EgoLM: Multi-Modal Language Model of Egocentric Motions –*Supplementary Material*

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

We provide more implementation details (Sec. 1) and qualitative results (Sec. 2) in this supplementary material. We also provide more discussion on the difference with other related works.

1. Implementation Details

1.1. Instruction Templates

Below, we show the instruction templates for motion tracking and motion narration tasks. Motions are encoded to tokens and filled in *<Motion_Placeholder>*. *<Narration_Placeholder>* is the placeholder for corresponding motion narration. The encoded three-points 6-DoF poses features are placed at *<TP_Placeholder>*. The placeholder for egocentric video features is *<CLIP_Placeholder>*.

Task: *Motion Tracking*
Instruction: *Perform motion tracking based on the given three-points and CLIP embeddings.*
Input: *Input CLIP embeddings:* `<CLIP_Placeholder>`. *Input three-points feature:* `<TP_Placeholder>`
Output: `<Motion_Placeholder>`

Task: *Motion Narration*
Instruction: *Describe the human motion based on the given three-points and CLIP embeddings.*
Input: *Input CLIP embeddings:* `<CLIP_Placeholder>`. *Input three-points feature:* `<TP_Placeholder>`
Output: `<Narration_Placeholder>`

1.2. Auto-regressive Inference for Motion Tracking

At inference time, motion understanding is the same as the language model inference. For motion tracking, it usually requires online inference over a long period. With a language model, which is an auto-regressive model, it is straight-forward to perform online motion tracking. As shown in Fig. 1, firstly, an initialization over the first t frames of data is required. When the new data frame $t+1$ comes in, the input conditions are updated accordingly. Then, it is not necessary to predict all the motion tokens from frame 2 to frame $t+1$. We take the previously generated motion tokens from frame 2 to frame t as inputs and prompt the network to generate one more token for frame $t+1$.

1.3. Evaluation Metrics

For motion tracking, we use joint position errors and joint angle errors to evaluate the performance. Specifically, for the joint position errors, we first align ground truth skeletons and generated skeletons by the head positions only by translation. Then full body, upper body and lower body joint position errors are calculated separately. Joint angle errors are calculated on full body and root joints. For the evaluation of motion VQ-VAE in main paper Tab. 4, we apply widely adopted metrics for motion regression, *i.e.*, Mean Per-Joint Position Error (MPJPE) [3], Procrustes-Aligned (PA-)MPJPE [6], and joint position acceleration (ACCL) error. For the motion understanding, we use standard NLP metrics, please kindly refer to corresponding papers for more details.

2. More Qualitative Results

2.1. Three-Points Motion Tracking

We show four more visual examples of three-points motion tracking in Fig. 2 and Fig. 3. AvatarPoser [5] and BoDiffusion [1] are solid baselines that perform well on easy walking cases, *e.g.*, upper example in Fig. 2. For the workout sequence, *i.e.*, lower example in Fig. 3, even only given three points of upper body, the distribution of lower body motion can be collapsed and generate reasonable motions that matches the ground truth. In Fig. 3, we demonstrate the effectiveness of including egocentric videos as inputs. Without any environment context, AvatarPoser and BoDiffusion often fail to distinguish standing and sitting down. We do not assume the knowledge of the head height over the floor, meaning that the three-points positions are normalized to the local coordinates of the first frame. Therefore,

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

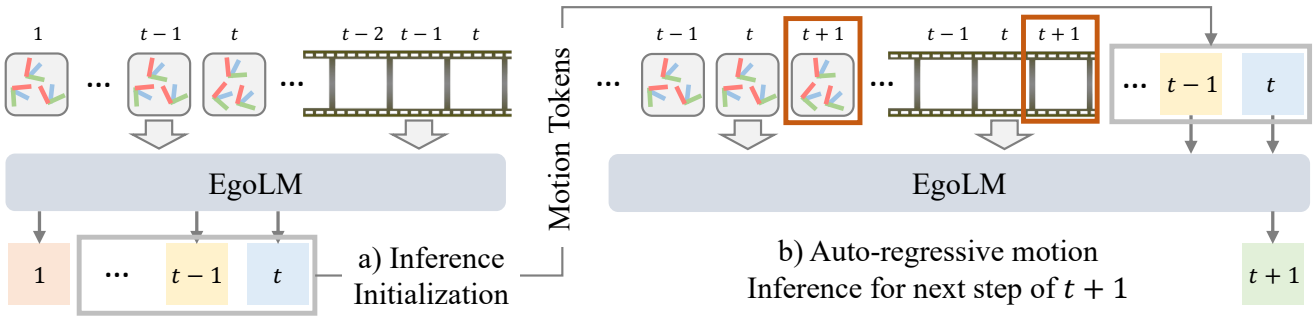


Figure 1. **Online Motion Tracking Inference.** For the new time step of $t+1$ with new data coming in, last motion tokens are combined with the new input tokens to decode the next motion token $t+1$.

it is hard for baseline methods to disambiguate certain scenarios. We propose to introduce contexts using egocentric videos, which contains rich information about the environment and how the person is interacting with it. Therefore, our model can generate the most accurate motions by utilizing these information. For more visualization of three-points motion tracking, please kindly refer to our supplementary videos.

