Ego4o: Egocentric Human Motion Capture and Understanding from Multi-Modal Input

Jian Wang[1,4] Rishabh Dabral[1,4] Diogo Luvizon[1,4] Zhe Cao[2]
Lingjie Liu[3] Thabo Beeler[2] Christian Theobalt[1,4]
[1]MPI Informatics & Saarland Informatics Campus [2]Google [3]University of Pennsylvania
[4]Saarbrücken Research Center for Visual Computing, Interaction and Artificial Intelligence

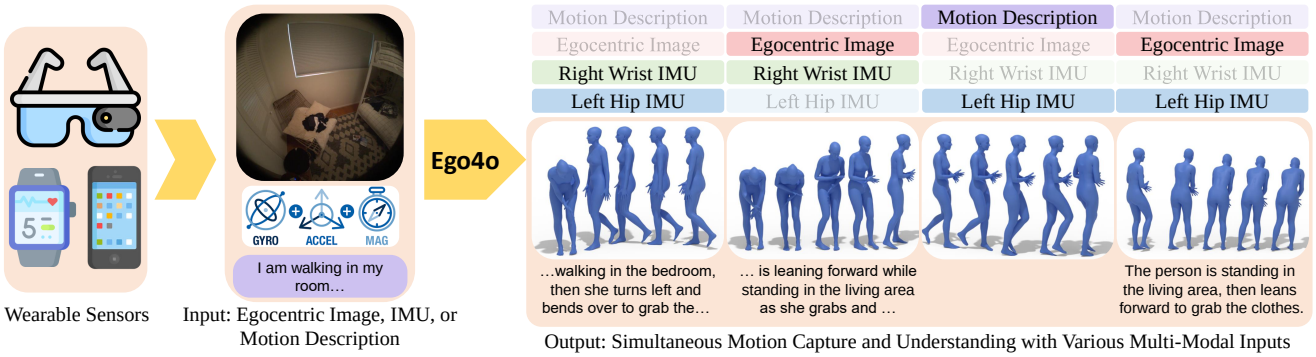

Figure 1. Our method can use an egocentric image and 1-3 IMU sensors from wearable devices to accurately predict human motion and generate motion descriptions. Motion descriptions, when available, can also enhance motion capture accuracy. Ego4o supports flexible input combinations, functioning with or without images, or with varied IMU placements.

Abstract

This work focuses on tracking and understanding human motion using consumer wearable devices, such as VR/AR headsets, smart glasses, cellphones, and smartwatches. These devices provide diverse, multi-modal sensor inputs, including egocentric images, and 1-3 sparse IMU sensors in varied combinations. Motion descriptions can also accompany these signals. The diverse input modalities and their intermittent availability pose challenges for consistent motion capture and understanding. In this work, we present Ego4o (o for omni), a new framework for simultaneous human motion capture and understanding from multi-modal egocentric inputs. This method maintains performance with partial inputs while achieving better results when multiple modalities are combined. First, the IMU sensor inputs, the optional egocentric image, and text description of human motion are encoded into the latent space of a motion VQ-VAE. Next, the latent vectors are sent to the VQ-VAE decoder and optimized to track human motion. When motion descriptions are unavailable, the latent vectors can be input into a multi-modal LLM to generate human motion descriptions, which can further enhance motion capture accuracy. Quantitative and qualitative evaluations demonstrate the effectiveness of our method in predicting accurate human motion and high-quality motion descriptions.

1. Introduction

Recently, more and more research has focused on human motion capture and understanding using widely available wearable devices, such as VR/AR headsets, smart glasses, cellphones, and smartwatches [34, 42, 43, 45, 59, 65, 71]. This interest is driven by broad application scenarios, including sports, healthcare, VR/AR, and personal assistants. These devices provide diverse, multi-modal sensor inputs related to human motion, such as inertial measurement unit (IMU) data, egocentric camera images, and even voice-enabled conversation data, where the text description of human motion can be extracted. However, existing works mostly focus on motion capture from one single input modality. Some methods [34, 42, 65] predict the human motion from egocentric cameras, while others capture the human motion from VR tracker [3, 27] or IMU signals [45, 75]. Each individual modality provides only a limited view of human motion, constraining the accuracy of both motion capture and understanding. For exam-

ple, the full human body motion is barely seen or significantly occluded from the egocentric camera. Text descriptions can offer information about human motion categories but lack the precision to detail specific movements. IMUs on VR/AR headsets, smartwatches, and cellphones are very sparse: they usually only track the movement of one or two limbs, as people rarely wear watches on both wrists. Moreover, IMU-based methods struggle with static pose estimation due to the absence of dynamic acceleration signals.

We observe that different input modalities serve complementary roles in motion analysis. Motion descriptions and egocentric images provide rich semantic context about both the activity being performed and the environmental setting. For example, when the egocentric view shows a desk in close proximity, it strongly indicates that the person is sitting. IMU signals from wearable devices capture precise kinematic data for specific body segments. For instance, a smartwatch's IMU sensors can capture detailed hand movements, enabling the system to differentiate between blocking and smashing motions in table tennis, which may be indiscernible from the egocentric camera perspective alone.

To fully leverage the information from wearable devices, we present **Ego4o** (o for omni), a novel framework that achieves 3D pose estimation (motion capture) and motion description generation (motion understanding) by fusing multi-modal inputs. These inputs may include 1–3 sparse IMU sensors, egocentric images, and motion descriptions from everyday wearable devices. As illustrated in Fig. 1, the Ego4o maintains robust performance with different combinations of input modalities.

The Ego4o method first employs a multi-modal transformer to encode diverse inputs into motion codes in a part-based discrete motion representation space, which is constructed with a VQ-VAE [58]. Since input availability can change during use—as users may disable egocentric cameras and microphones, or vary the number of IMU sensors by removing phones or wearing smartwatches—we implement a random masking strategy during training. This approach enables the model to adapt seamlessly to different combinations of input modalities. Finally, the motion codes are then decoded into human motion predictions, which are further refined through test-time optimization in the VQ-VAE's latent space.

Building on the previous step, we demonstrate that the obtained motion codes can be utilized as input for Large Language Models (LLMs) to generate detailed descriptions of human movement. We developed a multi-modal joint training approach that fine-tunes LLMs to simultaneously process both motion codes and egocentric images. Our work shows that LLMs' inherent strengths in in-context reasoning and image understanding can be effectively leveraged to generate high-quality motion descriptions.

Obtaining human-produced motion descriptions is typically challenging in real-world scenarios, which can limit the motion capture performance. Our insight is that the generated high-quality motion descriptions can serve as valuable conditioning signals that enhance the accuracy of our motion capture system when the human-provided motion description is absent. This introduces a feedback loop between motion capture and understanding, advancing the state-of-the-art in both tasks.

