EUROGRAPHICS 2023 / K. Myszkowski and M. Nießner
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 2

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

Anindita Ghosh[1,2,3] Rishabh Dabral[2,3] Vladislav Golyanik[2,3] Christian Theobalt[2,3] Philipp Slusallek[1,3]

[1] German Research Center for Artificial Intelligence (DFKI)
[2]Max-Planck Institute for Informatics
[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

$$\text{MPJPE} = \frac{1}{NT} \sum_{n \in N} \sum_{t \in T} \left\| J_t - \hat{J}_t \right\|_2 , \tag{1}$$

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 type of 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 [GCO*21]. The Average Variance Error (AVE) computes the L_2 error between the variance of the joint positions and that of the ground truth as

$$\text{AVE} = \frac{1}{N} \sum_{n \in N} \|\sigma - \hat{\sigma}\|_2 , \text{ with } \sigma = \frac{1}{T-1} \sum_{t \in T} \left(J_t - \widetilde{J} \right)^2 , \tag{2}$$

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

We also report four statistical metrics namely the Frechet 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 motions in our test split and the ground-truth motions in the test split, and calculate the feature distribution between the generated motions and the ground-truth motion. To extract the motion features, 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. The Diversity between these two set of motions is calculated as

$$\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} \left\| F_i - \hat{F}_i \right\|_2 , \tag{3}$$

where $F_1, F_2, ..., F_{S_d}$ and $\hat{F}_1, \hat{F}_2, ..., \hat{F}_{S_d}$ 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^{th} action, we randomly sample two subsets of same size and extract their respective motion feature vectors. Multimodality is then calculated as

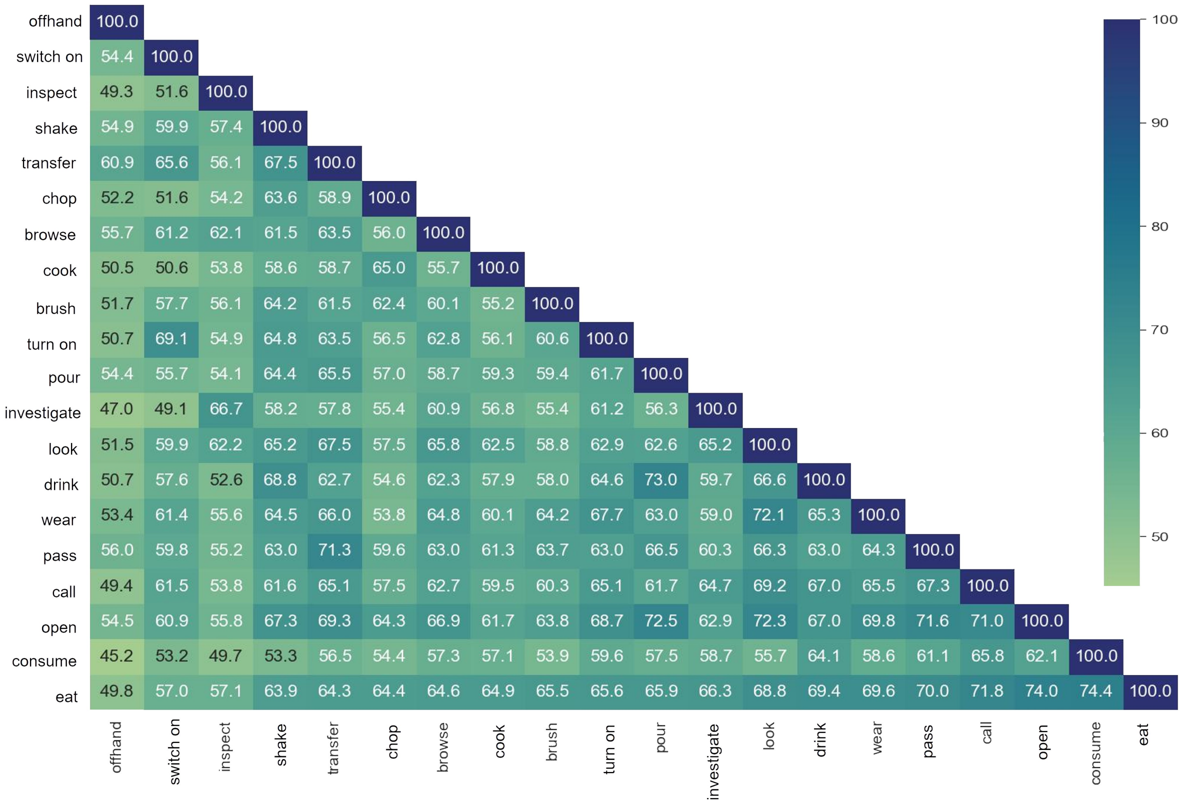
$$\text{Multimodality} = \frac{1}{CS_l} \sum_{c=1}^{C} \sum_{i=1}^{S_l} \left\| F_{c,i} - \hat{F}_{c,i} \right\|_2 , \tag{4}$$