2.2. One-Point Motion Tracking

We show four more examples of one-point motion tracking in Fig. 4 and Fig. 5. The introduction of egocentric videos has two advantages. Firstly, similar to the case in three-points body tracking, the environment contexts in egocentric videos can disambiguate cases like standing and sitting. Secondly, specifically for one-point motion tracking, egocentric videos provide clues of hand positions. As shown in all four examples, when the person raises the arms in front of the body, hands would be visible in the egocentric videos, which helps the hand position tracking. Admittedly, high-level semantic information provided by CLIP [7] encoders cannot accurately track hand positions. Therefore, as shown in the lower example in Fig. 4, our method correctly generates arms moving in the air, but lacks accuracy. For more visual examples of one-point motion tracking, please kindly refer to our supplementary video.

2.2.1. Multiple Samples.

Note that EgoLM is essentially a generative model. Therefore, our model is capable of generating different samples with the same inputs. In Fig. 6, we show three random samplings on the same input one-point and egocentric video. When hands are not visible in the frame, *i.e.*, the left highlighted frame, hand positions are not constrained, and therefore shows high diversity across different samples. For the other highlighted frames, hands are visible in the egocentric videos, which helps to collapse the distribution of possible positions of hands. But as discussed above, our way of encoding egocentric videos cannot accurately track the hand positions. Therefore, our model also shows some diversity of hand positions in these cases.

To further demonstrate the diversity of our model, we also show three random samples from our one-point motion tracking model that does not take egocentric videos as inputs in Fig. 7. Lack of any indication of the hand positions, the upper body generation is even less constrained than that of the lower body and shows high diversity across three samples.

2.3. Motion Narration

We show eight more examples of motion narration in Fig. 8 and Fig. 9. Similar to the main paper, we use green to highlight correct parts in the answers and red for wrong answers. Similar to the observation made in the main paper, even though TM2T [2] and MotionGPT [4] have access to the full body motion, the generated narrations are reasonable but completely wrong if consider the environment context. For example, in the upper right example in Fig. 9, given the simple walking sequence, both TM2T and MotionGPT can correctly understanding that the person is walking forward. But they all give the wrong answers about the places the person is walking in. Thanks to the egocentric videos, our model successfully produces the correct description as "walking towards the beds".

2.4. Motion Prediction

As a by-product of the second stage of our training pipeline, motion pre-training, we build a motion prediction network. Given leading motions as the prompts, our model is capable of auto-regressively sample motions that complete the motion prompts. As shown in Fig. 10, the first three samples show three different samples given the same motion prompt. We can increase the intensity of the generated motions by increasing the temperature. The last three samples show three random samples given various motion prompts, *e.g.*, bending forward, sitting down and standing.

3. Discussion

We summarized our differences with related works in the main paper Tab. 1. Below, we emphasize three key aspects that differentiate it from previous works, *e.g.*, MotionGPT

and T2M-GPT. a) Novel network architecture. EgoLM is the first work to show the effectiveness of PQ-VAE motion tokenizer and decoder-only LLM for motion-language learning. Both architecture designs contribute to the performance of downstream tasks (*c.f.* Tab. 3 and 5). In contrast, MotionGPT and T2M-GPT both use vanilla motion VQ-VAE and encoder-decoder transformers. In addition, our work also demonstrates how raw sensor data (*e.g.,* video and device motion) can drive multi-modal multi-task instruction tuning by learning the feature projections instead of tokenizing the input. While MotionGPT and T2M-GPT only involve tokenized motion and languages. b) Efficient motion pre-training. MotionGPT requires motion-text pairs for pre-training, while EgoLM only uses motions, which are easier to collect. T2M-GPT performs supervised training from scratch, leading to sub-optimal motion and language distribution learning. Moreover, we propose novel motion augmented pre-training which improves the motion distribution (*c.f.* Tab. 7). c) Different downstream motion tasks. We formulate different tasks than MotionGPT or T2M-GPT, *e.g.*, egocentric motion tracking and understanding, that involves three-point/one-point device motion and egocentric videos. While MotionGPT and T2M-GPT focus on text-to-motion and motion-to-text generation.

References

[1] 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. 1

[2] 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. 2

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1

[4] 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

[5] 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

[6] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[7] 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. 2

