EgoLM: Multi-Modal Language Model of Egocentric Motions

Fangzhou Hong[1,2†], Vladimir Guzov[1,3,4†], Hyo Jin Kim[1], Yuting Ye[1]
Richard Newcombe[1], Ziwei Liu[2 ✉], Lingni Ma[1 ✉]
[1]Meta Reality Labs Research, [2]S-Lab, Nanyang Technological University,
[3]Tübingen AI Center, University of Tübingen, [4]MPI for Informatics, Saarland Informatics Campus

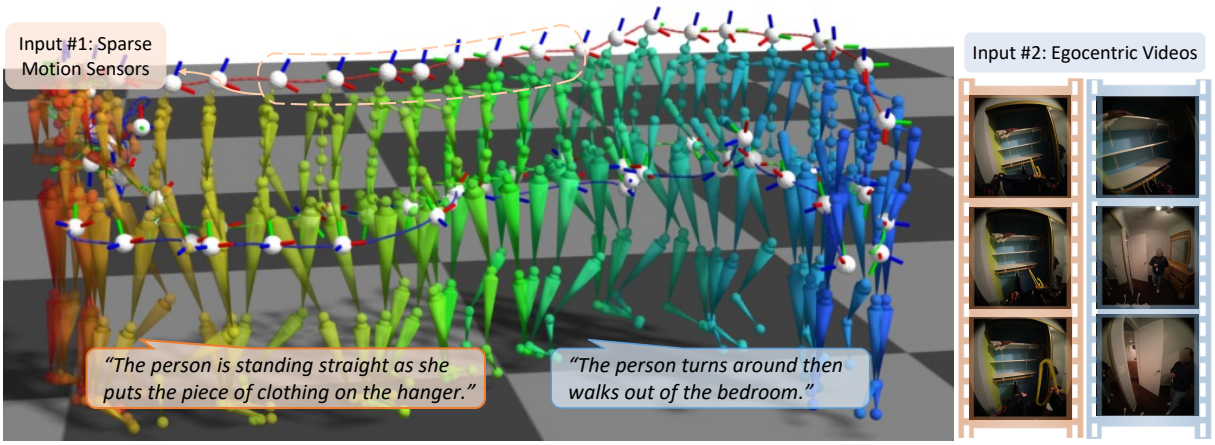


Figure 1. We propose **EgoLM**, a multi-modal language model that unifies egocentric motion tracking and understanding from wearable sensor data, *i.e.*, sparse motion sensors and egocentric videos.

Abstract

*As wearable devices become more prevalent, understanding the user's motion is crucial for improving contextual AI systems. We introduce **EgoLM**, a versatile framework designed for egocentric motion understanding using multi-modal data. EgoLM integrates the rich contextual information from egocentric videos and motion sensors afforded by wearable devices. It also combines dense supervision signals from motion and language, leveraging the vast knowledge encoded in pre-trained large language models (LLMs). EgoLM models the joint distribution of egocentric motions and natural language using LLMs, conditioned on observations from egocentric videos and motion sensors. It unifies a range of motion understanding tasks, including motion narration from video or motion data, as well as motion generation from text or sparse sensor data. Unique to wearable devices, it also enables a novel task to generate text descriptions from sparse sensors. Through extensive experiments, we validate the effectiveness of EgoLM in addressing the challenges of under-constrained egocentric motion learning, and demonstrate its capability as a generalist model through a variety of applications. Project page: https://hongfz16.github.io/projects/EgoLM.*

✉ Corresponding author.
† Work done during internships at Meta Reality Labs Research.

1. Introduction

Smart wearable devices, such as Ray-Ban Meta [33] and Spectacles [48], offer new opportunities for developing personal AI assistants by capturing the world from the user's perspective. They provide real-time egocentric observations about the user's environment and actions. On the other hand, large language models (LLMs) [2, 53] encode such context through text in their latent space, which can be leveraged for common-sense reasoning and human understanding. The fusion of egocentric perception and common-sense reasoning presents a unique and exciting opportunity for advancing contextual AI research, among which, egocentric motion understanding is an essential task [41].

However, a key challenge in utilizing egocentric perception is the **lack of direct observations of the wearer**. Two types of observations are available from wearable devices, *i.e.*, *1) egocentric videos* and *2) sparse motion sensors*. Egocentric videos, captured by cameras mounted on smart glasses, provide rich contextual information of the wearer's environment and interactions. But the wearer's body is rarely visible in the video, due to constrained cam-

Table 1. **Comparison with Related Works.** EgoLM uses novel techniques to effectively unify a wide range of multi-modal motion understanding tasks. "Vid.": egocentric videos. "Mot.": motions. "P.T.": pre-training.

Method	Motion Tokenizer	Backbone Type	Pre-Training	Instruction Tuning	Modalities				
					3pts	1pt	Mot.	Vid.	Text
T2M-GPT [62]	Vanilla VQ-VAE	Transformer	Text-to-Motion Gen	N/A			✓		✓
MotionGPT [21]	Vanilla VQ-VAE	Encoder-Decoder LM	Text-to-Motion Gen	Motion-Text Translation			✓		✓
LLaVA [28]	N/A	Decoder-Only LM	N/A	Image Understanding				✓	✓
EgoLM (Ours)	Motion Product Quantization(PQ) VAE	Decoder-Only LM	Motion-Aug LM P.T. w/ only Motion Data	3pts/1pt/Vid. Motion Tracking 3pts/Mot./Vid. Motion Narration	✓	✓	✓	✓	✓

era mounting position and angle. Sparse motion sensors provide low-level kinematic motion of a few important body parts, *i.e.*, head motions from glasses and wrists movements from smart watches. However, they are insufficient to inform the full body pose, especially for the lower body.

Our insight is that these **two types of indirect observations are complementary to each other**. Egocentric videos can provide strong clues of the environment, and help disambiguate the lower body motion. For example, a laptop placed on an office table is a strong indication that the wearer is sitting rather than squatting. Sparse motion sensors, on the other hand, offer precise tracking of important body parts, such as hand movements, which can help in scenarios where no body part is visible in the video. For example, sparse motion sensors can differentiate between jumping jacks and simple jumps, where egocentric video may appear identical.

