SAM 3D Body: Robust Full-Body Human Mesh Recovery

Supplementary Material

In this supplementary material, we first provide ablation studies on model inference with keypoint and mask conditioning. We then show qualitative results on the full-body inference strategy and hand pose estimation results. We also provide additional experiment results on the new 3D benchmarks. Then, we describe the technical details of our model training, annotation and mesh fitting pipeline. Finally, we share the limitations of our current model and some directions for future work.

1. Evaluating 3DB Prompt Following

3DB is a promptable model that supports conditioning on 2D keypoints or segmentation masks for controllable human pose estimation. In this section, we evaluate the model's ability to follow the provided prompts and analyze their impact on pose estimation performance.

2D Keypoint Prompt. The 2D keypoint prompt provides a user-friendly mechanism for adjusting pose estimation by specifying joint locations in the image [3, 7]. In Table 1, we present an ablation study on varying the number of keypoint prompts provided during inference, where the keypoint with the largest error is selected for prompting. We observe that the model effectively follows the prompts and both 2D and 3D performance improve as more prompts are provided (noise scale $= 0$). Notably, although the keypoint prompt is provided in the 2D image space, 3DB is able to leverage this information to infer a more accurate 3D body pose (4.4mm improvement on MPJPE on EMDB).

We further evaluate the sensitivity of our model to the quality of keypoint prompts, as shown in Table 1 for #Prompt $= 1$. The noise scale is defined relative to the bounding box size, as in PCK. We observe that the model is robust to small keypoint inaccuracy (noise scale < 0.05), as such noise naturally exists in annotations of in-the-wild datasets. When the noise level becomes larger, performance degrades because the model tends to follow the incorrect keypoint prompts.

Finally, our full-body inference pipeline leverages the 2D keypoint prompting capability to improve hand pose estimation quality. To illustrate the impact of this strategy, we provide a qualitative comparison in Figure 1. From this comparison, it is evident that without keypoint prompting, 2D keypoint alignment at the wrist and hand joints is significantly worse than that achieved by the default inference. On the other hand, without integrating the hand decoder during inference, the predicted wrist rotation is often suboptimal, leading to inferior 2D finger joint alignment.

Maks Prompt. The capability of mask conditioning is essential when handling multiple people with close interaction, where the standard bounding box information is insufficient to clearly specify the person of interest for the model [6, 8]. To assess the impact of incorporating masks as additional input to our model, we compare the model inference result with and without mask-conditioning on three multi-person (MP) datasets. We follow the prior work to provide ground-truth segmentation masks when available [6, 8], and extract SAM2 [5] masks for the ITW dataset SA1B using the bounding boxes. As shown in Table 2, conditioning our model with person-specific segmentation masks yields significant improvements – the same model (3DB-DINOv3) with mask conditioning improves PVE by 33.1 and MPJPE by 29.4 on Hi4D. Notably, both Hi4D [8] and Harmony4D [2] are multi-person dataset that captures close interactions between two individuals, featuring frames with significant occlusion, which poses a challenge for most HMR methods, especially in disambiguating between individuals. Using segmentation masks for each person as additional input, 3DB effectively addresses this challenge and accurately predicts the corresponding person. For the experiments on our SA1B-Hard dataset, we observe that the performance gain on the "Multi-person" subset (+4.4%) is more significant than that on the overall dataset (+0.9%), indicating the importance of mask-conditioning for multi-person scenarios.

Table 1. Ablation on 2D keypoint prompting with 3DB-H. We report results under varying numbers of prompts, as well as different noise scales for a single prompt.

# Prompts	0	1					2
Noise scale	0	0	0.01	0.03	0.05	0.1	0
COCO (PCK@0.05↑)	86.7	**90.2**	90.2	89.5	87.6	80.9	**93.0**
EMDB (MPJPE ↓)	63.3	**60.1**	60.3	61.5	63.3	67.8	**58.9**

Table 2. Comparison on mask-conditioned inference with 3DB-DINOv3 on multi-person datasets.

