

GIMO: Gaze-Informed Human Motion Prediction in Context

Yang Zheng^{1,2} Yanchao Yang^{1,*} Kaichun Mo¹ Jiaman Li¹
Tao Yu² Yebin Liu² C. Karen Liu¹ Leonidas J. Guibas¹

¹ Stanford University ² Tsinghua University

Abstract. Predicting human motion is critical for assistive robots and AR/VR applications, where the interaction with humans needs to be safe and comfortable. Meanwhile, an accurate prediction depends on understanding both the scene context and human intentions. Even though many works study scene-aware human motion prediction, the latter is largely underexplored due to the lack of ego-centric views that disclose human intent and the limited diversity in motion and scenes. To reduce the gap, we propose a large-scale human motion dataset that delivers high-quality body pose sequences, scene scans, as well as ego-centric views with the eye gaze that serves as a surrogate for inferring human intent. By employing inertial sensors for motion capture, our data collection is not tied to specific scenes, which further boosts the motion dynamics observed from our subjects. We perform an extensive study of the benefits of leveraging the eye gaze for ego-centric human motion prediction with various state-of-the-art architectures. Moreover, to realize the full potential of the gaze, we propose a novel network architecture that enables bidirectional communication between the gaze and motion branches. Our network achieves the top performance in human motion prediction on the proposed dataset, thanks to the intent information from eye gaze and the denoised gaze feature modulated by the motion. Code and data can be found at <https://github.com/y-zheng18/GIMO>.

1 Introduction

A large portion of the human brain cortex is devoted to processing visual signals collected by the optic nerve, and over half of the nerve fibers carry information from the fovea that is responsible for sharp central vision. When modulated through foveal fixation, or equivalently, *eye gaze*, important sensory input of fine details perceived with the fovea can inform future actions of the human agent [42, 8]. As shown in Fig. 1, a human agent intending to perform two tasks entails distinctive gaze patterns, even though the first few moves are not very distinguishable. Hence, it is beneficial to employ eye gaze when making human

* Corresponding author: yanchao@cs.stanford.edu

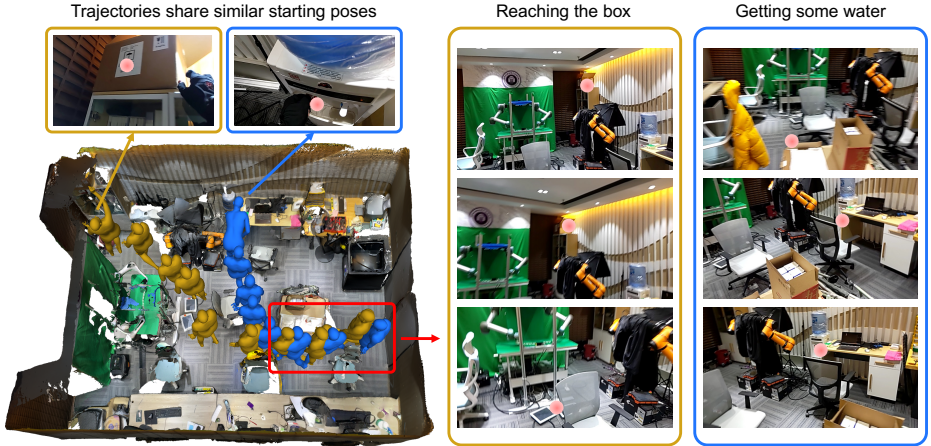


Fig. 1. Human motion driven by different intents look similar at the beginning. However, the scanning patterns of the eye gaze (red dots) during the starting phase are pretty distinctive, which suggests that we can leverage eye gaze to reduce uncertainties when predicting future body movements.

motion predictions in the 3D scene, which is of great importance for human-machine interactions [1,6]. For example, a human agent wearing an AR/VR headset may approach a chair to sit on it or just grab a cup on the table behind it. If the latter is true, we may want the headset to send out a warning for collision avoidance based on the forecast future. To resolve ambiguities for reliable human motion prediction, there is an increasing interest in leveraging eye gaze as it highly correlates to the underlying intent that motivates the consequent actions.

The key to understanding the role of gaze and how it can effectively inform human motion prediction lies in two folds. First, it is critical to have a dataset with high-quality 3D body pose annotations and corresponding eye gaze information. Besides data quality, the 3D scene and motion dynamics should be diverse to enable meaningful learning and evaluation of the gain when eye gaze is incorporated. Second, it is also crucial to have a network architecture that can efficiently utilize *sparse* eye gaze during predictions given the multi-modal setting (e.g., gaze, human motion and scene geometry) and the fact that *not* every single gaze is of the *same* significance regarding the agent’s intent (e.g., one may get distracted by a salient object in the scene that has nothing to do with the task at hand).

However, most existing human motion datasets *do not* support evaluating the effect of eye gaze due to the lack of ego-centric data annotated with both gaze and 3D body pose within the same scene. Recently, there are a few datasets proposed on ego-centric social interaction and object manipulation where gaze and the viewer’s 3D poses are available. Nevertheless, they are not suitable for ego-centric human motion prediction since the diversity of scenes and the variation in motion dynamics are very limited. To validate the benefits of eye gaze in *human motion prediction*, we propose a large-scale ego-centric dataset, which

contains the scene context, eye gaze, and accurate 3D body poses of the human actors. By employing an advanced motion capture system based on Inertial Measurement Units (IMUs), we can collect 3D pose data with high fidelity and avoid the limits of conventional multi-camera systems. For example, the actor can walk through any environment without performing a cumbersome setup of motion capture devices. Moreover, accurate poses can be recorded without any 2D-3D lifting, which could induce errors due to occlusions and noise in the detection. These advantages enable the actors to perform various long-horizon activities in a diverse set of daily living environments.

In order to check the effectiveness of eye gaze in improving human motion prediction, we perform an extensive study with multiple state-of-the-art architectures. However, we note that gaze and motion could both be inherently ambiguous in forecasting future movements. For example, the gaze may be allocated to a TV monitor while walking towards the dining table. In this case, the actor may simply follow the momentum, thus rendering the eye gaze uninformative about the body motion. To utilize the full potential of eye gaze in human motion prediction, we further propose a novel architecture that manifests cross-modal attention such that *not only* future motion can benefit from the eye gaze, *but also* the significance of gaze in predicting the future can be reinforced by the observed motion. In our experiments, better human motion predictions are observed across various architectures. Furthermore, the proposed architecture achieves the top performance measured under different criteria, verifying the effectiveness of our bidirectional fusion scheme.

In summary, we make the following contributions. First, we provide a large-scale human motion dataset that enables investigating the benefits of eye gaze under diverse scenes and motion dynamics. Second, we propose a novel architecture with a bidirectional multi-modal fusion that better suits gaze-informed human motion prediction through mutually disambiguating motion and gaze. Finally, we validate the usefulness of eye gaze for human motion prediction with multiple architectures and verify the effectiveness of the proposed architecture by showing top performance on the proposed dataset.