Another key challenge in egocentric human understanding is **aligning motion and language representations**, so that we can leverage the vast contextual knowledge embedded in LLMs to describe motion. While motion signals are continuous, low-level kinematic representations, natural language consists of discrete tokens. To bridge this gap, we **treat motion as a form of language**. By tokenizing motions and repurposing a pre-trained LLM to model the joint distribution of motion and language, we facilitate an effective alignment between these two distinct representations.

With the above insights, we introduce **EgoLM**, a versatile framework for egocentric motion understanding that leverages rich sensor observations and strong contextual understanding from LLMs. As shown in Fig. 1, EgoLM takes sparse motion sensor data and egocentric videos as inputs, and generates motion and natural languages as outputs. The framework unifies a range of motion understanding tasks, at both the *kinematic* and *semantic* levels. At the kinematic level, EgoLM can perform motion tracking from three-points [22] or one-point [25] sensor data, incorporating egocentric videos for disambiguation. At the semantic level, EgoLM can generate motion narration from various combinations of input modalities. More importantly, we highlight a novel task of motion narration from three-points and egocentric videos, unique to AR use cases.

Compared with recent VLMs [27, 28], our approach tackles a more complex and challenging problem involving **more modalities and tasks with greater disparities**. In particular, both our input modalities and output tasks encode information at varying levels of granularity. To tackle it, we employ **multi-modal multi-task joint training** through instruction tuning. Multiple input modalities are aligned to LLM latent space with rich contextual information, and interleaved between text instructions. Multi-task training exploits connections between tasks and benefits each other. For instance, three-points motion tracking bridges the gap between sparse motion sensors and natural languages, improving the performance of motion narration from three-points and videos. Moreover, the performance of motion tokenization and pre-training are crucial for motion tracking quality. Therefore, we propose the **Motion Product Quantization(PQ) VAE** to improve the motion reconstruction quality, and **Motion-Augmented LM Pre-Training** for better motion distribution modeling.

To validate the proposed framework, we perform extensive experiments on a large-scale motion dataset, Nymeria [31]. Compared with previous dedicated motion tracking and understanding models, we show better performance in both tasks, under different combinations of input modalities, proving EgoLM as a generalist model. Our contributions are summarized below.

1) We introduce a egocentric motion generalist model EgoLM, which integrates a variety of motion understanding tasks at both kinematic and semantic levels. By leveraging large language models (LLMs), we aim to enhance egocentric perception, thereby contributing to the advancement of contextual AI research. **2) We address the challenge of under-constrained egocentric motion learning** by combining two complementary modalities, *i.e.*, sparse motion sensors and egocentric videos. This new paradigm enables two unique applications for AR use cases: *motion tracking and narration from sparse motion sensors and egocentric videos*. **3) We employ multi-modal multi-task joint training to bridge substantial gaps between modalities and tasks.** Extensive experiments validate the effectiveness of this training strategy.

2. Related Work

Motion Regression. Many efforts are devoted to regress 2D or 3D keypoints from human images or videos [29, 32, 39, 52]. Wearable motion sensors are also used for motion capture [23, 34, 35, 42, 60]. Recent advancements in VR/AR have developed a new setup for motion tracking [3, 9, 10, 15, 22, 25, 69], *i.e.*, three-points and one-point

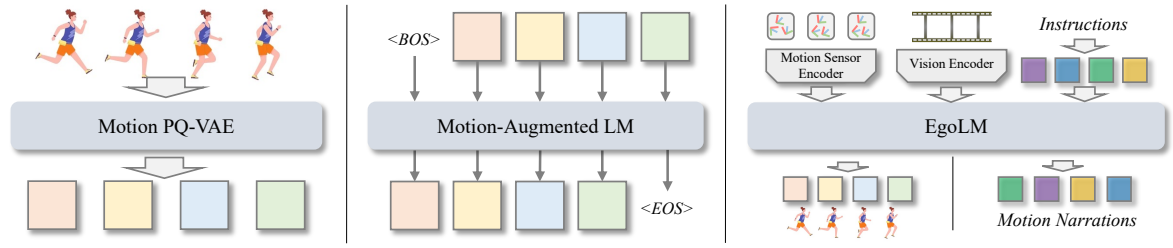


Figure 2. **Overview of EgoLM.** Three steps are designed for the training of EgoLM, *i.e.*, motion PQ-VAE training for motion tokenization, motion-augmented LM pre-training and multi-modal multi-task instruction tuning.

body tracking. In this work, we target motion tracking from sparse sensors and rich semantics in egocentric videos to disambiguate under-constrained cases.

Motion Generation. There have been many efforts in generating motions from various conditions, *i.e.*, action labels [12, 30, 40], natural languages [13, 14, 17, 43, 51, 62–65]. Recently, researchers use LLMs to model the joint motion-language distribution for text-to-motion generation [67, 70]. EgoLM also adopts the similar idea. But in comparison with MotionGPT [21], as listed in Tab. 1, EgoLM improves the motion tokenizer with PQ-VAE, employs the more scalable decoder-only LM with motion-augmented pre-training, does not rely on paired data for pre-training and supports more egocentric motion tasks and modalities.

Motion Understanding. There have been many setups in motion understanding. From the input side, human videos, either from third-person view [24, 50, 54, 55, 58] or first-person view [5–7, 59], are used for this task. From the output side, action recognition has been a classic task [5, 50]. More recently, with the development of LLMs, natural languages are used as output [4, 11, 20, 56, 57]. In EgoLM, we highlight a new setup of motion narration from sparse motion sensors and egocentric videos, unique to AR use cases.

Language Models. LLMs have been a huge success in recent years with the large-scale pre-training [2, 44] and alignment [1, 37]. To exploit the powerful text generation ability, image [27, 28] or video understanding [61] are defined as conditional text generation. LLaVA [28] proposes to encode images with pre-trained vision encoders [45] and perform visual instruction tuning. EgoLM adopts the similar idea to tackle the challenge of large modality and task gaps. As shown in Tab. 1, compared with LLaVA, EgoLM handles a more complex egocentric setup, with more modalities and tasks with larger disparities.