	Hi4D [8]		Harmony4D		SA1B-Hard	SA1B-MP
Models	PVE ↓	MPJPE ↓	PVE ↓	MPJPE ↓	Avg-PCK ↑	Avg-PCK ↑
3DB (w/o mask)	91.4	76.4	42.7	35.6	75.4	67.9
3DB (w/ mask)	**58.3**	**47.0**	**36.5**	**30.1**	**76.3**	**72.3**

2. Qualitative results on hand performance

During the inference stage of 3DB, we leverage two strategies to further improve hand pose estimation quality: unifying hand decoder prediction into body decoder and keypoint prompting (wrist + elbow). The benefit of the hand decoder comes from the hand-specific data used during training and

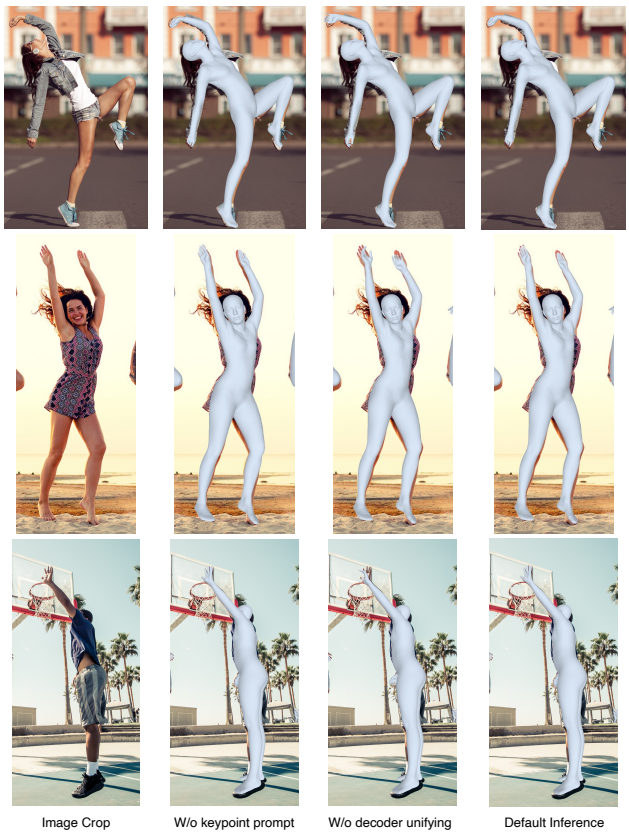

Figure 1. Qualitative comparison to show the impact from using keypoint prompting and unifying the predictions from hand decoder and body decoder.

the flexibility of a free-moving wrist due to the dedicated prediction head. By unifying the hand decoder's output to the body decoder, our model can provide full-body prediction with improved hand pose estimation. However, we noticed that simply integrating the hand decoder's output into the middle of the Momentum Human Rig kinematic tree can lead to errors in adjacent joints, particularly at the elbows. Therefore, we leverage the promptability of 3DB to mitigate these errors introduced by the hand decoder's output. Specifically, we use the wrist location from the hand decoder as well as the elbow location predicted by the body decoder to prompt the body decoder to generate a refined full-body pose estimation result.

To illustrate the impact of these two strategies on full-body pose estimation, we provide a qualitative comparison in Figure 1. It is evident that without keypoint prompting, the 2D keypoint alignment at the wrist and hand joints is significantly worse than the results from the default inference. On the other hand, without integrating the hand decoder during inference, the predicted wrist rotation can be suboptimal, leading to inferior 2D hand joint alignment.

Hand crop visualization. As shown in Figure 4 in the main paper, SAM 3D Body consistently achieves more accurate body pose and shape recovery, especially for fine details like limbs and hands. The 2D overlays further illustrate better alignment with input images, demonstrating the robustness of our approach even under difficult conditions. When we focus on hand-crop images where the human body is invisible or truncated out of images, we demonstrate the effectiveness of model as in Figure 2. Here, we only visualize the mesh output by the hand decoder for simplicity and clearness.

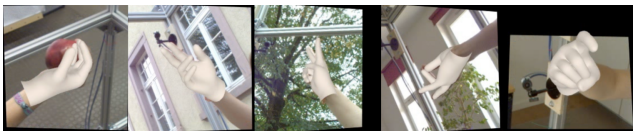

Figure 2. Qualitative results of hand estimation on Freihand [9]

3. Evaluating 3D Categorical Performance

Categorical 3D analysis using existing single view datasets is challenging as the underlying pseudo ground truth are low-fidelity approximations of the real geometry. In order to perform a more detailed categorical analysis of HMR methods, we constructed an evaluation dataset using a mix of synthetic and real data from multi-view datasets with high camera counts (more than 100 cameras).

