Supplementary Material

DPoser-X: Diffusion Model as Robust 3D Whole-body Human Pose Prior

Junzhe Lu[1,*] Jing Lin[2,*] Hongkun Dou[3] Ailing Zeng[4] Yue Deng[3]
Xian Liu[5] Zhongang Cai[6] Lei Yang[6] Yulun Zhang[7] Haoqian Wang[1,†] Ziwei Liu[2,†]

[1]Tsinghua University [2]Nanyang Technological University [3]Beihang University
[4]Anuttacon [5]NVIDIA Research [6]SenseTime Research [7]Shanghai Jiao Tong University

This appendix provides comprehensive details to supplement our main work. Section A presents an in-depth discussion of related studies. The parameterization of diffusion models and their connection to score functions are recapped in Section B, followed by the perspective of Score Distillation Sampling (SDS) to understand our DPoser regularization in Section C.

Section D examines the runtime and computational overhead introduced by DPoser, while Section E explores test-time timestep scheduling across both pose and image domains. The datasets used for training and evaluation are detailed in Section F. Section G covers detailed experimental setup including task-specific loss, evaluation metrics used in each task and implementation of comparative methods.

Further evaluations of DPoser-X on additional tasks and datasets are provided in Section H. Section I outlines the training process for DPoser, whereas Section J discusses extended optimization techniques. Section K outlines the method's limitations, presents failure cases, and suggests avenues for future research. Lastly, Section L showcases additional qualitative results.

A. Related Work

A.1. Human Pose Priors

Human body models like SMPL [41] and SMPL-X [48] serve as powerful tools for parameterizing both pose and shape, thereby offering a comprehensive framework for describing human gestures. Within the SMPL model, body poses are represented using rotation matrices or joint angles linked to a kinematic skeleton. Adjusting these parameters enables the representation of diverse human actions. Nonetheless, feeding unrealistic poses into these models can result in non-viable human figures, primarily because plausible human poses are confined within a complex, high-dimensional manifold due to biomechanical constraints.

Various strategies [1, 11, 23, 48, 59] have been put forward to build human pose priors. Generative frameworks like GMMs, VAEs [32], and Generative Adversarial Networks (GANs) [20] have shown promise in encapsulating the multifaceted pose distribution, facilitating advancements in tasks like human mesh recovery [19, 29]. Further, some studies have delved into conditional pose priors tailored to specific tasks, incorporating extra information such as image features [4, 51], 2D joint coordinates [8], or sequences of preceding poses [38, 52]. Our initiative leans towards an unconditional pose prior without relying on additional inputs like images or text, aiming for a versatile application across various pose-related scenarios.

A.2. Diffusion Models for Pose-centric Tasks

Diffusion models [25, 55–57] have emerged as powerful tools for capturing intricate data distributions, aligning well with the demands of multi-hypothesis estimation in ambiguous human poses. Notable works include DiffPose [26], which employs a Graph Convolutional Network (GCN) architecture conditioned on 2D pose sequences for 3D pose estimation by learned reverse process (*i.e.*, generation). Similarly, DiffusionPose [51] and GFPose [8] employ the generation-based pipeline but take different approaches in conditioning. Further, ZeDO [28] concentrates on 2D-to-3D pose lifting, while Diff-HMR [4] and DiffHand [35] explore estimating SMPL parameters and hand mesh vertices, respectively. EgoWholeBody [62] and RoHM [66] focus on refining noisy motion sequences via diffusion-based generation. BUDDI [46] stands out for using diffusion models to capture the joint distribution of interacting individuals and leveraging SDS loss [49, 61] for optimization during testing phases.

While DPoser shares similar optimization implementation with BUDDI, it sets itself apart by introducing a wider perspective of inverse problems and equipping an innovative timestep scheduling strategy tailored to human poses. Unlike other approaches [8, 26, 28, 51] that primarily focus on 3D location-based representation, DPoser takes on

* Equal contribution. † Corresponding authors.

the more demanding task of modeling SMPL-based rotation pose representation. Furthermore, DPoser-X improves whole-body modeling with detailed hand and facial expressions, making it a versatile choice for pose-centric tasks.