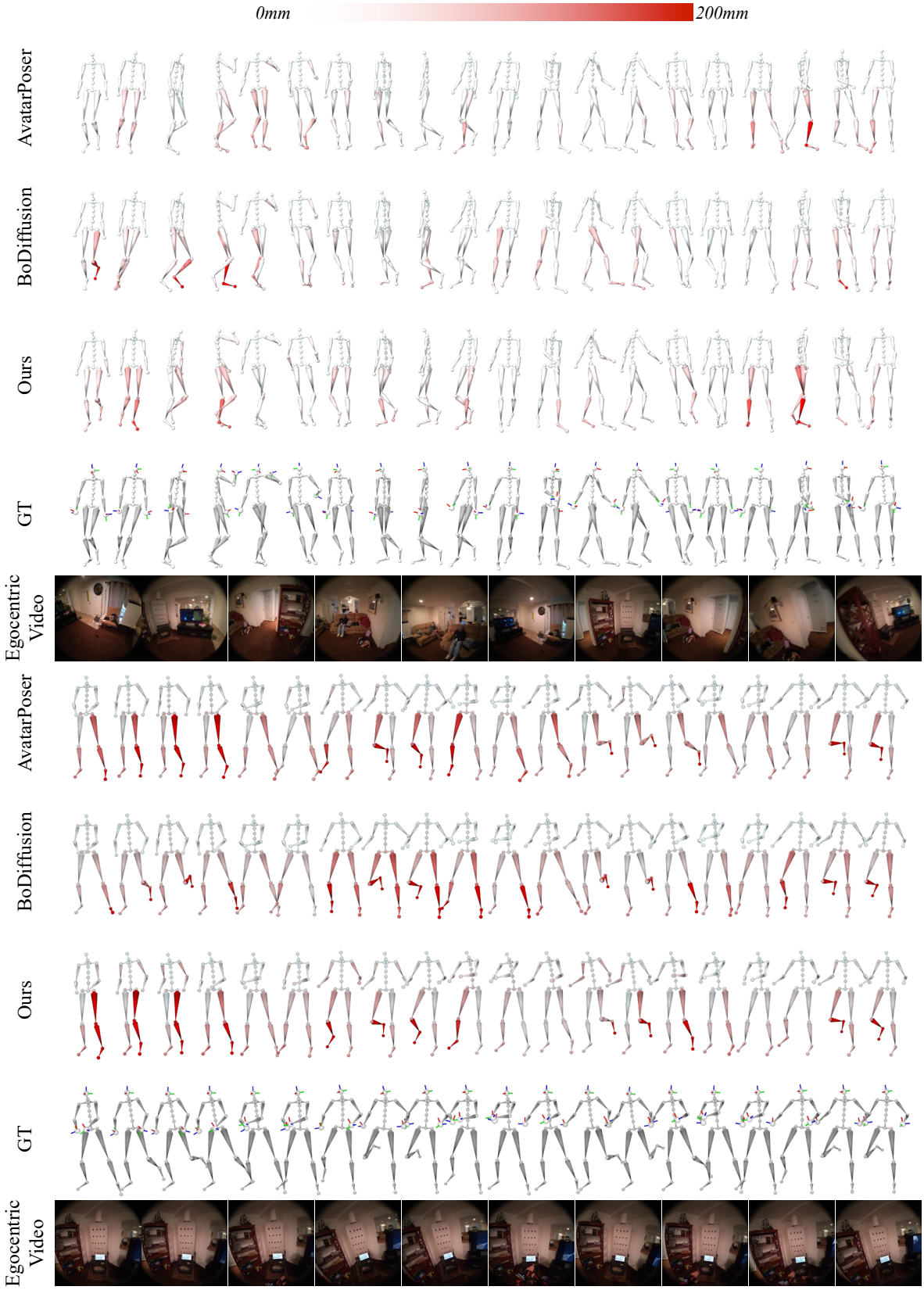


Figure 2. **Qualitative Results of Three-Points Motion Tracking.** Skeletons are color-coded by joint position errors.