3. Method

The overview of EgoLM is demonstrated in Fig. 2. There are three key steps in EgoLM training. In the first step, we train a motion PQ-VAE as the motion tokenizer (Sec. 3.2). The second step is motion-augmented LM pre-training for motion distribution learning (Sec. 3.3). The last step is multi-modal multi-task joint training to guide the model to perform various egocentric motion tasks (Sec. 3.4).

3.1. Preliminaries

Language Model. Language models (LMs) model the distribution of natural languages. It consists of three parts. The first is a codebook that stores the token embeddings. The second is the transformer backbone that takes text embeddings as inputs. Output features are mapped to probabilities of the next tokens by the third part of LM head.

Motion Representation. Human motions are represented as sequences of poses, global translations and rotations defined on the root joint. Each frame of pose is represented by joint angles, defined on a kinematic tree. For better learning of motion dynamics, we also include joint angle velocity in the representation. To avoid the normalization of global translation, we use the translation velocity $V_t^r \in \mathbb{R}^3$ for each frame, which can be integrated back to global translations. To ease the regression difficulty of rotation angles, we use 6D rotation representations [16] for the root rotation $R_t^r \in \mathbb{R}^6$, root rotation velocity $R_t^{rv} \in \mathbb{R}^6$, joint angles $R_t^j \in \mathbb{R}^{22\times6}$, and joint angle velocity $R_t^{jv} \in \mathbb{R}^{22\times6}$. Formally, we represent human motions with T frames as $M = \{P_t\}_{t=1}^T$, where $P_t = [V_t^r; R_t^r; R_t^{rv}; R_t^j; R_t^{jv}] \in \mathbb{R}^{279}$. Forward kinematics (FK) and integration of root velocity are used to recover the joint positions $J = \text{FK}(M) \in \mathbb{R}^{23\times3}$.

3.2. Motion PQ-VAE

To treat the motion as a form of a language and train with LMs, a motion tokenizer is in need, which is usually realized by VQ-VAE [36]. However, VQ-VAE often suffers from inferior reconstruction quality, leading to poor generation quality. Therefore, we propose Motion Product Quantization VAE (PQ-VAE) for improved motion tokenization and decoding quality.

As shown in Fig. 3 a), the motion PQ-VAE consists of a fully convolutional encoder $\mathcal{E}$ and decoder $\mathcal{D}$. The fully convolutional design enables processing motions with arbitrary lengths. The encoder embeds raw motion representation to latent features $f^m = \mathcal{E}(M)$, where $f^m \in \mathbb{R}^{T/r\times c}$, $M \in \mathbb{R}^{T\times279}$. r is the down-sample rate. Then, multiple product codebooks are learned to quantize the motion latent features. Product quantization [18] increases the codebook expressiveness by decomposing the latent space

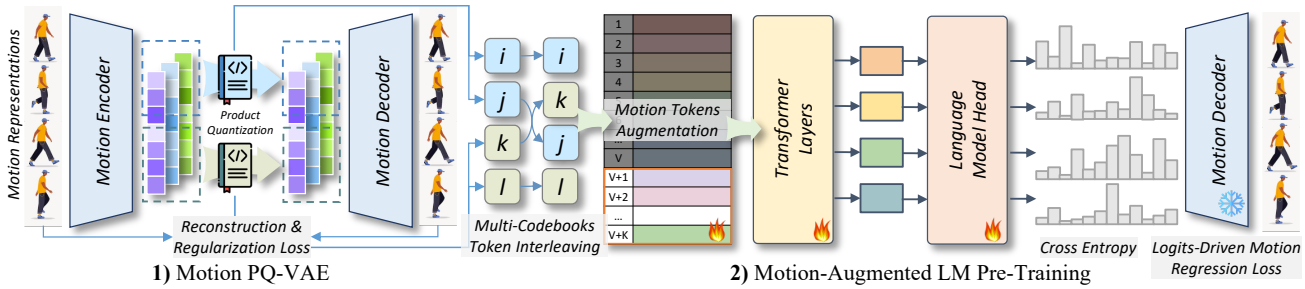


Figure 3. **Details of a) Motion Tokenizer (PQ-VAE) and b) Motion-Augmented LM Pre-Training.** Product quantization provides high-fidelity motion tokenization. It is used for motion pre-training with a decoder-only LM, where motion augmentation is implemented.

into a Cartesian product of sub-spaces with lower dimensions. Specifically, the latent feature f^m is split equally into N trucks $\{f_n^m\}_{n=1}^N$, which are quantized separately by N codebooks $\{Z_n\}_{n=1}^N$. Each codebook with K entries is defined as $Z_n = \{z_i\}_{i=1}^K$, where $z_i \in \mathbb{R}^{c/N}$. The quantization process for feature f_{tn}^m at frame t and trunk n is formulated as $i_{tn} = Q(f_{tn}^m) = \arg\min_{z_i \in Z_n} \|f_{tn}^m - z_i\|_2$. To further ensure high-fidelity motion tokenization, we also employ exponential moving average and codebook reset [8]. After quantization, we obtain the corresponding codebook entry for the motion latent feature $\hat{f}^m = \{\hat{f}_t^m\}_{t=1}^{T/r} = \{z_{i_t}\}_{t=1}^{T/r}$. It is input into the decoder $\mathcal{D}$ to decode raw motion representation $\hat{M} = \mathcal{D}(\hat{f}^m)$.