We validate Ego4o's effectiveness through quantitative and qualitative evaluations. Experimental results demonstrate that our proposed method achieves better motion capture accuracy, while simultaneously generating detailed descriptions of human movements. By integrating motion capture and motion description generation within a unified framework, Ego4o advances toward making motion analysis accessible and practical for everyday applications with consumer devices. In summary, our contributions are:

- We introduce Ego4o, a novel framework that flexibly integrates multi-modal egocentric inputs to enable simultaneous motion capture and description generation;
- We design a multi-modal encoder with a random masking training strategy to accommodate varying combinations of input modalities;
- We employ multi-modal joint fine-tuning of large language models to bridge modality gaps and support accurate motion description generation;
- We show that AI-generated motion descriptions can improve the accuracy of egocentric human motion capture.

2. Related Work

Egocentric Human Motion Capture. Recently, there has been growing interest in estimating egocentric 3D poses from body-worn devices. Some methods [26, 34, 42, 46, 77, 78] leverage head-mounted, front-facing cameras to infer motion from head movements, while others [10, 27, 28, 32, 67, 74] employ three-point trackers for motion capture. Additional approaches [1, 2, 39, 40, 49, 57, 62–65, 72] use down-facing fisheye cameras to capture full-body movement, while others [17, 18, 29, 75] rely on IMUs for body tracking. Similar to our method, recent works such as IMUPoser [45], MobilePoser [71], and Diffusion-Poser [59] use 1-3 IMU sensors to capture human motion. While these methods have significantly advanced the field, they primarily focus on single-modality solutions, leaving the potential of multi-modal integration largely unexplored.

Recent studies in egocentric motion capture have explored multi-modal approaches. EgoLocate [76] uses six IMUs and egocentric video for large-area motion capture. EMHI [11] introduced a dataset combining down-facing stereo cameras, 6DOF trackers, and IMUs, while HMD^2 [19] leverages a conditional diffusion model with egocentric video and head 6DOF pose. However, the input to this method is always fixed. The work most simi-

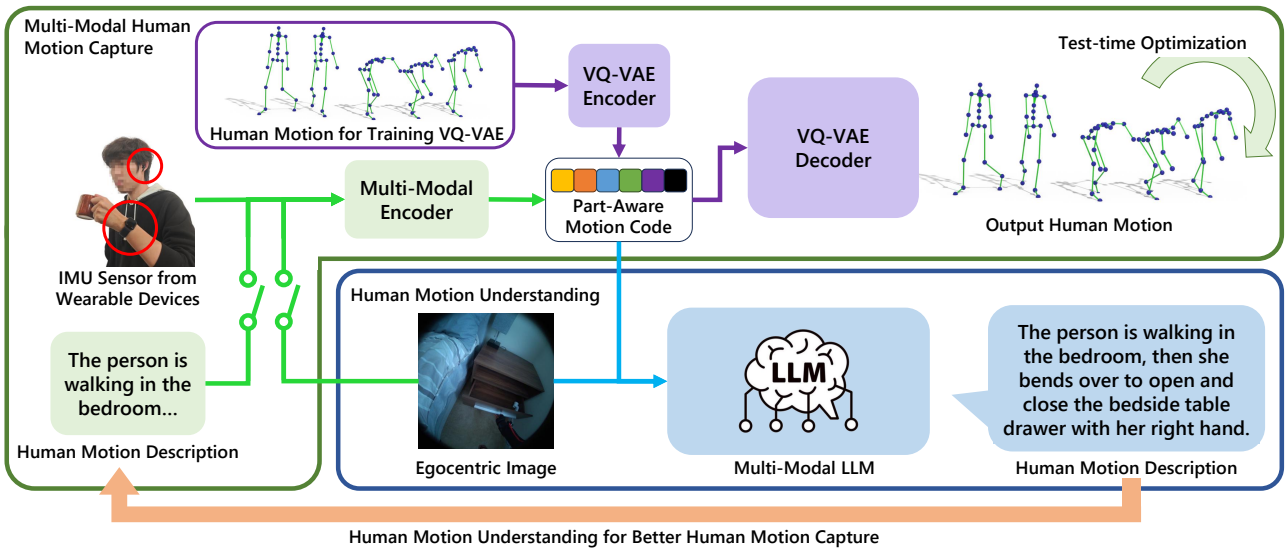

Figure 2. Overview of our Ego4o framework. We first train a VQ-VAE (purple blocks) to learn the part-aware motion codebook (Sec. 3.1). For motion capture (green blocks), the system processes IMU sensor data, egocentric images, and motion descriptions through a multi-modal encoder to generate motion codes in the codebook. These codes are then decoded to predict human motion (Sec. 3.2). For motion understanding (blue blocks), the system combines motion codes and egocentric images in a finetuned LLM to generate motion descriptions (Sec. 3.3), which can be fed back to enhance motion capture accuracy.

lar to ours is EgoLM [20], which uses head and hand 6D tracking data as input, whereas our method employs 1-3 IMUs. While EgoLM uses a LLM for motion capture, resulting in high computational costs and low accuracy, our approach is faster through a simple encoder-decoder architecture. Moreover, by finetuning on a larger-scale multi-modal LLM, our method enables multi-round conversation and generalization to out-of-distribution images.

Human Motion Generation. Human motion generation has been a long-standing challenge in computer vision and graphics. Some works [14, 50, 70] generate human motion from action labels. However, action labels provide only limited representational ability. Recently, numerous works [6, 15, 51, 53, 56, 79, 80, 82] have focused on generating human motion from text descriptions. Researchers have also leveraged powerful LLMs to model the joint motion-language distribution [4, 16, 25, 68], enabling both human motion generation from text input and text generation from motion. While we use a similar approach to enable human motion understanding with LLMs, our Ego4o framework differs by focusing on accurate human motion capture and supporting multiple egocentric modalities.

Egocentric Motion Understanding. Recently, many works have aimed at human motion understanding from the egocentric perspective. Previous works [7, 8, 35, 41, 54, 66] usually use egocentric head-mounted front-facing cameras for the human action recognition task. More recently, some researchers [5, 9, 13, 24, 55, 69, 73] have leveraged Large Language Models (LLMs) for egocentric human motion understanding and have used natural language as output. In contrast to these methods, our Ego4o approach leverages multi-modal egocentric information, including egocentric images, text descriptions, and IMU signals as input. Our framework can capture human motion and simultaneously generate descriptions about the human activity.