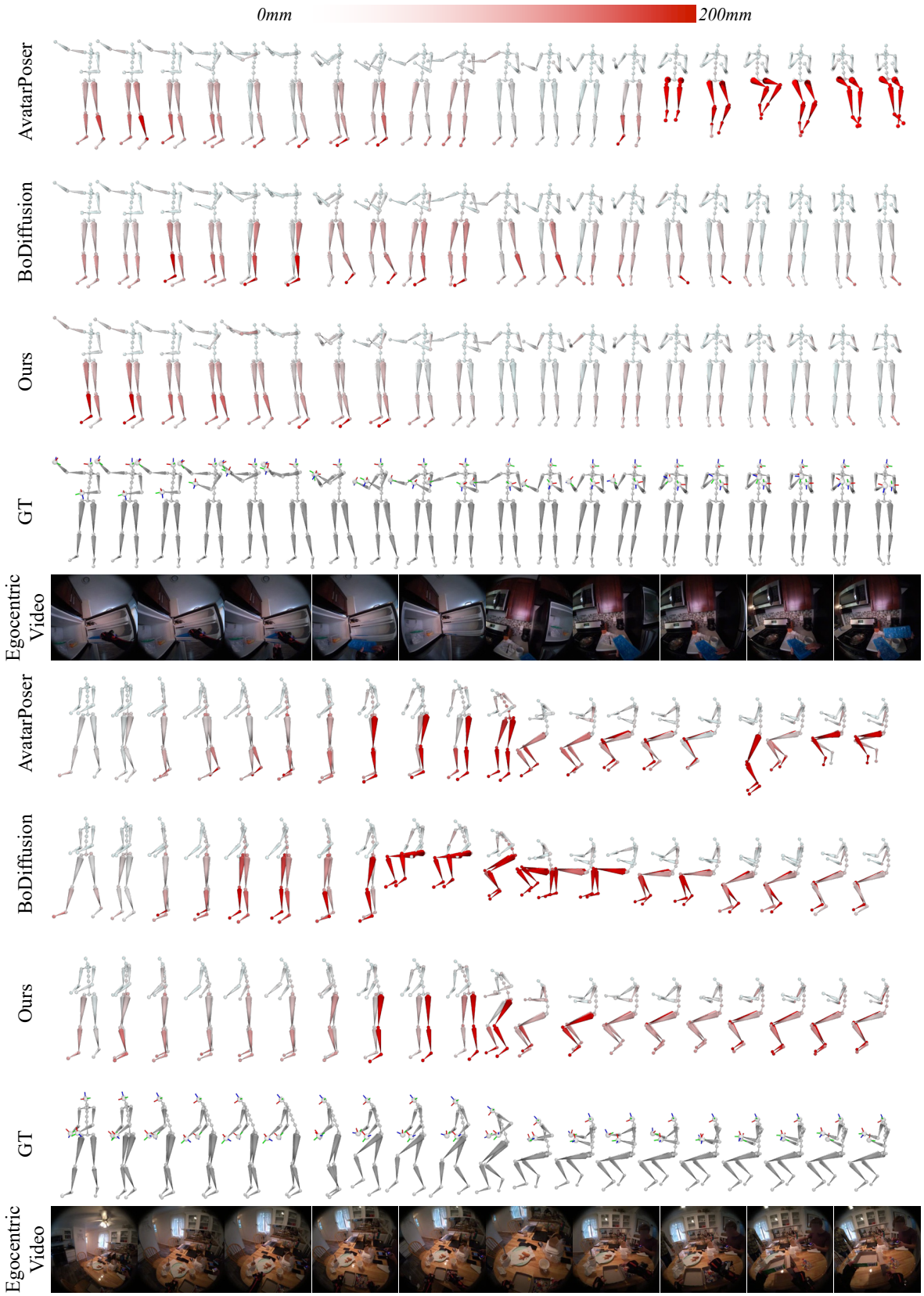


Figure 3. **Qualitative Results of Three-Points Motion Tracking.** Skeletons are color-coded by joint position errors.

Figure 4. **Qualitative Results of One-Point Motion Tracking.** Skeletons are color-coded by joint position errors.

