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

**Xitong Yang**[⋆], **Devansh Kukreja**[⋆], **Don Pinkus**[⋆], **Anushka Sagar**, **Taosha Fan**, **Jinhyung Park**[◦], **Soyong Shin**[◦], **Jinkun Cao**, **Jiawei Liu**, **Nicolas Ugrinovic**, **Matt Feiszli**[†], **Jitendra Malik**[†], **Piotr Dollar**[†], **Kris Kitani**[†]

Meta Superintelligence Labs
[⋆]Core Contributor, [◦]Intern, [†]Project Lead

We introduce SAM 3D Body (3DB), a promptable model for single-image full-body 3D human mesh recovery (HMR) that demonstrates state-of-the-art performance, with strong generalization and consistent accuracy in diverse in-the-wild conditions. 3DB estimates the human pose of the body, feet, and hands. It is the first model to use a new parametric mesh representation, Momentum Human Rig (MHR), which decouples skeletal structure and surface shape. 3DB employs an encoder–decoder architecture and supports auxiliary prompts, including 2D keypoints and masks, enabling user-guided inference similar to the SAM family of models. We derive high-quality annotations from a multi-stage annotation pipeline that uses various combinations of manual keypoint annotation, differentiable optimization, multi-view geometry, and dense keypoint detection. Our data engine efficiently selects and processes data to ensure data diversity, collecting unusual poses and rare imaging conditions. We present a new evaluation dataset organized by pose and appearance categories, enabling nuanced analysis of model behavior. Our experiments demonstrate superior generalization and substantial improvements over prior methods in both qualitative user preference studies and traditional quantitative analysis. Both 3DB and MHR are open-source.

**Demo:** https://www.aidemos.meta.com/segment-anything/editor/convert-body-to-3d
**Code:** https://github.com/facebookresearch/sam-3d-body
**Website:** https://ai.meta.com/sam3d

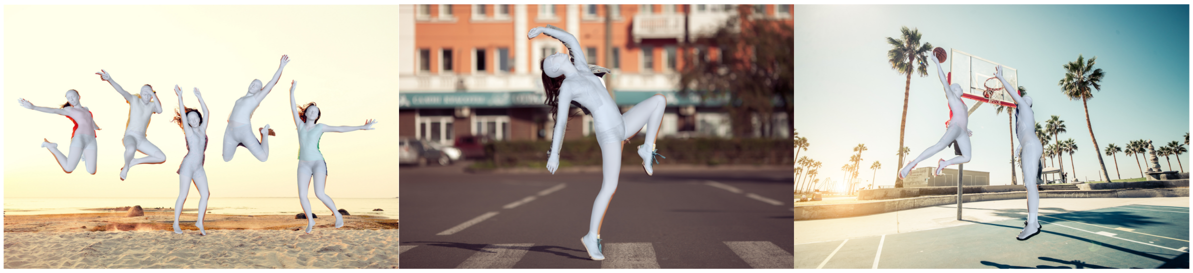


**Figure 1** Human mesh recovery results using SAM 3D Body (3DB). Our model demonstrates robust performance in estimating challenging poses across diverse viewpoints and produces accurate body and hand pose estimations within a unified framework.

## 1 Introduction

Estimating 3D human pose (skeleton pose and structure) and shape (soft body tissue) from images is an essential capability for vision and embodied AI systems to understand and interact with people. Despite notable progress in human mesh recovery (HMR) (9; 7; 33; 50; 51), existing approaches still exhibit unsatisfactory robustness when applied to in-the-wild images, which limits their applicability to real-world scenarios such as robotics (37; 32; 47) and biomechanics (36). In particular, current models often fail on individuals presenting challenging poses, severe occlusion, or captured from uncommon viewpoints. They also struggle to reliably estimate both the overall body pose and the fine details of the hands and feet in a unified full-body framework.

We argue that the primary challenges in developing a robust full-body human mesh recovery model stem from both the data and model aspects. First, collecting large-scale and diverse human pose datasets with high-quality mesh annotations is inherently difficult and computationally costly. Most existing datasets either suffer from low diversity due to laboratory capture settings (4; 12; 13) or from low mesh quality resulting from pseudo-labeling (48; 1). Second, current HMR architectures do not adequately address the distinct optimization mechanisms required for body and hand pose estimation, nor do they incorporate effective training strategies to handle uncertainty and ambiguity from monocular images.

In this work, we present SAM 3D Body (3DB), a robust full-body HMR model fueled by large-scale, high-quality human pose data curated by our data engine.

**Robust Full-body HMR Model.** We make three main contributions to improve model performance on both body and hand pose estimation. (i) We propose a novel promptable encoder–decoder architecture (17; 39) that enables the model to condition on optional 2D keypoints, masks or camera information for controllable pose estimation. This promptable design naturally facilitates interactive guidance in ambiguous or challenging scenarios during training, and provides a coherent approach to integrate hand and body predictions. (ii) Our model utilizes a shared image encoder and two separate decoders for the body and hands. This two-way-decoder design effectively alleviates conflicts in optimizing body and hand pose estimation, which arise from differences in input resolution, camera estimation, and supervision objectives. (iii) Unlike most prior work that relies on the SMPL (26) human mesh model, we build 3DB on a new parametric mesh representation, MHR (8), which decouples skeletal pose and body shape, providing richer control and interpretability for full-body reconstruction.

**Data Engine for Diverse Human Pose and High-quality Annotation.** HMR methods have increasingly turned to large-scale training data for higher performance (9; 3; 54). However, high-quality 3D supervision remains scarce, and existing in-the-wild datasets are still limited in scale and diversity. To this end, we design a new data creation pipeline that features: (i) *Data Quality*: Our annotation pipeline combines various combinations of components such as geometric constraints, parametric priors, and dense keypoint regression, which automatically yields high-quality 3D human mesh annotations. (ii) *Data Quantity*: We curate data from large licensed stock photo repositories, multiple multi-view capture datasets, and synthetic data. We create a large scale of **7 million** images with high-quality annotation. (iii) *Data Diversity*: Our data is diversified using a VLM-based data engine that mines for in-the-wild challenging images and routes them for annotation. This ensures coverage of rare poses, difficult viewpoints, and varied appearances, providing a more diverse dataset for supervision.

Together, the data engine and full-body HMR model enable 3DB to recover high-fidelity full-body human meshes from a single image. 3DB achieves state-of-the-art performance across both body and hand pose estimation. Extensive experiments demonstrate that 3DB consistently outperforms prior HMR methods on standard metrics, generalizes better to unseen datasets, and is preferred by users in a study of $7,800$ participants, achieving a significant $5:1$ win rate in visual quality. To our knowledge, it is the first single model that delivers the **best performance to body-specialized models and comparable performance to hand-specialized models**, while providing interactive control and strong robustness under challenging poses and in-the-wild scenarios.

## 2 Related Work

**Human Mesh Models:** The most widely used human mesh model is SMPL (26), which parameterizes human body into pose and shape. SMPL-X (34) goes further to include hands (MANO (40)) and faces (FLAME (21)). SMPL models intertwine the skeletal structure and soft-tissue mass within the *shape space*, which can limit interpretability (*e.g.*, the parameters do not always map directly to bone lengths) and controllability. Alternatively, Momentum Human Rig (8), an enhancement of ATLAS (31), explicitly decouples the skeletal structure and body shape, and we adopt it as our representation of the human body.