B. Parameterization of Score-based Diffusion Models

In the seminal work by Song *et al*. [57], it is demonstrated that both score-based generative models [56] and diffusion probabilistic models [25] can be interpreted as discretized versions of stochastic differential equations (SDEs) defined by score functions. This unification allows the training objective to be interpreted either as learning a time-dependent denoiser or as learning a sequence of score functions that describe increasingly noisy versions of the data.

We begin by revisiting the training objective for score-based models [56] to elucidate the link with diffusion models [25]. Consider the transition kernel of the forward diffusion process $p_{0t}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t\mathbf{x}_0, \sigma_t^2\mathbf{I})$. Our goal is to learn score functions $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ through a neural network $s_\theta(\mathbf{x}_t; t)$, by minimizing the L2 loss as follows (we omit the expectation operator for conciseness) :

$$\mathbb{E}\left[w(t)||s_\theta(\mathbf{x}_t; t) - \nabla_{\mathbf{x}_t} \log p_t\left(\mathbf{x}_t\right)||_2^2\right]. \tag{1}$$

Here, $\mathbf{x}_t = \alpha_t\mathbf{x}_0 + \sigma_t\epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Based on denoising score matching [60], we know the minimizing objective Eq. (1) is equivalent to the following tractable term:

$$\mathbb{E}\left[w(t)||s_\theta(\mathbf{x}_t; t) - \nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0)||_2^2\right]. \tag{2}$$

To link this with the noise predictor $\epsilon_\theta(\mathbf{x}_t; t)$ in diffusion models, we can employ the reparameterization $s_\theta(\mathbf{x}_t; t) = -\frac{\epsilon_\theta(\mathbf{x}_t;t)}{\sigma_t}$. Then, Eq. (2) can be simplified as follows:

$$\begin{aligned}
&w(t)||-\frac{\epsilon_\theta(\mathbf{x}_t;t)}{\sigma_t} - \nabla_{\mathbf{x}_t}\log p_{0t}(\mathbf{x}_t \mid \mathbf{x}_0)||_2^2 \\
=&w(t)||-\frac{\epsilon_\theta(\mathbf{x}_t;t)}{\sigma_t} + \frac{(\mathbf{x}_t - \alpha_t\mathbf{x}_0)}{\sigma_t^2}||_2^2 \\
=&w(t)||-\frac{\epsilon_\theta(\mathbf{x}_t;t)}{\sigma_t} + \frac{\sigma_t\epsilon}{\sigma_t^2})||_2^2 \\
=&\frac{w(t)}{\sigma_t^2}||\epsilon_\theta(\mathbf{x}_t;t) - \epsilon)||_2^2
\end{aligned} \tag{3}$$

The resulting form of Eq. (3) aligns precisely with the noise prediction form of diffusion models [25] (refer to Eq. (4) in the main text). This implies that by training $\epsilon_\theta(\mathbf{x}_t; t)$ in a diffusion model context, we simultaneously get a handle on the score function, approximated as $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \approx -\frac{\epsilon_\theta(\mathbf{x}_t;t)}{\sigma_t}$.

C. View DPoser as Score Distillation Sampling

Interestingly, the gradient of DPoser (Eq. (10) in the main text) coincides with Score Distillation Sampling (SDS) [49, 61], which can be interpreted as aiming to minimize the following KL divergence:

$$KL\big(p_{0t}\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \parallel p_t^{\texttt{SDE}}\left(\mathbf{x}_t; \theta\right)\big), \tag{4}$$

where $p_t^{\texttt{SDE}}\left(\mathbf{x}_t; \theta\right)$ denote the marginal distribution whose score function is estimated by $\epsilon_\theta(\mathbf{x}_t; t)$. For the specific case where $t \to 0$, this term encourages the Dirac distribution $\delta(\mathbf{x}_0)$ (*i.e*., the optimized variable) to gravitate toward the learned data distribution $p_0^{\texttt{SDE}}\left(\mathbf{x}_0; \theta\right)$, while the Gaussian perturbation like Eq. (4) softens the constraint. Building on this understanding, we can borrow advanced techniques from SDS—a rapidly evolving area ripe for methodological innovations [10]. To extend this, we experiment with a multi-step denoising strategy adapted from HiFA [68], substituting our original one-step denoising process. This alternative, however, yields suboptimal results across most evaluation metrics, as demonstrated in Table S-1. A plausible explanation could be that our proposed truncated timestep scheduling effectively manages low noise levels (*i.e*., small t), thus negating the need for more denoising steps. In our main experiments, we keep the efficient one-step denoiser.

