GIMO: Gaze-Informed Human Motion Prediction in Context Supplementary Document

1 Experimental Setup

1.1 Implementation Details

As demonstrated in the Fig. 3 of the main paper, our method is built on Pointnet++ [8] and a cross-modal transformer [2]. A 256D global feature F_O and a 256D per-point feature map F_P of the scene are extracted from the input point cloud. The feature of an arbitrary point e is computed through the inversed distance weighted interpolation on the 3 nearest neighbors of e from the scene point cloud (Eq. 2 of the main paper), where we query the 256D gaze feature f_g and obtain the 256D scene context feature f_{m_v} of the current motion from SMPL-X per-vertex features. The 32D motion parameter x is embedded into 256D motion feature f_m through a linear layer. The motion embedding is then fed to a motion-scene transformer with f_{m_v} as query and further fed to another motion-gaze transformer with gaze feature f_g as the query. The gaze feature is updated by a gaze-motion transformer queried by motion feature f_m. We then concatenate the global scene feature F_O, the updated motion feature f_{m_g} and gaze feature f_{g_m} to get the 768D multi-modal embedding, which is used to predict the 32D future motion parameter by a cross-modal transformer. All the transformers adopt a 6 layer architecture as proposed in [2] with 256D latent embedding. Note that here the input and output motion parameter x consists of a 3D global translation vector t, a 3D global orientation vector r (represented as axis angle), and a 32D pose embedding h obtained from VPoser [6]. We omit predicting the hand poses p and the shape parameter β since the global body pose can be well represented by parameter $\{t, r, h\}$, and we aim at future work to include hand poses and the body shape for more detailed motion prediction.

For the baseline methods, we re-implement spatio-temporal transformer [1], a RNN based network [4], and MultimodalNet [5] to adapt for our experimental settings. The 3D joint angle representation is used as motion input and output to train the spatio-temporal transformer and RNN as introduced in [4], while MultimodalNet is based on the 32D motion parameter the same as ours. An 8 layer transformer [9] with 512 embedding size and 8 heads attention is used in spatio-temporal transformer [1]. A three layer RNN with 1024 hidden size is deployed to predict future motion with simple motion input or motion and gaze input [3]. In MultimodalNet [5], the motion input is firstly embedded into 256D feature space through linear layers and then fed to a transformer encoder to get the motion embeddings. The gaze embedding is also obtained with linear layers and a transformer encoder. The global scene feature from PointNet [7], the

Fig. 1: An overview of the scanned scenes in our dataset.

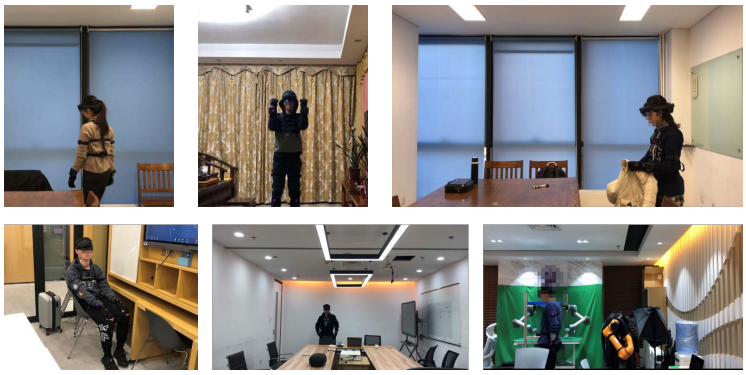

Fig. 2: Subjects in the scenes.

gaze embedding and the motion embedding are stacked and fed to a transformer decoder to generate future motion. The transformer encoders and decoder are all based on a 6 layer architecture with 256 latent size. Therefore, all the baselines share similar network capacity with our method.

1.2 Training Loss

We employ the L1 loss between the predicted motion parameter and ground truth to train our method. The full loss consists of translation loss, orientation loss and pose embedding loss. The translation loss is formulated as:

$$\mathcal{L}_{trans} = \frac{1}{T} \sum_{k=1}^{T} ||\widehat{t_k} - t_k||_1 \tag{1}$$

Fig. 3: Motion trajectories from our dataset. Better visualized in the supplementary video.

where T is the length of output pose, and $\widehat{t_k}$ is the predicted global translation parameter of the k-th pose in the T-length future motion, and t_k is the ground truth. We compute the orientation loss as:

$$\mathcal{L}_{ori} = \frac{1}{T} \sum_{k=1}^{T} ||\widehat{r_k} - r_k||_1 \tag{2}$$