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

2. CLIP based embedding vs. randomly initialized vector embeddings for the intent labels

We visualize the cosine similarities of the intent vectors embedded using CLIP [RKH*] in Fig. 1. We see embeddings 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 whereas embeddings 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 between intent embeddings where the intent labels are embedded using randomly initialized vectors of 512 dimension (as done in Sec. 5.4 Ablation 1). We see that for the randomly initialized vector embeddings, there is no semantic understanding between similar or dissimilar intents.



	offhand	switch on	inspect	shake	transfer	chop	browse	cook	brush	turn on	pour	investigate	look	drink	wear	pass	call	open	consume	eat
offhand	100.0																			
switch on	54.4	100.0																		
inspect	49.3	51.6	100.0																	
shake	54.9	59.9	57.4	100.0																
transfer	60.9	65.6	56.1	67.5	100.0															
chop	52.2	51.6	54.2	63.6	58.9	100.0														
browse	55.7	61.2	62.1	61.5	63.5	56.0	100.0													
cook	50.5	50.6	53.8	58.6	58.7	65.0	55.7	100.0												
brush	51.7	57.7	56.1	64.2	61.5	62.4	60.1	55.2	100.0											
turn on	50.7	69.1	54.9	64.8	63.5	56.5	62.8	56.1	60.6	100.0										
pour	54.4	55.7	54.1	64.4	65.5	57.0	58.7	59.3	59.4	61.7	100.0									
investigate	47.0	49.1	66.7	58.2	57.8	55.4	60.9	56.8	55.4	61.2	56.3	100.0								
look	51.5	59.9	62.2	65.2	67.5	57.5	65.8	62.5	58.8	62.9	62.6	65.2	100.0							
drink	50.7	57.6	52.6	68.8	62.7	54.6	62.3	57.9	58.0	64.6	73.0	59.7	66.6	100.0						
wear	53.4	61.4	55.6	64.5	66.0	53.8	64.8	60.1	64.2	67.7	63.0	59.0	72.1	65.3	100.0					
pass	56.0	59.8	55.2	63.0	71.3	59.6	63.0	61.3	63.7	63.0	66.5	60.3	66.3	63.0	64.3	100.0				
call	49.4	61.5	53.8	61.6	65.1	57.5	62.7	59.5	60.3	65.1	61.7	64.7	69.2	67.0	65.5	67.3	100.0			
open	54.5	60.9	55.8	67.3	69.3	64.3	66.9	61.7	63.8	68.7	72.5	62.9	72.3	67.0	69.8	71.6	71.0	100.0		
consume	45.2	53.2	49.7	53.3	56.5	54.4	57.3	57.1	53.9	59.6	57.5	58.7	55.7	64.1	58.6	61.1	65.8	62.1	100.0	
eat	49.8	57.0	57.1	63.9	64.3	64.4	64.6	64.9	65.5	65.6	65.9	66.3	68.8	69.4	69.6	70.0	71.8	74.0	74.4	100.0

Figure 1: Confusion matrix showing the cosine similarity percentage where intent labels are encoded using CLIP [RKH*]. 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

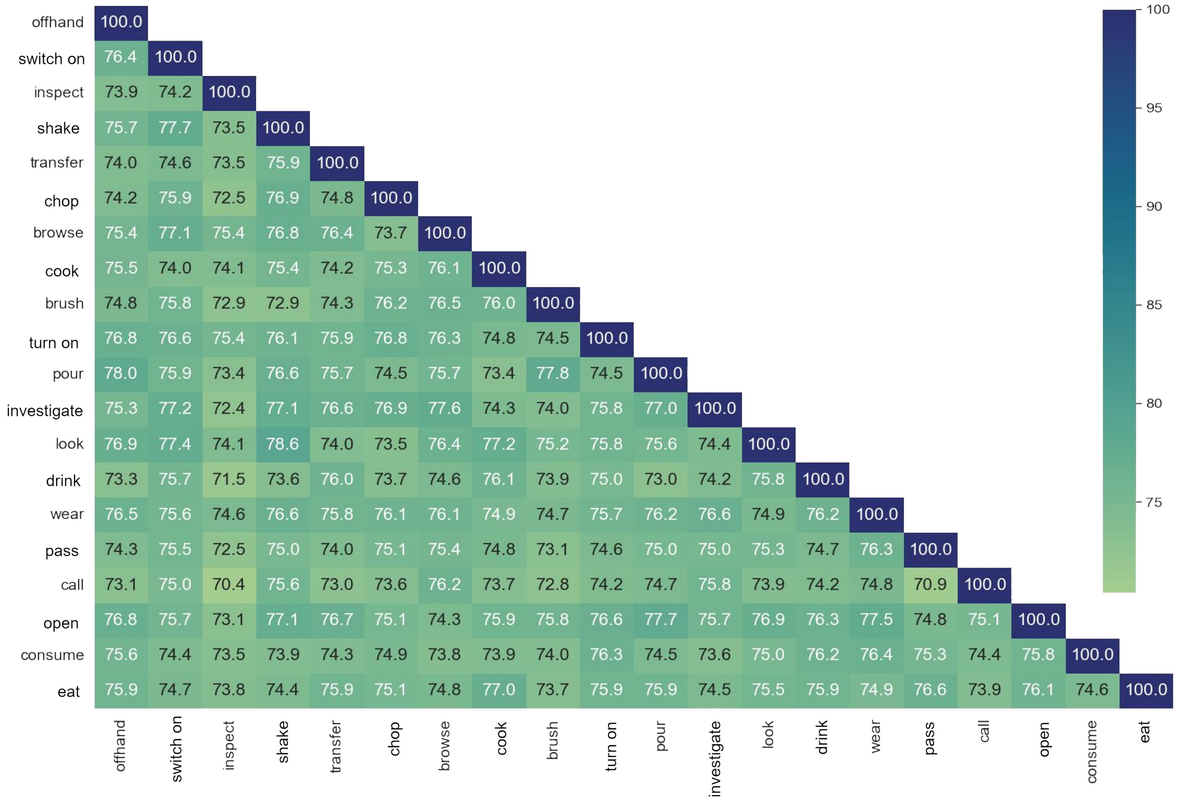
[GCO*21] Ghosh, Anindita, Cheema, Noshaba, Oguz, Cennet, et al. "Synthesis of Compositional Animations From Textual Descriptions". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 1.