D. Runtime Comparison

Diffusion models generally require iterative steps for gradual denoising, making them less efficient than VAEs and GANs in generation tasks. However, when applied to downstream optimization processes, DPoser introduces minimal additional computational overhead. This is due to two key factors: (1) DPoser regularization involves only a single-step denoising at each optimization step, and (2) the stop-gradient operator ensures that the regularization does not require backpropagation through the trained network.

To assess DPoser's efficiency, we benchmarked its runtime against various prior models (including a baseline without pose prior) for human mesh recovery across 100 images in a consistent execution environment. As shown in Table S-2, incorporating DPoser results in only a modest (10%) increase in optimization runtime compared to the baseline. In contrast, GAN-S [11] incurs a significant computational cost due to its required GAN-inversion phase, which converts initial poses into their latent representations.

E. Analysis of Test-time Timestep Scheduling

During optimization, the selection of timestep is crucial for downstream tasks. As discussed in Section 2.4, the key information of pose data emerges at small t values ($t \leq 0.3$), which serves as a coarse range. Moreover, the L2 loss format of our DPoser regularization gives an intuitive view of

Strategy	Whole-body Mesh Recovery		Body Pose Completion		Motion Denoising	
	PA-MPVPE (all) ↓	PA-MPVPE (hands) ↓	MPVPE ↓	APD ↑	MPVPE ↓	MPJPE ↓
1 step	**60.98**	**15.60**	**38.79**/78.31/27.13	6.53	**38.21**	**19.87**
5 step	61.39	15.70	40.15/85.01/31.96	7.72	40.22	21.21
10 step	61.52	15.74	41.04/87.36/32.51	**8.07**	40.69	21.34

Table S-1. Ablation of different denoising steps in DPoser's optimization.

w/o prior	GMM [1]	VPoser [48]	Pose-NDF [59]	GAN-S [11]	DPoser
15.64	16.14	16.83	21.88	74.60	17.34

Table S-2. Runtime comparison (in seconds) of different prior models for human mesh recovery on 100 images, evaluated using an RTX 3090Ti GPU.

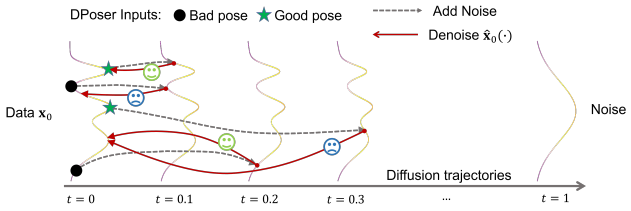


Figure S-1. Visualization of the impact of different timestep values in DPoser regularization. A larger t effectively corrects undesirable poses but may excessively alter well-posed inputs, resulting in plausible yet unrelated poses. Conversely, a smaller t better preserves the original pose but struggles to correct implausible ones.

Noise std	$[0.15, 0.05]$	$[0.2, 0.05]$	$[0.2, 0.1]$	$[0.25, 0.1]$
40 mm	**19.83**	<u>19.87</u>	21.68	22.14
100 mm	36.13	34.15	**33.18**	<u>33.83</u>

Table S-3. Ablation of timestep range for motion denoising on the AMASS dataset [42]. MPJPE is reported as the metric.