3. Method

Our method (Fig. 2) processes a combination of egocentric images I, textual motion descriptions X_a, and data from one, two, or three IMU sensors, including device acceleration A and rotation R. The IMU sensors may be placed at up to five locations: the head, the wrists, or the hips, reflecting typical placements for devices like VR headsets, smartwatches, and cellphones. From these multi-modal inputs, we achieve accurate motion capture and can generate textual description of the motion when they are absent. To achieve this, we first train a part-based motion VQ-VAE (Sec. 3.1) to learn the discrete motion representation for IMUs, then use the multi-modal encoder to project the inputs to the motion representation space(Sec. 3.2) The discrete motion codes, while designed for IMU-based motion capture, can also be reused for generating motion descriptions (Sec. 3.3). Finally, we show that motion descriptions generated by our multi-modal LLM can further enhance motion capture performance. (Sec. 3.4).

3.1. Learning Part-Aware Motion Representation

In this section, we describe how to learn discrete human motion representation with VQ-VAE [58] and further en-

able the projection of multi-modal inputs to the motion representation space. Most previous works [4, 16, 25, 79, 84] treat the human body as a holistic entity, encoding the full human body motion into a single VQ-VAE codebook. Though this holistic encoding is effective, it presents limitations for our use case.

Our method aims to support flexible IMU sensor configurations, ranging from a single head-mounted IMU in smart glasses to various combinations of sensors embedded in smartwatches and smartphones. To achieve this adaptability, we implement the part-aware VQ-VAE architecture introduced in TLControl [61], which establishes separate motion codebooks for individual body segments. These separate motion codebooks enable the direct projection of available IMU signals into their corresponding part-specific motion codebooks, while simultaneously facilitating the generation of latent features for body segments lacking sensor coverage. For example, when processing data from a wrist-mounted IMU, the system not only projects this information into the arm-specific motion codebook but also generates leg movements. This projection mechanism operates analogously to a text-infilling task [33], where the system infers motion patterns for unmonitored body segments based on the available sensor data. By employing this part-aware architecture, our system achieves better motion capture accuracy across diverse IMU configurations, offering a versatile solution compared to conventional holistic approaches.

Next, we discuss how to learn the motion representation with part-aware VQ-VAE. Specifically, all the joints are first divided into six joint groups, including head, left arm, right arm, root, left leg, and right leg. The input ground truth human motion is first encoded to the HumanML3D [15] representation $J \in \mathbb{R}^{T\times M}$, where T is the motion length and $M = 263$ corresponds to the motion representation dimensions. Next, J in each time step is split into six groups according to the correlated human body part: $J = [J_{\text{Head}}, J_{\text{LArm}}, J_{\text{RArm}}, J_{\text{LLeg}}, J_{\text{RLeg}}, J_{\text{Root}}]$. For each body part i, we train a separate encoder $\mathcal{E}_i$ to learn an independent codebook $C_i \in \mathbb{R}^{N_{\text{code}}\times d}$, where N_{code} is the size of codebook while d is the dimension of each codebook. The encoder first encodes the human motion into features $Q_i \in \mathbb{R}^{T'\times d}$, where $T' = T/4$. Next, the Q_i is quantized with the codebook C_i, obtaining the quantized feature $\hat{Q}_i$. The quantized features from all of the body parts are finally concatenated and sent to the VQ-VAE decoder $\mathcal{D}$ to get the reconstructed motion $\hat{J}_{\text{recon}}$. The training for the part-aware VQ-VAE is detailed in the suppl. mat. In the next section, we project multi-modal inputs into this representational space for motion capture and understanding.

3.2. Multi-Modal Human Motion Capture

In this section, we introduce our multi-modal human motion capture method. The process begins with a transformer-based multi-modal encoder that projects IMU signals, egocentric images, and motion descriptions into the motion representation space learned by the VQ-VAE. These motion features are then processed by the VQ-VAE decoder to reconstruct human motion. Additionally, we offer an optional test-time optimization procedure that can further enhance the accuracy of the motion capture results.

3.2.1. Multi-Modal Encoder

Our transformer-based multi-modal encoder processes three input types: motion description T_m, egocentric image I, and IMU signal sequences. The egocentric image I and motion description T_m are encoded into image features F_I and textual features F_T respectively using CLIP [52]. The IMU signal sequence comprises acceleration vectors $A \in \mathbb{R}^{T\times N_{imu}\times 3}$ and rotation matrices $R \in \mathbb{R}^{T\times N_{imu}\times 3\times 3}$, where T represents the sequence length and $N_{imu} = 5$ is the *maximum* number of IMU locations. In practice, our method is designed to work with arbitrary number of IMUs. The rotation matrices R are converted to 6D representations [83] $R_{6d} \in \mathbb{R}^{T\times N_{imu}\times 6}$. The IMU acceleration and rotation data are then concatenated and reshaped into an input IMU sequence F_{imu} of length $T' \times N_{imu}$, where $T' = T/4$. These IMU sequences F_{imu}, along with the image features F_I and the textual motion description features F_T, are processed through an embedding layer before being fed into a transformer encoder [60]. The encoder predicts the logits of the motion code IDs $L_{t,i}$ for the i^{th} IMU at each time step t of the input sequence. Finally, we employ Gumbel Softmax [23] to get the motion code index $\delta_i \in \{0, 1, 2, ..., N_{\text{code}}\}$ of the corresponding i^{th} IMU and select the quantized motion feature $\hat{Q}_{t,i}$ from the motion code book. The quantized motion features are sent to the VQ-VAE decoder $\mathcal{D}$ following the same way in Sec. 3.1 to get the human motion prediction $\hat{J}$. The multi-modal encoder can be trained with the following loss function, which includes the motion code classification loss and human motion reconstruction loss:

$$\mathcal{L} = \mathbb{E}_{\hat{L}}\left(-\log P(\hat{L}|A, R, I, T_m)\right) + \lambda \left\|\hat{J} - J\right\|_2 \quad (1)$$

where λ is the weight of reconstruction loss. To simulate real-world scenarios where certain input modalities may be unavailable, we implement a masking strategy during training. This involves randomly masking egocentric images and textual descriptions. We also simultaneously select random combinations of one to three IMU sensors as active inputs. The remaining IMU sensors are masked to ensure our model learns to operate effectively with varying sensor availability.

3.2.2. Test-Time Optimization

Limb movement in the motion prediction $\hat{J}$ may not fully align with the corresponding IMU's acceleration and orien-

tation. This can be refined through optional test-time optimization. The task is to find a motion feature Q in the VQ-VAE latent space such that the reconstructed human motion $J = \mathcal{D}(Q)$ minimizes the energy function:

$$Q^* sch = \underset{Q}{\operatorname{argmin}} \lambda_a E_a(J, A) + \lambda_r E_r(J, R) \quad (2)$$

where $E_a(\cdot)$, $E_r(\cdot)$ are the IMU acceleration term and IMU orientation term respectively. For simplicity, we assume IMUs are positioned near their corresponding body joints—for instance, a smartphone's IMU approximates hip joint motion. To compute the IMU acceleration term, we first calculate the acceleration of each joint position that corresponds to an IMU sensor placement. For the joint associated with the i^{th} IMU at time step t, the acceleration is calculated using: $\hat{\mathbf{a}}_t^i = (J_{t+2}^i - 2J_{t+1}^i + J_t^i)/(\Delta t^2)$, where $\Delta t = 1$ in our experiment. The overall IMU acceleration term is calculated as: $E_a(J, A) = \sum_i \sum_t \|\hat{\mathbf{a}}_t^i - \mathbf{a}_t^i\|_2$, where $t = 0, 1, 2, ...$ represents time steps in the motion sequence, i indexes the available IMU sensors.