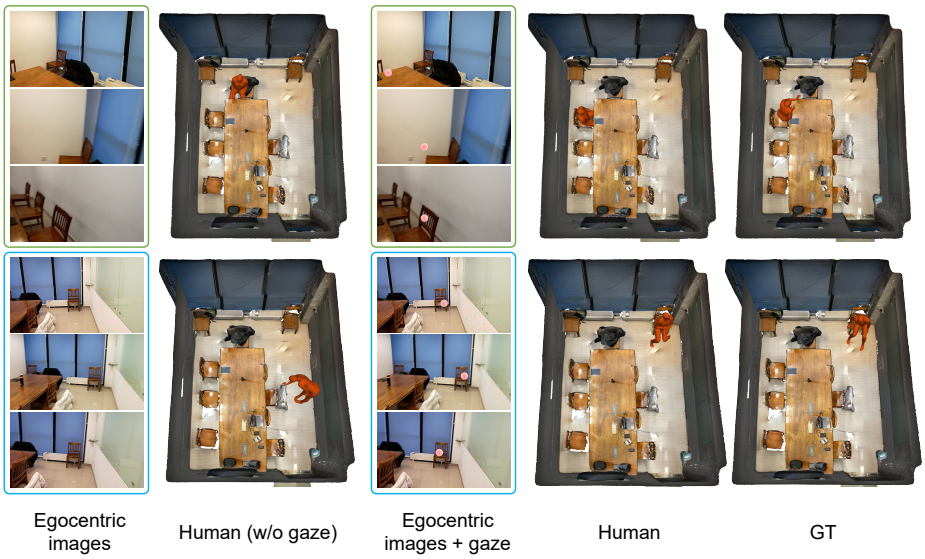

Fig. 4: Human evaluation. Two human subjects are required to watch a egocentric video (without gaze or with gaze) and infer the final pose of the trajectory. The subjects choose a pose from a pose database which comes from the training set, and put the pose into the 3D scene as the final position of the motion according to the egocentric video. We show that humans can easily solve the task with the intention clues extracted from gaze, while without the gaze information even human intelligence can be confused.

where $\widehat{r_k}$ is the predicted global orientation parameter. The pose embedding loss is designed as:

$$\mathcal{L}_p = \frac{1}{T} \sum_{k=1}^{T} ||\widehat{h_k} - h_k||_1 \tag{3}$$

where $\widehat{h_k}$ is the predicted pose embedding. Finally, the full loss is formulated as:

$$\mathcal{L} = \lambda_t \mathcal{L}_{trans} + \lambda_o \mathcal{L}_{ori} + \lambda_p \mathcal{L}_p \tag{4}$$

where we set $\lambda_t, \lambda_o, \lambda_p$ to 1 during training.

2 GIMO Dataset

Our dataset consists of 217 motion trajectories collected in 19 scenes by 11 subjects. Fig. 1 provides an overview of the scanned scenes in our dataset, which cover a wide range of daily indoor environments, including living rooms, meeting rooms, library, lab, etc. Fig. 2 shows the recruited subjects collecting data in the scenes. More motion trajectories are demonstrated in Fig. 3. For better visualization, please refer to the supplementary video.

Fig. 5: More Qualitative results.

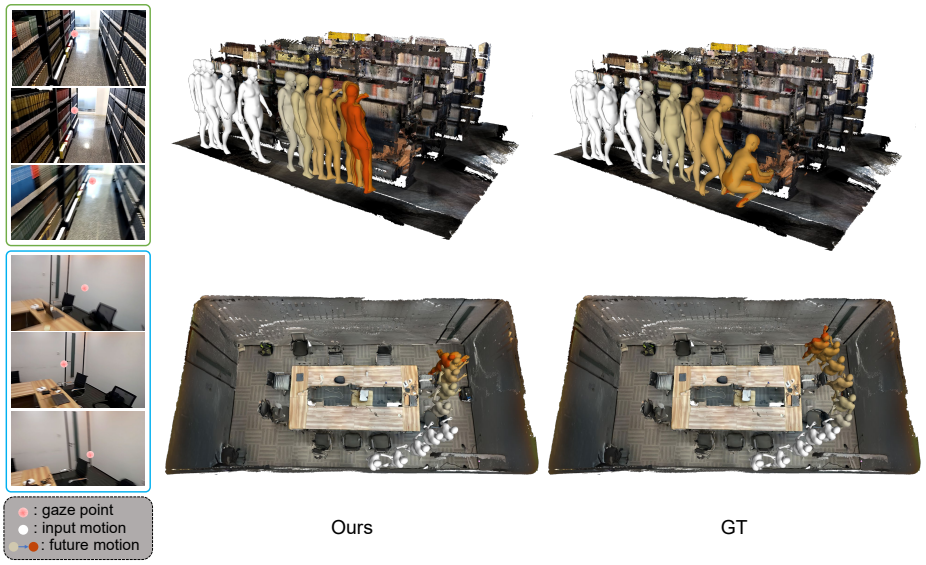

Fig. 6: Failure cases of our method. When the noisy gazes account for a large portion of the input, our method is confused to interpret the subject's intention.

3 More Results

3.1 Human Evaluation

We conduct a human evaluation experiment to validate the function of gaze in disambiguating future motion prediction. For simplification, the subjects predict

the final pose of the motion instead of the full motion trajectory. To this end, two human subjects are recruited and required to watch an ego-centric video (without gaze or with gaze) and infer the final pose of the trajectory. The subjects first choose a pose from a pose database which is constructed by poses from the training set, and then put the pose into the 3D scene as the final position of the motion according to the ego-centric video they have seen. Fig. 4 shows that humans can easily extract the intention clues from the gaze and solve the problem accurately, while without the gaze information even human intelligence can be confused.

3.2 More Results of Baselines and Failure cases

Fig. 5 provides more results of the baseline methods, further demonstrating the superiority of our method in predicting future motion from the multi-modal gaze, motion and scene information. However, we find that when the input gazes are quite noisy which convey little intention clues, our method can fail to interpret the subject's goal and generate inaccurate results, as shown in 6. Since our method predicts future motion from sparse inputs (2fps), the uninformative gazes can account for a large portion of the input. The problem might be mitigated by leveraging high fps inputs since we find that in the recorded sequences the most attention is paid to objects related to the destination of the motion.

References

1. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction. In: 2021 International Conference on 3D Vision (3DV). pp. 565–574. IEEE (2021)
2. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
3. Kratzer, P., Bihlmaier, S., Midlagajni, N.B., Prakash, R., Toussaint, M., Mainprice, J.: Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. IEEE Robotics and Automation Letters **6**(2), 367–373 (2020)
4. Kratzer, P., Toussaint, M., Mainprice, J.: Prediction of human full-body movements with motion optimization and recurrent neural networks. In: 2020 ICRA. pp. 1792–1798 (2020)
5. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021)
6. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
7. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)

8. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)