For PQ-VAE training, two types of training losses are used. The first is the commitment loss $\mathcal{L}_c = \|f^m - \hat{f}^m\|_2$ for the codebook learning. The second is motion reconstruction loss $\mathcal{L}_r$, which consists of raw representation loss $\mathcal{L}_m$, joint position loss $\mathcal{L}_j$, rotation velocity loss $\mathcal{L}_v$, which are defined as

$$\begin{aligned}\mathcal{L}_r &= \lambda_m \mathcal{L}_m + \lambda_j \mathcal{L}_j + \lambda_v \mathcal{L}_v \\ &= \lambda_m \|M - \hat{M}\|_1 + \lambda_j \|\text{FK}(M) - \text{FK}(\hat{M})\|_1 \\ &+ \lambda_v \|R_{1:T-1}^{rv} - (R_{1:T-1}^{r})^{-1} R_{2:T}^{r}\|_1 \\ &+ \lambda_v \|R_{1:T-1}^{jv} - (R_{1:T-1}^{j})^{-1} R_{2:T}^{j}\|_1,\end{aligned} \tag{1}$$

where $\|\cdot\|_1$ is smoothed L1 loss. The total loss is $\mathcal{L}_{pq} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r$, where λ_* are manually adjusted weights.

3.3. Motion-Augmented LM Pre-Training

EgoLM aims to empower egocentric motion learning with strong prior in pre-trained LMs. However, the pre-trained LM only models the distribution of natural languages. Therefore, to facilitate motion generation, we perform motion-augmented LM pre-training to learn motion distributions. The motion augmentation on LM pre-training is in two ways. The first is to augment LM networks for new motion tokens. The second is to enforce motion awareness with motion-augmented next-token prediction training.

Firstly, since the pre-trained LM is designed for text tokens only, LM network augmentation for motion tokens is in need, as shown in Fig. 3. Firstly, we expand the LM codebook in accordance with the size of motion codebook. The output shape of the LM head is also expanded accordingly.

To accommodate tokens produced by multiple product codebooks in motion PQ-VAE, we employ token interleaving to arrange orders of motion tokens. Specifically, tokens from the n-th codebook is defined as $W^n = \{i_{tn}\}_{t=1}^{T/r}$. The interleaving operation will rearrange the tokens to $W = \{[i_{tn}]_{n=1}^N\}_{t=1}^{T/r} = \{w_t\}_{t=1}^{L_W}$. They are fed into the LM to learn the motion distribution by next-token prediction [44].

Motion-Augmented Next-Token Prediction. As part of the next-token prediction loss, we use the common cross entropy loss $\mathcal{L}_{ce}$ to maximize the log-likelihood of the next-token probability given network parameter Θ, which is formulated as $\mathcal{L}_{ce} = -\sum_{i=2}^{L_W} \mathbb{P}(w_i | w_1 ... w_{i-1}; \Theta)$. Additionally, to bridge the gap between the motion tokens and raw representations, we further enforce motion awareness with motion regression loss. However, the token sampling process is not differentiable. As a circumvention, we propose to use predicted logits $l_n \in \mathbb{R}^{T/r \times K}$ to blend motion features $f_b = \{Z_n \cdot \text{softmax}(l_n)\}_{n=1}^N$, which can be decoded to motion representations for regression loss. Specifically, it is defined as $\mathcal{L}_{reg} = \|\mathcal{D}(f_b) - M\|_1$. In summary, the motion-augmented next-token prediction loss is $\mathcal{L}_{nt} = \lambda_{ce} \mathcal{L}_{ce} + \lambda_{reg} \mathcal{L}_{reg}$.

As the by-product of this stage training, we obtain an auto-regressive motion generator. Given a leading motion sequence as the prompt, it can sample an arbitrary length of human motions that continues the given motion. More importantly, the LM learns human motion distributions and has the ability of sampling human motions with high quality, which lays a solid foundation for the next stage.

3.4. Multi-Modal Multi-Task Instruction Tuning

As discussed above, EgoLM addresses a more challenging problem, involving multiple modalities and tasks with significant disparities. On the modality side, in addition to motion and natural languages, we need to integrate data from sparse motion sensors and egocentric videos, which capture information at varying levels of granularity. Furthermore, EgoLM approaches egocentric motion understanding tasks from both kinematic and semantic perspectives. To tackle the challenge, we propose to employ multi-modal multi-task joint training to bridge the gaps between modalities and uncover the inherent connections between tasks.

Recent research on multi-modal LLMs has demonstrated

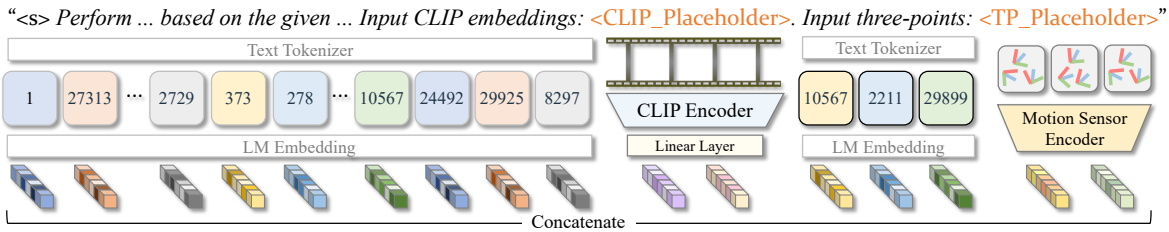


Figure 4. **Details of Multi-Modal Multi-Task Instruction Tuning.** Different modalities are encoded separately. Their features are concatenated in the order of the instruction template and input into the transformer layers of the language model.