To compute the IMU orientation term, we first calculate the orientation of each limb in predicted motion $\hat{J}$ that corresponds to an IMU sensor placement. For the limb associated with the ith IMU at time step t, the orientation vector is: $\hat{\mathbf{r}}_t^i = (J_{t,\text{child}}^i - J_{t,\text{parent}}^i)/\|J_{t,\text{child}}^i - J_{t,\text{parent}}^i\|_2$ where $J_{t,\text{child}}^i$ and $J_{t,\text{parent}}^i$ represent the child and parent joint positions of the limb segment associated with the ith IMU.

Next, we calculate the orientation of each available IMU sensor with $\mathbf{r}_t^i = M^i \cdot R_t^i \cdot [0, 1, 0]^T$ where R_t^i represents the rotation matrix of the ith IMU at time step t, $[0, 1, 0]^T$ denotes the initial orientation vector, and M^i represents the calibration rotation matrix between the IMU sensor and its corresponding limb segment, determined through prior calibration. The IMU orientation term is then computed as: $E_r(J, R) = \sum_i \sum_t \|\hat{\mathbf{r}}_t^i - \mathbf{r}_t^i\|_2$, where i indexes the available IMU sensors.

3.3. Egocentric Human Motion Understanding

In this section, we present our approach to human motion understanding through multi-modal LLM fine-tuning. While existing pre-trained multi-modal LLMs excel at modeling language and image distributions, they lack the capability to process data related to human motion. To natively enable such understanding of human motion, we extend LLaVA (Vicuna-7B) [38] by incorporating a new motion modality and fine-tuning it using the multi-modal egocentric dataset Nymeria [43].

3.3.1. Architecture

The architecture of our egocentric LLM is shown in Fig. 3. Given an input image I and IMU sensor data (A, R), we first employ the multi-modal encoder described in Sec. 3.2

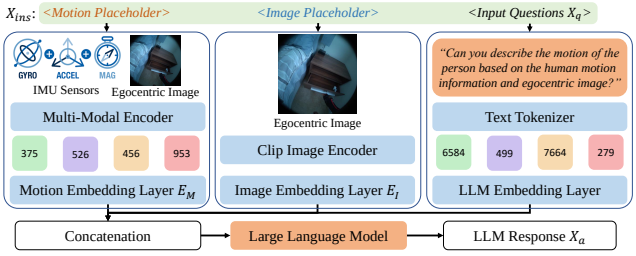

Figure 3. Egocentric Human Motion Understanding. Each modality is encoded separately and then concatenated in the order specified by the input instruction X_{ins} before being fed into the LLM.

to generate human motion codes, which serve as discrete motion tokens. These motion codes are then processed through a linear embedding layer $\mathbf{E}_M$ to produce human motion features H_M, aligned with the language model's word embedding dimensionality. For image processing, we utilize a pretrained CLIP [52] image encoder E_I to map image features F_I into the word embedding space, resulting in H_I. Finally, the image features H_I, motion features H_M, and text encodings H_T are concatenated and fed into the LLM to generate the response.

3.3.2. Training

For each IMU signal sequence (A, R) and the corresponding egocentric image I, we generate conversation pairs (X_q, X_a). The questions X_q (see an example in Fig. 3) are prompts requesting human motion descriptions, randomly sampled from a pre-defined list. The answers X_a are drawn from the fine-grained motion descriptions in the Nymeria [43] dataset. The input instruction set is constructed as $X_{\text{ins}} = \text{RandomSelect}\{[I, X_q], [A, R, X_q], [A, R, I, X_q]\}$.

The LLM can be fine-tuned on prediction tokens using an auto-regressive training objective. The probability of generating the answer X_a is computed as:

$$p(X_a|X_{\text{ins}}) = \prod_{i=1}^{L} p_\theta(x_i|X_v, X_{\text{ins},<i}, X_{a,<i}) \quad (3)$$

where L represents the token sequence length, θ denotes the trainable parameters, and $X_{\text{ins},<i}$ and $X_{a,<i}$ are the instruction and answer tokens preceding the current prediction token x_i. The trainable parameters are optimized using the negative log-likelihood loss.

The training process consists of two stages: in the first stage, we conduct motion pre-training to achieve motion feature alignment. In the second stage, we use multi-modal fine-tuning to enable egocentric motion understanding.

Motion Pre-Training. For the pre-training phase, we restrict input instructions to those containing only IMU signals $[A, R, X_q]$. To ensure proper alignment between motion features and the pre-trained LLM's word embeddings,

we exclusively train the motion embedding layer E_M while keeping all other architectural components frozen.

Multi-Modal Finetuning. In this phase, we maintain frozen weights for both the CLIP encoder and multi-modal encoder, while continuing to update three components: the image embedding layer E_I, motion embedding layer E_M, and LLM parameters using LoRA [21] finetuning.

3.4. Ego4o-LLM Descriptions for Better MoCap

While our motion capture module functions effectively without verbal descriptions, incorporating high-quality motion descriptions can significantly enhance its accuracy, especially in disambiguating the challenging cases caused by self-occlusion. However, obtaining such descriptions in real-world scenarios presents a challenge, as users are typically reluctant to narrate their actions in real-time. To address this limitation, we leverage our system's ability to generate accurate motion descriptions through the Ego4o LLM. Though these generated descriptions may not perfectly match human-provided reference descriptions, they prove valuable inductive bias for enhancing motion capture performance. To further bridge the gap between generated and ground truth descriptions, we finetune our multi-modal encoder using generated descriptions for only 300 iterations, with results detailed in our ablation study (Sec. 4.4).

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets In our experiments, we evaluate our method on two datasets: the DIP-IMU dataset [22] for assessing human motion capture accuracy from IMU devices, and the Nymeria dataset [43] for evaluating both motion capture accuracy and motion description generation quality. For the results on DIP-IMU dataset, we first train the VQ-VAE on the AMASS dataset [44], then train it on synthetic IMU-based motion capture data generated from AMASS following IMUPoser [45]. The network is subsequently fine-tuned on the DIP-IMU training split before evaluation.

