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 \( \mathcal{F}_{IL} (\cdot) \), the initial geometric feature is equivariant and the initial pattern feature is invariant:

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

**Proof:** For the \( i \)-th agent, we show its initial geometric feature is equivariant to the input motion under Euclidean transformation. When transforming the past motion, we have

\[
\phi_{\text{init},g}(X_i R + \mathbf{t}) - \overline{X} R + \mathbf{t} = (W_{\text{init},g}(X_i - \overline{X}) + \overline{X}) R + \mathbf{t} = \mathcal{G}^{(0)} R + \mathbf{t}
\]

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,

\[
\begin{align*}
\Delta (X_i R + \mathbf{t}) &= \Delta(X_i) R = V_i R, \\
||V^T_i R||^2_2 &= V^T_i R R^T V^T_i = V^T_i V^T_i = ||V^T_i||^2_2 = \rho_i, \\
\text{angle}(V^T_i R, V^T_i R) &= \frac{V^T_i R (V^T_i R - 1) R^T}{||V^T_i R||_2 ||V^T_i R - 1||_2} = \frac{V^T_i V^T_i - 1}{||V^T_i||_2 ||V^T_i - 1||_2} = \phi_{\text{init},h}([\rho_i; \theta_i]) = h_i^{(0)}.
\end{align*}
\]

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

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

\[
\{ c_{ij} \} = \mathcal{F}_{IRM}(\mathcal{G}^{(0)} R + \mathbf{t}, \mathbf{H}^{(0)}).
\]

**Proof:** We first show the column-wise \( \ell_2 \)-distance of geometric feature \( ||\mathcal{G}_i^{(0)} - \mathcal{G}_j^{(0)}||_{2,\text{col}} \) is invariant since for the \( c \)-th column \( (c = 1, \cdots, C) \), we have \( ||\mathcal{G}_i^{(0)} R + \mathbf{t} - (\mathcal{G}_j^{(0)} R + \mathbf{t})||_2 = ||\mathcal{G}_i^{(0)} R - \mathcal{G}_j^{(0)} R||_2 = ||\mathcal{G}_i^{(0)} c - \mathcal{G}_j^{(0)} c||_2 \). 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 \( c_{ij} \) vector being invariant since \( \mathbf{h}_i \) and \( ||\mathcal{G}_i^{(0)} - \mathcal{G}_j^{(0)}||_{2,\text{col}} \) are invariant.

\[
\text{sm} \left( \phi_{\text{rc}}([h_i^t; h_j^t]; ||\mathcal{G}_i^{(0)} - \mathcal{G}_j^{(0)}||_{2,\text{col}}) / \tau \right) = c_{ij}.
\]

3. The \( \ell_2 \)-distance of geometric feature learning layer \( \mathcal{F}_{\text{EGFL}}^{(t)}(\cdot) \) is equivariant:

\[
\mathcal{G}^{(t+1)} R + \mathbf{t} = \mathcal{F}_{\text{EGFL}}^{(t)}(\mathcal{G}^{(t)} R + \mathbf{t}, \mathbf{h}^{(t)}, \{ c_{ij} \}).
\]

**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 \( i \)-th agent \( (i = 1, 2, \cdots, M) \),

\[
\phi_{\text{att}}^{(t)}(h_i^{(t)}) \cdot (\mathcal{G}_i^{(t)} R + \mathbf{t} - (\mathcal{G}_j^{(t)} R + \mathbf{t})) + \overline{G}_i^{(t)} R + \mathbf{t} = \phi_{\text{att}}^{(t)}(h_i^{(t)}) \cdot (\mathcal{G}_i^{(t)} - \bar{G}_i^{(t)}) R + \overline{G}_i^{(t)} R + \mathbf{t}
\]

\[
= (\phi_{\text{att}}^{(t)}(h_i^{(t)})) \cdot (\mathcal{G}_i^{(t)} - \bar{G}_i^{(t)}) R + \overline{G}_i^{(t)} R + \mathbf{t} 
\]

\[
 \rightarrow \mathcal{G}_i^{(t)} R + \mathbf{t}
\]

Thus the inner-agent attention is equivariant. We then show the interaction-aggregation is equivariant. When transforming the input geometric feature, we show the column-wise \( \ell_2 \)-distance of geometric feature \( ||\mathcal{G}_i^{(t)} - \mathcal{G}_j^{(t)}||_{2,\text{col}} \) is invariant since for the \( c \)-th column \( (c = 1, \cdots, C) \), we have \( ||\mathcal{G}_i^{(t)} R + \mathbf{t} - (\mathcal{G}_j^{(t)} R + \mathbf{t})||_2 = ||\mathcal{G}_i^{(t)} c - \mathcal{G}_j^{(t)} c||_2 \).