**Human Mesh Recovery (HMR):** Early HMR methods like HMR 2.0 (9) were *body-only* methods that predicted the body without articulated hands or feet  (18; 22; 7). Instead, 3DB follows the more recent paradigm of full-body methods (2; 5; 41; 3; 51) that estimate *body+hands+feet*. There are also part-specific hand mesh recovery methods (35; 38) that only estimate the pose and shape of the hands, which usually have more

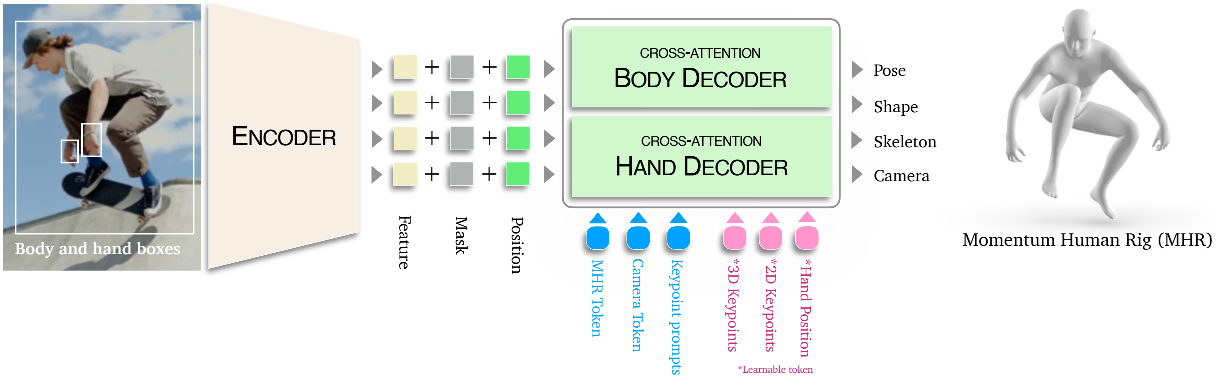


**Figure 2** SAM 3D Body Model Architecture. We employ a promptable encoder–decoder architecture with a shared image encoder and separate decoders for body and hand pose estimation.

accurate performance compared to full-body methods. In contrast, 3DB shows strong performance on both hand and full-body estimation.

**Promptable Inference:** Promptable inference, popularized by the SAM family (17; 39), enables user or system-provided prompts (such as 2D keypoints or masks) to guide model predictions. Similarly to (51), our approach supports various prompt types, including 2D keypoints and masks, and by integrating prompt tokens directly into the transformer architecture, enables user-guided mesh recovery.

**Data Quality and Annotation Pipelines:** A major bottleneck in HMR is the quality of training data. Many datasets rely on pseudo-ground-truth (pGT) meshes obtained from monocular fitting (18; 14), which often contain systematic errors in pose, shape, and camera parameters (33). Recent work (7; 50) highlights the impact of annotation noise on reported metrics and generalization. To address this, multi-view datasets (28; 16; 30) and synthetic data have been used in our work to provide higher-fidelity supervision. Our method builds on these insights by employing a scalable data engine that mines challenging cases using vision-language models, and by leveraging a multi-stage annotation pipeline that combines dense keypoint detection, strong parametric priors, and robust optimization.

# 3 SAM 3D Body Model Architecture

Our goal is to recover 3D human meshes (*i.e.*, MHR parameters) accurately, robustly and interactively from a single image. To this end, we design 3DB as a promptable encoder–decoder architecture (see Figure 2) with a rich set of prompt tokens. 3DB is designed to be *interactive* as it can accept 2D keypoints or masks, allowing users or downstream systems to guide inference.

## 3.1 Image Encoder

The human-cropped image $I$ is normalized and passed through a vision backbone to produce a dense feature map $F$. An optional set of hand crops $I_{\text{hand}}$ can also be provided to obtain hand crop feature maps $F_{\text{hand}}$:

$$F = \text{ImgEncoder}(I), \tag{1}$$

$$F_{\text{hand}} = \text{ImgEncoder}(I_{\text{hand}}). \tag{2}$$

3DB considers two optional prompts: 2D keypoints and segmentation masks. Keypoint prompts are encoded by positional encodings summed with learned embeddings and are provided as additional tokens for the pose decoder. Mask prompts are embedded using convolutions and summed element-wise with the image embedding (17).

### 3.2 Decoder Tokens

3DB has two decoders: The body decoder outputs the full-body human rig and an optional hand decoder can provide enhanced hand pose results. The pose decoders take a set of *query tokens* as input to predict the parameters of MHR and camera parameters. There are four types of query tokens: MHR+camera, 2D keypoint prompt, auxiliary 2D/3D keypoint tokens and optional hand position tokens.

**MHR+Camera Token:** The initial estimate of MHR and (optionally) camera parameters is embedded as a learnable token for MHR parameter estimation:

$$T_{\text{pose}} = \text{RigEncoder}(E_{\text{init}}) \in \mathbb{R}^{1\times D}, \tag{3}$$

$$E_{\text{init}} \in \mathbb{R}^{d_{\text{init}}}. \tag{4}$$

**2D Keypoint Prompt Tokens:** If 2D keypoint prompts $K$ are provided (*e.g.*, from a user or detector), they are encoded as:

$$T_{\text{prompt}} = \text{PromptEncoder}(K) \in \mathbb{R}^{N\times D}, \tag{5}$$

$$K \in \mathbb{R}^{N\times 3}, \tag{6}$$

where each keypoint is represented by $(x, y, \text{label})$.

**Hand Position Tokens:** The hand token, $T_{\text{hand}} \in \mathbb{R}^{2\times D}$, is used in the body decoder to locate the hand positions inside the human images. This set of tokens is optional, without which 3DB can still produce a full-body human rig because the output from body decoder already includes hands.

**Auxiliary Keypoint Tokens:** To further enhance interactivity and model capacity, we include learnable tokens for all 2D and 3D keypoints.

$$T_{\text{keypoint2D}} \in \mathbb{R}^{J_{2D}\times D}, \tag{7}$$

$$T_{\text{keypoint3D}} \in \mathbb{R}^{J_{3D}\times D}. \tag{8}$$

These tokens allow the model to reason about specific joints and support downstream tasks such as keypoint prediction or uncertainty estimation.

### 3.3 MHR Decoder

All tokens are concatenated to form the full set of queries:

$$T = [T_{\text{pose}},\ T_{\text{prompt}},\ T_{\text{keypoint2D}},\ T_{\text{keypoint3D}},\ T_{\text{hand}}] \tag{9}$$

This flexible assembly enables the model to operate in both fully automatic and user-guided modes, adapting to the available prompts. The body decoder attends to both the query tokens $T$, the full-body image features $F$,

$$O = \text{Decoder}(T, F) \in \mathbb{R}^{(3+N+J_{2D}+J_{3D})\times D}. \tag{10}$$

Through cross-attention, the body decoder fuses prompt information with visual context, enabling robust and editable mesh recovery. Optionally, the hand decoder can take the same prompt information while attends to the hand crop features $F_{\text{hand}}$ to provide another output token $O_{\text{hand}}$.

