eth/hotel/univ/zara1/zara2 subsets, respectively. The model is trained for 50 epochs.

E. Further Experiment Results

Different numbers of layers Table 1 shows the effect of different numbers of feature learning layers L on the H3.6M dataset. We find that i) initially increasing L leads to better performance as a more comprehensive geometric feature and pattern feature will be learned; and ii) when the number of layers is sufficient, the performance tends to be stable.

Table 2. Comparisons of short-term prediction on Human3.6M. Results at 80ms, 160ms, 320ms, 400ms in the future are shown.

Motion	Walking				Eating				Smoking				Discussion			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res-sup.	29.4	50.8	76.0	81.5	16.8	30.6	56.9	68.7	23.0	42.6	70.1	82.7	32.9	61.2	90.9	96.2
Traj-GCN	12.3	23.0	39.8	46.1	8.4	16.9	33.2	40.7	7.9	16.2	31.9	38.9	12.5	27.4	58.5	71.7
DMGNN	17.3	30.7	54.6	65.2	11.0	21.4	36.2	43.9	9.0	17.6	32.1	40.3	17.3	34.8	61.0	69.8
MSRGCN	12.2	22.7	38.6	45.2	8.4	17.1	33.0	40.4	8.0	16.3	31.3	38.2	12.0	26.8	57.1	69.7
PGBIG	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7
SPGSN	10.1	19.4	34.8	41.5	7.1	14.9	30.5	37.9	6.7	13.8	28.0	34.6	10.4	23.8	53.6	67.1
EqMotion(Ours)	**9.0**	**17.5**	**32.6**	**39.2**	**6.3**	**13.6**	**28.9**	**36.5**	**5.5**	**11.3**	**23.0**	**29.3**	**8.2**	**18.8**	**42.1**	**53.9**
Motion	Directions				Greeting				Phoning				Posing			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res-sup.	35.4	57.3	76.3	87.7	34.5	63.4	124.6	142.5	38.0	69.3	115.0	126.7	36.1	69.1	130.5	157.1
Traj-GCN	9.0	19.9	43.4	53.7	18.7	38.7	77.7	93.4	10.2	21.0	42.5	52.3	13.7	29.9	66.6	84.1
DMGNN	13.1	24.6	64.7	81.9	23.3	50.3	107.3	132.1	12.5	25.8	48.1	58.3	15.3	29.3	71.5	96.7
MSRGCN	8.6	19.7	43.3	53.8	16.5	37.0	77.3	93.4	10.1	20.7	41.5	51.3	12.8	29.4	67.0	85.0
PGBIG	7.2	17.6	40.9	51.5	15.2	34.1	71.6	87.1	8.3	18.3	38.7	48.4	10.7	25.7	60.0	76.6
SPGSN	7.4	17.2	39.8	50.3	14.6	32.6	70.6	86.4	8.7	18.3	38.7	48.5	10.7	25.3	59.9	76.5
EqMotion(Ours)	**6.3**	**15.8**	**38.9**	**50.1**	**12.7**	**30.1**	**68.3**	**85.2**	**7.4**	**16.7**	**36.9**	**47.0**	**8.2**	**18.9**	**43.4**	**57.5**
Motion	Purchases				Sitting				Sittingdown				Takingphoto			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res-sup.	36.3	60.3	86.5	95.9	42.6	81.4	134.7	151.8	47.3	86.0	145.8	168.9	26.1	47.6	81.4	94.7
Traj-GCN	15.6	32.8	65.7	79.3	10.6	21.9	46.3	57.9	16.1	31.1	61.5	75.5	9.9	20.9	45.0	56.6
DMGNN	21.4	38.7	75.7	92.7	11.9	25.1	44.6	50.2	15.0	32.9	77.1	93.0	13.6	29.0	46.0	58.8
MSRGCN	14.8	32.4	66.1	79.6	10.5	22.0	46.3	57.8	16.1	31.6	62.5	76.8	9.9	21.0	44.6	56.3
PGBIG	12.5	28.7	**60.1**	**73.3**	8.8	19.2	42.4	53.8	13.9	27.9	57.4	71.5	8.4	18.9	42.0	53.3
SPGSN	12.8	28.6	61.0	74.4	9.3	19.4	42.3	53.6	14.2	27.7	56.8	**70.7**	8.8	18.9	**41.5**	**52.7**
EqMotion(Ours)	**11.2**	**26.8**	60.5	75.2	**8.1**	**18.0**	**41.2**	**52.9**	**13.0**	**26.5**	**56.2**	**70.7**	**7.9**	**17.7**	40.9	52.8
Motion	Waiting				Walking Dog				Walking Together				Average			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res-sup.	30.6	57.8	106.2	121.5	64.2	102.1	141.1	164.4	26.8	50.1	80.2	92.2	34.7	62.0	101.1	115.5
Traj-GCN	11.4	24.0	50.1	61.5	23.4	46.2	83.5	96.0	10.5	21.0	38.5	45.2	12.7	26.1	52.3	63.5
DMGNN	12.2	24.2	59.6	77.5	47.1	93.3	160.1	171.2	14.3	26.7	50.1	63.2	17.0	33.6	65.9	79.7
MSRGCN	10.7	23.1	48.3	59.2	20.7	42.9	80.4	93.3	10.6	20.9	37.4	43.9	12.1	25.6	51.6	62.9
PGBIG	8.9	20.1	43.6	54.3	18.8	39.3	73.7	86.4	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
SPGSN	9.2	19.8	43.1	54.1	17.8	37.2	**71.7**	**84.9**	8.9	18.2	33.8	40.9	10.4	22.3	47.1	58.3
EqMotion(Ours)	**7.6**	**17.4**	**39.9**	**51.1**	**16.6**	**36.4**	72.5	86.2	**7.8**	**16.1**	**30.6**	**37.1**	**9.1**	**20.1**	**43.7**	**55.0**