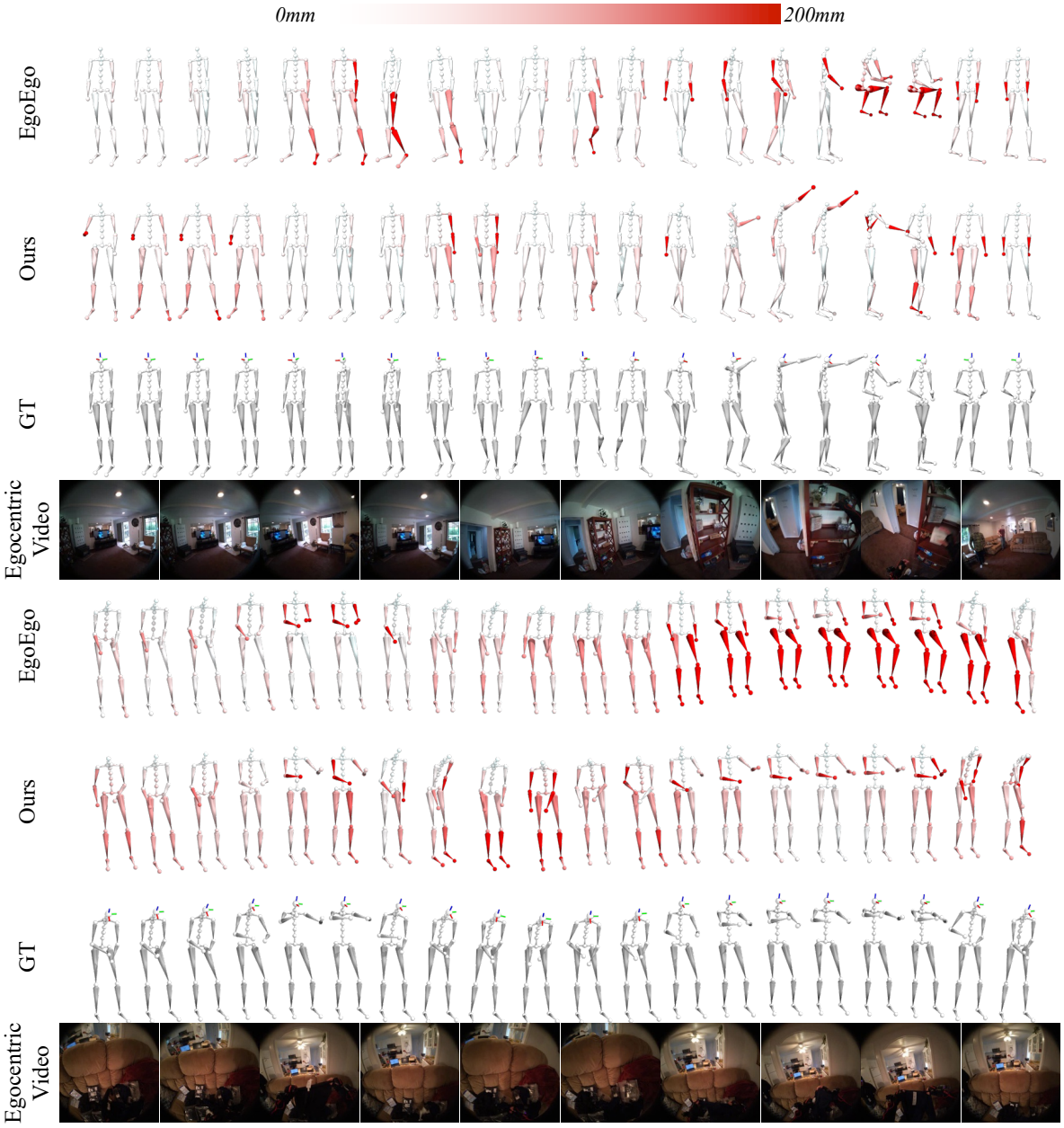


Figure 5. **Qualitative Results of One-Point Motion Tracking.** Skeletons are color-coded by joint position errors.

Figure 6. **Three Random Samples of One-Point Motion Tracking with Egocentric Videos as Inputs.** Since we use language models as our backbone, EgoLM has the ability to randomly sample outputs given the same inputs. Egocentric videos provide strong clues for hand positions, leading to less diversity as shown in the highlighted areas.

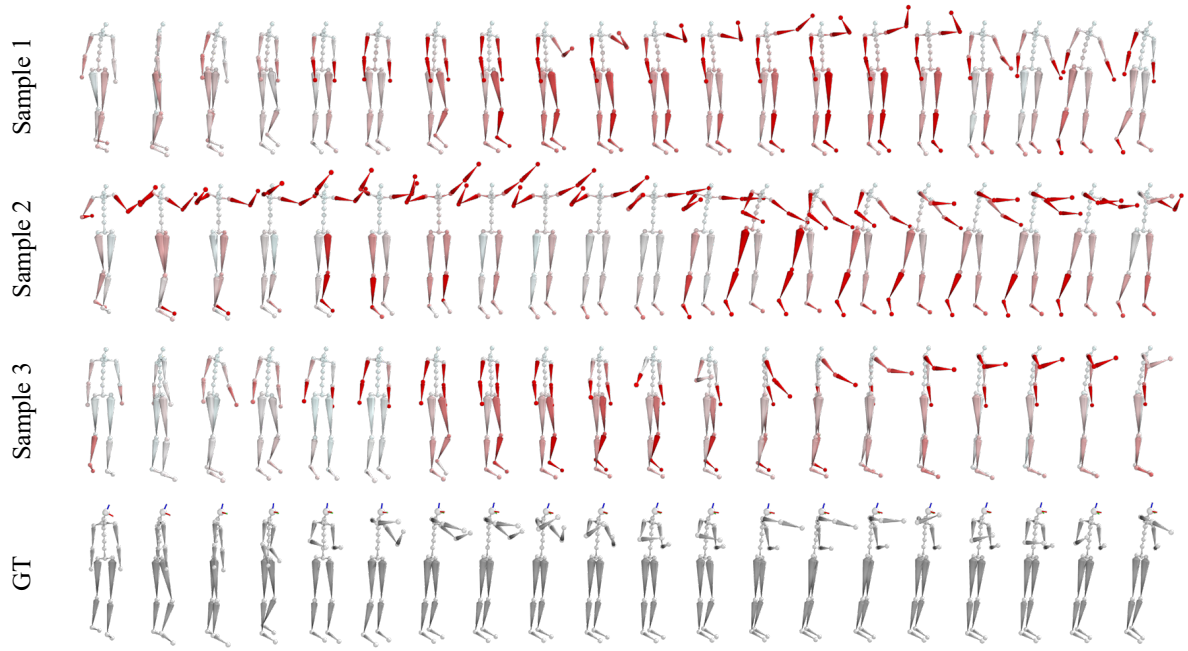
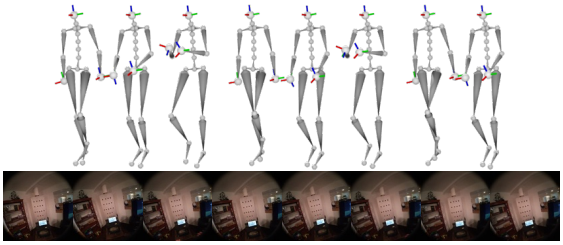
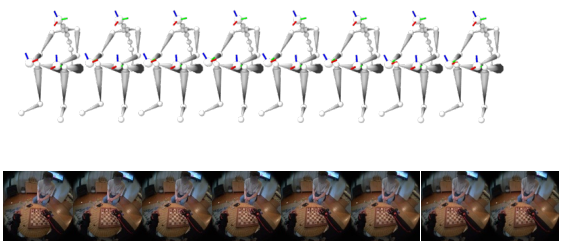


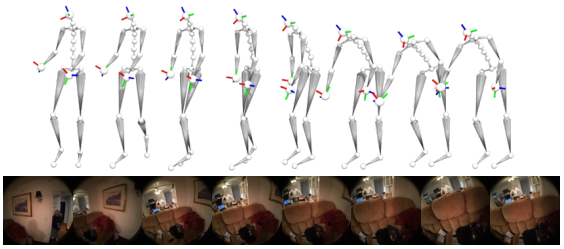
Figure 7. **Three Random Samples of One-Point Motion Tracking without Egocentric Videos as Inputs.** With only head poses as inputs, the generation of full body motion, especially upper body motions, is less constrained.