The first output token of $O$ is passed through an MLP to regress the final mesh parameters: $\theta = \text{MLP}(O_0) \in \mathbb{R}^{d_{\text{out}}}$, where $\theta = \{\mathbf{P}, \mathbf{S}, \mathbf{C}, \mathbf{S}_k\}$ are the predicted MHR parameters: pose, shape, camera pose and skeleton, respectively. Another set of outputs can be computed from $O_{\text{hand}}$ for a pair of MHR hands, which can be merged to the body output to improve the estimation of the hand.

## 4 Model Training and Inference

**Model Training.** 3DB is trained with a comprehensive multi-task loss terms, $\mathcal{L}_{\text{train}} = \sum_i \lambda_i \mathcal{L}_i$, where each $\mathcal{L}_i$ is a task-specific loss targeting a specific prediction head or anatomical structure. $\lambda_i$ are hyper-parameters set empirically. To stabilize training, certain loss terms (*e.g.*, 3D keypoints) are introduced with a warm-up schedule, gradually increasing their weights over the course of training. We also simulate an interactive setup (17; 46) for training by randomly sampling prompts in multiple rounds per sample. This multi-task, prompt-aware loss design provides strong supervision across all outputs. We describe the losses in details below.

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

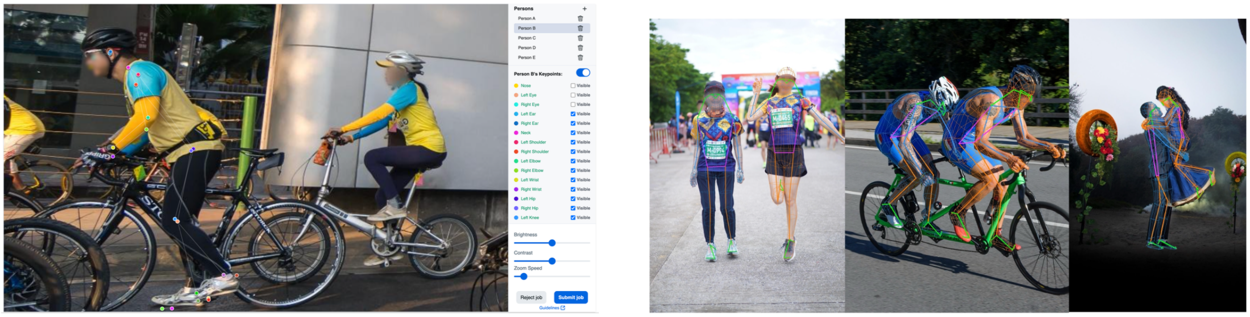
**Model Inference.** For the model inference, we use the output from the body decoder as the default. However, if the hand detector head finds hands inside the input image, we can choose to merge the output from the hand decoder to body output to enhance the hand pose estimation quality. At this stage, we typically use the wrist and elbow keypoint estimation from the hand decoder and the body decoder separately as prompting keypoints to align and combine the predictions from the two decoders. Finally, the predicted local MHR parameters are merged to a full-body configuration following the kinematic tree of the mesh model.

## 5 Data Engine for Diversity

Obtaining highly accurate human mesh annotations paired with the images can be computationally costly. Instead, one common strategy is to annotate a large video collection and leveraging temporal constraints to get more reliable pseudo annotations. While it is possible to get a large number of training images from videos, the poses, appearance, imaging conditions, and background might be very similar. In order to increase the diversity of our training dataset, we implemented an automated data engine that selectively routes difficult images for annotation, enabling scalable and efficient dataset curation.

At the core of our data engine is a Vision-Language Model (VLM) driven mining strategy. Rather than relying on simple heuristics or random sampling, we leverage VLMs to automatically generate and update mining rules that identify high-value images for annotation. The VLM identifies images exhibiting challenging scenarios for pose estimation, including occlusion (where the human subject is partially hidden by objects or other people), unusual poses (rare or complex body configurations such as acrobatics or dance), interaction (human-object or human-human activities like holding tools or group actions), extreme scale (subjects appearing at atypical distances from the camera), low visibility (poor lighting, motion blur, or partial visibility), and hand-body coordination (tight coupling of hand and body poses, as in sign language or sports).

Mining rules are automatically updated iteratively based on failure analysis of the current model, allowing the engine to adaptively focus on the most challenging or informative samples. Failure analysis is performed semi-manually, by evaluating 3DB on the current set of annotated images, visualizing the most challenging images using keypoint location error, and then manually annotating the image with a few words. These words and images are used to create text prompt for the VLM. New images selected by the VLM are then routed for manual annotation. By focusing annotation efforts on the most informative samples, our data engine enables efficient search through tens of millions of images, while maximizing the value and diversity of each annotated image. By collecting a highly diverse dataset, it provides the basis on which to build a very robust HMR model that works on a wide range of in-the-wild images.



Persons +
Person A
Person B
Person C
Person D
Person E

Person B's Keypoints:
Nose □ Visible
Left Eye □ Visible
Right Eye □ Visible
Left Ear ☑ Visible
Right Ear ☑ Visible
Neck ☑ Visible
Left Shoulder ☑ Visible
Right Shoulder ☑ Visible
Left Elbow ☑ Visible
Right Elbow ☑ Visible
Left Wrist ☑ Visible
Right Wrist ☑ Visible
Left Hip ☑ Visible
Right Hip ☑ Visible
Left Knee ☑ Visible

**Figure 3** Left: GUI of our annotation tool for annotating 2D keypoints. Right: Comparison of the dense (thin) and sparse (thick) keypoints for pseudo annotation.

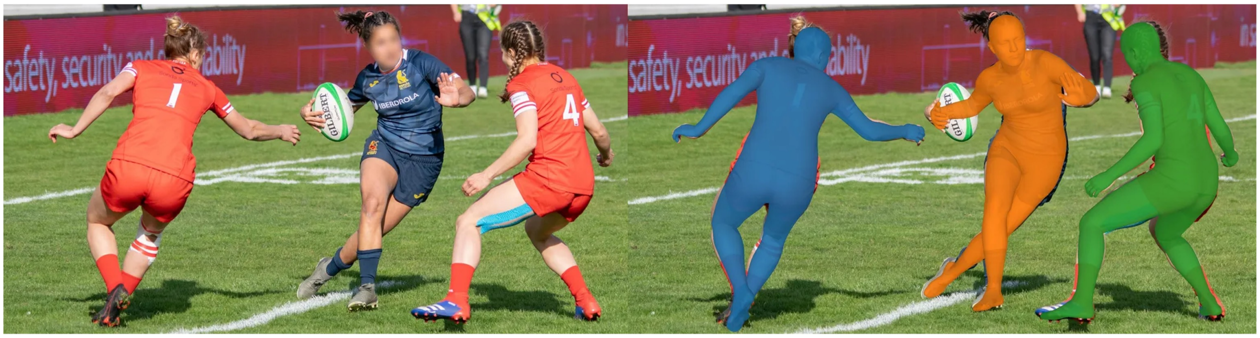


**Figure 4** Example of single-image MHR mesh fitting for ITW datasets. Source: SA-1B (17).

# 6 Data Annotation and Mesh Fitting

In addition to the robustness enabled by the data diversity derived from our data engine, the accuracy of our model depends heavily on the quality of our annotations. To this end, we designed a multi-stage annotation pipeline that produces accurate 3D mesh pseudo-ground truth from both in-the-wild single image datasets and a variety of multi-view datasets, using various combinations of manual 2D keypoint annotation, sparse and dense keypoint detection, geometric constraints, temporal constraints, strong parametric priors, and robust optimization methods.

