Masked Modeling for Human Motion Recovery Under Occlusions

Supplementary Material

In this supplementary material, we provide further details on our training and inference strategies, as well as the evaluation setups for WHAM [14] and GVHMR [13]. We then present additional quantitative results on the EgoBody-Occ dataset to factor out the influence of our random cropping augmentation. In addition, please refer to our Project Website for more video results and visual comparisons of our method against the baselines.

A. Training Details

A.1. SMPL-X Body Model.

Our training pipeline integrates motion, image, and video datasets, which requires a unified parametric body representation across all data sources. We adopt the SMPL-X model [11] as our common annotation format due to its expressiveness in modeling full-body human meshes. For image datasets [1, 6, 8, 10] that provide only SMPL [9] annotations, we convert SMPL parameters to SMPL-X using a fast model-fitting algorithm [15].

Since our tokenizer operates directly on mesh inputs via mesh convolution layers, we remove the two disconnected eyeball meshes from SMPL-X to enforce a fully connected mesh topology. This modified SMPL-X mesh is used consistently for tokenization and model training.

A.2. Motion Pretraining.

The motion encoder takes masked local pose token $\bar{\boldsymbol{Z}}$ and noisy canonical trajectory $\bar{\boldsymbol{R}}_{\text{cano}}$ as input and reconstructs the global motion in a canonicalized coordinate. To compute the canonical trajectory from any given sequence, we offset each frame by inverse rigid transformation with the orientation and translation of the first frame.

As described in Sec. 4.2 of the main paper, we manually perturb the canonical trajectory $\bar{\boldsymbol{R}}_{\text{cano}}$ to simulate the distribution of noisy trajectories regressed from the per-frame image encoder. Specifically, during training we adopt a progressive perturbation strategy, where we begin with clean trajectories for the first 30% of the training iterations. In the next 30%, we add Gaussian noise independently to the global orientation $\boldsymbol{\Phi}$ and translation $\boldsymbol{t}$ of each frame, with a standard deviation of 1° for orientation and 10,mm for translation. In the final 40% of training, the noise level is increased to 2° for orientation and 30,mm for translation. This progressive strategy enables the model to gradually adapt to increasing levels of noise in the global trajectory.

We find that directly training on clips of 60 frames from AMASS leads to suboptimal performance due to the high correlation between adjacent frames, which may cause the model to get stuck in local minima. Experiments show that the model often fails to produce a valid human mesh when training on the multi-frame clips directly. Thus, we first pretrain the model on meshes of single frame to first let the network learn a valid pose prior over the tokens, with the trajectory set to zero. Finetuning on this checkpoint with multi-frame clips then successfully allows the model to learn the temporal dynamics of human motion on top.

A.3. Image Pretraining.

During image pretraining on image datasets [1, 6, 8, 10], we perform random rotation, flipping and photometric augmentation to improve generalization. As these datasets lack global translation ground-truth, we then perform a quick finetuning on images of EgoBody [17] and BEDLAM [2] datasets.

We follow the aggressive random cropping augmentation strategy from [3] for robustness on truncated bounding boxes. Given the projected ground-truth joints on the image, we sample from a set of predefined cropping modes that retain only a small subset of joints and compute a bounding box based on them. During image pretraining, random cropping augmentation is applied with a probability of 0.1.

A.4. Video Fine-tuning.

In the final stage of training, we fine-tune on EgoBody and BEDLAM using video clips split into 60 frames, while keeping the weights of the image encoder, including backbone, frozen. To adapt random cropping to video data, we ensure that the cropping mode is consistent across all frames within a clip. Cropping is applied to a clip with 0.2 probability, and when enabled, between 20% and 100% of the frames in that clip are cropped.

A.5. Training loss.

As described in Sec. 4.5 of the main paper, our training loss consists of the cross entropy loss $\mathcal{L}_{\text{ce}}$ for the local pose token classification, local 3D mesh vertex loss $\mathcal{L}_{V_{3D}}$, global trajectory loss $\mathcal{L}_{\text{traj}}$, global 3D joint position loss $\mathcal{L}_{J_{3D}}$ and velocity loss $\mathcal{L}_{\dot{j}_{3D}}$, 2D keypoint reprojection loss $\mathcal{L}_{J_{2D}}$ and foot skating loss $\mathcal{L}_{\text{fs}}$:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{V_{3D}} + \mathcal{L}_{\text{traj}} + \mathcal{L}_{J_{3D}} + \mathcal{L}_{\dot{j}_{3D}} + \mathcal{L}_{J_{2D}} + \mathcal{L}_{\text{fs}}, \quad (1)$$