the impact of timestep. As shown in Fig. S-1, while t is small, since the adding noise and denoising path is short, the denoised pose is close to the origin and the DPoser guidance is weak. Specifically, considering the extreme case where $t \to 0$, in $\hat{\mathbf{x}}_0(t) = \frac{\mathbf{x}_t - \sigma_t \epsilon_\phi(\mathbf{x}_t; t)}{\alpha_t}$, the coefficient $\sigma_t \to 0$ while $\alpha_t \to 1$, causing $\hat{\mathbf{x}}_0(t)$ to approach $\mathbf{x}_0$, which leads to a near-zero DPoser loss. On the contrary, suitably large t means strong DPoser guidance and can correct implausible poses better. Thus, we tailor $[t_{max}, t_{min}]$ intervals to specific tasks based on their noise scales. To verify this, we conduct ablation of the timestep range on motion denoising. As evidenced in Table S-3, to achieve the best performance, larger t values are required for noisier inputs. Based on the above analyses, we select task-specific timestep intervals $[t_{max}, t_{min}]$ as follows: $[0.2, 0.05]$ for motion denoising (40 mm noise), $[0.15, 0.05]$ for pose completion and inverse kinematics, and $[0.12, 0.08]$ for mesh recovery. All the experiments, including body-only, hand-only, face-only, and whole-body, share the same timestep hyperparameters without more tuning.

It is also noteworthy that our truncated timestep scheduling is designed for human poses and does not work well on images. In image domains, the initial timesteps play a crucial role in generating foundational perceptual content. In our study, we employed a 256x256 unconditional diffusion model [13] trained on ImageNet [12] with variational diffusion sampling [43] for image inpainting. This model employs 1000 discrete timesteps during training. We compared standard scheduling (timesteps 990 to 0) with truncated scheduling (timesteps 495 to 0), both using 100 steps. The results, shown in Fig. S-2, indicate that truncation negatively affects image quality. While the standard approach preserved perceptual content, the truncated method produced disjointed patches that were misaligned with the original context. These results affirm that truncated timestep scheduling excels in pose data where key information emerges in later stages but falls short in image tasks where early timesteps are essential. This scheduling is thus bespoke to the characteristics of human poses and is unsuitable for image processes that rely on the full diffusion timeline for content fidelity.

F. Dataset Description

This section provides a detailed overview of the datasets used in our experiments, categorized based on the body part they focus on. We describe each dataset's specific use case along with the number of samples available for each dataset.

F.1. Body-only Dataset

AMASS The AMASS dataset [42] is a large-scale collection of high-quality 3D human body meshes derived from multiple motion capture sources. It provides motion sequences and human poses in a SMPL-based format, covering a broad range of activities such as walking, sitting, dancing, and running. Following the same splits as VPoser [48], and after sampling to de-duplicate the data, we use approximately 55 million body poses in the SMPL-X [48] format to train our DPoser-body model. The test split consists of 54,000 body poses, which are used to evaluate model per-