## 6.1 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. Then, a team of trained annotators review and manually correct the estimated joint locations if needed, as shown in Figure 3(a). 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.*

## 6.2 Single-Image Mesh Fitting

For each image, we predict 595 dense 2D keypoints using a high-capacity keypoint detector that is conditioned on the sparse 2D keypoints obtained from the manual annotation step described above, as illustrated in Figure 3(b). Building upon prior dense keypoint detection framework that did not exploit cues other than pixels (33; 11; 6), our approach predicts accurate 2D dense keypoints from in-the-wild images (Figure 5b) by jointly leveraging image cues and sparse keypoint guidance. We then initialize MHR using a current version of 3DB's predictions for pose, shape, and camera intrinsics, which is used as an initialization for mesh optimization. MHR fitting is then performed via gradient-based refinement of the model parameters, minimizing a composite fitting loss $\mathcal{L}_{\text{fit}} = \sum_j \lambda_j \mathcal{L}_j$, where each $\mathcal{L}_j$ is a task-specific loss including 2D keypoint

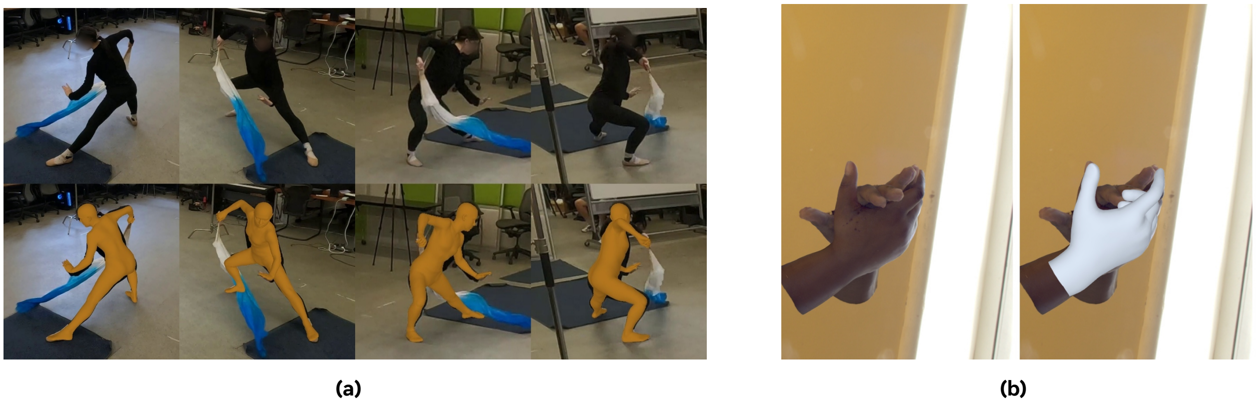


**Figure 5** Examples of MHR mesh fitting results. (a) Multi-view mesh fitting. Source: EgoExo4D (10). (b) Scan-based mesh fitting. Source: Re:Interhand (29).

loss, initialization-anchored regularization and priors. Hyper-parameters $\lambda_j$ are set via cross-validation. We apply several loss terms and priors to make the fitting goal: **2D Keypoint Loss** is the L2 distance between projected and detected dense 2D keypoints, to ensure minimal 2D reprojection error. **Initialization-Anchored Regularization** penalizes deviation from the initial prediction by applying L2 losses on both the Momentum Human Rig parameters and their corresponding 3D keypoints, thereby preventing model drift. **Pose and Shape Prior** enforces anatomical plausibility via a learned Gaussian Mixture prior and L2 regularization. Following the pipeline above, we derive the image to MHR fittings as training supervision as in Figure 4.

### 6.3 Multi-View Mesh Fitting

Though single-view mesh fitting is effective for a large and diverse set of images, the annotation quality tends to be lower fidelity due to the depth ambiguities and natural occlusion. Therefore, we also exploit multi-view mesh fitting on suitable datasets. For multi-view video datasets, we further extend the pipeline to jointly fit 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 camera parameters and refined via second-order optimization-based update of the model parameters, minimizing a composite fitting loss, $\mathcal{L}_{\text{multi}} = \sum_k \lambda_k \mathcal{L}_k$, where each $\mathcal{L}_k$ is a task-specific loss including **the 2D keypoint loss and the regularization and priors as single-view mesh fitting**, together with additional 3D keypoint loss and temporal smoothness: **3D Keypoint Loss** is the L2 distance between mesh joints and triangulated 3D keypoints obtained from multi-view geometry, providing strong spatial supervision. **Temporal Smoothness Loss** encourages estimated pose parameters to temporally smooth, penalizing abrupt changes in motion and promoting realistic temporal dynamics. $\lambda_k$ are set via cross-validation. Optimization alternates between updating camera parameters, shape, skeleton, and pose, with robust keypoint filtering (*e.g.*, robust losses, RANSAC, smoothing). Body specific parameters (*e.g.*, shape, skeleton parameters) are optimized jointly across frames. The mesh fitting happens on body full-body data and hand data as shown in Figure 5.

## 7 Training Datasets

We train our model on a mix of single-view, multi-view, and synthetic datasets listed in Table 1, covering general body pose, hands, interactions, and “in-the-wild’ conditions to ensure the quality, quantity and diversity of training data.

**Single-view in-the-wild:** We utilize datasets that captures people in unconstrained environments with diverse appearance, pose, and scene conditions. For this, we use AIChallenger (53), MS COCO (25), MPII (1), 3DPW (48), and a subset of SA-1B (17).

**Multi-view consistent:** To incorporate geometric consistency for more reliable annotations, we use multi-view

**Table 1** List of 3DB training datasets. ⋆ denotes the datasets providing samples to train the hand decoder.

| Dataset | # Images/Frames | # Subjects | # Views |
|---|---|---|---|
| MPII human pose (1) | 5K | 5K+ | 1 |
| MS COCO (25) | 24K | 24K+ | 1 |
| 3DPW (48) | 17K | 7 | 1 |
| AIChallenger (53) | 172K | 172K+ | 1 |
| SA-1B (17) | 1.65M | 1.65M+ | 1 |
| Ego-Exo4D (10) | 1.08M | 740 | 4+ |
| DexYCB (4) | 291K | 10 | 8 |
| EgoHumans (15) | 272K | 50+ | 15 |
| Harmony4D (16) | 250K | 24 | 20 |
| InterHand (30)⋆ | 1.09M | 27 | 66 |
| Re:Interhand (29)⋆ | 1.50M | 10 | 170 |
| Goliath (28)⋆ | 966K | 120+ | 500+ |
| Synthetic⋆ | 1.63M | – | – |

data from Ego-Exo4D (10), Harmony4D (16), EgoHumans (15), InterHand2.6M (30), DexYCB (4) and Goliath  (28).