To comprehensively evaluate 3D human mesh reconstruction performance for HMR, we define a set of 34 distinct categories based on interpretable scene and subject attributes, such as occlusion, truncation, viewpoint, pose difficulty, shape, and interaction. Unlike the manual classification used for 2D categories, these 3D categories are automatically generated using rule-based criteria applied to metadata and geometric cues. This systematic approach enables consistent, scalable, and objective analysis of model performance across diverse real-world conditions.

Based on results from Table 3, 3DB demonstrates superior performance in challenging scenarios. Particularly within the *very hard* pose categories, 3DB consistently outperforms both CameraHMR and PromptHMR in the *pose_3d:very_hard* category and in *pose_2d:very_hard*. These results indicate that 3DB possesses inherent strengths in accurately estimating poses under the most challenging conditions. Additionally, 3DB exhibits a significant advantage in handling the *truncation:severe* scenario in comparison to CameraHMR and achieves better performance in the *viewpoint:topdown_view* compared to PromptHMR.

4. Model Training Details

We describe the losses we used for model training in detail as follows.

2D/3D Keypoint Loss: We supervise 2D/3D joint locations

Table 3. 3D categorical performance analysis.

	CameraHMR [4]			PromptHMR [6]			3DB		
	PVE	MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE
aux:depth_ambiguous	126.25	102.25	81.33	109.58	91.77	69.24	**64.38**	**52.72**	**39.85**
aux:orient_ambiguous	84.26	71.77	45.07	83.79	72.93	46.17	**42.35**	**36.64**	**25.16**
aux:scale_ambiguous	118.18	104.77	50.93	112.95	102.28	47.26	**58.64**	**51.16**	**27.67**
fov:medium	82.88	68.81	46.86	76.31	64.84	42.85	**43.58**	**36.97**	**25.57**
fov:narrow	82.15	69.82	49.73	90.41	77.95	53.49	**52.14**	**43.89**	**36.18**
fov:wide	71.55	60.05	38.66	74.98	64.55	42.87	**37.97**	**33.06**	**22.44**
interaction:close_interaction	107.59	90.95	57.62	115.19	98.12	64.87	**54.23**	**44.98**	**29.76**
interaction:mild_interaction	89.98	75.28	52.93	106.55	90.38	62.74	**42.63**	**34.65**	**27.16**
pose_2d:hard	117.91	107.74	77.16	117.73	110.64	79.16	**62.93**	**57.50**	**45.58**
pose_2d:very_hard	150.20	140.61	92.66	150.15	145.07	95.40	**62.22**	**56.84**	**42.39**
pose_3d:hard	133.89	121.11	84.21	129.30	118.59	81.82	**71.42**	**63.68**	**49.10**
pose_3d:very_hard	213.66	206.34	143.23	186.35	179.46	129.51	**114.20**	**110.62**	**86.43**
pose_prior:average_pose	68.52	56.70	37.22	70.32	59.73	39.42	**36.06**	**30.95**	**21.35**
pose_prior:easy_pose	57.83	47.31	29.92	62.85	53.58	32.80	**29.53**	**24.66**	**17.20**
pose_prior:hard_pose	94.64	80.04	54.53	88.12	76.19	51.15	**51.65**	**44.24**	**31.09**
shape:average_bmi	70.35	58.07	38.08	71.01	60.25	39.90	**36.58**	**31.41**	**21.31**
shape:high_bmi	84.52	69.96	47.55	79.49	67.83	43.04	**43.33**	**36.49**	**22.45**
shape:low_bmi	80.93	65.70	42.71	69.92	58.76	37.30	**38.74**	**32.73**	**21.82**
shape:very_high_bmi	87.18	72.91	47.54	81.17	69.05	44.03	**48.51**	**41.11**	**24.80**
shape:very_low_bmi	108.16	91.25	47.26	94.16	81.12	38.64	**51.76**	**45.69**	**22.97**
truncation:left_body	135.30	113.17	87.98	127.53	110.67	91.33	**91.28**	**76.46**	**62.23**
truncation:lower_body	127.81	97.84	75.82	151.52	118.65	83.79	**92.87**	**67.10**	**60.77**
truncation:right_body	110.28	91.58	71.17	115.71	98.43	72.15	**75.04**	**62.84**	**50.62**
truncation:severe	230.51	213.64	124.01	186.57	168.22	122.70	**126.53**	**113.66**	**88.42**
truncation:upper_body	85.59	79.68	56.36	86.06	80.88	56.94	**50.83**	**48.79**	**38.39**
viewpoint:average_view	75.61	62.69	41.90	74.17	62.80	41.81	**41.25**	**35.22**	**24.41**
viewpoint:bottomup_view	89.83	72.25	53.00	95.46	78.87	55.57	**56.50**	**47.07**	**34.03**
viewpoint:topdown_view	101.69	91.13	59.15	104.29	97.92	63.39	**42.84**	**38.78**	**27.90**