2 Related Work

Datasets for human motions. Human motion modeling is a long-standing problem and is extensively explored with high-quality motion capture datasets, ranging from small-scale CMU Graphics Lab Motion Capture Database [5] to large-scale ones like AMASS [31]. Human3.6M [13] captures high-quality motions using a multi-view camera system and serves as a standard benchmark for motion prediction and 3D pose estimation. While these datasets provide adequate data to learn motion dynamics, the constraints from the 3D environment are usually not included. Later, more datasets containing the 3D scene are proposed, and scene-aware motion prediction can be studied using GTA-1M dataset [4]. PROX [11] includes both 3D scene and human interaction motions which can be used to explore scene-aware motion generation [51] task and the problem of placing human to the scene [59,60]. As the data is always collected with a

human agent, ego-centric videos are provided in EgoPose [55,54], Kinpoly [30] and HPS [9] to study how the motion estimation and prediction can benefit from these ego-centric observations. Moreover, social interaction is considered in You2Me [36] and EgoBody [57]. However, existing datasets do not contain diverse 3D scenes and human motions with intentions, we collect a large-scale dataset for gaze-guided human motion prediction, and it consists of high-quality human motions, 3D scene, ego-centric video and corresponding eye gaze information.

Human motion prediction. RNNs have proven successful in modeling human motion dynamics [7,34,27,3]. [32] proposes an attention-based model to guide the future prediction with motion history. To effectively exploit both spatial and temporal dependencies in human pose sequences, ST-Transformer [2] designs a spatial temporal transformer architecture to model the human motions. Pose Transformers [35] investigates a non-autoregressive formulation using transformer model and shows superior performance in terms of both efficiency and accuracy. As human motions are tightly correlated with the scene context, scene-aware motion prediction is also actively studied [4,10]. A three-stage pipeline is established to predict long-term human motions conditioned on the scene context [4]. SAMP [10] further includes object geometry to estimate interaction positions and orientations, and generates motions following a calculated collision-free trajectory. Besides the scene constraints, other modalities such as gaze and music also provide clues for future motion prediction. Transformer [48] is applied to generate dance movements conditioned on music [24,25,47]. MoGaze [21] verifies the effectiveness of eye gaze information for motion prediction with an RNN model in a full-body manipulation scenario. Our work aims to predict long-term future motions with both 3D scene and gaze constraints. We differ from existing motion prediction works, as their focus is the dense motion predictions, while we are predicting long-term sparse motions to understand human intentions.

Human motion estimation. 3D pose estimation is extensively studied in third-person view images or videos [43,56,19,20,18,29,12]. VIBE [18] propose a sequential model to estimate human poses and shapes from videos, along with a motion discriminator to constrain the predictions in a plausible motion manifold. TCMR [12] explicitly enforces the neural nets to leverage past and future frames to eliminate jitters in predictions. Motion priors are founded effective in improving the temporal smoothness and tackling the occlusion issues [40,23,58]. More attentions are received in ego-centric pose estimation recently. Pose estimation from images captured using a fish eye camera is explored in [45,53,41,44,50]. [15] deploy a chest-mounted camera and predict motions based on an implicit motion graph. Following the chest-mounted camera setting, You2Me [36] introduces the motions of the visible second person as an additional signal to constrain the motion estimation of the camera wearer. [55,54,30] explores motion estimation and prediction with head-mounted front-facing camera. In this work, we are addressing the ego-centric motion prediction task where past motions are given. Our proposed dataset can benefit the ego-centric motion estimation problem.

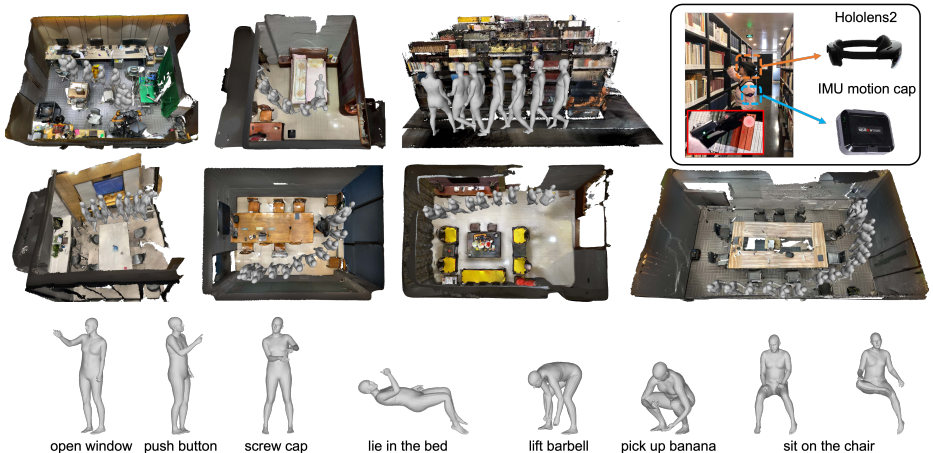


Fig. 2. We collect human motion data in various indoor environments (1st, 2nd rows), allowing the human subject to perform a diverse range of daily activities exhibiting rich dynamics (bottom). Top-right: motion and gaze capture devices.

3 GIMO Dataset: Gaze and Motion with Scene Context

Human motion is affected by the scene, which provides physical constraints and the agent’s psychological demand that drives body movements. To have a concrete assessment of the benefits induced by eye gaze, we need both ego-centric views, and 3D body poses of the agent. Particularly, they should be temporally synchronized and spatially aligned within the 3D scenes. Current datasets for human motion prediction are either collected in a virtual environment risking being unrealistic or captured by an array of cameras with limited scene diversity and motion dynamics. Moreover, eye gaze is usually not available.

Therefore, we propose a real-world large-scale dataset that provides high-quality human motions, ego-centric views with eye gaze, as well as 3D environments. Next, we describe our data collection pipeline.

3.1 Hardware Setup

We employ a commercialized IMU-based motion capture system to record high-quality 3D body poses of the human agent, whose eye gaze in 3D is detected using an AR device mounted on the head. The 3D scenes are scanned by a smartphone equipped with lidar sensors (please see Fig. 2, top-right).

Motion capture. To capture daily activities in various indoor environments, we resort to motion capture from IMU signals following HPS [9]. While HPS only provides SMPL [28] models with body movements, we take advantage of an advanced commercial product *Noitom PERCEPTION NEURON STUDIO*,¹ which can record at 96 fps 3D body and hand joint movement of the subject.

¹ <https://noitom.com/perception-neuron-series>

Table 1. Statistics of existing and our datasets. * means the 3D scene is virtual, e.g., from game engine [4] or CAD models [10]. *Ego* denotes egocentric images are available, and *Intent* indicates whether the motions have clear semantic intentions, e.g., fetching a book.