Thus the learned aggregation weights \( e_{ij}^{(t)} \) is invariant. We have the inter-agent aggregation’s equivari-
Thus the inner-agent attention is equivariant. We then show that the non-linear function is equivariant. When transforming the criterion is invariant and two equations under two different conditions, we can have the equivalence of the two equations under different conditions, and

\[
\mathbf{q}_{i,c}^{(t)} \mathbf{R} + \mathbf{G}_{i}^{(t)} + \mathbf{R} + t = (\mathbf{q}_{i,c}^{(t)} + \mathbf{G}_{i}^{(t)}) \mathbf{R} + t = \mathbf{g}_{i,c}^{(t+1)} \mathbf{R} + t
\]

and

\[
\mathbf{q}_{i,c}^{(t)} \mathbf{R} - \left( \mathbf{q}_{i,c}^{(t)} \mathbf{R}, \mathbf{k}_{i,c}^{(t+1)} \mathbf{R} \right) \frac{\mathbf{k}_{i,c}^{(t+1)} \mathbf{R}}{\|\mathbf{k}_{i,c}^{(t+1)} \mathbf{R}\|_2} + \mathbf{G}_{i}^{(t)} \mathbf{R} + t
\]

\[
= \mathbf{q}_{i,c}^{(t)} \mathbf{R} - \left( \mathbf{q}_{i,c}^{(t)} \mathbf{R}, \mathbf{k}_{i,c}^{(t)} \mathbf{R} \right) \frac{\mathbf{k}_{i,c}^{(t)} \mathbf{R}}{\|\mathbf{k}_{i,c}^{(t)} \mathbf{R}\|_2} + \mathbf{G}_{i}^{(t)} \mathbf{R} + t
\]

\[
= \mathbf{q}_{i,c}^{(t)} \mathbf{R} - \left( \mathbf{q}_{i,c}^{(t)} \mathbf{R}, \mathbf{k}_{i,c}^{(t)} \mathbf{R} \right) \frac{\mathbf{k}_{i,c}^{(t)} \mathbf{R}}{\|\mathbf{k}_{i,c}^{(t)} \mathbf{R}\|_2} + \mathbf{G}_{i}^{(t)} \mathbf{R} + t
\]

\[
= \left( \mathbf{q}_{i,c}^{(t)} - \left( \mathbf{q}_{i,c}^{(t)} \mathbf{R}, \mathbf{k}_{i,c}^{(t)} \mathbf{R} \right) \frac{\mathbf{k}_{i,c}^{(t)} \mathbf{R}}{\|\mathbf{k}_{i,c}^{(t)} \mathbf{R}\|_2} + \mathbf{G}_{i}^{(t)} \mathbf{R} + t
\]

\[
= \mathbf{g}_{i,c}^{(t+1)} \mathbf{R} + t
\]

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 \( t \)th pattern feature learning layer \( \mathcal{F}_{\text{HPL}}^{(t)}(\cdot) \) is invariant:

\[
\mathcal{H}_{t}^{(t+1)} = \mathcal{F}_{\text{HPL}}^{(t)}(\mathcal{G}^{(t)} \mathbf{R} + t, \mathcal{H}^{(t)}).
\]

Proof: Similar with the invariance of reasoning module, we first have the column-wise \( \ell_2 \)-distance of geometric feature

\[
||\mathbf{g}_{i}^{(t)} - \hat{\mathbf{g}}_{i}^{(t)}||_{l_2}\text{col} \text{ is invariant Thus we have the variable in the message passing } \mathbf{m}_{i}^{(t)} , \mathbf{p}_{i}^{(t)} \text{ all invariant. Finally the next layer’s pattern feature } \mathbf{h}_{i}^{(t+1)} \text{ is invariant.}
\]

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

\[
\mathcal{Y} \mathbf{R} + t = \mathcal{F}_{\text{EOL}}(\mathcal{G}^{(t)} \mathbf{R} + t).
\]