using an L_1 loss, incorporating learnable per-joint uncertainty to modulate the loss based on prediction confidence. For 3D body and hand keypoints, we normalize them with their respective pelvis and wrist locations before computing the loss. Hand keypoints are weighted according to annotation availability. 2D keypoints are supervised in the cropped image spaces, and we upweight the loss for the user-provided keypoint to encourage prompt consistency when keypoint prompts are available.

Parameter Losses: MHR parameters (pose, shape) are supervised with L_2 regression losses, and joint limit penalties are imposed to discourage anatomically implausible poses.

Hand Detection Loss: 3DB can localize the hand position by a built-in hand detector. We apply GIoU loss and L_1 loss to supervise the hand box regression. We also predict the uncertainty of hand boxes and turn off the hand decoder on hand-occluded samples during inference.

5. Annotation Details

5.1. Manual Annotation

Given a set of images selected by the data engine, we use a current version of 3DB to estimate initial 2D joint positions. A team of trained annotators correct the estimated joint locations, if needed, using the GUI shown in Figure 3. The annotators also assign a per-joint visibility label according to a strict rubric. Joints with substantial occlusion or other factors that would prevent accurate placement (*e.g.*, 50% occlusion, motion blur) are marked as *not visible*.

Figure 3. GUI for annotating 2D keypoints.

5.2. Single-Image Auto-annotation

Given an RGB image, we first predict 595 dense 2D keypoints using a high-capacity keypoint detector that is conditioned on the sparse 2D keypoints provided by the datasets. This conditioning architecture reformulates dense keypoints detection as a sparse-to-dense lifting task, which is inherently easier than detection from the scratch. By leveraging both the RGB image and sparse keypoints guidance, the

model is capable of predicting accurate 2D dense keypoints from in-the-wild images (see Figure 3 of main paper). We initialize the parametric Momentum Human Rig model using the 3DB's predictions for pose, shape, and camera intrinsics, providing a strong prior for optimization. Mesh fitting is then performed via gradient-based refinement of the model parameters, minimizing a composite loss:

$$\mathcal{L} = \lambda_{\text{KP}}\mathcal{L}_{\text{KP}} + \lambda_{\text{3D}}\mathcal{L}_{\text{3D}} + \lambda_{\text{param}}\mathcal{L}_{\text{param}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}}, \quad (1)$$

where:

- $\mathcal{L}_{\text{KP}}$: Dense keypoint reprojection loss, aligning projected mesh keypoints to detected 2D keypoints.
- $\mathcal{L}_{\text{3D}}$: 3D keypoint loss, regularizing to the initial model-predicted 3D keypoints.
- $\mathcal{L}_{\text{param}}$: Parameter regularization, constraining optimization to remain close to the initial parameters.
- $\mathcal{L}_{\text{prior}}$: Pose and shape priors, enforcing anatomical plausibility via a Gaussian Mixture prior and L2 regularization.

Stage-specific weights λ are used to balance these terms throughout the optimization.

This approach combines strong priors with dense 2D supervision, enabling robust and accurate mesh fitting from single images, even in challenging cases.

Figure 4. Image to single-image mesh fitting. Source: SA-1B.