where $\mathcal{L}_{\text{traj}}$ consists of multiple global trajectory predictions from the model: the coarse trajectory $\bar{\boldsymbol{R}}$ predicted from the image encoder, the reconstructed canonical trajectory $\hat{\boldsymbol{R}}_{\text{cano}}$ from the motion encoder and the refined trajectory $\hat{\boldsymbol{R}}$

from the decoder. During training, the sampling operation from the predicted token logits is not trivially differentiable. We adopt the Gumbel-Softmax trick [7] to enable the back-propagation through the sampling operation, which allows us to train the model end-to-end. Specifically, we sample from the softmax distribution with temperature τ cosine annealed from 1.0 to 0.01 during training. The joint-related losses $\mathcal{L}_{J_{3D}}, \mathcal{L}_{j_{3D}}, \mathcal{L}_{J_{2D}}$ are based on global joint positions, which are also computed with the aforementioned trajectory predictions, where $\hat{\boldsymbol{R}}_{\text{cano}}$ is rolled out with the ground-truth orientation and translation for the first frame. Note that each loss term is only applied when it is applicable: $\hat{\boldsymbol{R}}_{\text{cano}}$ during motion pretraining, $\bar{\boldsymbol{R}}$, $\hat{\boldsymbol{R}}$ during image pretraining, and all three during video fine-tuning. The loss terms on vertices and global joints are supervised with the ground-truth by an RMSE loss, which proves to be more stable than the L2 loss, generating notably smoother motion. The trajectory losses are supervised with L2 loss. The loss weights throughout the whole training process are set to $\lambda_{\text{ce}} = 1, \lambda_{V_{3D}} = 10, \lambda_{\text{traj}} = 1, \lambda_{J_{3D}} = 1, \lambda_{j_{3D}} = 20, \lambda_{J_{2D}} = 1, \lambda_{\text{fs}} = 1$.

B. Inference Details

During inference, we iteratively recover the tokens based on their confidence scores, defined as the softmax probabilities of the logits over the codebook. In our experiments, we employ greedy sampling with top-$k = 1$, opting for the most confident token predictions at each position. This choice is motivated by our observation that sampling from the full codebook introduces limited variation. At each iteration of the inference, the number of tokens to be remasked is determined by a mask scheduling function $\gamma(\frac{t}{T})$, where t is the current iteration and T is the total number of iterations. Following prior practice [5, 12], we adopt a cosine function $\gamma(\frac{t}{T}) = \frac{1+\cos(\frac{\pi t}{T})}{2}$ to gradually recover the local pose tokens. Additionally, the canonicalized global trajectory $\bar{\boldsymbol{R}}_{\text{cano}}$ is also iteratively updated. It is initialized using the coarse trajectory $\bar{\boldsymbol{R}}$ predicted by the image encoder in the first iteration, and subsequently refined using the output trajectory $\hat{\boldsymbol{R}}$ from the previous step.

By combining the canonical mesh decoded from pose tokens $\hat{\boldsymbol{Z}}$ and predicted global trajectory $\hat{\boldsymbol{R}}$, we obtain the global mesh vertices $\hat{\boldsymbol{V}}$, which represent the reconstructed motion. While it is possible to regress joint positions and directly render this non-parametric mesh, we find that due to its incompleteness (refer to Sec. A.1), the predicted meshes do not seamlessly integrate with the evaluation protocol, which requires a complete SMPL or SMPL-X mesh. In addition, the non-parametric mesh often lacks visual quality in regions with fine details (e.g., hands) and provides no explicit mechanism for enforcing consistent shape across the entire sequence. To address these issues, we employ *smplfitter* [15] to fit the incomplete SMPL-X meshes to full parametric SMPL-X meshes. This fitting process is fast, yields low vertex error, and solves for shared shape parameters across the sequence.

C. Evaluation Details of WHAM and GVHMR

For both methods, test-time flip augmentation is disabled for fairness. As explained in Sec. 5.4 of the main paper, WHAM and GVHMR output two sets of SMPL parameters: one in the camera coordinate system, denoted as $\boldsymbol{\gamma}_c, \boldsymbol{\Phi}_c, \boldsymbol{\theta}, \boldsymbol{\beta}$, and another in a ground-aware world coordinate system, denoted as $\boldsymbol{\gamma}_w, \boldsymbol{\Phi}_w, \boldsymbol{\theta}, \boldsymbol{\beta}$. While the local pose and shape parameters $\boldsymbol{\theta}, \boldsymbol{\beta}$ are shared between the two, the global trajectory components are predicted independently by the network, and there is no single rigid transformation that aligns the two coordinate systems consistently across the sequence.