**High-fidelity synthetic:** We use a photorealistic synthetic extension of the Goliath dataset (28). It provides millions of frames with ground-truth MHR parameters across diverse identities, clothing, and contexts. Synthetic data ensures accurate supervision for human mesh recovery, complementing real-world datasets that prioritize diversity over quality.

**Hand datasets:** These datasets (marked with ⋆ in Table 1), such as Re:Interhand (29), are used to train both the body and hand decoder. We provide wrist-truncated hand samples to train the hand decoder.

# 8 Evaluation

We follow prior HMR work and report standard pose and shape evaluation metrics: MPJPE (27), PA-MPJPE (55), PVE (20), and PCK (55). To evaluate on SMPL-based datasets, a MHR mesh is mapped to the SMPL mesh format.

## 8.1 Evaluating Performance on Common Datasets

We first evaluate 3DB on five standard benchmark datasets in Table 2, comparing with a wide variety of state-of-the-art (SoTA) mesh recovery methods. We present results with two variants of the model; 3DB-H leverages the commonly used ViT-H (632M) backbone, and 3DB-DINOv3 uses the recent DINOv3 (840M) (45) encoder. We use an off-the-shelf field-of-view (FOV) estimator (49) to provide camera intrinsics for model inference. 3DB outperforms all other single-image methods and is even competitive with video-based approaches that additionally leverage temporal information.

In particular, our model achieves superior results in the EMDB and RICH datasets, which are *out-of-domain* (*i.e.*, not included in the training set), indicating better generalization than previous SoTA methods. 3DB exceeds the second best model, NLF, on all datasets in terms of 3D metrics except for RICH which dataset NLF uses in training while our model does not. 3DB is also state-of-the-art on PCK for 2D evaluation on the COCO and LSPET datasets, demonstrating strong 2D alignment.

## 8.2 Evaluating Performance on New Datasets

Throughout our experiments, we found that mesh recovery models are particularly fragile in out-of-domain settings due to camera, appearance, and pose differences. To understand how methods perform on new, unseen data distributions, we additionally evaluate on five new datasets (38.6K images) in Table 3. The five new datasets include (1) Ego-Exo4D (10), (2) Harmony4D (16), (3) Goliath (28), (4) in-house synthetic data and (5) SA1B-Hard. Ego-Exo4D captures humans in diverse, skilled activities, divided into physical (EE4D-Phys) and procedural (EE4D-Proc) domains. Harmony4D focuses on close multi-human interaction in dynamic sports settings. Goliath offers diverse motions in a precise, studio environment. The synthetic

**Table 2** Comparison on five common benchmarks. The best results are highlighted in bold, while the second-best results are underlined. Results evaluated using publicly released checkpoint denoted by $^\dagger$. Models trained using RICH denoted by $^*$.

| | | 3DPW (14) | | | EMDB (24) | | | RICH (24) | | | COCO | LSPET |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Models | PA-MPJPE ↓ | MPJPE ↓ | PVE ↓ | PA-MPJPE ↓ | MPJPE ↓ | PVE ↓ | PA-MPJPE ↓ | MPJPE ↓ | PVE ↓ | PCK@0.05 ↑ | PCK@0.05 ↑ |
| IMAGE | HMR2.0b (9) | 54.3 | 81.3 | 93.1 | 79.2 | 118.5 | 140.6 | 48.1$^\dagger$ | 96.0$^\dagger$ | 110.9$^\dagger$ | 86.1 | 53.3 |
| | CameraHMR (33) | 35.1 | 56.0 | 65.9 | 43.3 | 70.3 | 81.7 | 34.0 | 55.7 | 64.4 | 80.5$^\dagger$ | 49.1$^\dagger$ |
| | PromptHMR (51) | 36.1 | 58.7 | 69.4 | 41.0 | 71.7 | 84.5 | 37.3 | 56.6 | 65.5 | 79.2$^\dagger$ | 55.6$^\dagger$ |
| | SMPLerX-H (3) | 46.6$^\dagger$ | 76.7$^\dagger$ | 91.8$^\dagger$ | 64.5$^\dagger$ | 92.7$^\dagger$ | 112.0$^\dagger$ | 37.4$^\dagger$ | 62.5$^\dagger$ | 69.5$^\dagger$ | – | – |
| | NLF-L+fit* (43) | 33.6 | 54.9 | 63.7 | 40.9 | 68.4 | 80.6 | 28.7$^\dagger$ | 51.0$^\dagger$ | 58.2$^\dagger$ | 74.9$^\dagger$ | 54.9$^\dagger$ |
| VIDEO | WHAM (44) | 35.9 | 57.8 | 68.7 | 50.4 | 79.7 | 94.4 | – | – | – | – | – |
| | TRAM (52) | 35.6 | 59.3 | 69.6 | 45.7 | 74.4 | 86.6 | – | – | – | – | – |
| | GENMO (19) | 34.6 | **53.9** | 65.8 | 42.5 | 73.0 | 84.8 | 39.1 | 66.8 | 75.4 | – | – |
| | 3DB-H (Ours) | **33.2** | 54.8 | 64.1 | 38.5 | 62.9 | 74.3 | 31.9 | 55.0 | 61.7 | **86.8** | **68.9** |
| | 3DB-DINOv3 (Ours) | 33.8 | 54.8 | **63.6** | **38.2** | **61.7** | **72.5** | **30.9** | **53.7** | **60.3** | 86.5 | 67.8 |

**Table 3** Comparison on five new benchmark datasets. The best results are highlighted in bold, while the second-best results are underlined. MPJPE is computed on 24 SMPL keypoints.

| | EE4D-Phy | | EE4D-Proc | | Harmony4D | | Goliath | | Synthetic | | SA1B-Hard |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Models | PVE ↓ | MPJPE ↓ | PVE ↓ | MPJPE ↓ | PVE ↓ | MPJPE ↓ | PVE ↓ | MPJPE ↓ | PVE ↓ | MPJPE ↓ | Avg-PCK ↑ |
| CameraHMR (33) | 71.1 | 58.8 | 70.3 | 60.2 | 84.6 | 70.8 | 66.7 | 54.5 | 102.8 | 87.2 | 63.0 |
| PromptHMR (51) | 74.6 | 63.4 | 72.0 | 62.6 | 91.9 | 78.0 | 67.2 | 56.5 | 92.7 | 80.7 | 59.0 |
| NLF (43) | 75.9 | 68.5 | 85.4 | 77.7 | 97.3 | 84.9 | 66.5 | 58.0 | 97.6 | 86.5 | 66.5 |
| 3DB-H Leave-one-out (Ours) | **49.7** | **44.3** | **52.9** | **47.4** | **63.5** | **54.0** | **54.2** | **46.5** | **85.6** | **75.5** | **73.1** |
| 3DB-H Full dataset (Ours) | 37.0 | 31.6 | 41.9 | 36.3 | 41.0 | 33.9 | 34.5 | 28.8 | 55.2 | 47.2 | 76.6 |

dataset consists of single-human images with diverse camera angles and parameters. SA1B-Hard is a subset of 2.6K images extracted from SA1B using our data engine. Together, these five new datasets present a challenging new testbed for mesh recovery methods.