Dataset	Frame	Sub.	3D scene	Ego	Gaze	3rd-person	Human pose from	Parametric model	Intent	Task
EGTEA Gaze+ [26]	2419k	32		✓	✓				✓	Action recognition
TIA [52]	330k	-			✓	✓			✓	Attention prediction
Human3.6M [13]	3600k	11				✓	Marker-based			Pose estimation
TNT15 [49]	13k	4				✓	RGB+IMU			Pose estimation
3DPW [33]	51k	7				✓	RGB+IMU	SMPL		Pose estimation
Panoptic [16]	297k	180+				✓	Multi-RGB			Pose estimation
TotalCapture [17]	1,900k	5				✓	Multi-RGB	Frank		Pose estimation
HPS [9]	300k	7	✓	✓			IMU	SMPL		Pose estimation
EgoBody [57]	153k	20	✓	✓	✓	✓	Multi-RGB-D	SMPL-X		Pose estimation
EgoMoCap [30]	148k	3		✓			Marker-based			Pose estimation
PROX [11], [60]	100k	20	✓			✓	RGB	SMPL-X		Human generation
GTA-IM [4]	1000k		✓*			✓	Game engine			Motion prediction
SAMP [10]	1	1	✓*				Marker-based	SMPL-X	✓	Motion prediction
GIMO (ours)	129k	11	✓	✓	✓		IMU	SMPL-X	✓	Motion prediction

To obtain the full-body pose and hand gesture of the subject, we apply SMPL-X [37] model to fit the recorded IMU signals from multiple joints. Compared to human motion datasets like PROX [11], where the 3D body pose is estimated from monocular RGB videos, the pose obtained using the above procedure is free from estimation errors caused by noisy detection and occlusions. Fitting parametric human body models for poses from multi-view RGB(D) streams or with marker-based systems is also commonly used to collect human motion data [17, 57, 13], however, our pipeline requires much less effort in presetting the environment; thus, we can collect human motion data in any indoor scene. These characteristics endow us with the capability to ensure the diversity of the scene and motion dynamics in our dataset.

Gaze capture. Following [57], we use Hololens2² and its Research Mode API [46] to capture the 3D eye gaze. It also records ego-centric video at 30 fps in 760×428 resolution, long-throw depth streams at 1-5 fps in 512×512 , and 6D poses of the head-mounted camera. The 3D scene is reconstructed through TSDF fusion given the recorded depth, which is used for the subsequent global alignment. The eye gaze is recorded as a 3D point in the coordinate system of the headset.

3D scene acquisition. To obtain high-quality 3D geometry of the scene (the reconstructed TSDF result from Hololens2 is usually noisy), we use an iPhone13 Pro Max equipped with LiDAR sensors to scan the environment through 3D Scanner APP³. The output mesh contains about 500k vertices and photorealistic texture, providing sufficient details to infer the affordance of the scene. The data collection process involving human agents and the alignment of different coordinate frames to the scanned meshes are described in the following.

² <https://www.microsoft.com/en-us/hololens>

³ <https://apps.apple.com/us/app/3d-scanner-app/id1419913995>

Table 2. Activities performed by our subjects.

Category		Activities
Resting		Sitting or laying on objects
Interacting with objects	Touching, holding, stepping on, reaching to objects	
	Opening, pushing, transferring, throwing,	
Changing the state of objects	picking up, lifting, connecting, screwing,	
	grabbing, swapping objects	

3.2 Data Collection with Human in Action

One distinct feature of our dataset is that it captures long-term motions with clear intentions. Different from prior datasets for motion estimation purposes where the subjects are performing random actions such as jumping and waving hands, we aim at collecting motion trajectories with semantic meaning, e.g., walk to open the door. Thus, we focus on collecting data from various daily activities in indoor scenes. The full statistics of our dataset are listed in Tab. 1.

To this end, we recruit 11 university students (4 female and 7 male) and ask them to perform the activities defined in Tab. 2. The subjects are instructed to start from a distant location to the goal object and then move to the destination to act. Therefore, long-term motion with clear intention can be obtained. Especially, the collection progress includes the following steps: (i) the subject wears the head-mounted Hololens2, the IMU-based motion capture clothes, and gloves, where calibration is performed to set up the motion capture system; (ii) the subject chooses the action from the activities in Tab. 2 according to the affordance of the scene; (iii) the 3D scene is scanned; (iv) the subject starts to carry out the planned activities in the scene while data are collected; (v) the scene is reset for the following subjects to perform their activities. Note, if the subject changes the scene geometry, we reset the objects to their original states to avoid rescanning the whole environment.

As a result, our dataset contains 129k ego-centric images, 11 subjects, and 217 motion trajectories in 19 scenes, manifesting enough capacity and diversity for gaze-informed human motion prediction. As illustrated in Fig. 2, the motions are smooth and convey clear semantic intentions.

3.3 Data Preparation

Synchronization. Given compatibility issues, it is difficult to synchronize the motion capture system with Hololens2 without modifying their commercialized software. Instead, we use a hand gesture that can be observed in the ego-centric view as a starting signal. Once the pose and ego-centric image of the hand gesture are aligned, the rest frames can be synchronized according to the timestamps.

Parametric model fitting. To obtain the 3D body pose and shape of the subject, we fit SMPL-X [37] model to the 3D joints (23 body joints, 15 left-hand joints, and 15 right-hand joints), which are computed from the recorded IMU signals by the provided commercial software. In addition, the 6D head pose is used to determine the head position and orientation of the SMPL-X model.

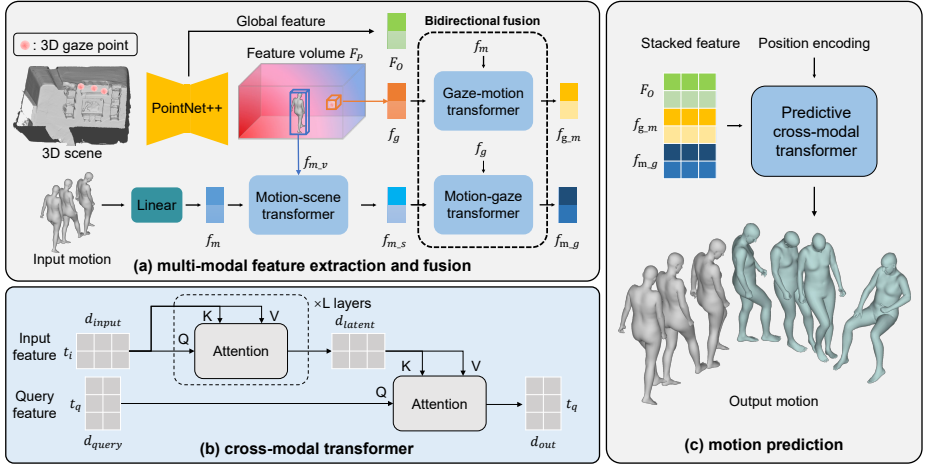


Fig. 3. Our gaze-informed human motion prediction architecture. Multi-modal features, i.e., gaze feature, human motion feature, and global scene feature, are extracted and then fused through the proposed *bidirectional* fusion scheme (a). The fused features are then stacked into a holistic representation and used for future motion prediction (c). The cross-modal transformer component [14] is illustrated in (b). Please refer to Sec. 4 for more details.