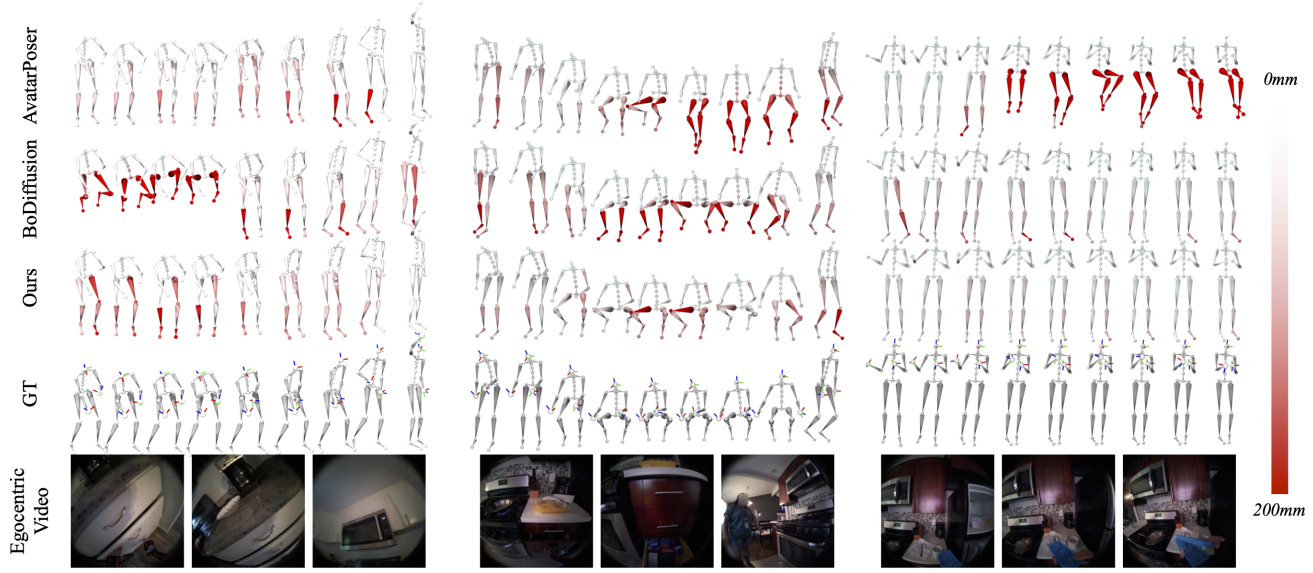


Figure 5. **Qualitative Results of Three-Points Motion Tracking.** Skeletons are color-coded by the joint position errors. Baseline methods use 3pts as inputs. Ours uses 3pts and videos as inputs.

that instruction tuning [1, 28, 37, 68] effectively aligns different modalities and integrates multiple tasks. In our approach, various modalities are encoded differently. For motions and natural languages, both serve as inputs and outputs; thus, they are tokenized for auto-regressive modeling. Sparse motion sensors and egocentric videos are used exclusively as inputs. It is more efficient to encode these into continuous features that align with the LM latent space. Different tasks are differentiated by text instructions. Specifically, the instruction template typically includes: 1) text instructions specifying the tasks to perform; 2) inputs relevant to the task; and 3) expected outputs. Please refer to supplementary material for instruction examples and explanations.

A detailed illustration of how we organize different modalities of data is shown in Fig. 4. The encoded three-points 6-DoF poses features are placed at *<TP_Placeholder>*. The placeholder for egocentric video features is *<CLIP_Placeholder>*. Texts are tokenized and embedded to feature vectors through LM embedding. Egocentric videos are sampled to sequences of frames and encoded by CLIP image encoder [45], which are further projected by linear layers to the LM feature space. Similarly, sparse motion sensor data, *e.g.*, sequences of three-points 6-DoF poses, is encoded by a fully convolutional encoder. Lastly, all the encoded features are concatenated in an interleaved way and input into LM transformer layers.

With instruction templates defined, we can facilitate joint training across the following tasks: a) motion tracking with three-points and egocentric videos, b) motion narration using three-points and egocentric videos, c) text-to-motion generation, and d) motion-to-text generation. During training, these four tasks are randomly sampled with equal probability. The loss function is the motion-augmented next-token prediction loss $\mathcal{L}_{nt}$, defined in Sec.3.3.

During inference of motion narration, natural language is sampled in the same manner as regular LMs. For motion tracking, our auto-regressive modeling offers the advantage of online inference. At each new time step, the incoming data is concatenated with historical data and fed into EgoLM. A single feed-forward inference is then performed to obtain the motion token for the current time step. For further details, please refer to the appendix.

4. Experiments

4.1. Experiment Setup

Dataset. We use the Nymeria dataset [31] to train and validate our method. The dataset includes: **a)** full-body motions captured by the Xsens Mocap system [47], **b)** egocentric videos recorded with Aria glasses [49], and **c)** motion narrations by human annotators. Three-point 6-DoF poses are derived from ground truth joints for comparison with prior

Table 2. **Quantitative Results of Motion Tracking.** EgoLM performs comparably with task-specific algorithms. Incorporating video input can outperform methods without. "Full", "Upper", "Lower" are joint position errors in mm. "J.A.", "Root" are joint angle errors for full body and root joint in degree. *We concatenate CLIP embeddings with three-points input to adapt AvatarPoser. †We replace three-points with one-point to train AvatarPoser. We highlight the first and second scores.

Method	Input Modality 3pts	1pt	Video	Full	Upper	Lower	J.A.	Root
AvatarPoser [22]	✓			85.89	52.78	165.18	12.41	14.78
Bodiffusion [3]	✓			79.80	52.79	152.68	12.74	13.09
Ours	✓			83.88	54.06	148.37	13.31	14.13
AvatarPoser* [22]	✓		✓	127.08	100.02	190.32	18.90	21.80
Ours	✓		✓	73.38	49.67	124.58	12.48	13.23
AvatarPoser† [22]		✓		129.23	94.19	192.34	16.55	21.60
EgoEgo [25]		✓		132.16	100.02	190.32	18.90	21.80
Ours		✓		127.45	97.87	174.92	16.97	20.57
Ours		✓	✓	106.95	83.73	141.26	14.67	19.04

work. The motion tracking training set comprises 147.89 hours of data, with a test set of 41.93 hours. For motion understanding, the training set includes 16,673 segments (totaling 15.77 hours), while the test set contains 7,468 segments (totaling 6.76 hours).

Training Details. Motion PQ-VAE utilizes two codebooks, each containing 8,192 entries with a code dimension of 64. The down-sample rate is set to $r = 4$. For motion tracking, all experiments use a batch size of 60 frames (equivalent to 1 second), with random rotation augmentations. We employ GPT-2 Medium [44] as the language backbone.