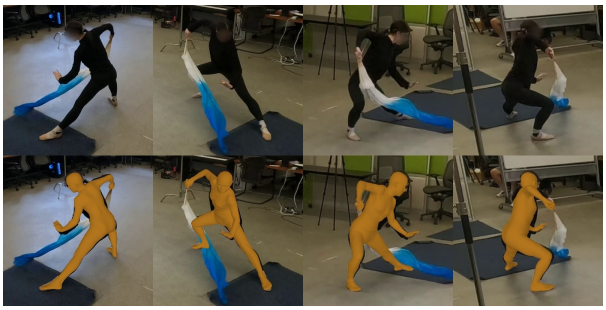

Figure 5. Multi-view MHR mesh fitting. Source: EgoExo4D [1]

5.3. Multi-View Video Annotation

For multi-view video, we extend the pipeline to jointly fit the 3D mesh across all frames and camera views, leveraging both spatial and temporal cues. Synchronized 2D keypoints are extracted for each camera and frame, then triangulated to obtain sparse 3D keypoints. The mesh model is initialized from these triangulated points and available camera calibration.

The optimization minimizes a composite loss over all frames and views:

$$\mathcal{L}_{\text{multi}} = \lambda_{\text{2D}}\mathcal{L}_{2D} + \lambda_{\text{3D}}\mathcal{L}_{3D} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}, \quad (2)$$

where:

$$\mathcal{L}_{2D} = \sum_{t,c,i} w_{i,c,t} \left\| \pi_c \left(J_i(\theta_t, \beta) \right) - k^{2D}_{i,c,t} \right\|^2, \quad (3)$$

$$\mathcal{L}_{3D} = \sum_{t,i} w_{i,t} \left\| J_i(\theta_t, \beta) - k^{3D}_{i,t} \right\|^2, \quad (4)$$

$$\mathcal{L}_{\text{smooth}} = \sum_{t} \left(\alpha_v \|\Delta\theta_t\|^2 + \alpha_a \|\Delta^2\theta_t\|^2 + \alpha_j \|\Delta^3\theta_t\|^2 \right). \quad (5)$$

Here, $J_i(\theta_t, \beta)$ denotes the 3D position of joint i at time t, π_c is the projection for camera c, and $k^{2D}_{i,c,t}$, $k^{3D}_{i,t}$ are observed 2D and triangulated 3D keypoints. Temporal smoothness is enforced via finite differences.

Optimization alternates between updating camera parameters, mesh shape, and pose, with robust keypoint filtering (e.g., robust losses, RANSAC, smoothing) applied to mitigate outliers. Shared parameters (e.g., shape and scale parameters) are optimized jointly across frames, ensuring temporal and spatial consistency.

Our annotation pipeline fuses dense keypoint detection, strong parametric priors, and robust multi-stage optimization to generate high-fidelity 3D mesh supervision. This process, applied to both single images and multi-view video, is foundational to the data quality that underpins our model's strong generalization and performance.

We show some mesh fitting results from our training dataset are shown in Figure 4 and Figure 5.

6. Limitations

We discuss some limitations of 3DB as presented in this paper and suggest possible next steps to address these limitations. First, 3DB processes each individual separately, without taking multi-person or human-object interactions into account. This limits its ability to accurately interpret relative positions and physical interactions. A natural next step would be to incorporate interactions among humans, objects, and the environment into the model's training process. Second, while our model has achieved significant improvements in hand pose estimation as part of the full-body estimation task, its accuracy does not surpass that of specialized hand-only pose estimation methods. Additionally, due to the limited availability of high-quality full-body data during training, the hand estimation performance from the body decoder alone is also suboptimal. This limitation can be addressed by incorporating more diverse full-body data

into the training of 3DB. Third, both 3DB and the underlying mesh model MHR fall short in modeling human body shapes across all age groups. As a result, they may produce suboptimal pose estimations and shape modeling for children.

References

[1] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 4

[2] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37:107270–107285, 2024. 1

[3] Jingyuan Liu, Li-Yi Wei, Ariel Shamir, and Takeo Igarashi. ipose: Interactive human pose reconstruction from video. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2024. 1

[4] Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pages 1562–1571. IEEE, 2025. 3

[5] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[6] 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. 1, 3

[7] Jie Yang, Ailing Zeng, Feng Li, Shilong Liu, Ruimao Zhang, and Lei Zhang. Neural interactive keypoint detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15122–15132, 2023. 1

[8] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023. 1

[9] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 813–822, 2019. 2