Alignment. The Hololens2 coordinate system and the fitted SMPL-X models need to align with the high-quality 3D scene scans. The former is aligned through ICP between the TSDF fusion result of the depth recorded by Hololens2 and the 3D scene. The SMPL-X motion sequence is first transformed to the Hololens2 coordinate system via human annotations, i.e., the start and end shapes of the human body are scanned by Hololens2 and visible in the TSDF reconstruction, which serves as anchor shapes for aligning the fitted models. The pose can then be aligned to the 3D scene using the global transformation obtained from the previous ICP alignment between the scene scans. Our dataset is named GIMO, and we describe our method for gaze-informed motion prediction in the following.

4 Gaze-Informed Human Motion Prediction

Gaze conveys relevant information about the subject’s intent, which can be used to enhance long-horizon human motion prediction. On the other hand, past motions [4,2], ego-centric views [55,10], or 3D context [10,51] could provide helpful constraints on human motion, yet, the prediction is still challenging and suffers from uncertainties in the future. Here, we aim at gaze-informed long-term human motion prediction. Specifically, given the past motion, 3D scene, and 3D eye gaze as inputs, we study how they can be integrated to resolve the ambiguities in future motion and generate intention-aware motion predictions.

To fully utilize the geometry information provided by the 3D scene and intention clues from past motions and gaze, we propose a novel framework with a

bidirectional fusion scheme that facilitates the communication between different modalities. As shown in Fig. 3, we use PointNet++ [39] as the encoding backbone to extract per-point features of the 3D scene, followed by several cross-modal transformers to transcend information from multi-modality embeddings.

4.1 Problem Definition

We represent a motion sample as a parametric sequence $X_{i:j} = \{x_i, x_{i+1}, \dots, x_j\}$ where $x_k = (t_k, r_k, h_k, \beta_k, p_k)$ is a pose frame at time k . Here $t \in \mathbb{R}^3$ is the global translation, $r \in SO(3)$ denotes the global orientation, $h_k \in \mathbb{R}^{32}$ refers to the body pose embedding, $\beta \in \mathbb{R}^{10}$ is the shape parameter, and $p \in \mathbb{R}^{24}$ is the hand pose, where SMPL-X body mesh $M = \mathcal{M}(t_k, r_k, h_k, \beta_k, p_k)$ can be obtained using VPoser [37]. The 3D scene is represented as a point cloud $S \in \mathbb{R}^{n \times 3}$, and the 3D gaze point $g \in \mathbb{R}^3$ is defined as the intersection points between the gaze direction and the scene. Thus, given the inputs of a motion sequence $X_{1:t}$ along with the corresponding 3D gaze $G_{1:t} = \{g_1, g_2, \dots, g_t\}$ and the 3D scene S , we aim to predict the future motion $X_{t:t+T} = \Phi(X_{1:t}, G_{1:t}, S|\theta)$ where θ represents the network parameters.

4.2 Multi-modal Feature Extraction

Instead of extracting the multi-modal embeddings independently [25], we propose a novel scheme to integrate the motion, gaze, and scene features. The gist is to let the motion and gaze features communicate to each other, so their uncertainties regarding the future can be mutually decreased, resulting in more effective utilization of the gaze information.

Scene feature extraction. To learn the constraints from the 3D scene and guide the network to pay attention to local geometric structures, we apply PointNet++ to extract both global and local scene features. Specially, we obtain the per-point feature map and a global descriptor of the scene as follows:

$$F_P, F_o = \Phi_{scene}(S|\theta_s) \quad (1)$$

where $S \in \mathbb{R}^{n \times 3}$ is the input point cloud, $F_P \in \mathbb{R}^{n \times d_p}$ is the per-point d_p dimensional feature map, and $F_o \in \mathbb{R}^{d_o}$ is the global descriptor of the scene. Given the per-point feature F_P , the feature of an arbitrary point e can be computed through the inversed distance weighted interpolation [39]:

$$F_{P|e} = \frac{\sum_{i=1}^{n_e} w_i F_{P|p_i}}{\sum_{i=1}^{n_e} w_i}, w_i = \frac{1}{\|p_i - e\|_2} \quad (2)$$

where $\{p_1, p_2, \dots, p_{n_e}\}$ are the nearest neighbors of e in the scene point cloud.

Gaze feature extraction. We query the gaze point feature f_g from the per-point scene feature map F_P according to Eq. 2, i.e., $f_g = F_{P|g}$. Thus, the interpolated gaze feature contains relevant scene information that provides cues to infer the subject’s intention.

Motion feature extraction. A linear layer is used to extract the motion embedding f_m from the input motion parameter x . To endow the embedding awareness of the 3D scene, we further query the scene features of the SMPL-X vertices using Eq. 2. These SMPL-X per-vertex features are then fed to PointNet [38] to get the ambient scene context feature f_{m_v} of the current motion pose:

$$f_{m_v} = \text{PointNet}(\{F_{P|v}, v \in \mathcal{M}(x)\}) \quad (3)$$

where $\mathcal{M}(x)$ is the SMPL-X vertex set with motion parameter x .

4.3 Attention-Aware Multi-modal Feature Fusion

Given the multi-modal nature of the gaze, scene, and motion, an efficient feature fusion module is necessary to leverage the information from different modalities. Instead of directly concatenating the features [25], we propose a more effective scheme by deploying a cross-modal transformer [14] to fuse the gaze, motion, and scene features (Fig. 3). We explain our design in the following.

Cross-modal transformer. The cross-modal transformer [14] is used to capture the correlations between input embedding sequences and to establish communications between the multi-modal information. It is largely based on attention mechanism [48]. An attention function [14] maps a query and key-value pairs to an output as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V, Q = qW_q, K = kW_k, V = vW_v \quad (4)$$

where $q \in R^{l_q \times d_q}$, $k \in R^{l_{kv} \times d_k}$, $v \in R^{l_{kv} \times d_v}$ are input query, key and value vectors, and $W_q \in R^{d_q \times d_K}$, $W_k \in R^{d_k \times d_K}$, $W_v \in R^{d_v \times d_V}$ embed the inputs. Here d denotes the dimension of the input vector and l is the sequence length.

As shown in Fig. 3 (b), the cross-modal transformer is built on a stack of attention layers, which maps a t_i -length input into a t_q -length output by querying a t_q -length feature:

$$\phi_{out} = \text{cross.trans}(\phi_{query}, \phi_{input}) \quad (5)$$

It is proved to be efficient in processing multi-modal signals, e.g., text, & audio.

Motion feature fusion. The motion feature should be aware of the 3D scene context and the subject’s intention inferred from the gaze information, so that it can guide the prediction network to generate more reasonable motion trajectories (e.g., free from penetration and collision) and accurate estimations of the ending position or pose of the subject. For this purpose, we first use the scene context feature f_{m_v} acquired from the ambient 3D environment (Eq. 3) as the query to update the motion feature f_m through a motion-scene transformer:

$$f_{m_s} = \text{cross.trans}(f_{m_v}, f_m) \quad (6)$$

Thus, the output motion embedding f_{m_s} is expected to be aware of the 3D scene. We then feed f_{m_s} to the next motion-gaze transformer where the gaze feature f_g is the query input:

$$f_{m_g} = \text{cross_trans}(f_g, f_{m_s}) \quad (7)$$

The final motion embedding f_{m_g} is expected to integrate both the 3D scene information and the intention clues from the gaze features.