Proof: When transforming the input geometric feature,

\[
\mathcal{F}_{\text{EOL}}(\mathcal{G}^{(t)} \mathbf{R} + t)
\]

\[
= (\mathbf{W}_{\text{out}} \mathbf{g}_{i}^{(t)} \mathbf{R} + t - (\mathbf{G}_{i}^{(t)} \mathbf{R} + t)) + \mathbf{G}_{i}^{(t)} \mathbf{R} + t
\]

\[
= (\mathbf{W}_{\text{out}} \mathbf{g}_{i}^{(t)} - \mathbf{G}_{i}^{(t)} + \mathbf{G}_{i}^{(t)} \mathbf{R} + t + \mathbf{G}_{i}^{(t)} \mathbf{R} + t + \mathbf{G}_{i}^{(t)} \mathbf{R} + t)
\]

\[
= \mathcal{Y} \mathbf{R} + t
\]

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 \( \mathbf{X}_i \leftarrow \mathbf{W}_{\text{DCT}} \mathbf{X}_i - \mathbf{X} \) where \( \mathbf{W}_{\text{DCT}} \in \mathbb{R}^{T_s \times T_p} \) is the DCT coefficients matrix. Correspondingly, we transform the predicted motion by an inverse DCT (iDCT) operation: \( \hat{\mathbf{Y}}_i \leftarrow \mathbf{W}_{\text{iDCT}} \mathbf{\hat{Y}}_i + \mathbf{X} \) and \( \mathbf{W}_{\text{iDCT}} \in \mathbb{R}^{T_s \times T_p} \) is the iDCT coefficients matrix. The remove-and-add operation about the mean location \( \mathbf{X} \) 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

\[
\mathbf{G}_{i}^{(t)} \leftarrow \mathbf{\phi}_{\rho}(\mathbf{p}_{i}) + \mathbf{G}_{i}^{(t)}
\]

where \( \mathbf{\phi}_{\rho}(\cdot) \) is the velocity magnitude sequence and function \( \mathbf{\phi}_{\rho}(\cdot) \) 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 \( i \)th output produced by the \( i \)th prediction head is \( \mathbf{\hat{Y}}_i \), we use a minimum \( \ell_2 \) prediction loss formulated by

\[
\mathcal{L} = \min_i ||\mathbf{\hat{Y}}_i - \mathbf{\hat{Y}}_i||_2^2.
\]
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 historical observations 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.