The Nymeria dataset contains approximately 170k human motion sequences, each 5 seconds in duration. We split the sequences into training ($\sim$ 119k sequences) and test ($\sim$ 51k sequences) sets based on different scenes and motion capture identities. For the evaluation on the Nymeria dataset, we train the VQ-VAE and multi-modal encoder and fine-tune the Ego4o LLM on the Nymeria training dataset. More implementation and training details are in suppl. mat.

Evaluation Metrics For evaluating human motion capture accuracy, we calculate joint position errors using MPJPE and PA-MPJPE (with Procrustes alignment). We also evaluate joint jitter error to assess the smoothness of predicted motion. For motion understanding, which generates natural language outputs, we employ NLP metrics including BERT score [81], BLEU [48], and ROUGE-L [36]. Details of the evaluation metrics are provided in the suppl. mat.

Method	MPJPE (mm)	PA-MPJPE (mm)	Jitter (km/s^3)
DIP-IMU Dataset			
DIP (6 IMU)	73	–	3.01
TransPose (6 IMU)	59	–	0.14
IMUPoser	97	–	0.19
Ego4o-IMU	**84.06**	**63.95**	**0.076**
Nymeria Dataset			
IMUPoser	105.7	72.94	0.054
Ego4o-IMU	<u>95.86</u>	<u>69.03</u>	<u>0.049</u>
Ego4o	**84.82**	**62.33**	**0.048**

Table 1. Quantitative results for human motion capture on the DIP-IMU and Nymeria datasets: DIP-IMU results use 1-3 IMUs. Nymeria results include 1-3 IMUs for IMUPoser and Ego4o-IMU, while Ego4o uses 1-3 IMUs, a single egocentric image, and ground truth human motion descriptions.

4.2. Comparisons on IMU-Based Human Mocap

In this section, we present our egocentric human motion capture results. Since no publicly available method supports as many modalities, we compare Ego4o to the most relevant IMU-based human motion capture IMUPoser [45] on the DIP-IMU [22] and Nymeria [43] datasets. For a fair comparison, we disable the egocentric image and motion description inputs and use only 1-3 IMUs, naming this setup Ego4o-IMU. We also evaluate the full multimodal Ego4o.

The results in Fig. 5 show the performance of Ego4o, Ego4o-IMU, and IMUPoser under different IMU setups on the Nymeria dataset, following the same evaluation protocol as IMUPoser. The results in Tab. 1 present the average performance across these various IMU configurations. These results demonstrate that Ego4o outperforms IMUPoser when using only IMU inputs. Furthermore, incorporating egocentric images and motion descriptions further enhances Ego4o's performance. Unfortunately, we were unable to compare against some related works [59, 71] due to the lack of available code. For a qualitative comparison, we visualize the body poses estimated by Ego4o and IMUPoser on the DIP-IMU and Nymeria datasets in Fig. 4. Results show that Ego4o method can accurately predict human pose from not only IMU sensor inputs but also the multi-modal inputs of egocentric images and motion descriptions.

4.3. Comparisons on Motion Understanding

This section highlights Ego4o's motion description generation capabilities. In this experiment, we do not use the motion description as input and instead rely solely on random 1-3 IMU sensors and egocentric images. We compare Ego4o's performance against previous motion description generation methods, TM2T [16] and MotionGPT [25], where we first predict the human motion using Ego4o and

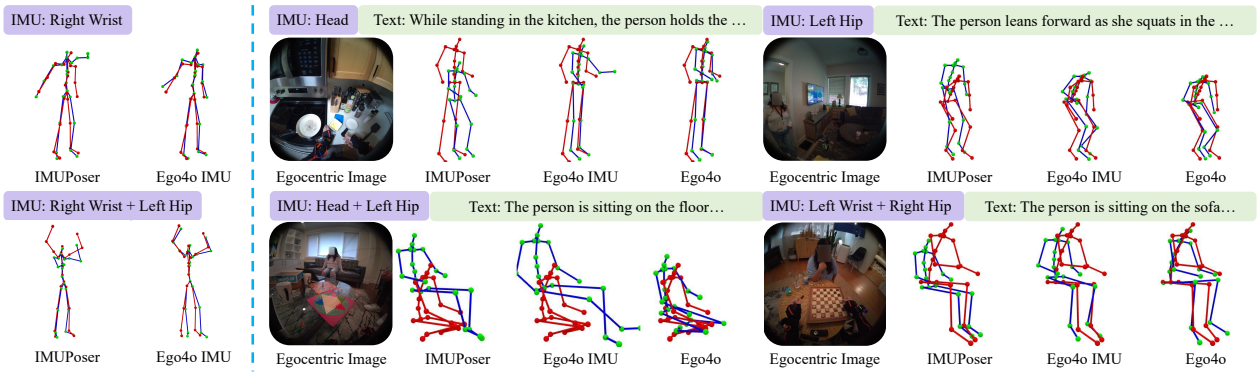

Figure 4. Comparison of human motion capture results between Ego4o, Ego4o-IMU and IMUPoser [45] on the DIP-IMU [22] (left) and Nymeria dataset [43] (right). The red skeleton is the ground truth, while the green skeleton is the predicted pose. Our predictions are more accurate than the baselines when only using IMU input, and using egocentric images and motion descriptions improves the performance.

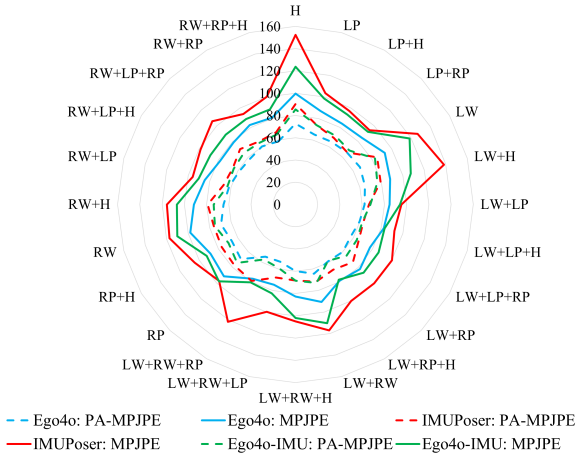

Figure 5. Quantitative results of human motion capture on Nymeria dataset. The result compares our method with IMUPoser under different IMU setups. H, LP, RP, LW, and RW indicate the IMU located on different body parts. H: head, LP: left hip, RP: right hip, LW: left wrist, RW: right wrist.

Method	Bert(idf)	Bleu@1	Bleu@4	RougeL
TM2T	11.08	40.11	8.99	30.70
MotionGPT	14.09	42.22	**10.31**	32.33
Ego4o	**30.13**	**53.83**	7.46	**38.95**

Table 2. Quantitative results of motion description generation.

then use those predictions as input to the other networks. The comparison results are shown in Tab. 2, where Ego4o outperforms the previous methods across most metrics, particularly in terms of BERTScore [81] and RougeL [36]. Ego4o did not surpass the prior methods in Bleu@4, as the Bleu score [48] focuses solely on n-gram overlap and cannot fully capture the quality of semantic understanding [81].