Gaze feature fusion. While gaze can help generate intention-aware motion features, the motion could also provide informative guidance to mitigate the randomness of gaze since not every gaze point reveals meaningful user intent. Therefore, we treat the gaze embedding in a bidirectional manner, i.e., the motion embedding f_m is also used as the query to update the gaze features such that the network can learn which gaze features contribute more to the future motion:

$$f_{g_m} = \text{cross_trans}(f_m, f_g) \quad (8)$$

The bidirectionally fused multi-modal features are then composed into holistic temporal representations of the input to perform human motion prediction. As illustrated in Fig. 3 (c), the updated gaze feature f_{g_m} , motion feature f_{m_g} and the global scene feature F_O are used to predict the future motion by:

$$X_{t:t+T} = \text{cross_trans}(h_{\text{position}}, \text{cat}(f_{g_m}, f_{m_g}, F_O)_{1:t}) \quad (9)$$

where cat denotes the concatenation operation, and h_{position} is the latent vector that contains temporal positional encodings for the output [14]. We verify the effectiveness of our design in utilizing gaze information through experiments.

5 Experiments

In this part, we explain our experimental setup and results. Our goal is to examine the following questions:

1. Does gaze help disambiguate human motion prediction?
 2. How do state-of-the-art methods perform on our dataset?
 3. What is the contribution of each part of our design to the final performance?
- Overall, is the proposed architecture effective?

5.1 Experimental Setup

In our experiments, we predict the future motion in 5 seconds from 3 seconds input, where the first 3 seconds of a trajectory is just about to start an activity (i.e., beginning to move for fetching a book) in our dataset, and in the next 5 seconds the trajectory proceeds to finish the activity. We set the motion frame rate to 2 fps, i.e., 6 pose input and 10 pose output. Note that once the waypoints are predicted, a full motion sequence with high fps can be easily generated [51].



Fig. 4. Qualitative results. Top row: results on a known scene from the training set. Bottom row: results in a new environment. We compare our method with MultimodalNet [25] and ours without gaze. Please zoom in for details.

Since we aim to explore the effect of gaze in disambiguating motion prediction, high-frequency motion is not necessary in our experiments.

Baselines. We implement several state-of-the-art motion prediction and generation baselines including spatio-temporal transformer [2] and an RNN network [22] for full motion prediction from the past motion input, and MultimodalNet [25] based on transformer for motion synthesis from multi-modal data (i.e., gaze, motion, and the 3D scene feature in our experiments). We build our pipeline by incorporating 6 cross-modal transformer layers [14] to extract 256-dimensional gaze and motion features. L1 loss between the predicted motion and the ground truth is used to train the network. More details about the network architecture and training are available in the supplementary material.

5.2 Evaluation

To evaluate, we divide the 217 trajectories of our dataset into 180 trajectories for training and 37 for testing. The 37 motions consist of 27 trajectories (different from the training ones) performed in known scenes from the training set and 10 in 2 new environments scanned only for evaluation purposes.

Evaluation metrics. We employ the destination error and the path error as our evaluation metrics. The destination error refers to the global translation, rotation error and the mean per-joint position error (MPJPE) [13] of the last pose in the predicted motion. The destination pose contains essential information about the subject’s goal, which is our experiments’ primary focus. The path error

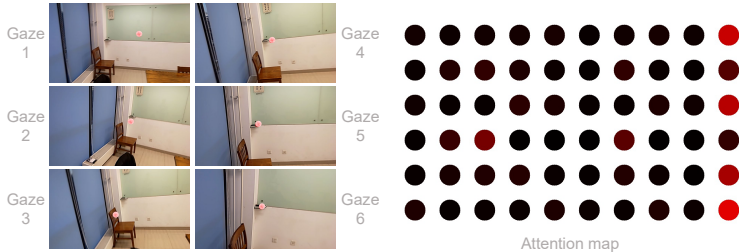


Fig. 5. The attention map of the 6 input gaze for the 10 output motion. The gaze influences the ending output most (brighter means larger weight), indicating that the gaze features reveal the subject’s final goals.

is computed as the mean error of the predicted poses in 5s [4]. We treat the global translation and rotation error as the $l1$ distance between the predicted SMPL-X translation and orientation parameter with the ground truth [51].

Quantitative evaluation. As shown in Tab. 3 and Tab. 4, while the state-of-the-art methods based on spatio-temporal transformer [2] suffer from ambiguities since the prediction is simply from the past motion, a simple RNN method with motion and gaze input [21] can significantly reduce the ambiguity, indicating the effectiveness of gaze in guiding the prediction of motion. Our method achieves promising results in predicting reasonable future motion with small destination and translation errors. Compared to MultimodalNet [25] built on the vanilla transformers [47], our method outperforms in recognizing the subject’s intent from the gaze and thus predicts more accurate destination poses.

Qualitative evaluation. Fig. 4 shows that in a ”going to sit” activity performed in one scene from the training set (top row), our method manages to generate accurate destination poses, i.e., sitting on the sofa. In the new environment, the subject first grabs a blackboard eraser and then starts wiping. While all the methods generate walking actions, ours without gaze input fails to predict the correct motion. When given gaze, results from MultimodalNet [25] and our method both reach out the hand and try to grab something. Our prediction successfully arrives at the destination point where the eraser lies; however, the **Table 3.** Destination accuracy. We report the global translation and orientation error and mean per-joint position error (*MPJPE*).

Method	Known scenes			New scenes		
	Trans	Ori	MPJPE	Trans	Ori	MPJPE
ST-Transformer [2]	0.587	0.864	279.9	0.516	0.682	236.8
RNN [22]	0.538	0.822	272.5	0.547	0.894	230.4
MultimodalNet [25]	0.442	0.699	260.0	0.389	0.658	236.0
RNN+gaze [21]	0.389	0.882	264.2	0.345	0.611	230.0
MultimodalNet+gaze [25]	0.316	0.743	266.6	0.300	0.583	204.9
Ours (w/o gaze)	0.393	0.656	262.1	0.389	0.709	228.7
Ours (pointnet)	0.310	0.659	240.6	0.394	0.563	234.5
Ours (vanilla)	0.353	0.739	249.0	0.365	0.602	220.4
Ours	0.245	0.579	237.8	0.280	0.556	209.0

Table 4. Path errors of the predicted motions.

Method	Known scenes			New scenes		
	Trans	Ori	MPJPE	Trans	Ori	MPJPE
ST-Transformer [2]	0.329	0.503	201.4	0.339	0.537	201.7
RNN [22]	0.308	0.476	195.2	0.324	0.495	180.3
MultimodalNet [25]	0.273	0.383	190.0	0.294	0.425	177.0
RNN+gaze [21]	0.235	0.457	190.1	0.278	0.288	182.6
MultimodalNet+gaze [25]	0.246	0.424	193.1	0.250	0.374	183.7
Ours (w/o gaze)	0.305	0.412	180.1	0.315	0.403	182.5
Ours (pointnet)	0.218	0.360	180.5	0.267	0.403	184.5
Ours (vanilla)	0.238	0.399	182.7	0.286	0.348	180.3
Ours	0.213	0.340	177.1	0.261	0.322	160.3

results of MultimodalNet [25] reach out to the wrong place. More visualizations and failure cases are included in the supplementary material.

