Supplementary materials for Head2Body: Body pose generation from Multi-sensory Head-mounted Inputs

Minh Tran[1][*], Hongda Mao[2], Qingshuang Chen[2], Yelin Kim[2]
[1]Department of Computer Science, University of Southern California, Los Angeles, CA, USA
[2]Amazon, Santa Cruz, CA, USA
minhntra@usc.edu

1. Head-IMU pretraining

The IMU encoder in our framework is pre-trained using a self-supervised approach (BestRQ) [1] on a large-scale dataset of head-IMU signals from the Ego4D dataset [2], consisting of over 1700 hours of unlabeled motion data. The goal of pre-training is to capture temporal dynamics and motion patterns in head-IMU signals, enabling the encoder to generalize well across downstream tasks with limited labeled data.

Data Preparation The IMU signals are down-sampled from 800 Hz to 200 Hz, and normalized to ensure consistency across devices. Each input sample consists of 4-second windows of 6-DoF motion data, capturing both accelerometer and gyroscope readings. We do not apply any augmentation to the IMU signals.

Architecture and Pre-Training Objective The IMU encoder is based on a Conformer model [3] with six layers, each comprising eight attention heads and a hidden dimension of 512 units. To process the raw 200 Hz IMU signals, we developed a custom feature extractor that downsamples the input to 25 Hz while producing feature representations compatible with the encoder's hidden size. The feature extractor consists of a GroupNorm layer, followed by three 1D convolutional layers, and a final GroupNorm layer. The pre-training objective employs a masked motion prediction strategy, where 15% of the input sequences are randomly selected for masking, along with their subsequent four frames, resulting in approximately 60% of the input data being masked. To produce the pseudo-labels for SSL, we utilize a Random VQ module comprising 8192 codebooks, each with a codebook size of 16. The model is trained to reconstruct the missing segments using contextual information from the unmasked portions of the sequence, as detailed in Section 3.1.

Training Configuration Pre-training is conducted on 8 L40S GPUs with a batch size of 16 for $400K$ epochs. The learning rate is set at $8e^{-4}$, using a linear warmup for $25K$ steps, and weight decay is set to 0.05. Gradient clipping with a maximum norm of 0.5 is applied to ensure stable training. The total pre-training process takes approximately 3 days.

	AMASS [10]	Kinpoly [9]	Exo-ego4d [2]
# samples	54K	266	1.5K
total duration	60hr	1.3hr	13hr
modality	Hp	Hp+V+I	Hp+V+I
environment	indoor/outdoor	indoor	indoor/outdoor
annotations	SMPL [8]	SMPL [8]	3D COCO [7]

Table 1. Details of the datasets used in this work.

2. Implementation Details

All models are trained using the AdamW optimizer with a learning rate of $1e^{-4}$ and a weight decay of $1e^{-5}$. Training spans 100 epochs with a batch size of 32. The vision encoder utilizes a ResNet-50 [4] backbone pre-trained on ImageNet and fine-tuned during training. For the Transformer model responsible for discrete motion token prediction, we employ a 6-layer architecture (3 in encoder and 3 in decoder) with 8 attention heads and a hidden dimension of 512 units. Positional encodings are added to the input tokens to preserve temporal ordering. The Vector Quantization (VQ) module comprises 512 codebooks, each containing 512 entries. During training, the discrete motion tokens are predicted using teacher-forcing, while inference employs an auto-regressive decoding strategy. To process egocentric video frames, we use a frame resolution of 512×512, downscaled from the original resolution using bicubic interpolation. Data augmentations are not applied. For IMU data integration, input sequences are synchronized with video frames at 25 Hz.

3. Datasets

The datasets used in this work vary in size, modalities, and environmental diversity, as summarized in Table 1. For the splittings, we follow the setup provided in *EgoEgo* [6]. For the EgoExo4D dataset, we follow the official Body Pose estimation challenging training and evaluation settings.

[*]This work was done during an internship at Amazon.

	Time (ms)	GFLOPS
Egoego [6]	651	653
VQ-Poser (ours)	96.4	9.15

Table 2. Performance analysis of Body Pose Generation module for processing 1-second (30 frames) of data, with times reported in milliseconds (ms).

	Time (ms)	GFLOPS
NExT-Chat	2.34E+04	1.46E+06
HeadNet	83.5	1.02
IMU-encoder	1.54	2.48
Image-encoder	35.0	109

Table 3. Performance analysis of Feature Extraction modules for processing 1-second (30 frames) of data, with times reported in milliseconds (ms).

- **AMASS** is a large-scale dataset comprising 54,000 samples, with a total duration of approximately 60 hours. It focuses on head pose (Hp) data captured in various environments. Ground-truth annotations are provided in the form of SMPL body parameters, covering a wide range of pose types. Because the dataset does not come with real head-IMU data, we extract the IMU information following the pipeline in *IMUGPT* [5].
- **KinPoly** consists of 266 samples with a total duration of 1.3 hours. This dataset includes head pose (Hp), vision (V), and IMU (I) data captured in indoor settings. The annotations are provided using SMPL body parameters and are specific to five defined actions.
- **Exo-Ego4D** contains 1,500 samples spanning approximately 13 hours of recording. Like KinPoly, it features head pose, vision, and IMU data (Hp+V+I), but the environment varies across different scenarios. Annotations include 3D body poses, and the dataset covers procedural (*e.g.,* cooking) and physical (*e.g.,* basketball) pose types.

This diversity in datasets enables comprehensive evaluation of our approach across different modalities, environments, and pose types.

4. Runtime Analysis

Tables 2 and 3 present a comprehensive analysis of the computational efficiency of various components in our framework, including the base models and feature extraction (FE) modules. The performance is evaluated in terms of frames per second (FPS) and giga floating-point operations per second (GFLOPS) for processing 1-second (30 frames) of input data. The results highlight that VQ-Poser effectively balances computational efficiency and performance, demonstrating superior resource usage compared to the diffusion-based EgoEgo model [6]. However, a significant performance bottleneck is observed in the vision-language segmentation mask extraction process, which imposes a substantial computational burden. This bottleneck overshadows the inference efficiency gains introduced by VQ-Poser.

References

[1] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2022. 1

[2] 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. 1

[3] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *Interspeech 2020*, 2020. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Zikang Leng, Hyeokhyen Kwon, and Thomas Plötz. Generating virtual on-body accelerometer data from virtual textual descriptions for human activity recognition. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*, pages 39–43, 2023. 2

[6] 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

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[8] 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. 1

[9] 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

[10] 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. 1