Table 3. Comparisons of long-term prediction on Human3.6M. Results at 560ms and 1000ms in the future are shown.

Motion	Walking		Eating		Smoking		Discussion		Directions		Greeting		Phoning		Posing	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res-Sup.	81.7	100.7	79.9	100.2	94.8	137.4	121.3	161.7	110.1	152.5	156.3	184.3	143.9	186.8	165.7	236.8
Traj-GCN	54.1	59.8	53.4	77.8	50.7	72.6	91.6	121.5	71.0	101.8	115.4	148.8	69.2	103.1	114.5	173.0
DMGNN	71.4	85.8	58.1	86.7	50.9	72.2	81.9	138.3	102.1	135.8	144.5	170.5	71.3	108.4	125.5	188.2
MSRGCN	52.7	63.0	52.5	77.1	49.5	71.6	88.6	117.6	71.2	100.6	116.3	147.2	68.3	104.4	116.3	174.3
PGBIG	48.1	56.4	51.1	76.0	46.5	69.5	87.1	118.2	**69.3**	**100.4**	110.2	143.5	65.9	102.7	106.1	164.8
SPGSN	46.9	53.6	49.8	73.4	46.7	68.6	89.7	118.6	70.1	100.5	111.0	143.2	66.7	102.5	110.3	165.4
EqMotion(Ours)	**43.4**	**52.8**	**48.4**	**73.0**	**41.0**	**63.4**	**75.3**	**105.6**	70.4	101.3	**108.7**	**142.0**	**64.7**	**101.0**	**84.9**	**139.4**
Motion	Purchases		Sitting		Sitting Down		Taking Photo		Waiting		Walking Dog		Walking Together		Average	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res-Sup.	119.4	176.9	166.2	185.2	197.1	223.6	107.0	162.4	126.7	153.2	173.6	202.3	94.5	110.5	129.2	165.0
Traj-GCN	102.0	143.5	78.3	119.7	100.0	150.2	77.4	119.8	79.4	108.1	111.9	148.9	55.0	65.6	81.6	114.3
DMGNN	104.9	146.1	75.5	115.4	118.0	174.1	78.4	123.7	85.5	113.7	183.2	210.2	70.5	86.9	93.6	127.6
MSRGCN	101.6	139.2	78.2	120.0	102.8	155.5	77.9	121.9	76.3	106.3	111.9	148.2	52.9	65.9	81.1	114.2
PGBIG	95.3	**133.3**	**74.4**	**116.1**	**96.7**	**147.8**	**74.3**	118.6	**72.2**	**103.4**	104.7	139.8	51.9	64.3	76.9	110.3
SPGSN	96.5	133.9	75.0	116.2	98.9	149.9	75.6	**118.2**	73.5	103.6	**102.4**	**138.0**	49.8	60.9	77.4	109.6
EqMotion(Ours)	**93.5**	134.5	74.7	116.6	98.1	149.9	76.7	122.0	71.4	104.6	104.8	141.2	**44.5**	**56.0**	**73.4**	**106.9**

F. Limitation and Future Work

This work focuses on a generally applicable motion prediction method. In the future, we plan to expand the method by adding specific designs for different tasks to further improve the model performance. We also expect the method can use more types of data to assist prediction, such as images and videos that contain map information.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 3

[2] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017. 3

[3] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020. 3

[4] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and

Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 3

[5] 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, 2013. 3

[6] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pages 2688–2697. PMLR, 2018. 3

[7] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 3

[8] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020. 3

[9] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 3

[10] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 3

[11] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021. 3

[12] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021. 3