5.3 Ablation Study

In this part, we aim to answer question 3 by finding the factors that contribute to the superior performance of our method.

Variant 1: gaze. We evaluate the baseline’s performance with and without gaze input to explore how gaze could influence the motion prediction results. As clearly demonstrated in Tab. 3 and Tab. 4, the RNN network [21] and the MultimodalNet [25] both gain significant accuracy improvement given gaze inputs. Fig. 4 shows that without gaze, our method is confused about the future destination. To find more intuitions about the role of gaze in motion prediction, we visualize the attention weights of gaze feature query over the motion feature as depicted in Fig. 5. Interestingly, we find the gaze feature does influence the ending poses in the predicted motion, implying that the gaze can serve as a strong indicator of the destination of a motion, which reveals the user’s intent.

Variant 2: pointnet++ for scene feature query. We propose to use Pointnet++ [39] to extract the per-point feature of the scene such that the gaze feature and scene-aware motion feature can be obtained (section 4.2). We replace it with Pointnet to extract the global scene feature and use a linear layer to get the gaze feature. Results in Tab. 3 and Tab. 4 demonstrate that the variant can act well on scenes from the training set, but lose its competitiveness when generalized to new environments with different 3D structures.

Variant 3: cross-modal transformer. The cross-modal transformer architecture proves to be effective in bridging multi-modal information [14]. We replace it with the vanilla transformer [48] as used in [25]. Results in Tab. 3 and Tab. 4 (*Ours (vanilla)*) demonstrate the loss of accuracy compared to the full design. Note that the path error of the variant on the new scenes is even larger than the results without gaze input, indicating that the vanilla transformer might not be efficient enough to capture the correlations between multi-modal inputs. Thus a more sophisticated design such as a cross-modal transformer is needed.

6 Conclusion, Discussion and Future Work

We present the GIMO dataset, a real-world dataset with ego-centric images, 3D gazes, 3D scene context, and ground-truth human motions. With the collected dataset, we define a new task, i.e., gaze-informed human motion prediction, which leverages eye gaze to infer the subject’s potential intention to minimize the ambiguities in motion prediction. We further contribute a novel framework, which achieves promising results in predicting long-term future motions. While our method only relies on 3D inputs, we aim as future work to incorporate visual information from ego-centric images to further boost the accuracy.

Instead of the proposed task, our dataset can benefit various applications, e.g., intention-aware motion synthesis and gaze-guided ego-centric pose estimation. We believe our work not only opens new directions for motion prediction but will have foreseeable impacts on ego-centric vision topics.

Acknowledgments The authors are supported by a grant from the Stanford HAI Institute, a Vannevar Bush Faculty Fellowship, a gift from the Amazon Research Awards program, the NSFC grant No.62125107, and No.62171255. Also, Toyota Research Institute provided funds to support this work.

References

- Admoni, H., Scassellati, B.: Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* **6**(1), 25–63 (2017) [2](#)
- 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) [4](#), [8](#), [12](#), [13](#), [14](#), [19](#)
- Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3d human motion modelling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7144–7153 (2019) [4](#)
- Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: European Conference on Computer Vision. pp. 387–404. Springer (2020) [3](#), [4](#), [6](#), [8](#), [13](#)
- CMU Graphics Lab: (2000), <http://mocap.cs.cmu.edu/> [3](#)
- Duarte, N.F., Raković, M., Tasevski, J., Coco, M.I., Billard, A., Santos-Victor, J.: Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters* **3**(4), 4132–4139 (2018) [2](#)
- Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE international conference on computer vision. pp. 4346–4354 (2015) [4](#)
- Gottlieb, J., Oudeyer, P.Y., Lopes, M., Baranes, A.: Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences* **17**(11), 585–593 (2013) [1](#)
- Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human positioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4318–4329 (2021) [4](#), [5](#), [6](#)

10. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11374–11384 (2021) [4](#), [6](#), [8](#)
11. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2282–2292 (2019) [3](#), [6](#)
12. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 68–84 (2018) [4](#)
13. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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](#), [12](#)
14. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Kop-pula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795* (2021) [8](#), [10](#), [11](#), [12](#), [14](#), [19](#)
15. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from ego-centric video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3501–3509. IEEE (2017) [4](#)
16. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3334–3342 (2015) [6](#)
17. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8320–8329 (2018) [6](#)
18. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5253–5263 (2020) [4](#)
19. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11127–11137 (2021) [4](#)
20. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2252–2261 (2019) [4](#)
21. 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](#), [13](#), [14](#), [19](#)
22. 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) [12](#), [13](#), [14](#), [19](#)
23. Li, J., Villegas, R., Ceylan, D., Yang, J., Kuang, Z., Li, H., Zhao, Y.: Task-generic hierarchical human motion prior using vaes. In: 2021 International Conference on 3D Vision (3DV). pp. 771–781. IEEE (2021) [4](#)
24. Li, J., Yin, Y., Chu, H., Zhou, Y., Wang, T., Fidler, S., Li, H.: Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171* (2020) [4](#)
25. 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) [4](#), [9](#), [10](#), [12](#), [13](#), [14](#), [19](#)

26. Li, Y., Liu, M., Rehg, J.: In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) **6**
27. Li, Z., Zhou, Y., Xiao, S., He, C., Huang, Z., Li, H.: Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363* (2017) **4**
28. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* **34**(6), 1–16 (2015) **5**
29. Luo, Z., Golestaneh, S.A., Kitani, K.M.: 3d human motion estimation via motion compression and refinement. In: *Proceedings of the Asian Conference on Computer Vision* (2020) **4**
30. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems* **34** (2021) **4, 6**
31. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5442–5451 (2019) **3**
32. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: *European Conference on Computer Vision*. pp. 474–489. Springer (2020) **4**
33. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 601–617 (2018) **6**
34. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2891–2900 (2017) **4**
35. Martínez-González, A., Villamizar, M., Odobez, J.M.: Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2276–2284 (2021) **4**
36. Ng, E., Xiang, D., Joo, H., Grauman, K.: You2me: Inferring body pose in egocentric video via first and second person interactions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9890–9900 (2020) **4**
37. 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) **6, 7, 9, 19**
38. 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) **10, 20**
39. 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, 14, 19**
40. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11488–11499 (2021) **4**
41. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)* **35**(6), 1–11 (2016) **4**