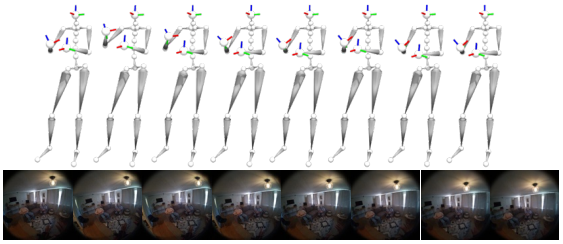
TM2T: *The person still marches in place while facing his peers. The person still swings both of his hands up and down. The person still marches in place with his left foot and right foot alternately. The person still repeatedly bends both of his legs alternately. The person still marches in place with his left foot and right foot alternately.*
MotionGPT: *The human swings his body to the right and swings back to the left while standing, hunching his back and doing some exercise in the living area with his colleagues. The human slightly swings both of his arms back and forth on his side. The human raises his right leg to his waist level then stretches and lowers it while his left foot is fixed on the floor.*
V2T: *The human is standing in the living room while watching the television. The person is resting both arms on his sides. The human has both feet fixed on the floor.*
Ours: *The person is swaying her body side to side while exercising in the living area. The person repeatedly swings and bends both of her arms in front of her then lowers it down on her side. The person repeatedly raises both of her feet in front of her then lowers them down on the floor alternately.*
GT: *The person is walking in place in front of the laptop. The human repeatedly bends both of her arm in front of her them lowers them down on her side. The human repeatedly steps both of her feet alternately.*



TM2T: *The person is sitting on a chair and leaning backward on the table while talking to her peers. The person is resting both of her arms on the table, lifts and bends her left arm as she touches the table with her left hand. The person is sitting with both legs bent and with both feet flat on the floor widely apart.*
MotionGPT: *The person is still sitting on the chair with a hunched back while playing arcade and eating some chips. The person's both arms are bent forward while holding and sliding the joystick with his left hand to the left then his right hand is on top of the buttons and clicks them with his right fingers. The person's both legs are still bent while sitting on the chair with both feet flat on the floor and slightly apart.*
V2T: *The human is sitting on the sofa and leaning forward while arranging the chess pieces on the chessboard. The person has both of her arms extended forward while picking up the chess pieces with her left hand and puts down the chess piece with her right hand on the chess board. The human is sitting with both feet fixed on the floor and shoulder-width apart.*
Ours: *The person is sitting in front of the checkerboard. The person is extending his right arm toward the checkerboard while keeping his left arm on top of his leg. The human is bending both of his knees while keeping both of his feet flat on the floor.*
GT: *The human is sitting in front of the table as he plays chess. The person is moving the knight with his right hand while his left hand remains resting on his leg. The human is bending both of his knees while keeping both of his feet flat on the floor.*

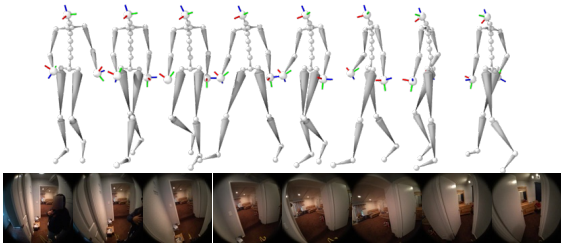


TM2T: *The person walks towards the cabinet then bends forward to pick up and reach for the clothes. The person extends his right arm to pick up the clothes from the cabinet then bends his left arm to hold the clothes.*
MotionGPT: *The person bends forward while standing in the living room. The person extends her right arm to open the cabinet and extends her left arm to grab the keys on the right. The person slightly bends both of her legs then steps her right foot forward while her left foot is fixed on the floor.*
V2T: *The human walks towards the couch and bends down while putting down the piece of clothing. The person extends both of her arms to pick up and put down the piece of clothing with her right hand while holding the clothes with her left hand. The human steps both of her feet forward alternately.*
Ours: *The human walks towards the sofa then slightly leans forward to put down the folded piece of clothing. The person extends her right arm to put down the folded piece of clothing on the sofa, then extends her left arm to pick up another piece of clothing on the sofa. The human is stepping both of her feet forward alternately then bends both of her legs to support her body.*
GT: *The person bends his body to get another clothes on the sofa. The person extends his right arm to get the clothes with his right hand then raises his left arm to hold the clothes with his left hand. The person steps both feet forward towards the sofa.*