In the original paper, the authors compute accuracy metrics (e.g., MPJPE) using $\boldsymbol{\gamma}_c, \boldsymbol{\Phi}_c$, while motion realism metrics are derived from $\boldsymbol{\gamma}_w, \boldsymbol{\Phi}_w$. However, we argue that it is only fair to evaluate all metrics on a single, consistent set of predictions. In order to compute joint position errors and render outputs in the image plane using the world-grounded predictions $\boldsymbol{\gamma}_w, \boldsymbol{\Phi}_w$, we need to project them into the camera space.

This projection is achieved by solving a rigid alignment problem in the least squares sense: we estimate a constant world-to-camera transformation across the entire sequence. The rotation is obtained via Orthogonal Procrustes alignment of the per-frame orientations, and the translation is computed by minimizing the residual displacement between camera-space and transformed world-space pelvis positions. Using this estimated transformation, we can compute MPJPE and other joint-based metrics in the camera coordinate frame for the world-grounded outputs.

Judging from Fig. 3, while the camera-space predictions $\boldsymbol{\gamma}_c, \boldsymbol{\Phi}_c$ align well with the image observations, they exhibit noticeable motion jitter. In contrast, the world-grounded predictions $\boldsymbol{\gamma}_w, \boldsymbol{\Phi}_w$ are refined to produce smoother and more realistic trajectories. However, when projected back into the camera space, these predictions may suffer from significant misalignment. We refer readers to the supplementary video for additional visual comparisons.

D. Evaluation with GT Bounding Boxes

We present additional quantitative results on the EgoBody-Occ dataset in Tab. 1, under a modified evaluation setup for image cropping. Specifically, we project the ground-truth SMPL-X mesh onto the image plane and compute a bounding box that fully encloses the projected joints, with a scale factor of 1.2. This ensures that the cropped image contains the full body of the subject, even when the subject is severely occluded. We observe that the performance of MEGA [4], WHAM [14] and GVHMR [13] is significantly impacted by

	Method	PA-MPJPE↓	MPJPE↓			PVE↓	GMPJPE↓	RTE↓	Accel↓	G-accel↓	Jitter↓	Sliding↓
			-all	*-vis*	*-occ*							
per-frame	MEGA [4]	35.96	81.41	79.23	97.21	100.33	-	-	-	-	-	-
	TokenHMR [3]	36.94	75.25	72.02	98.63	93.77	-	-	-	-	-	-
	PromptHMR [16]	34.85	48.50	45.23	72.17	59.78	-	-	-	-	-	-
temporal-based	RoHM [18]	54.53	79.01	75.85	101.7	105.18	308.8	2.23	2.81	3.78	12.74	3.28
	WHAM [14] -Cam	35.21	65.21	62.82	82.51	80.33	210.72	0.99	3.05	6.38	27.47	8.01
	WHAM [14] -World	35.21	78.43	75.86	97.06	98.58	404.32	3.40	2.99	3.09	8.70	2.50
	GVHMR [13] -Cam	35.68	59.07	55.64	83.93	73.29	430.33	1.24	2.51	8.60	41.80	8.29
	GVHMR [13] -World	35.68	61.02	57.43	87.06	75.20	478.55	2.25	2.72	3.23	11.35	**1.68**
	Ours	**26.88**	**39.59**	**37.91**	**51.72**	**51.05**	**117.48**	**0.54**	**2.17**	**1.98**	**2.30**	2.43

Table 1. **Quantitative evaluation results on EgoBody-occ, evaluated with ground-truth bounding box.** The best / second best results are in **boldface**, and underlined, respectively.

the bounding box, which tend to predict unrealisticly small human figures proportional to the bounding box size on the image plane. This implies the depth estimation in the camera space is overestimated, leading to a large error in the global space metrics. Our method mitigates this issue by adopting the random cropping augmentation in TokenHMR [3], which enhances robustness to truncated bounding boxes.

While the inability to handle variable bounding boxes is a valid limitation of these methods, we further demonstrate that our approach outperforms the baselines even when provided with correctly positioned and scaled bounding boxes. This is enabled by our masked modeling framework and effective cross-modality learning. The results support our claim that MoRo remains robust under occlusion, generating more accurate and plausible motion in comparison to the baselines, after ruling out the factor of random cropping augmentation.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1

[2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 1

[3] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *CVPR*, 2024. 1, 3

[4] Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Mega: Masked generative autoencoder for human mesh recovery, 2025. 2, 3

[5] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pages 1900–1910, 2024. 2

[6] 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 TPAMI*, 2014. 1

[7] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. 2

[8] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 1

[10] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. 2017. 1

[11] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1

[12] Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Pu Wang, Hongfei Xue, Srijan Das, and Chen Chen. Genhmr: Generative human mesh recovery, 2024. 2

[13] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 3

[14] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[15] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2

[16] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1148–1159, 2025. 3

[17] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. 1

[18] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *CVPR*, 2024. 3