[GZW*20] Guo, Chuan, Zuo, Xinxin, Wang, Sen, et al. "Action2motion: Conditioned generation of 3d human motions". *Proceedings of the 28th ACM International Conference on Multimedia*. 2020 1.

[HRU*17] Heusel, Martin, Ramsauer, Hubert, Unterthiner, Thomas, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". *Advances in neural information processing systems* (2017) 1.

[PBV21] Petrovich, Mathis, Black, Michael J., and Varol, Gül. "Action-Conditioned 3D Human Motion Synthesis with Transformer VAE". *International Conference on Computer Vision (ICCV)*. 2021 1.

[RKH*] Radford, Alec, Kim, Jong Wook, Hallacy, Chris, et al. "Learning Transferable Visual Models From Natural Language Supervision". *Proceedings of the 38th International Conference on Machine Learning* 1, 2.



	offhand	switch on	inspect	shake	transfer	chop	browse	cook	brush	turn on	pour	investigate	look	drink	wear	pass	call	open	consume	eat
offhand	100.0																			
switch on	76.4	100.0																		
inspect	73.9	74.2	100.0																	
shake	75.7	77.7	73.5	100.0																
transfer	74.0	74.6	73.5	75.9	100.0															
chop	74.2	75.9	72.5	76.9	74.8	100.0														
browse	75.4	77.1	75.4	76.8	76.4	73.7	100.0													
cook	75.5	74.0	74.1	75.4	74.2	75.3	76.1	100.0												
brush	74.8	75.8	72.9	72.9	74.3	76.2	76.5	76.0	100.0											
turn on	76.8	76.6	75.4	76.1	75.9	76.8	76.3	74.8	74.5	100.0										
pour	78.0	75.9	73.4	76.6	75.7	74.5	75.7	73.4	77.8	74.5	100.0									
investigate	75.3	77.2	72.4	77.1	76.6	76.9	77.6	74.3	74.0	75.8	77.0	100.0								
look	76.9	77.4	74.1	78.6	74.0	73.5	76.4	77.2	75.2	75.8	75.6	74.4	100.0							
drink	73.3	75.7	71.5	73.6	76.0	73.7	74.6	76.1	73.9	75.0	73.0	74.2	75.8	100.0						
wear	76.5	75.6	74.6	76.6	75.8	76.1	76.1	74.9	74.7	75.7	76.2	76.6	74.9	76.2	100.0					
pass	74.3	75.5	72.5	75.0	74.0	75.1	75.4	74.8	73.1	74.6	75.0	75.0	75.3	74.7	76.3	100.0				
call	73.1	75.0	70.4	75.6	73.0	73.6	76.2	73.7	72.8	74.2	74.7	75.8	73.9	74.2	74.8	70.9	100.0			
open	76.8	75.7	73.1	77.1	76.7	75.1	74.3	75.9	75.8	76.6	77.7	75.7	76.9	76.3	77.5	74.8	75.1	100.0		
consume	75.6	74.4	73.5	73.9	74.3	74.9	73.8	73.9	74.0	76.3	74.5	73.6	75.0	76.2	76.4	75.3	74.4	75.8	100.0	
eat	75.9	74.7	73.8	74.4	75.9	75.1	74.8	77.0	73.7	75.9	75.9	74.5	75.5	75.9	74.9	76.6	73.9	76.1	74.6	100.0

Figure 2: Confusion matrix showing the cosine similarity percentage where action labels are encoded using randomly initialized vectors of 512 dimension (as done in Sec. 5.4 Ablation 1). We see that there is no semantic understanding between similar or dissimilar intents when using random initialization.