As it is difficult to compare methods using the exact same training data and methodology due to prohibitive data usage licenses, unclear descriptions of training data, and lack of training code (CameraHMR, PromptHMR, and NLF are trained on 6, 9, and 48 datasets, respectively), we fairly test the generalization ability of 3DB by using a leave-one-out training procedure. This ensures a fair comparison with prior work which have also not seen these datasets. To serve as an in-domain, upper bound comparison, we also show the performance of 3DB when trained on the *full dataset* (*i.e.*, training data is also sampled from these new datasets). For both the baselines and our model, we use ground truth camera intrinsics for model inference for all 3D datasets, except for SA1B-Hard which we used FOV estimated by MoGe-2 (49).

We present the results in Table 3. Despite being trained on a large number of datasets, we find that prior work still struggle with these five domains, incurring a significant drop in performance. In contrast, our leave-one-out model shows strong generalization, owing to our more diverse data distribution and stronger training framework. Interestingly, we notice that existing methods constantly trade places for second across different datasets, reflecting strong dataset-specific biases. This indicates that each baseline overfit to a narrow slice of the underlying data distribution.

**Table 4** Comparison on Freihand for hand pose estimation. Methods using Freihand for training are denoted by $^\dagger$.

| Method | PA-MPVPE ↓ | PA-MPJPE ↓ | F@5 ↑ | F@15 ↑ |
|---|---|---|---|---|
| LookMa (11) | 8.1 | 8.6 | 0.653 | - |
| METRO (24)$^\dagger$ | 6.3 | 6.5 | 0.731 | 0.984 |
| HaMeR (35)$^\dagger$ | 5.7 | 6.0 | 0.785 | 0.990 |
| MaskHand (42)$^\dagger$ | 5.4 | 5.5 | 0.801 | 0.991 |
| WiLoR (38)$^\dagger$ | 5.1 | 5.5 | 0.825 | 0.993 |
| 3DB-H (Ours) | 6.3 | 5.5 | 0.735 | 0.988 |
| 3DB-DINOv3 (Ours) | 6.2 | 5.5 | 0.737 | 0.988 |

**Table 5** 2D categorical performance analysis on the SA-1B Hard dataset.

| | CameraHMR (33) | | PromptHMR (51) | | 3DB | |
|---|---|---|---|---|---|---|
| | aPCK(body) | aPCK(feet) | aPCK(body) | aPCK(feet) | aPCK(body) | aPCK(feet) |
| Body_shape - In-the-wild | 87.64 | 78.56 | 85.73 | 77.87 | **90.76** | **92.12** |
| Camera_view - Back or side view | 59.69 | 46.64 | 61.92 | 47.74 | **76.27** | **66.81** |
| Camera_view - Bottom-up view | 55.18 | 34.84 | 46.56 | 29.25 | **69.62** | **55.35** |
| Camera_view - Others | 51.48 | 33.80 | 54.39 | 38.55 | **76.62** | **71.52** |
| Camera_view - Overhead view | 55.08 | 39.46 | 43.65 | 24.63 | **73.33** | **66.94** |
| Hand - Crossed or overlapped fingers | 73.20 | 62.85 | 72.48 | 62.43 | **81.36** | **84.04** |
| Hand - Holding objects | 76.73 | 72.11 | 73.57 | 68.92 | **83.40** | **85.92** |
| Hand - Self-occluded hands | 73.22 | 58.06 | 72.43 | 56.19 | **80.07** | **80.82** |
| Multi_people - Contact or interaction | 63.23 | 51.65 | 61.77 | 47.60 | **74.81** | **69.92** |
| Multi_people - Overlapped | 53.11 | 41.88 | 57.17 | 41.43 | **70.82** | **64.71** |
| Pose - Contortion or bending | 47.08 | 32.78 | 42.61 | 20.98 | **65.20** | **53.04** |
| Pose - Crossed legs | 63.95 | 32.24 | 56.15 | 27.35 | **76.40** | **58.80** |
| Pose - Inverted body | 46.12 | 30.01 | 39.83 | 24.64 | **78.18** | **72.19** |
| Pose - Leg or arm splits | 57.51 | 31.43 | 54.76 | 33.11 | **83.69** | **72.49** |
| Pose - Lotus pose | 63.19 | 14.38 | 54.85 | 12.87 | **74.53** | **57.97** |
| Pose - Lying down | 51.29 | 35.88 | 44.59 | 26.88 | **71.35** | **66.53** |
| Pose - Sitting on or riding | 79.66 | 71.65 | 70.15 | 61.16 | **84.85** | **81.51** |
| Pose - Sports or athletic activities | 78.93 | 69.34 | 73.62 | 60.37 | **85.10** | **82.80** |
| Pose - Squatting or crouching or kneeling | 62.74 | 41.47 | 54.41 | 33.84 | **72.85** | **61.85** |
| Visibility - Occlusion (foot cues) | 62.93 | 26.83 | 58.00 | 30.81 | **75.43** | **54.74** |
| Visibility - Occlusion (hand cues) | 61.01 | 53.89 | 58.55 | 51.13 | **76.04** | **72.01** |
| Visibility - Truncation (lower-body truncated) | 39.27 | - | 46.50 | - | **61.95** | - |
| Visibility - Truncation (others) | 79.18 | 74.82 | 77.06 | 74.99 | **84.23** | **86.72** |
| Visibility - Truncation (upper-body truncated) | 62.37 | 54.90 | 56.01 | 49.28 | **64.49** | **70.99** |

## 8.3 Evaluating Hand Pose Estimation Performance

One significant characteristic of 3DB is its strong performance in estimating hand shape and pose. Previous full-body human pose estimation methods (3; 2; 23) revealed a notable gap in hand pose accuracy compared to *hand-only* pose estimation methods (35; 38). This performance gap arises from two main factors. First, hand-only methods can leverage large-scale datasets of hand poses, whereas full-body methods cannot utilize these datasets because of the absence of full-body images and annotations. Second, a free-moving wrist allows hand pose models to more easily fit finger poses with 2D and 3D alignment, while for full-body methods, wrist rotation and position are highly constrained by the body's pose and position. Despite these challenges, 3DB demonstrates strong hand pose accuracy. 3DB benefits from the flexible model training design that incorporates both hand and body data and the hand decoder. Additionally, being promptable, 3DB provides a natural mechanism to align the wrists of the body prediction with those of the hands. We evaluate 3DB's hand estimation on the representative FreiHand (56) benchmark in Table 4. For fair comparison against hand-only models, we use the output from our hand decoder for evaluation. Despite not training on the Freihand dataset, which gives a strong in-domain boost, 3DB's hand pose estimation accuracy is already comparable to SoTA hand pose estimation methods that include Freihand alongside many other hand-centric datasets

## 8.4 Evaluating 2D Categorical Performance

To better understand the strengths and weaknesses of models on a variety of image types, we compare the performance across our 24 categories defined over SA1B-Hard (17). Our proposed evaluation set is designed to capture a broad spectrum of human appearance and activity in images, ensuring robust evaluation across real-world scenarios. It consists of 24 total categories, which are organized under several high-level groups: Body Shape, Camera View, Hand, Multi-person, Pose and Visibility.