Evaluation Protocols. For motion tracking, we calculate joint position errors (for full, upper and lower body), joint angle errors (for full body and root joint). For motion narration, the outputs are natural languages. Therefore, we adopt NLP metrics, including BERT [66], BLEU [38], and ROUGE [26] scores. For more details about the evaluation protocols, please kindly refer to the appendix.

4.2. Motion Tracking

Quantitative Results. We present the quantitative results of motion tracking in Tab. 2. All methods are evaluated using batch inference, with the size of 60 frames. We assess various input combinations from three modalities, *i.e.*, three-points 6-DoF poses ("3pts"), one-point 6-DoF poses ("1pt") and egocentric videos ("Vid"). In the 3pts-only and 1pt-only settings, EgoLM demonstrates competitive performance compared with task-specific algorithms, with large advantages in lower body tracking performance, highlighting the effectiveness of LMs for precise motion tracking. Additionally, we incorporate egocentric videos to provide contextual information for motion tracking. For three-points tracking, this additional modality results in a 10 mm improvement in full-body joint error. The adapted AvatarPoser* fails to exploit the video input, highlighting the challenge in using modalities with large disparities. For one-point tracking, the inclusion of egocentric videos leads to a 20 mm reduction in joint error, underscoring their effectiveness in disambiguating the ill-posed problem.

Table 3. **Quantitative Results of Motion Narration.** Different input modality combinations are tested. All metrics are higher the better. "Mot" stands for motion. "Vid" stands for videos. We highlight the first scores for different settings.

Method	Input Modality 3pts	Mot.	Vid.	Bert	Bleu@1	Bleu@4	RougeL
TM2T [14]		✓		11.08	40.11	8.99	30.70
MotionGPT [21]		✓		14.09	42.22	10.31	32.33
Ours (M2T&T2M)		✓		15.90	42.68	11.06	33.71
Ours (MV2T&T2M)		✓	✓	20.32	45.33	12.80	35.31
Ours (TP2T)	✓			11.94	41.70	9.85	31.47
Ours (V2T)			✓	16.62	43.03	11.34	33.13
Ours (TPV2M + MV2T)	✓		✓	19.97	45.41	12.81	35.04
Ours (TPV2T)	✓		✓	18.38	44.55	12.12	33.80
Ours (Joint Training)	✓		✓	19.40	45.45	12.74	34.82

Qualitative Results. The results and comparisons for three-point motion tracking are presented in Fig. 5. Due to the inherent ambiguity, AvatarPoser incorrectly generates standing poses for squatting sequences (second example). BoDiffusion, while capable of producing correct results in some instances (*e.g.*, the squatting example), also faces ambiguity issues, as demonstrated in the bending-down sequence (first example). These examples highlight the importance of contextual consideration in motion tracking for effective disambiguation. Our full model reliably performs three-point body tracking in these challenging scenarios.

The results for one-point motion tracking are presented in Fig. 6. This task is particularly challenging for upper body tracking. In the first example, the upper body motions generated by EgoEgo significantly diverge from the ground truth. In the second example, EgoEgo mistakenly produces sitting poses for standing frames and vice versa, showing the ambiguity issue. In contrast, egocentric videos not only help to resolve ambiguity issues but also provide clues about hand positions. In the first example, when hands are visible in the frames, EgoLM leverages vision clues to generate accurate arm movements. More visual results are provided in the appendix.

4.3. Motion Narration

Quantitative Results. We report the quantitative results of motion narration in Tab. 3. This task involves three input modalities, *i.e.*, three-points ("3pts"), motions, and egocentric videos ("Vid"), with various combinations evaluated. We first compare EgoLM with two existing motion narration methods that utilize motion as their sole input, *i.e.*, TM2T [14] and MotionGPT [21]. TM2T trains language generation from scratch and consequently exhibits poor performance. MotionGPT leverages a pre-trained T5 model [46]. EgoLM(M2T&T2M) outperforms these methods, benefiting from the scalability of its decoder-only architecture. When we combine egocentric videos with mo-

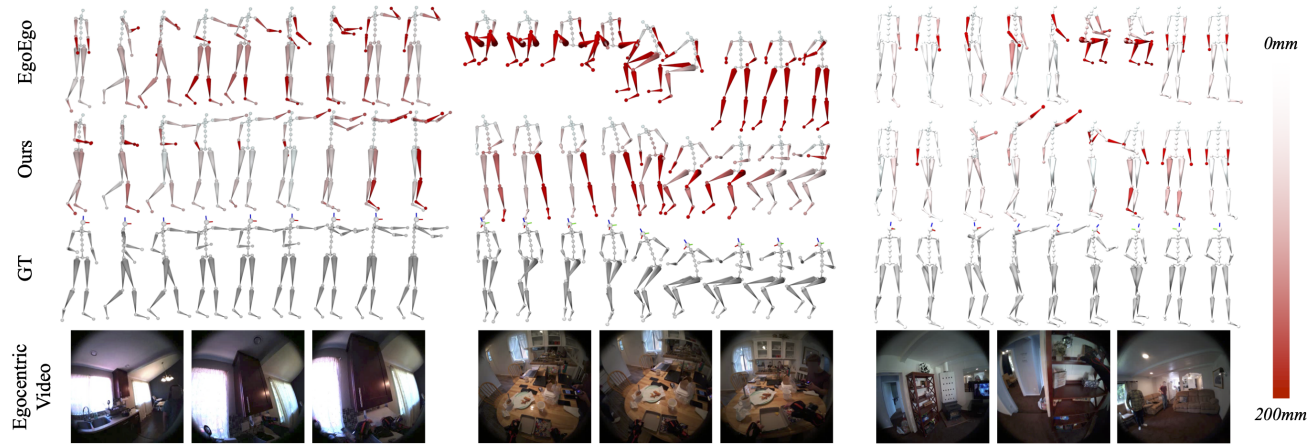


Figure 6. **Qualitative Results of One-Point Motion Tracking.** Skeletons are color-coded by joint position errors. EgoEgo only uses one-point as inputs. Ours includes egocentric videos as inputs.