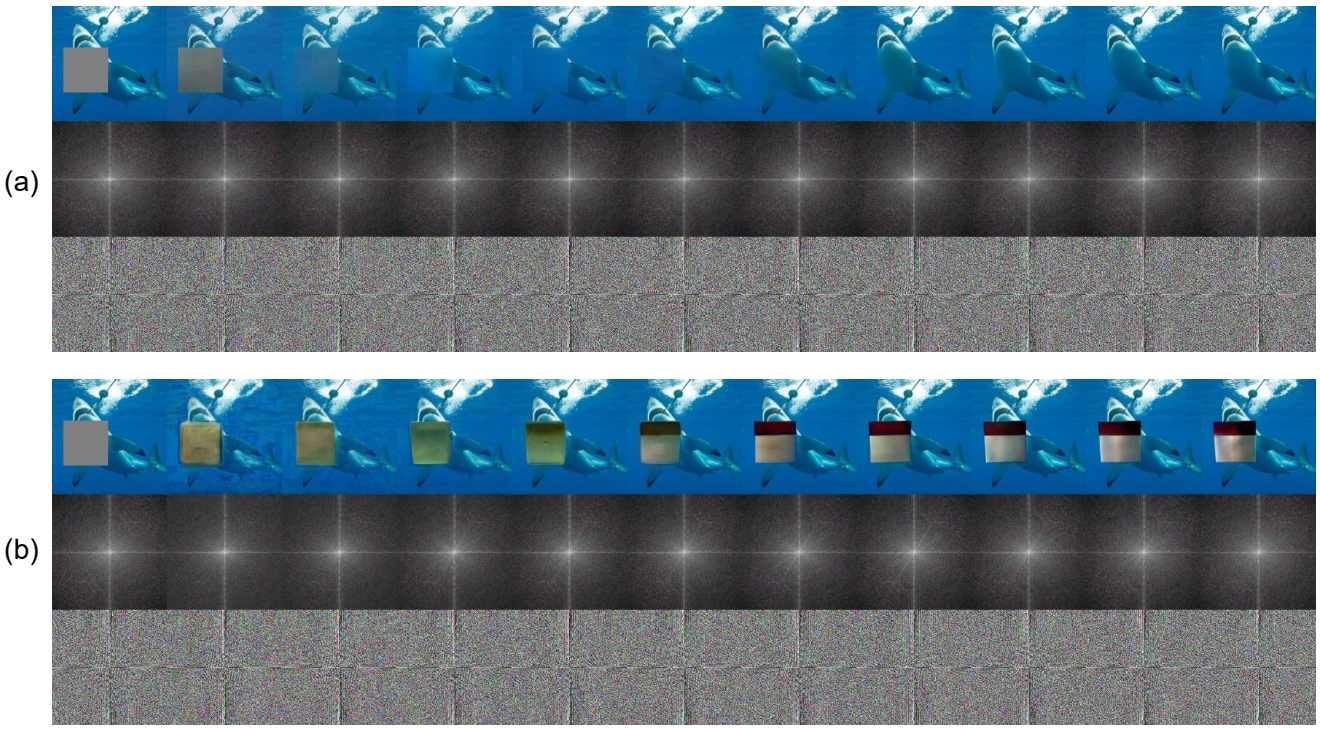


Figure S-2. Image inpainting using standard and truncated timestep scheduling. The process evolution is shown over iterations with the middle row depicting the log-magnitude spectrum and the bottom row the phase spectrum. (a) The standard scheduling exhibits cohesive restoration with detail fidelity. (b) The truncated scheduling results in detail-rich patches that are perceptually incongruent with the original image context.

formance on tasks including body pose completion and motion denoising.

HPS The HPS dataset [21] contains over 300K synchronized RGB images, paired with reference 3D poses and locations, captured from seven people interacting with large-scale 3D scenes. The dataset includes motion sequences of various activities such as exercising, reading, eating, lecturing, using a computer, making coffee, and dancing. Following Pose-NDF [59], we use the HPS dataset to evaluate the motion denoising task without training on it. After sampling, we got 350 sequences, each consisting of 60 frames for testing.

F.2. Hand-only Dataset

FreiHAND FreiHAND [70] is a large-scale dataset for 3D hand pose estimation, focusing on single-hand poses. It includes 130,240 training samples (4×32560) and 3,960 evaluation samples. Each training hand pose is accompanied by 4 RGB images, providing diverse data for training robust models. We use 32,560 hand poses, represented in the MANO format [53], for training the DPoser-hand model. The remaining 3,960 evaluation samples are used to assess hand mesh recovery performance during testing.

DexYCB DexYCB [3] is a dataset for capturing 3D hand poses during hand-object interactions, focusing on single-hand poses. We use only the hand poses for training the DPoser-hand model. The training set includes 407,000 single hand poses.

HO3D HO3D [22] is a dataset that provides 3D annotations for both hand poses and object interactions. Similar to DexYCB, we utilize the hand poses for training the DPoser-hand model. The training set contains 83,000 hand poses.

H2O The H2O dataset [33] provides 3D pose annotations for two-hand and object interactions. For the purpose of training DPoser-hand, we use only the right-hand poses. The training set contains 58,000 hand poses.

ReInterHand ReInterHand [45] is a high-quality synthetic dataset designed for 3D hand pose estimation, specifically focusing on interacting hands. It includes annotations for both hands. For training DPoser-hand, we flip the left-hand pose as right-hand to unify the format. The dataset is split into training, validation, and test sets with an 8:1:1 ratio. We use approximately 186,000 hand poses for training, and 23,000 poses for testing. The test set is used to evaluate hand inverse kinematics tasks.