Method	MPJPE (mm)	PA-MPJPE (mm)	Jitter (km/s^3)
Ego4o-IMU	95.86	69.03	0.049
w/o optim	85.93	64.02	0.039
only gt text	86.22	63.14	0.048
only image	90.81	66.04	0.049
w/ gen text	88.65	64.79	0.048
image&gen text	87.00	63.67	0.049
Ego4o	84.82	62.33	0.048

Table 3. Ablation study of the IMU-based human motion capture on the Nymeria Dataset.

4.4. Ablation Study

4.4.1. Ablation Study on Human Motion Capture

Test-time optimization. To assess the performance impact of our test-time optimization (Sec. 3.2.2), we include results by not using the test-time optimization as "w/o optim" in Tab. 3. The MPJPE scores are higher, demonstrating the effectiveness of this module. We notice that the jitter is smaller when not using the test-time optimization. This is caused by the optimization with noisy IMU signals.

Multi-modal input. We evaluate the impact of different input modalities on motion capture performance. When evaluating the "only gt text" case in Tab. 3, which only uses the ground truth motion description and IMUs as input, and the "only image" case in Tab. 3, which only uses the egocentric image and IMUs, the results show a significant drop in performance compared to the full multi-modal setup. This highlights the complementary value that both the egocentric image and the ground truth motion description bring to enhancing the model's motion capture capabilities.

Generated text for better motion capture. In Sec. 3.4, we claim that if the ground truth motion description is unavailable, utilizing the generated text as input to the human motion capture module could enhance performance. To demonstrate this, in this experiment, we evaluate the mo-

Figure 6. Comparison of motion description generation between Ego4o and MotionGPT [25]. The egocentric image and ground truth human motion (for reference) are shown. Highlight predictions are marked in green, while incorrect predictions are marked in red. Ego4o's descriptions are more accurate, demonstrating the benefits of its joint modeling of multimodal inputs.

Method	Bert(idf)	Bleu@1	Bleu@4	RougeL
w/o image	22.44	48.63	5.81	36.48
w/o motion	25.55	50.90	6.32	35.71
gt motion	31.38	54.78	9.44	39.86
Ego4o	30.13	53.83	7.46	38.95

Table 4. Ablation study of imu-based motion understanding on the Nymeria dataset.

tion capture performance under two settings: First, with the generated motion description and IMUs as input, the results are shown as "w/ gen text". This performance is better than the "Ego4D-IMU" result, which only uses IMU data, and slightly worse than the "only gt text" case, which takes the ground truth text and IMU as input. Second, with the generated description, IMUs, and an egocentric image as input, the results are shown as "image&gen text". This performs better than the "only image" case, which uses the egocentric image and IMUs, and slightly worse than our full Ego4o method, which leverages IMU, egocentric image, and ground truth text.

The results demonstrate that using the generated descriptions, even if they do not perfectly match the ground truth, still leads to notable improvements in motion capture performance compared to the no-text baseline. This highlights the value of utilizing generated text to enhance the system's capabilities when ground truth descriptions are unavailable.

4.4.2. Ablation Study on Human Motion Understanding

Multi-modal input. In this experiment, we evaluate the performance of motion description generation without the egocentric image or without a human motion token as input. The results without image input, labeled as "w/o image" in Tab. 4 show a noticeable decline across all metrics. Without image context, the language model loses key contextual information, leading to reduced accuracy in motion description. The results without the human motion token input are shown as "w/o motion" in Tab. 4. The absence of human motion information causes a performance drop, as the egocentric image cannot see the human body, making it difficult to generate an accurate motion description.

Ground truth human motion. From "gt motion" row in Tab. 4, the accuracy of motion description generation is enhanced by using the ground truth motion codes (encoded by VQ-VAE encoder), compared to our Ego4o method that uses encoded motion tokens from IMUs and egocentric images. This suggests motion information is important in the model's understanding and generation capabilities.

5. Discussion

Limitations. Despite outperforming the state-of-the-art in various evaluation scenarios, our method has a few practical limitations. First, it requires motion sequences as input, which introduces latency in online applications. Second, the system's capacity for multi-round conversational interaction remains limited. To address this, future work can use better instructional fine-tuning of the large language model to generate multi-round conversational datasets.

Conclusion. In this paper, we introduced Ego4o, a framework for egocentric human motion capture and understanding that combines multi-modal inputs from wearable devices. Our versatile design allows us to operate not only with a variable number of IMU sensors, but also can optionally incorporate text and image modalities. By integrating kinematic data and semantic information, Ego4o achieves high accuracy in motion capture while providing detailed motion descriptions. We also showed the text descriptions generated by a motion-aware LLM can in turn be used to perform better text-assisted motion capture.

Our experiments demonstrate significant improvements in both tracking accuracy and description quality compared to existing methods. We envision future extensions of this work toward a long-desired human foundational model that adapts to various sensing modalities, incorporates common-sense reasoning about human physical attributes, and interacts naturally with users via text or audio.

References

[1] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2

[2] Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 3d human pose perception from egocentric stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 767–776, 2024. 2

[3] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4221–4231, 2023. 1

[4] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. 3, 4

[5] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *arXiv preprint arXiv:2312.06722*, 2023. 3

[6] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9760–9770, 2023. 3

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 3

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020. 3

[9] Eadom Dessalene, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Leap: Llm-generation of egocentric action programs. *arXiv preprint arXiv:2312.00055*, 2023. 3

[10] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023. 2

[11] Zhen Fan, Peng Dai, Zhuo Su, Xu Gao, Zheng Lv, Jiarui Zhang, Tianyuan Du, Guidong Wang, and Yang Zhang. Emhi: A multimodal egocentric human motion dataset with hmd and body-worn imus. *arXiv preprint arXiv:2408.17168*, 2024. 2

[12] Tamar Flash and Neville Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience*, 5(7):1688–1703, 1985. 2

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3

[14] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 3

[15] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 3, 4, 1

[16] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 3, 4, 6

[17] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning 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*, pages 4318–4329, 2021. 2

[18] Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Yunus Saracoglu, Torsten Sattler, and Gerard Pons-Moll. Interaction replica: Tracking human–object interaction and scene changes from human motion. In *2024 International Conference on 3D Vision (3DV)*, pages 1006–1016. IEEE, 2024. 2

[19] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C Karen Liu, Yuting Ye, and Lingni Ma. Hmd^2: Environment-aware motion generation from single egocentric head-mounted device. *arXiv preprint arXiv:2409.13426*, 2024. 2

[20] Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. Egolm: Multi-modal language model of egocentric motions. *arXiv preprint arXiv:2409.18127*, 2024. 3