We use the PCK (Percentage of Correct Keypoints) metric for 17 body keypoints and 6 feet keypoints. Results are reported using Avg-PCK, which is PCK averaged over a range of thresholds (*i.e.* 0.01, 0.025, 0.05, 0.075, 0.1 of the human bounding box size). Results in Table 5 show that 3DB outperforms all baselines on all categories. Qualitative examples are given in Figure 6.

One notable significance is for categories of *Visibility - Truncation* where the model shows significant advantages than CameraHMR or PromptHMR. Essentially, 3DB has learned a much stronger pose prior when dealing with body truncation in images. Other rows with the large improvements are *Pose - Inverted body* and *Pose - Leg or arm splits*. We largely attribute these improvements to the increased distribution of hard poses selected by the data engine.

**Table 6** 3D categorical performance analysis.

| | CameraHMR (33) | | | PromptHMR (51) | | | 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** |

## 8.5 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 6, 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* category in comparison to PromptHMR.

## 8.6 Qualitative Results

In addition to quantitative gains, our model shows clear qualitative improvements over baselines. Figure 6 compares SAM 3D Body to six state-of-the-art methods on the SA1B-Hard dataset, highlighting challenging cases with complex poses, shapes, and occlusions. As shown, SAM 3D Body consistently achieves more accurate body pose and shape recovery, especially for fine details like limbs and hands. The 2D overlays in Figure 6 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 7. Here, we only visualize the mesh output by the hand decoder for simplicity and clearness.

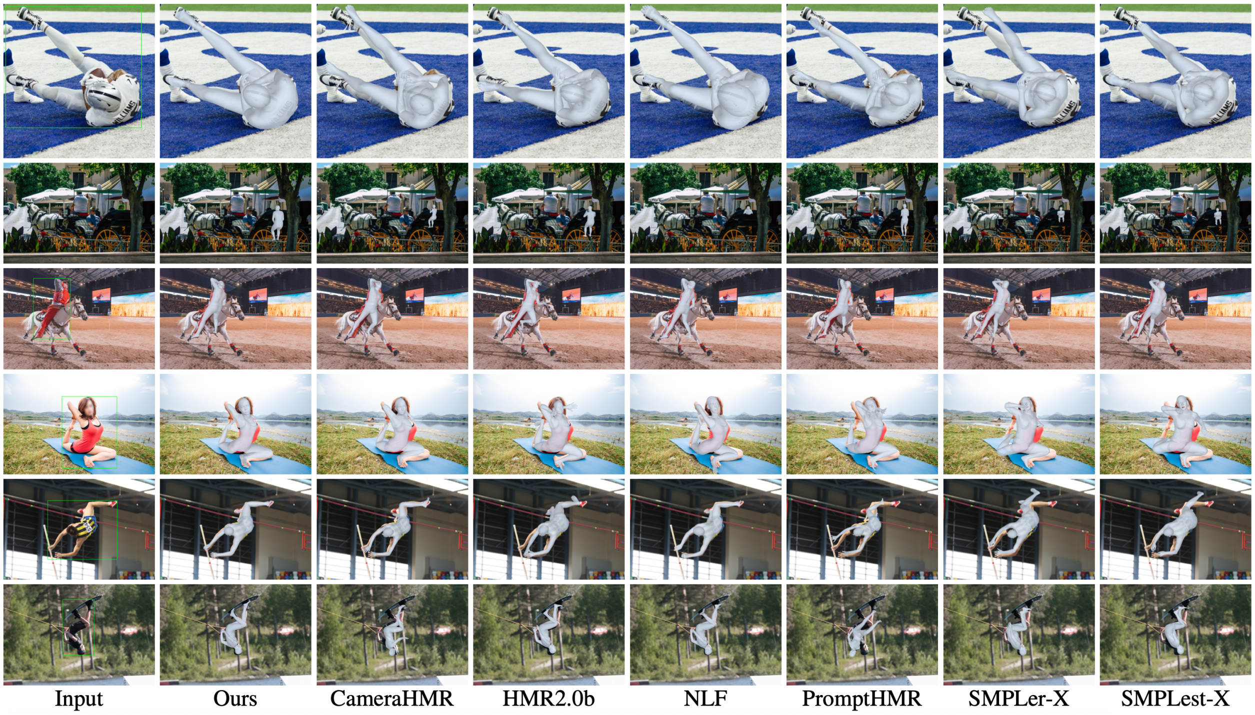


**Figure 6** Qualitative comparison of 3DB against state-of-the-art HMR methods. Source: SA-1B (17).

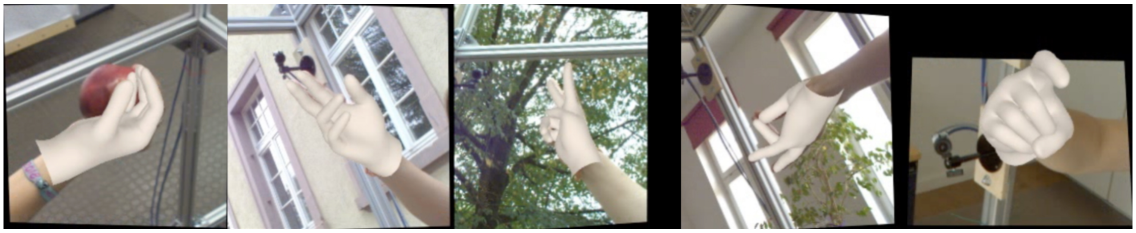


**Figure 7** Qualitative results of hand estimation using the hand decoder of 3DB. Source: Freihand (56).

## 8.7 Human Preference Study

We conducted a large-scale user preference study to evaluate the perceptual quality of human reconstructions produced by 3DB compared with existing approaches on the SA1B-Hard dataset. While quantitative metrics capture geometric and numeric accuracy, they do not always align with the human perception accuracy.

We designed six independent pairwise comparison studies, each comparing 3DB against one baseline method: HMR2.0b (9), CameraHMR (33), NLF (43), PromptHMR (51), SMPLer-X (3), and SMPLest-X (54). The study encompassed $7,800$ unique participants ($1,300$ unique per comparison) resulting in over $20,000$ total responses. Each participant was presented with a video stimuli. The left and right sides of the video displayed reconstructions from the two methods, and a video transition effect as used to fade-in the reconstruction result over the image. Participants were instructed to choose which 3D reconstruction better matched the original image by answering: *"Which 3D model of the person better matches the original image, left or right?"*. We quantify results using win rate and vote share. Win rate is the percentage of stimuli for which 3DB received more votes than the baseline. As summarized in Figure 8, 3DB consistently outperforms all baselines. Focusing on the strongest baseline, NLF, 3DB achieves a win rate of 83.8%.