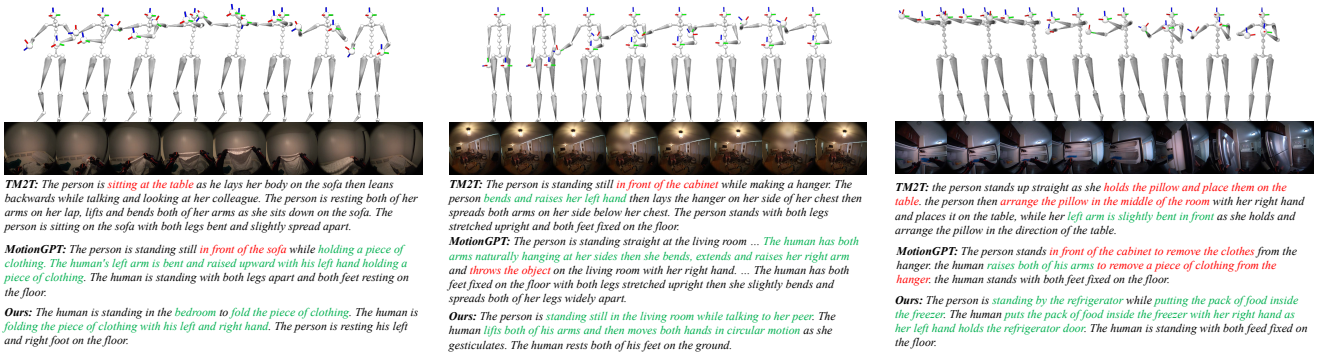


Figure 7. **Qualitative Results of Motion Narration.** We use green to highlight correct parts and red for mistakes.

tion inputs (MV2T&T2M), we achieve the best overall performance.

Using motion as input requires precise motion tracking, which is not always feasible, prompting us to explore sensor inputs instead. We tested two variants: three-points-only (TP2T) and egocentric videos only (V2T). The TP2T variant demonstrated a noticeable drop in performance compared to the motion-only version, as three-points provide limited information about body motion. Conversely, the V2T variant outperformed the motion-only version because egocentric videos capture relevant environmental context for our motion narrations. This underscores the significance of egocentric videos in understanding motion.

We then evaluate our highlighted setup of combining three-points and egocentric videos for motion narration. There are three approaches to achieve this. The first involves integrating two existing setups: 1) three-points motion tracking and 2) motion-to-text generation (TPV2M + MV2T). This variant shows a slight performance drop compared to MV2T due to error accumulation and requires a time-consuming two-pass inference. The second approach directly trains a three-points plus egocentric videos to text generation model (TPV2T) using our proposed multi-modal instruction tuning. While this outperforms using only egocentric videos or motions, it still lags behind the MV2T variant due to missing lower body information. To address this, we propose joint training of four tasks to establish connections between three-point poses and motion narrations, achieving optimal performance in a single forward pass for this new task.

Qualitative Results. We show three examples of motion narration in Fig. 7. TM2T and MotionGPT use full body motions as inputs, while our model incorporates three-points and egocentric videos. TM2T's language generation is trained from scratch, leading to frequent errors and nonsensical outputs. MotionGPT generates reasonable descriptions; for instance, in the third example, it correctly identifies the motion as "removing a piece of clothing from the hanger". However, our target motion narration is closely tied to environmental context, which TM2T and MotionGPT struggle with due to the absence of visual signals. In contrast, although EgoLM does not directly use motions as inputs, it jointly models the distributions of different modalities, enabling it to generate accurate narrations based on varying scenarios. Please refer to the appendix for more qualitative results results.

4.4. Ablation Study

Window Size of Motion Tracking. As shown in Tab. 4, increasing the window size for three-points motion track-

Table 4. Ablation Study on Window Size for Motion Tracking.

Win	Vid	Full	Upper	Lower	J.A.
60		83.88	54.06	148.37	13.31
120		79.61	52.66	138.87	13.01
60	✓	73.38	49.67	124.58	12.48
120	✓	72.76	49.20	123.09	12.52

Table 5. Ablation Study on Reconstruction Results of Motion PQ-VAE. $[mm]$

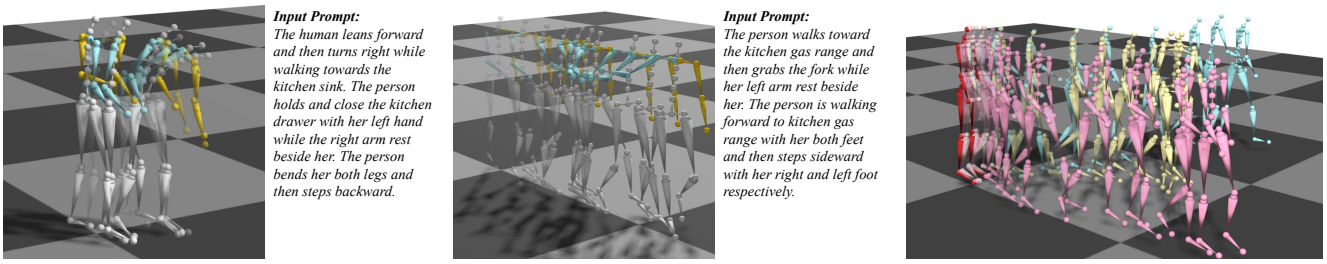
PQ	CB	Dim	MPJPE	PA-MPJPE	ACCEL
✗	2048	512	51.60	37.55	1.09
✓	2048	512	39.63	29.77	0.71
✓	16384	256	39.13	29.78	1.08
✓	16384	64	34.49	26.83	0.67

Table 6. Ablation on the size of LM.

GPT-2 Size	Medium	Large
Bert↑	18.38	19.56
Bleu@1↑	44.55	44.48
Bleu@4↑	12.12	12.49
RougeL↑	33.80	35.21

Table 7. Ablation on next-token prediction loss $\mathcal{L}_{nt}$.