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6, 2

[22] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 6, 7

[23] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4, 1

[24] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360, 2022. 3

[25] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 3, 4, 6, 8

[26] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017. 2

[27] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022. 1, 2

[28] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time ego-body pose estimation in large scenes. *arXiv preprint arXiv:2308.06493*, 2023. 2

[29] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[30] David G Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–99, 1989. 2

[31] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[32] Sunmin Lee, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler. Questenvsim: Environment-aware simulated motion tracking from sparse sensors. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 2

[33] M Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 4

[34] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 1, 2

[35] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):246–252, 2021. 3

[36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6, 7, 2

[37] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. 1

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5

[39] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yao Guo, and Guang-Zhong Yang. Ego+ x: An egocentric vision system for global 3d human pose estimation and social interaction characterization. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5271–5277. IEEE, 2022. 2

[40] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yijun Chen, Yao Guo, and Guang-Zhong Yang. Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. *IEEE Transactions on Multimedia*, 25:8880–8891, 2023. 2

[41] Minlong Lu, Ze-Nian Li, Yueming Wang, and Gang Pan. Deep attention network for egocentric action recognition. *IEEE Transactions on Image Processing*, 28(8):3703–3713, 2019. 3

[42] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34:25019–25032, 2021. 1, 2

[43] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024. 1, 5, 6, 7

[44] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 6

[45] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2023. 1, 2, 6, 7

[46] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. 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*, pages 9890–9900, 2020. 2

[47] J. Nocedal and S. Wright. *Numerical Optimization*. Springer New York, 2000. 2

[48] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6, 7, 2

[49] Jinman Park, Kimathi Kaai, Saad Hossain, Norikatsu Sumi, Sirisha Rambhatla, and Paul Fieguth. Domain-guided spatio-temporal self-attention for egocentric 3d pose estimation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1837–1849, 2023. 2

[50] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 3

[51] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. 3

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5, 1

[53] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3

[54] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9954–9963, 2019. 3

[55] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L Part, Ioannis Papaionnou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. Alanavlm: A multimodal embodied ai foundation model for egocentric video understanding. *arXiv preprint arXiv:2406.13807*, 2024. 3

[56] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[57] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7728–7738, 2019. 2

[58] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[59] Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu. Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2513–2523, 2024. 1, 2, 6

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 4

[61] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 4

[62] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. *ICCV*, 2021. 2

[63] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13157–13166, 2022.

[64] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13031–13040, 2023.

[65] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 777–787, 2024. 1, 2

[66] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021. 3

[67] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 2

[68] Qi Wu, Yubo Zhao, Yifan Wang, Yu-Wing Tai, and Chi-Keung Tang. Motionllm: Multimodal motion-language learning with large language models. *arXiv preprint arXiv:2405.17013*, 2024. 3

[69] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024. 3

[70] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2228–2238, 2023. 3

[71] Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–11, 2024. 1, 2, 6

[72] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. 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. 2

[73] Zihui Xue, Yale Song, Kristen Grauman, and Lorenzo Torresani. Egocentric video task translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2310–2320, 2023. 3

[74] Dongseok Yang, Jiho Kang, Lingni Ma, Joseph Greer, Yuting Ye, and Sung-Hee Lee. Divatrack: Diverse bodies and

motions from acceleration-enhanced three-point trackers. In *Computer Graphics Forum*, page e15057. Wiley Online Library, 2024. 2

[75] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13167–13178, 2022. 1, 2

[76] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics (TOG)*, 42(4):1–17, 2023. 2

[77] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018. 2

[78] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019. 2

[79] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 3, 4

[80] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[81] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 6, 7, 2

[82] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2025. 3

[83] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 4

[84] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1357–1366, 2024. 4

Ego4o: Egocentric Human Motion Capture and Understanding from Multi-Modal Input

Supplementary Material

6. Implementation Details

In this section, we describe the implementation details of our method.

6.1. Part-Aware VQ-VAE

6.1.1. Network Structure

We first introduce the network structure of the encoder $\mathcal{E}$ for each human body part. For each joint encoder, we utilize a codebook containing 4096 code vectors, each with a dimension of 64. The input human motion for a specific body part is $J_i \in \mathbb{R}^{T\times D}$, where T represents the motion length and D denotes the dimension of HumanML3D [15] representation. The input first traverses through a 1D convolutional layer (kernel size=3, stride=1, padding=1), followed by a ReLU activation function, producing a feature with 512 channels.

The motion feature then passes through two down-sampling blocks. Each down-sampling block comprises a 1D convolutional layer (kernel size=4, stride=2, padding=1) and three Resnet blocks. A Resnet block consists of a sequential structure: a convolutional layer, followed by a ReLU activation function, and another convolutional layer. The output from this sequence is combined with the input through addition to form the Resnet block's output.

A final 1D convolutional layer (kernel size=3, stride=1, padding=1) is applied to generate the feature $Q_i \in \mathbb{R}^{T'\times d}$, where d equals 64 (matching the code dimension) and $T' = T/4$. Before quantization, the encoded feature undergoes normalization. The full-body latent code $\hat{Q}_i$ is constructed by combining the quantized codes from all six joint encoders.

The decoder mirrors the encoder's architecture, with one key modification: convolutional layers having stride=2 are replaced with upsampling layers using nearest neighbor interpolation. This process finally yields the reconstructed human motion $\hat{J}_{\text{recon}}$.

6.1.2. Training Details

For training the part-aware VQ-VAE, we use standard loss terms including quantization, commitment, and reconstruction losses.

$$\mathcal{L} = \sum_i \Big(\beta\|\text{sg}[\hat{Q}_i] - Q_i\|_2 + \|\hat{Q}_i - \text{sg}(Q_i)\|_2\Big) + \|J - \hat{J}_{\text{recon}}\|_2 \tag{4}$$

where β is a balancing term, $\text{sg}[\cdot]$ denotes the stop-gradient operator. During training, we employ the Adam optimizer [31] with a batch size of 128 and a learning rate of 1×10^{-4}.

6.2. Multi-Modal Encoder

6.2.1. Network Structure

In implementing the masked trajectory transformer, we utilize a pre-trained CLIP-ViT-B/32 [52] model to extract features from the egocentric image and motion description. The IMU signals (A, R) are grouped into single feature tokens, with each token spanning 4 time steps. We do this grouping since it aligns with our downsampling rate of 4 in the part-based VQ-VAE framework. Each token transforms into a 512-dimensional feature through a linear projection layer.

The image features, motion description features, and IMU tokens are then concatenated and processed through a 4-layer transformer encoder to obtain the latent space representation. In each transformer encoder layer, the attention head number is 4, the dimension of the feed-forward network is 2048, and the dropout rate is 0.1. Subsequently, a 3-layer transformer encoder transforms this latent space into a sequence of logits. In each transformer encoder layer, the attention head number is 4, the dimension of the feed-forward layer is 1024 and the dropout rate is 0.1. We employ GumbelSoftmax [23] to convert these logits into motion code indices δ_i for each possible IMU location i. The final motion features $\hat{Q}_i$ are obtained by selecting from the corresponding VQ-VAE codebook C_i.