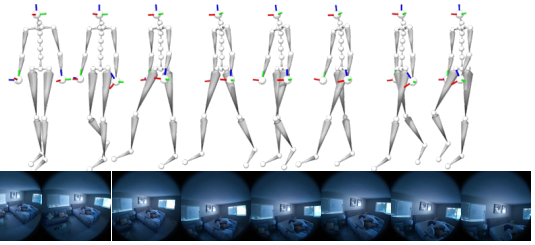


TM2T: *The person is still standing straight in front of the table while playing the board game with his peer. The person's both arms are still bent forward while both hands are still holding the edge of the knife.*
MotionGPT: *The human still stands near the closet. the human still holds the hanger with his left hand and his right hand holding the hanger. The person still stands with his feet slightly apart.*
V2T: *The person is standing straight in the living area with his colleagues while doing some exercise. The person raises both of his arms straight above his head from the back then lowers them in front and rests them on his side. The person is standing with both feet apart and fixed on the floor.*
Ours: *The person is standing in the living area. The human repeatedly swings both of his arms in front of him and in front of his stomach. The person is standing with both feet fixed on the floor.*
GT: *The human is standing still in front of his colleagues in the living room while playing charades. The person is slightly raising and lowering both of his arms to gesticulate. The human is resting both of his legs fixed to the floor.*

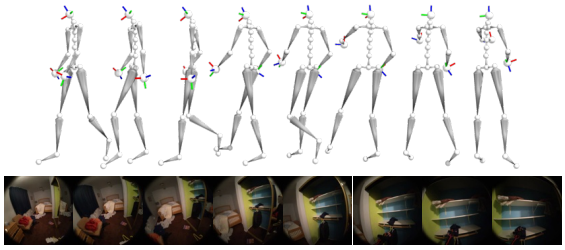
Figure 8. **Qualitative Results of Motion Narration.** We use green to highlight correct parts in the answers while red for wrong ones.



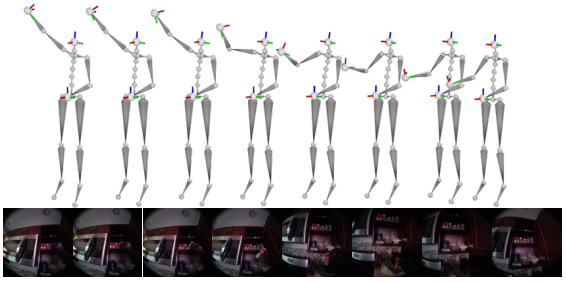
TM2T: *The person walks out of the bedroom then turns to the left to enter another bedroom. The person rests both her arms on her sides. the person is stepping forward with her right and left legs alternately. The person is stepping forward with her right and left legs alternately. The person then steps forward with her right feet.*
MotionGPT: *The person walks in the garage. the person sways his hands on the sides. The human extends both legs forward alternately.*
V2T: *The human walks towards the bedroom. The human slightly sways her hands on her sides. The human takes four steps towards the bedroom.*
Ours: *The human walks towards the door. The human puts down her right arm and sways both hands on the side. The person extends both legs forward alternately.*
GT: *The person walks towards the door. The person walks towards the door. The person rests his left arm on the side and he raises his right arm while holding the hanger with his right hand. The human extends both legs forward alternately.*



TM2T: *The person is walking forward in the pathway then she slightly leans forward as she sits on the pathway. The person alternately swings both hands on her sides while both arms hang naturally at her sides.*
MotionGPT: *The human is walking forward while looking at the office surrounding. The human has her both arms swaying them back and forth. The human extends both legs forward alternately.*
V2T: *The person is walking forward towards the bed. the person rests both arms on her sides. The person is extending both her legs forward alternately.*
Ours: *The human is walking towards the bed. The person is resting both of her arms beside her. The person is extending both of her legs forward alternately.*
GT: *The person walks towards the bed. The person slightly swings both of her arms back and forth. The person steps both of her legs forward alternately.*



TM2T: *The person walks towards the door then leans forward as he tucks in the chair and stands in front of the door to open it. The person's right arm is swinging back and forth on his side while his left arm is bent and his left hand holding the top railings then pushes the door open with his left hand.*
MotionGPT: *The person is walking forward towards the shower room, pauses on the shower room and then leans forward to put down the towel on the shower curtain holder. The person is bending both of his arms and then extends his left arm forward to put down the towel on the shower curtain holder. The person is alternately stepping both of his feet forward.*
V2T: *The person straightens up as she slightly turns to the left while walking towards the closet. The person keeps holding the clothes with her bent left arm as she lowers down and slightly raises her right arm and then she bends it back. The person steps both of her feet forward alternately.*
Ours: *The human turns clockwise as she walks towards the closet to put the clothes on the top shelf in the bedroom. The human is holding the clothes hanger with both of her bent arms in front of her then she extends her left arm froward and grabs the clothes hanger with her left hand. The human turns her right foot to the right, steps her left foot forward then slightly moves her right foot forward.*
GT: *The human walks towards the closet. The human raises his left arm to grab the clothes while he holds the hanger with his right hand. The person extends both legs forward alternately.*