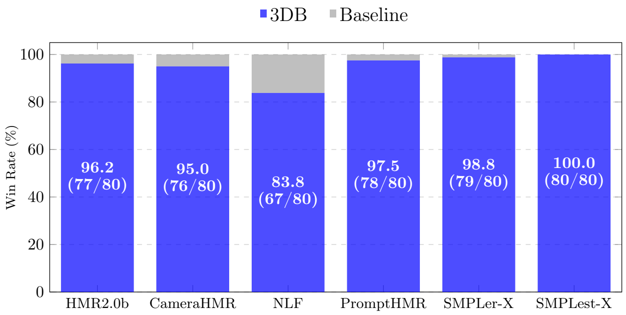


**Figure 8** Comparison of 3DB win rate against baselines for human preference study. Win rate (%) and number of wins out of 80.

# 9 Conclusion

We have presented 3DB, a robust HMR model for body and hands. Our approach leverages the Momentum Human Rig parametric body model, employs a flexible encoder–decoder architecture, and supports optional prompts such as 2D keypoints or masks to guide inference. A central advance of our work is in the supervision pipeline. Instead of relying on noisy monocular pseudo-ground-truth, we leverage multi-view capture systems, synthetic sources, and a scalable data engine that actively mines and annotates challenging samples. This strategy yields cleaner and more diverse training signals, supporting generalization beyond curated benchmarks. At the same time, 3DB employs a separate hand decoder to enhance the hand pose estimation with hand crops as input which makes it comparable to SoTA hand pose estimation methods.

# Acknowledgements

We gratefully acknowledge the following individuals for their contributions and support: Vivian Lee, George Orlin, Nikhila Ravi, Andrew Westbury, Jyun-Ting Song, Zejia Weng, Xizi Zhang, Yuting Ye, Federica Bogo, Ronald Mallet, Ahmed Osman, Rawal Khirodkar, Javier Romero, Carsten Stoll, Shunsuke Saito, Jean-Charles Bazin, Sofien Bouaziz, Yuan Dong, Su Zhaoen, Alexander Richard, Michael Zollhoefer, Roman Radle, Sasha Mitts, Michelle Chan, Yael Yungster, Azita Shokrpour, Helen Klein, Mallika Malhotra, Ida Cheng, Eva Galper.

# References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] Fabien Baradel, Matthieu Armando, Thomas Lucas, Romain Brégier, Philippe Weinzaepfel, and Grégory Rogez. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision (ECCV)*, 2024.

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[4] Yu-Wei Chao, Wei-Cheng Yang, Yu Xiang, et al. DexYCB: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention (ExPose). In *European Conference on Computer Vision (ECCV)*, 2020.

[6] Hanz Cuevas-Velasquez, Anastasios Yiannakidis, Soyong Shin, Giorgio Becherini, Markus Höschle, Joachim Tesch, Taylor Obersat, Tsvetelina Alexiadis, and Michael J. Black. Mamma: Markerless automatic multi-person motion action capture, 2025.

[7] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] Aaron Ferguson, Ahmed A. A. Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Federica Bogo, Igor Santesteban, Javier Romero, Jenna Zarate, Jeongseok Lee, Jinhyung Park, Jinlong Yang, John Doublestein, Kishore Venkateshan, Kris Kitani, Ladislav Kavan, Marco Dal Farra, Matthew Hu, Matthew Cioffi, Michael Fabris, Michael Ranieri, Mohammad Modarres, Petr Kadlecek, Rinat Abdrashitov, Romain Prévost, Roman Rajbhand ari, Ronald Mallet, Russel Pearsall, Sand y Kao, Sanjeev Kumar, Scott Parrish, Te-Li Wang, Tony Tung, Yuan Dong, Yuhua Chen, Yuanlu Xu, Yuting Ye, and Zhongshi Jiang. Mhr: Momentum human rig. *arXiv Preprint*, 2025. arXiv preprint; identifier to be added.

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. Includes HMR 2.0.

[10] Kristen Grauman et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

[11] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6), 2024.

[12] 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*, 2014.

[13] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[14] 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)*, pages 7122–7131, 2018.

[15] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. EgoHumans: An egocentric 3D multi-human benchmark. In *International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[16] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4D: A video dataset for in-the-wild close human interactions. *NeurIPS Datasets and Benchmarks*, 2024.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019.

[19] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *International Conference on Computer Vision (ICCV)*, 2025.

[20] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021.

[21] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans (FLAME). In *ACM Transactions on Graphics (TOG), SIGGRAPH Asia*, 2017.

[22] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022.

[23] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21159–21168, 2023.

[24] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics (TOG), SIGGRAPH Asia*, pages 248:1–248:16, 2015.

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017.

[28] Meta. Goliath dataset, 2025. Partial Release.

[29] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Matthew Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A Dataset of Relighted 3D Interacting Hands. In *NeurIPS 2023 Datasets and Benchmarks Track*, 2023.

[30] Gyeongsik Moon, Shoou-i Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*, 2020.

[31] Jinhyung Park, Javier Romero, Shunsuke Saito, Fabian Prada, Takaaki Shiratori, Yichen Xu, Federica Bogo, Shoou-I Yu, Kris Kitani, and Rawal Khirodkar. Atlas: Decoupling skeletal and shape parameters for expressive parametric human modeling. In *International Conference on Computer Vision (ICCV)*, pages 6508–6518, 2025.

[32] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022.

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

[34] 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 (SMPL-X). In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[35] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

[36] Owen Pearl, Soyong Shin, Ashwin Godura, Sarah Bergbreiter, and Eni Halilaj. Fusion of video and inertial sensing data via dynamic optimization of a biomechanical model. *Journal of biomechanics*, 155:111617, 2023.

[37] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018.

[38] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12254, 2025.

[39] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025. Oral.

[40] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Modeling and capturing hands and bodies together (MANO). In *ACM Transactions on Graphics (TOG), SIGGRAPH Asia*, 2017.

[41] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *International Conference on Computer Vision (ICCV) Workshops*, 2021.

[42] Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Mayur Jagdishbhai Patel, Hongfei Xue, Ahmed Helmy, Srijan Das, and Pu Wang. Maskhand: Generative masked modeling for robust hand mesh reconstruction in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8372–8383, 2025.

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

[44] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, June 2024.

[45] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[46] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE international conference on image processing (ICIP)*, pages 3141–3145. IEEE, 2022.

[47] Vasileios Vasilopoulos, Georgios Pavlakos, Sean L Bowman, J Diego Caporale, Kostas Daniilidis, George J Pappas, and Daniel E Koditschek. Reactive semantic planning in unexplored semantic environments using deep perceptual feedback. *IEEE Robotics and Automation Letters*, 5(3):4455–4462, 2020.

[48] Timo von Marcard, Gerard Pons-Moll, Michael J. Black, et al. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018.

[49] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025.

[50] Shengze Wang, Jiefeng Li, Tianye Li, Ye Yuan, Henry Fuchs, Koki Nagano, Shalini De Mello, and Michael Stengel. BLADE: Single-view body mesh estimation through accurate depth estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21991–22000, 2025.

[51] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *Computer Vision and Pattern Recognition (CVPR)*, 2025.

[52] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *European Conference on Computer Vision (ECCV)*, 2024.

[53] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE, 2019.

[54] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2025.

[55] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3d human pose estimation. *Advances in neural information processing systems (NeurIPS)*, 33:2408–2419, 2020.

[56] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2019.