6.2.2. Training Details

For training the multi-modal encoder, we optimize the encoder network while keeping the VQ-VAE decoder and CLIP [52] model frozen. The network is trained for 25 epochs using the Adam optimizer [31] with a learning rate of 1×10^{-4} and a batch size of 128. During training, the weighting parameter λ in Eq. (1) is set to 0.001.

6.2.3. Optimization Details

In the energy function Eq. (2) in Sec. 3.2.2, we set the weights $\lambda_a = 0.01$, and $\lambda_r = 1$, respectively. We use smaller weights for the IMU accelerations since they are noisy. During the run-time optimization stage, we employ the L-BFGS [37] method with a learning rate of 1 and a convergence tolerance of 1×10^{-6}. The optimization process runs for a maximum of 1,000 iterations, maintains a history

Setups	w/o optim		only gt text		only image		w/ gen text		image & gen text	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
H	99.69	72.60	95.09	70.52	106.62	78.69	104.92	75.24	96.53	71.56
LP	86.02	65.11	94.56	70.67	92.82	73.83	91.60	69.34	89.30	66.33
LP+H	85.92	65.10	84.66	67.52	87.47	67.50	85.93	65.49	86.41	66.64
LP+RP	85.59	65.95	87.82	66.25	88.50	66.87	87.41	66.43	92.53	67.57
LW	93.76	67.06	98.22	69.93	105.46	72.68	99.22	69.62	102.66	72.69
LW+H	89.24	66.81	84.50	61.30	96.04	67.86	96.04	69.23	90.30	64.79
LW+LP	85.73	64.21	86.83	65.24	88.02	63.25	83.79	61.25	82.68	61.52
LW+LP+H	83.25	61.76	78.71	56.10	83.20	61.32	81.27	61.88	84.40	62.98
LW+LP+RP	78.45	59.53	76.78	57.25	80.98	61.06	81.55	62.33	79.40	58.54
LW+RP	83.03	62.19	82.08	61.53	88.56	65.89	89.21	69.14	84.13	63.20
LW+RP+H	80.18	61.04	79.53	58.29	86.23	63.73	78.52	58.45	80.57	60.19
LW+RW	91.78	66.39	86.97	60.79	100.60	70.83	97.83	66.98	95.32	66.37
LW+RW+H	83.75	61.86	85.38	59.33	88.71	62.52	85.05	59.64	87.50	62.34
LW+RW+LP	75.67	55.33	86.49	61.75	83.67	59.28	81.56	59.34	79.21	57.23
LW+RW+RP	78.08	56.14	82.35	58.28	83.31	56.80	82.68	56.54	79.83	57.62
RP	92.25	71.45	95.32	73.19	91.01	68.26	87.94	66.35	88.91	69.01
RP+H	89.33	68.12	84.98	64.51	84.52	64.21	90.82	69.30	83.56	65.76
RW	99.01	71.26	97.44	69.03	112.56	74.99	103.29	72.26	104.32	71.15
RW+H	92.42	66.75	95.17	66.01	99.42	69.64	90.04	64.11	94.27	66.67
RW+LP	85.64	63.77	80.35	59.87	91.61	65.31	90.86	66.36	81.71	60.36
RW+LP+H	79.42	60.85	79.57	57.55	85.03	64.21	74.61	56.42	83.25	59.26
RW+LP+RP	79.90	59.93	77.60	57.26	83.17	59.37	84.70	60.58	78.54	56.40
RW+RP	83.42	62.29	86.20	62.59	86.47	63.91	88.75	64.07	82.57	61.11
RW+RP+H	81.84	60.93	82.52	60.57	86.44	62.98	84.56	60.68	80.13	58.59

Table 5. Ablation study of the IMU-based human motion capture on the Nymeria Dataset under different IMU setups. H, LP, RP, LW, and RW indicate the IMU located on different body parts. H: head, LP: left hip, RP: right hip, LW: left wrist, RW: right wrist. The results are shown in millimeters.

size of 200, and utilizes the strong Wolfe [47] conditions for line search.

6.3. Training Details of Multi-Modal LLM for Motion Understanding

During the pre-training phase, we train only the motion embedding layer E_M while keeping all other modules frozen. The embedding layer is trained for 1 epoch using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 16. In the multi-modal fine-tuning phase, we keep the CLIP model and multi-modal encoder frozen while fine-tuning both the image and motion embedding layers along with the large language model. We employ LoRA [21] with a rank of 128 and an alpha value of 256. The language model is fine-tuned for 4 epochs using the Adam optimizer with a learning rate of 2×10^{-5} and a batch size of 16.

7. Ablation Study on Different IMU Setups

In this section, we present ablation study results for human motion capture across different IMU configurations in Table 5.

8. Evaluation Metrics

We evaluate our method using three standard metrics for human motion capture accuracy: Mean Per Joint Position Error (MPJPE), Procrustes-aligned Mean Per Joint Position Error (PA-MPJPE) and Jitter. MPJPE measures the average Euclidean distance between predicted and ground truth joint positions. To compute PA-MPJPE, we first perform rigid alignment of the predicted pose to the ground truth using Procrustes analysis [30], then calculate the MPJPE. The Procrustes alignment helps evaluate pose accuracy independent of global position and orientation. Jitter [12] quantifies motion smoothness by measuring the mean jerk (third-time derivative of position) of all body joints in global space, expressed in km/s^3.

We evaluate our method with three metrics for motion description accuracy: BERT score [81], BLEU [48], and ROUGE-L [36]. BLEU measures the precision of n-gram matches between generated and reference texts, indicating how well the generated descriptions align with ground truth at the phrase level. BERT score leverages pre-trained BERT embeddings to compute semantic similarity

between generated and reference descriptions, providing a more contextually-aware evaluation than traditional n-gram based metrics. ROUGE-L computes the longest common subsequence between generated and reference descriptions, capturing the fluency and sequential consistency of the generated text.

9. W/O Part-Aware VQ-VAE

In this section, we evaluate the effectiveness of our part-aware VQ-VAE by comparing its reconstruction accuracy with that of the traditional VQ-VAE.

Our part-aware VQ-VAE achieves a Mean Per Joint Position Error (MPJPE) of 44.93 mm and a Procrustes-aligned MPJPE (PA-MPJPE) of 32.72 mm. In contrast, the traditional VQ-VAE yields higher errors with an MPJPE of 47.73 mm and a PA-MPJPE of 36.71 mm. These results demonstrate that our part-aware approach reduces the reconstruction error, indicating superior performance in preserving motion details and overall pose structure.