42. Tatler, B.W., Hayhoe, M.M., Land, M.F., Ballard, D.H.: Eye guidance in natural vision: Reinterpreting salience. *Journal of vision* **11**(5), 5–5 (2011) [1](#)
43. Tian, Y., Zhang, H., Liu, Y., Wang, I.: Recovering 3d human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923* (2022) [4](#)
44. Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., De la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. *arXiv preprint arXiv:2011.01519* (2020) [4](#)
45. Tome, D., Peluse, P., Agapito, L., Badino, H.: xr-egopose: Egocentric 3d human pose from an hmd camera. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7728–7738 (2019) [4](#)
46. Ungureanu, D., Bogu, F., Galliani, S., Sama, P., Duan, X., Meekhof, C., Stühmer, J., Cashman, T.J., Tekin, B., Schönberger, J.L., et al.: Hololens 2 research mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239* (2020) [6](#)
47. Valle-Pérez, G., Henter, G.E., Beskow, J., Holzapfel, A., Oudeyer, P.Y., Alexander-son, S.: Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)* **40**(6), 1–14 (2021) [4](#), [13](#)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017) [4](#), [10](#), [14](#), [19](#)
49. Von Marcard, T., Pons-Moll, G., Rosenhahn, B.: Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence* **38**(8), 1533–1547 (2016) [6](#)
50. Wang, J., Liu, L., Xu, W., Sarkar, K., Theobalt, C.: Estimating egocentric 3d human pose in global space. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11500–11509 (2021) [4](#)
51. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9401–9411 (2021) [3](#), [8](#), [11](#), [13](#)
52. Wei, P., Liu, Y., Shu, T., Zheng, N., Zhu, S.C.: Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 6801–6809 (2018) [6](#)
53. Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.P., Theobalt, C.: Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics* **25**(5), 2093–2101 (2019) [4](#)
54. Yuan, Y., Kitani, K.: 3d ego-pose estimation via imitation learning. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 735–750 (2018) [4](#)
55. Yuan, Y., Kitani, K.: Ego-pose estimation and forecasting as real-time pd control. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10082–10092 (2019) [4](#), [8](#)
56. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021) [4](#)
57. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Pollefeys, M., Bogu, F., Tang, S.: Egobody: Human body shape, motion and social interactions from head-mounted devices. *arXiv preprint arXiv:2112.07642* (2021) [4](#), [6](#)
58. Zhang, S., Zhang, Y., Bogu, F., Pollefeys, M., Tang, S.: Learning motion priors for 4d human body capture in 3d scenes. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11343–11353 (2021) [4](#)

59. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: Place: Proximity learning of articulation and contact in 3d environments. In: 2020 International Conference on 3D Vision (3DV). pp. 642–651. IEEE (2020) [3](#)
60. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6194–6204 (2020) [3](#), [6](#)

Appendix

A Experimental Setup

A.1 Implementation Details

As demonstrated in the Fig. 3 of the main paper, our method is built on Pointnet++ [\[39\]](#) and a cross-modal transformer [\[14\]](#). 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 [\[14\]](#) 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 [\[37\]](#). 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 [\[2\]](#), a RNN based network [\[22\]](#), and MultimodalNet [\[25\]](#) 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 [\[22\]](#), while MultimodalNet is based on the 32D motion parameter the same as ours. An 8 layer transformer [\[48\]](#) with 512 embedding size and 8 heads attention is used in spatio-temporal transformer [\[2\]](#). A three layer RNN with 1024 hidden size is deployed to predict future motion with simple motion input or motion and gaze input [\[21\]](#). In MultimodalNet [\[25\]](#), 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



Fig. 6. An overview of the scanned scenes in our dataset.

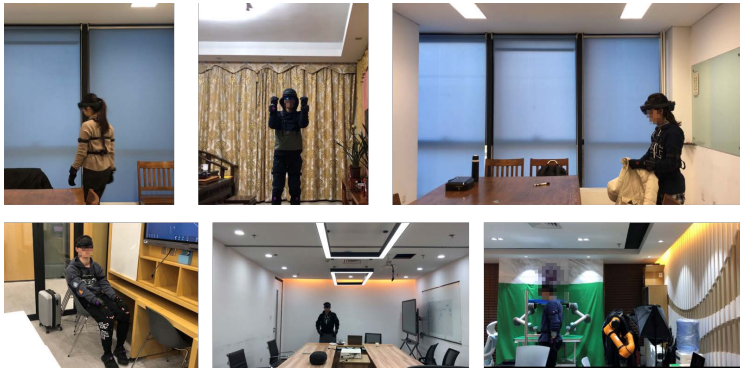


Fig. 7. Subjects in the scenes.

and a transformer encoder. The global scene feature from PointNet [38], the 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.

A.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:

$$\mathcal{L}_{trans} = \frac{1}{T} \sum_{k=1}^T \|\hat{t}_k - t_k\|_1 \quad (10)$$

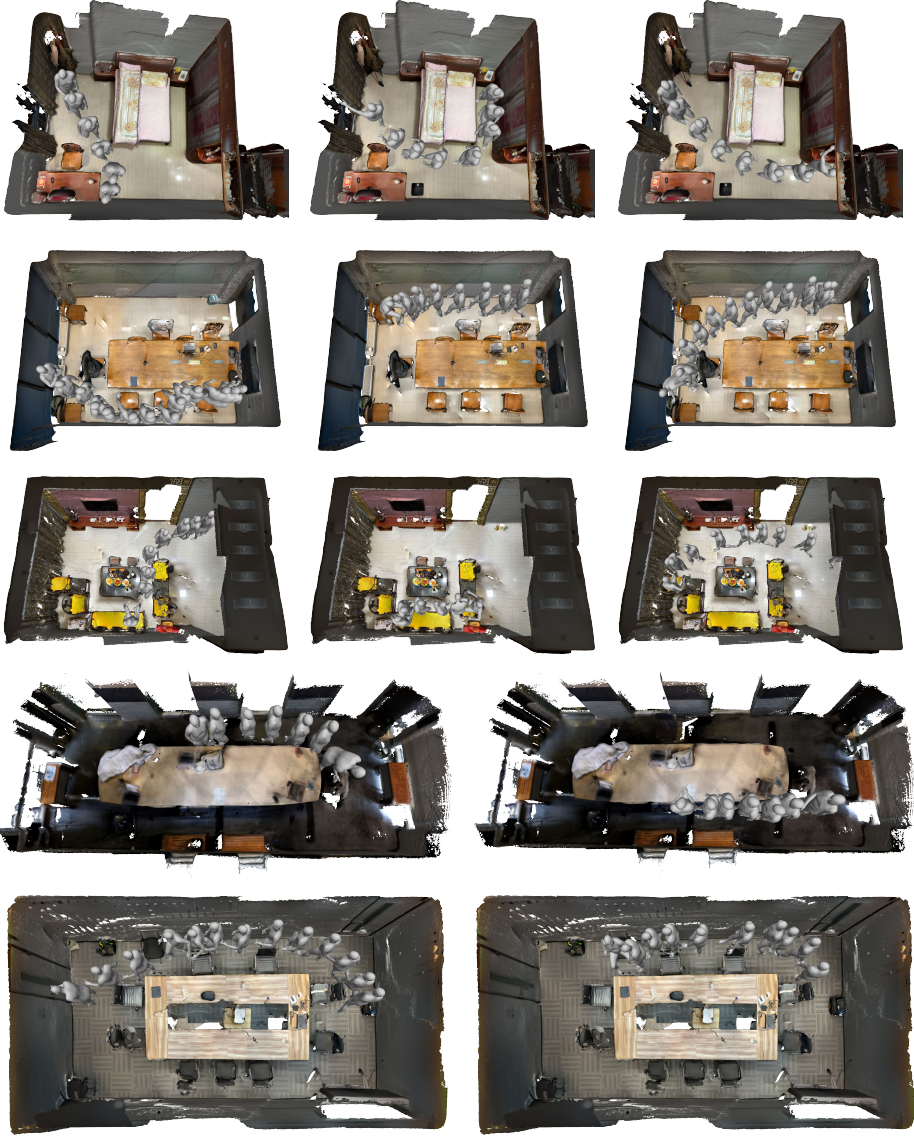


Fig. 8. Motion trajectories from our dataset. Better visualized in the supplementary video.

