IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions - Appendix

Anindita Ghosh1,2,3 Rishabh Dabral2,3 Vladislav Golyanik2,3 Christian Theobalt2,3 Philipp Slusallek1,3

1 German Research Center for Artificial Intelligence (DFKI)
2 Max-Planck Institute for Informatics (MPII)
3 Saarland Informatics Campus
https://vcai.mpi-inf.mpg.de/projects/IMoS

1. Evaluation Metrics

We evaluate our method using the Mean Per Joint Positional Error (MPJPE). It measures the mean joint error over all time steps \( T \) as

\[
\text{MPJPE} = \frac{1}{NT} \sum_{n \in N} \sum_{t \in T} \| J_t - \hat{J}_t \|_2,
\]

where \( \hat{J} \) are the joint positions computed from the synthesized SMPL-X parameters and \( N \) is the size of our test set. However, our task requires that the models synthesize a diverse set of plausible motions for any intent. Therefore, only calculating the Euclidean error with the ground-truth motion does not provide a complete picture of their synthesis quality.

Therefore, to understand the overall motion distribution statistics, we use the Average Variance Error (AVE) [GCO*21]. The Average Variance Error computes the \( L_2 \) error between the variance of the joint positions and that of the ground truth as

\[
\text{AVE} = \frac{1}{N} \sum_{n \in N} \| \sigma - \hat{\sigma} \|_2, \text{ with } \sigma = \frac{1}{T-1} \sum_{t \in T} (J_t - \hat{J})^2,
\]

where \( \hat{J} \) is the mean pose over \( T \) time steps, \( \sigma \) is the ground-truth variance, and \( \hat{\sigma} \) is the variance of the synthesized sequence.

We also report four statistical metrics, namely the Fréchet Inception Distance (FID) [HRU*17], Recognition Accuracy, Diversity, and Multimodality for a better comparison with the existing methods of Action2Motion [GZW*20] and ACTOR [PBV21].

For calculating FID, we extract features from the generated and the ground-truth motions in our test split and calculate the feature distribution between them. We train a standard RNN action recognition classifier for GRAB dataset, and use the final layer of this classifier as the motion features. A lower FID score means better quality of generated results.

Recognition accuracy indicates the correlation of the generated motions with their action types. We use the pre-trained RNN action recognition classifier to classify the motions in our test split, and calculate recognition accuracy.

Through Diversity, we measure variation in the motion features across all action categories. We sample two same-sized subsets of generated motions from various action types and extract the respective set of motion features. We calculate the Diversity between these two sets of motions as

\[
\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} \| F_i - \hat{F}_i \|_2,
\]

where \( F_1, F_2, \ldots, F_{S_d} \) and \( \hat{F}_1, \hat{F}_2, \ldots, \hat{F}_{S_d} \) are the motion feature vectors of the two subsets and \( S_d \) is the sample size.

Multimodality measures how generated motion’s features diversify within each action type. Given motion sequences from \( C \) different action types, for any \( c^t \) action, we randomly sample two subsets of the same size and extract their respective motion feature vectors. We then calculate Multimodality as

\[
\text{Multimodality} = \frac{1}{CN} \sum_{c=1}^{C} \sum_{i=1}^{S_l} \| F_{c,i} - \hat{F}_{c,i} \|_2,
\]

where \( F_{c,1}, F_{c,2}, \ldots, F_{c,S_l} \) and \( \hat{F}_{c,1}, \hat{F}_{c,2}, \ldots, \hat{F}_{c,S_l} \) are the motion feature vectors of the two subsets and \( S_l \) is the sample size.

2. CLIP-based embedding vs. random initialized vector embedding for the intent labels

We visualize the cosine similarities of the intent vectors embedded using CLIP [RKH*21] in Fig. 1. We see the embedding of intents with similar meanings such as “drink” and “pour”, “turn on” and “switch on”, “eat” and “consume”, “pass” and “transfer” have a higher cosine similarity. In contrast, the embedding of intents with different meanings such as “offhand” and “inspect”, or “switch on” and “inspect” have low similarity values between them.

In Fig. 2, we visualize the cosine similarities of the intent vectors embedded using 512 dimensional random initialized vectors (as done in Sec. 5.4 Ablation 1). We see that there is no semantic understanding between similar or dissimilar intents for the randomly initialized vector embeddings.
Figure 1: Confusion matrix showing the cosine similarity percentage where intent labels are encoded using CLIP [RKH*21]. We see embeddings of intents with similar meanings have a higher cosine similarity whereas embeddings of intents with different meanings have low similarity values.

References


Figure 2: Confusion matrix showing the cosine similarity percentage where intent labels are encoded using 512 dimensional random initialized vectors (as done in Sec. 5.4 Ablation 1). We see no semantic relation between similar intents when using a randomly initialized embedding.