<table>
<thead>
<tr>
<th>Layers</th>
<th>80ms</th>
<th>160ms</th>
<th>320ms</th>
<th>400ms</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.5</td>
<td>21.4</td>
<td>46.7</td>
<td>58.3</td>
<td>34.0</td>
</tr>
<tr>
<td>2</td>
<td>9.3</td>
<td>20.7</td>
<td>45.4</td>
<td>56.5</td>
<td>33.0</td>
</tr>
<tr>
<td>3</td>
<td>9.1</td>
<td>20.3</td>
<td>44.3</td>
<td>55.7</td>
<td>32.4</td>
</tr>
<tr>
<td>4</td>
<td>9.1</td>
<td>20.1</td>
<td>43.7</td>
<td>55.0</td>
<td>32.0</td>
</tr>
<tr>
<td>5</td>
<td>9.1</td>
<td>20.2</td>
<td>43.9</td>
<td>55.2</td>
<td>32.1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Motion</th>
<th>Walking</th>
<th>Eating</th>
<th>Smoking</th>
<th>Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>millisecond</td>
<td>80 160 320</td>
<td>400</td>
<td>80 160 320</td>
<td>400</td>
</tr>
<tr>
<td>Res-sup.</td>
<td>29.4 50.8</td>
<td>76.0 81.5</td>
<td>16.8 30.6</td>
<td>56.9 68.7</td>
</tr>
<tr>
<td>Traj-GCN</td>
<td>12.3 23.0</td>
<td>39.8 46.1</td>
<td>8.4 16.9</td>
<td>33.2 40.7</td>
</tr>
<tr>
<td>DMGNN</td>
<td>17.3 30.7</td>
<td>54.6 65.2</td>
<td>11.0 21.4</td>
<td>36.2 43.9</td>
</tr>
<tr>
<td>MSRCGN</td>
<td>12.2 22.7</td>
<td>38.6 45.2</td>
<td>8.4 17.1</td>
<td>33.0 40.4</td>
</tr>
<tr>
<td>PGBIG</td>
<td>10.2 19.8</td>
<td>34.5 40.3</td>
<td>7.0 15.1</td>
<td>30.6 38.1</td>
</tr>
<tr>
<td>SPGSN</td>
<td>10.1 19.4</td>
<td>34.8 41.5</td>
<td>7.1 14.9</td>
<td>30.5 37.9</td>
</tr>
<tr>
<td>EqMotion(Ours)</td>
<td>9.0 17.5</td>
<td>32.6 39.2</td>
<td>6.3 13.6</td>
<td>28.9 36.5</td>
</tr>
<tr>
<td>Directions</td>
<td>10.7 25.3</td>
<td>59.9 76.5</td>
<td>20.7 42.9</td>
<td>80 160</td>
</tr>
<tr>
<td>Greeting</td>
<td>8.9 18.2</td>
<td>33.8 40.9</td>
<td>8.0 16.3</td>
<td>31.3 38.2</td>
</tr>
<tr>
<td>EqMotion(Ours)</td>
<td>6.3 15.8</td>
<td>38.9 50.1</td>
<td>12.7 30.1</td>
<td>68.3 85.2</td>
</tr>
<tr>
<td>Motion</td>
<td>7.4 17.2</td>
<td>39.8 50.3</td>
<td>14.6 32.6</td>
<td>70.6 86.4</td>
</tr>
<tr>
<td>Poses</td>
<td>8.9 19.7</td>
<td>43.6 54.3</td>
<td>8.7 18.3</td>
<td>38.7 48.5</td>
</tr>
<tr>
<td>Poses</td>
<td>11.0 21.4</td>
<td>36.2 43.9</td>
<td>10.7 25.3</td>
<td>60.7 76.6</td>
</tr>
<tr>
<td>Taking-photo</td>
<td>12.2 25.8</td>
<td>59.7 76.5</td>
<td>13.9 27.9</td>
<td>57.4 71.5</td>
</tr>
<tr>
<td>NMS</td>
<td>8.9 19.7</td>
<td>43.6 54.3</td>
<td>8.7 18.3</td>
<td>38.7 48.5</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Motion</th>
<th>Walking</th>
<th>Eating</th>
<th>Smoking</th>
<th>Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>millisecond</td>
<td>560ms 1000ms</td>
<td>560ms 1000ms</td>
<td>560ms 1000ms</td>
<td>560ms 1000ms</td>
</tr>
<tr>
<td>Res-sup.</td>
<td>54.1 59.8</td>
<td>53.4 77.8</td>
<td>50.7 72.6</td>
<td>91.6 121.5</td>
</tr>
<tr>
<td>Traj-GCN</td>
<td>71.4 85.8</td>
<td>58.1 86.7</td>
<td>50.9 72.2</td>
<td>81.9 130.3</td>
</tr>
<tr>
<td>DMGNN</td>
<td>52.7 63.0</td>
<td>52.5 77.1</td>
<td>49.5 71.6</td>
<td>88.6 117.6</td>
</tr>
<tr>
<td>MSRCGN</td>
<td>48.1 56.4</td>
<td>51.1 76.0</td>
<td>46.5 69.5</td>
<td>87.1 118.2</td>
</tr>
<tr>
<td>PGBIG</td>
<td>46.9 53.6</td>
<td>49.8 73.4</td>
<td>46.7 68.6</td>
<td>89.7 118.6</td>
</tr>
<tr>
<td>SPGSN</td>
<td>43.4 52.3</td>
<td>48.4 73.0</td>
<td>41.0 63.4</td>
<td>75.3 105.6</td>
</tr>
<tr>
<td>EqMotion(Ours)</td>
<td>7.6 17.4</td>
<td>39.9 51.1</td>
<td>16.6 36.4</td>
<td>72.5 86.2</td>
</tr>
<tr>
<td>Directions</td>
<td>10.0 18.1</td>
<td>21.0 30.0</td>
<td>13.7 21.1</td>
<td>29.9 38.9</td>
</tr>
<tr>
<td>Greeting</td>
<td>9.5 17.6</td>
<td>20.1 28.8</td>
<td>12.7 21.1</td>
<td>23.4 30.8</td>
</tr>
<tr>
<td>EqMotion(Ours)</td>
<td>6.8 15.6</td>
<td>39.9 51.1</td>
<td>16.6 36.4</td>
<td>72.5 86.2</td>
</tr>
</tbody>
</table>

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


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