where T is the length of output pose, and \hat{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:

$$\mathcal{L}_{ori} = \frac{1}{T} \sum_{k=1}^T \|\hat{r}_k - r_k\|_1 \quad (11)$$

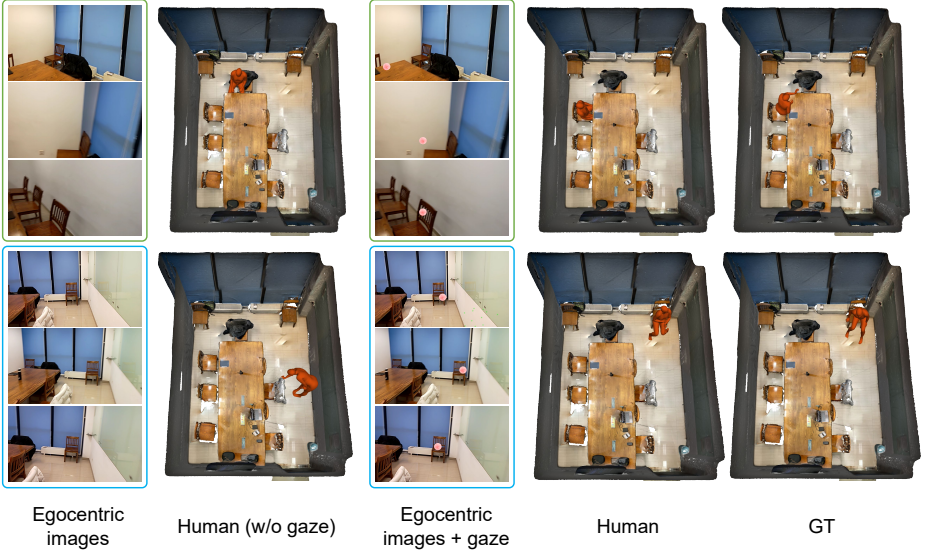


Fig. 9. Human evaluation. Two human subjects are required to watch a egocentric video (without gaze or with gaze) and infer the final pose of the trajectory. The subjects choose a pose from a pose database which comes from the training set, and put the pose into the 3D scene as the final position of 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:

$$\mathcal{L}_p = \frac{1}{T} \sum_{k=1}^T \|\widehat{h}_k - h_k\|_1 \quad (12)$$

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

$$\mathcal{L} = \lambda_t \mathcal{L}_{trans} + \lambda_o \mathcal{L}_{ori} + \lambda_p \mathcal{L}_p \quad (13)$$

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

B GIMO Dataset

Our dataset consists of 217 motion trajectories collected in 19 scenes by 11 subjects. Fig. 6 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. 7 shows the recruited subjects collecting data in the scenes. More motion trajectories are demonstrated in Fig. 8. For better visualization, please refer to the supplementary video.

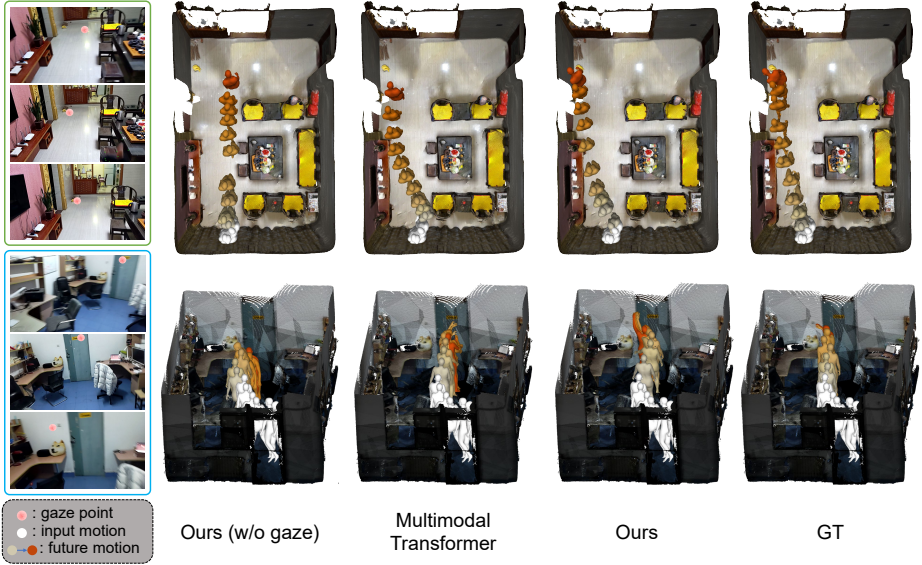


Fig. 10. More Qualitative results.

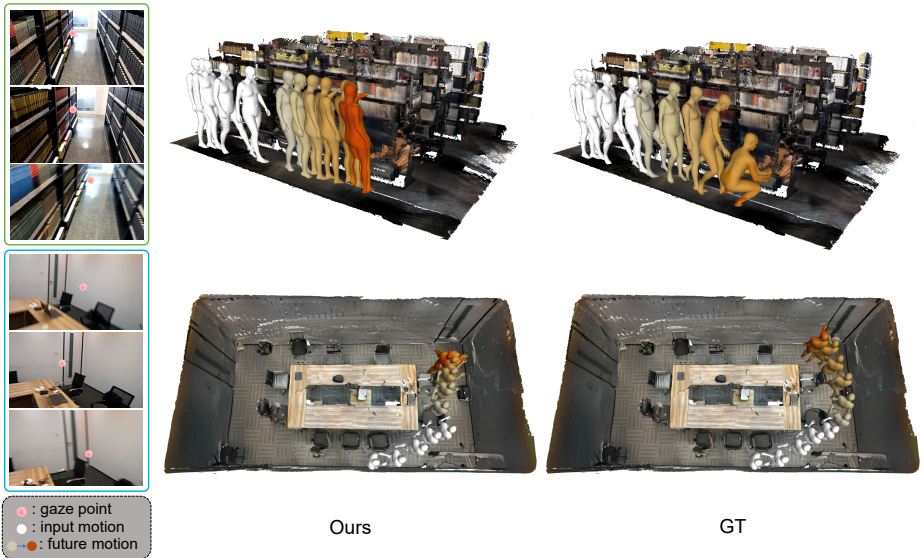


Fig. 11. Failure cases of our method. When the noisy gazes account for a large portion of the input, our method is confused to interpret the subject’s intention.

C More Results

C.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. 9 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.

C.2 More Results of Baselines and Failure cases

Fig. 10 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 11. 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.