Metrics	w/o $\mathcal{L}_{reg}$	w/ $\mathcal{L}_{reg}$
Full	74.10	73.38
Upper	50.38	49.67
Lower	125.89	124.58
J.A.	12.50	12.48



a) Text-to-Motion Generation Results **b)** Motion Prediction Results

Figure 8. **More Applications of EgoLM. a)** Qualitative results of text-to-motion generation. **b)** Qualitative results of motion prediction.

ing from 60 to 120 frames results in an improvement of 4.2 mm in joint position errors. This enhancement is expected, as a larger window size provides more context, aiding disambiguation. When egocentric videos are included, further improvements are observed. Notably, using 60 frames with egocentric video outperforms using 120 frames alone, suggesting that the context provided by egocentric videos is more effective than simply increasing the window size.

Motion PQ-VAE. Ablation studies on motion PQ-VAE are reported in Tab. 5. "PQ" indicates whether product quantization is used. "CB" denotes the number of codebook entries. The first two lines indicate that significant improvements can be achieved with product quantization. Additionally, increasing the number of codes and reducing code dimensions yields further enhancements.

Language Model Size. We use GPT-2 Medium (345M) for most of our experiments to maintain efficiency. To further assess the potential of EgoLM in scaling up to larger LMs, we train with GPT-2 Large (1.5B) and report performance on TPV2T in Tab. 6. The improved scores indicate EgoLM is a scalable and versatile framework.

Motion-Augmented Next-Token Prediction. To justify the usage of motion regression loss $\mathcal{L}_{reg}$ in next-token prediction training, as introduced in Sec. 3.3, we report the performance of not using $\mathcal{L}_{reg}$ in three-points motion tracking in Tab. 7. The improved score indicates the necessity and effectiveness of such design.

4.5. More Applications

Text-to-Motion Generation. As part of our joint training, EgoLM is capable of generating motions from texts, as shown in Fig. 8 a). Even with lengthy prompts describing the upper and lower body separately, our model successfully generates motions that align with the inputs.

Motion Prediction. As a by-product of the motion pretraining, EgoLM can function as a motion predictor. As shown in Fig. 8 b), given motion prompts (the red skeleton in the left), subsequent motions can be randomly sampled. We show three different samples in different colors.

5. Discussion

We propose EgoLM, an egocentric motion generalist model, that empowers egocentric motion understanding using LLMs. To address the challenge of limited wearer observation in egocentric perception, EgoLM integrates two complementary modalities to disambiguate the under-constrained scenarios. We also introduce multi-modal multi-task joint training to bridge gaps between different modalities and tasks, thereby implicitly connecting them and improving individual task performance. We hope our exploration of the fusion between egocentric perception and LLMs will inspire future research in contextual AI.

Limitations. Firstly, the motion tokenizer introduces reconstruction errors and bounds motion tracking performance. Secondly, for motion narration, each video frame is compressed by the CLIP encoder into a one-dimensional vector, making it difficult for the model to accurately identify the objects the person is interacting with. Furthermore, as commonly observed in language models [19], EgoLM also experience the hallucination problem.

Potential Societal Impact. While contextual AI presents opportunities for societal advancement, the collection and analysis of human data may raise privacy concerns.

Acknowledgements We thank the Surreal team, especially Yifeng Jiang, and Renzo De Nardi, for their valuable discussions and help in the project. This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221-0012, MOE-T2EP20223-0002), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 5

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 3

[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. *arXiv preprint arXiv:2304.11118*, 2023. 2, 6

[4] 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

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 3

[8] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 4

[9] 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

[10] Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, and Huijuan Xu. Stratified avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 153–163, 2024. 2

[11] 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

[12] 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

[13] 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 (CVPR)*, pages 5152–5161, 2022. 3

[14] 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, 6

[15] 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

[16] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500. IEEE, 2022. 3

[17] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 3

[18] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1): 117–128, 2010. 3

[19] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 8

[20] 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

[21] 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, 2024. 2, 3, 6

[22] 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. 2, 6

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

[24] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video

database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 3

[25] 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. 2, 6

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

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 3

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3, 5

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[30] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, pages 417–435. Springer, 2022. 3

[31] 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. 2, 5

[32] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2

[33] Meta. Ray-ban meta smart glasses. https://www.meta.com/smart-glasses, 2024. Accessed: 2024-09-30. 1

[34] Nicholas Milef, Shinjiro Sueda, and N Khademi Kalantari. Variational pose prediction with dynamic sample selection from sparse tracking signals. In *Computer Graphics Forum*, pages 359–369. Wiley Online Library, 2023. 2

[35] 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. 2

[36] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 3

[37] OpenAI. Introducing chatgpt, 2022. 3, 5

[38] 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

[39] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[40] 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

[41] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *arXiv preprint arXiv:2308.07123*, 2023. 1

[42] Jose Luis Ponton, Haoran Yun, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1):1–14, 2023. 2

[43] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2021. 3

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3, 4, 6

[45] 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. 3, 5

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 6

[47] Daniel Roetenberg, Henk Luinge, and Per Slycke. Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. *Xsens Motion Technol. BV Tech. Rep.*, 3, 2009. 5

[48] Snap. Spectacles. https://www.spectacles.com/, 2024. Accessed: 2024-09-30. 1

[49] Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 5

[50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[51] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[52] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 2

[53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al.

Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[54] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[55] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 3

[56] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. *arXiv preprint arXiv:2401.00789*, 2024. 3

[57] 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

[58] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[59] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. Egolife: Towards egocentric life assistant. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[60] 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. *arXiv preprint arXiv:2305.01599*, 2023. 2

[61] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3

[62] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023. 2, 3

[63] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023.

[64] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[65] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *European Conference on Computer Vision*, pages 397–421. Springer, 2024. 3

[66] 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

[67] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. *arXiv preprint arXiv:2306.10900*, 2023. 3

[68] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 5

[69] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14678–14688, 2023. 2

[70] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding, planning, generation and beyond. *arXiv preprint arXiv:2311.16468*, 2023. 3