TM2T: *The person is standing in front of the door. the person is raising his left arm and is resting his right arm on his side. The person bends both of his legs while resting on the floor.*
MotionGPT: *The person stands in the bedroom while talking to her colleague. The human is resting and bending her left arm in front while she lowers down her right hand before touching the wall with her right hand. The person stands with both feet fixed on the floor.*
V2T: *The human is standing straight while picking a condiment jar in the hanging cabinet. The human grabs a condiment jar with her right hand and flips up the other condiment jar in front of her with right hand and then she bends and slightly lowers down her right arm. The person is standing with both feet fixed on the ground.*
Ours: *The person is standing in front of the hanging cabinet and slightly leaning forward while picking up a condiment jar. The person is extending her right hand forward, picks up the condiment jar cover then puts it down again on the top of the hanging cabinet while resting her left arm on her side. The human is standing with both of her legs parallel to each other and both of her feet spread slightly apart.*
GT: *The person is standing on tiptoes while checking inside the cupboard. The human grabs and places the bottle down on the countertop with her right hand while her left hand is resting on the countertop. The human is standing on tiptoes with both feet as she reaches inside the cupboard.*

Figure 9. **Qualitative Results of Motion Narration.** We use green to highlight correct parts in the answers while red for wrong ones.

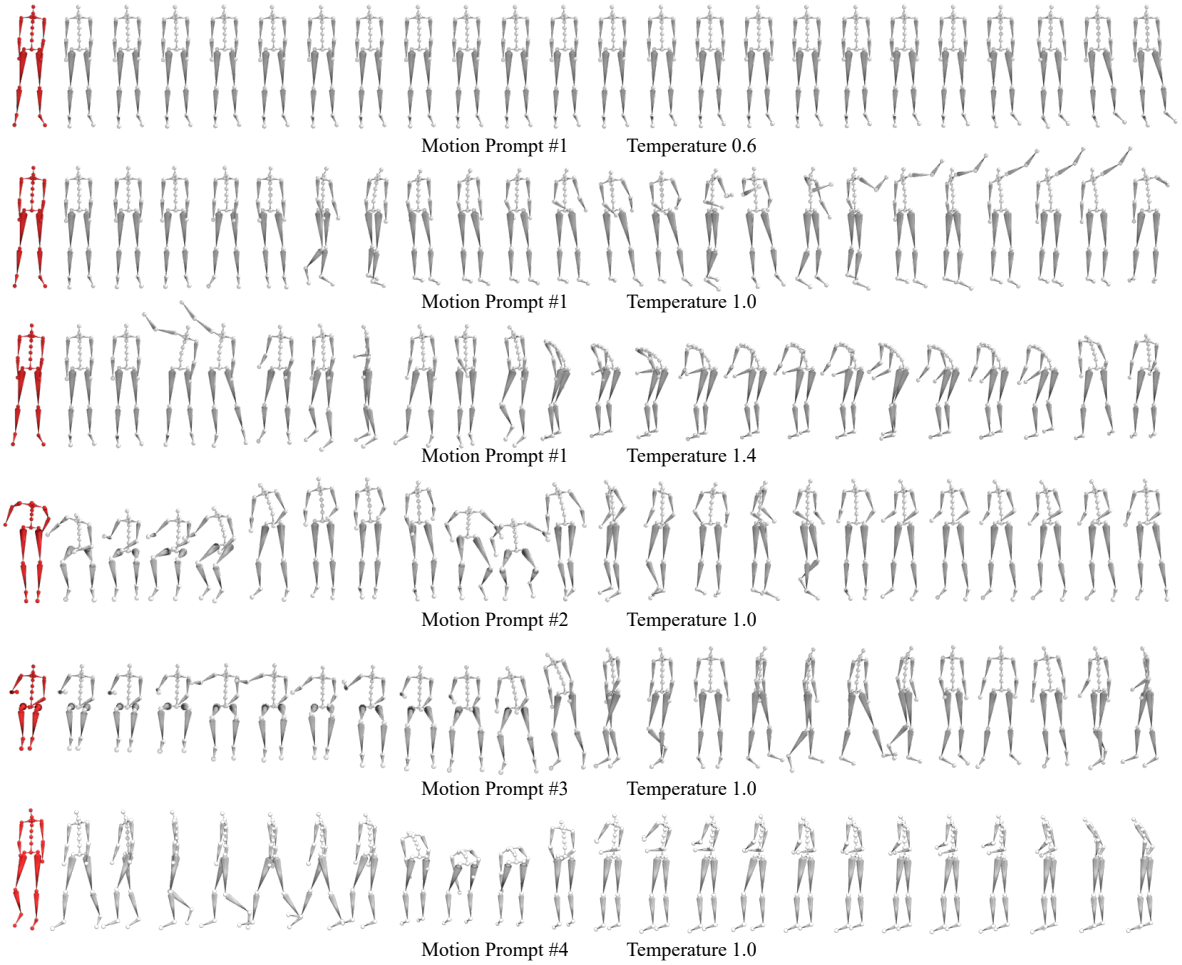


Figure 10. **Qualitative Results of Motion Prediction.** The first skeletons in red are input motion prompts. The following motions are randomly sampled auto-regressively from our motion pre-training network.