F.3. Face-only Dataset

MICA The MICA dataset [69] consists of eight smaller datasets that were unified to represent about 2315 subjects using the FLAME [36] model. It contains only shape geometry. We use the MICA dataset to train the shape component of DPoser-face, focusing on high-quality 3D face shapes.
WCPA WCPA [30] is a large-scale dataset focusing on 3D face reconstruction under perspective projection. It contains 200 subjects and 356,640 training instances, with detailed annotations for facial expressions. We use WCPA to train the expression component of DPoser-face, with 1/10 of the dataset reserved for testing. The test set is used to evaluate face reconstruction, considering both shape and expressions.
NOW NOW [54] is a widely-used benchmark for face reconstruction. It introduces standard evaluation metrics for assessing the accuracy and robustness of 3D face reconstruction methods, especially under variations in viewing angle, lighting, and occlusions. The validation set containing 352 images is employed for our face reconstruction task. We focus on the non-metrical evaluation of face shape, as the ground truth (GT) only includes shape. As in previous works such as DECA [17], expressions are set to zero in the FLAME [36] model to obtain a neutral face mesh for final evaluation.

F.4. Whole-body Dataset

BEAT2 BEAT2 [39] is a holistic co-speech dataset that combines the MoShed SMPL-X [48] body with FLAME head parameters. It refines the modeling of head, neck, and finger movements, providing high-quality 3D motion-captured data. After de-duplication, we have 1.48 million whole-body poses for training DPoser-X, and the test set contains 172,000 whole-body poses used for whole-body pose completion.
GRAB GRAB [58] is a dataset containing full 3D shape and pose sequences of 10 subjects interacting with 51 everyday objects of varying shapes and sizes. We use this dataset for training DPoser-X, with 391,000 whole-body poses. The data helps train the model for whole-body mesh recovery during grasping actions.
ARCTIC ARCTIC [15] is a dataset focused on two-hand object manipulation, with 2.1 million video frames paired with 3D hand and object meshes. After de-duplication, we have 77,000 whole-body poses for training DPoser-X. The validation set, which contains 10,000 whole-body poses, is used for testing whole-body mesh recovery and pose completion. Note that the face expressions are not annotated, so we set models' output expressions as zeros for evaluation. Face-related metrics in whole-body mesh recovery are only influenced by the human shape.
EgoBody EgoBody [65] is a large-scale dataset that captures 3D human motions during social interactions in 3D scenes. It provides SMPL-X [48] annotations for 3D whole-body pose, shape, and motion for both the interactee and the camera wearer. The training set contains 38,896 instances of whole-body poses (x2 for each subject), and the test set contains 24,665 instances. The test set is used for the whole-body pose completion task.
Fit3D Fit3D [18] is a dataset with over 3 million images and corresponding 3D human shape and motion capture ground truth data, covering 37 exercises performed by instructors and trainees. We take the subject *s04* which consists of 612 images after sampling, for whole-body mesh recovery tasks to test DPoser-X's generalization, without training the model on this dataset.
EHF EHF [48] is a curated dataset comprising 100 images with pseudo whole-body poses. Following Pose-NDF [59], we use this dataset to evaluate body mesh recovery performance, specifically calculating PA-MPJPE for body joints.

G. Experimental Details

In this section, we provide detailed descriptions of the experimental setups and specific loss functions for various tasks. These tasks include pose completion, motion denoising, inverse kinematics, face reconstruction, hand mesh recovery, and whole-body mesh recovery. In addition, we explain the evaluation metrics used in each task and implementation details of comparative methods.

G.1. Pose Completion

For partial observations $\mathbf{y}$, the measurement operator $\mathcal{A}$ is modeled as a known mask matrix $M \in \mathbb{R}^{d\times n}$. Based on our optimization framework denoted in Alg. 1, we define the task-specific loss, L_{comp}, as follows:

$$L_{\text{comp}} = ||M\mathbf{x}_0 - \mathbf{y}||_2^2. \tag{5}$$

Here, $\mathbf{x}_0$ denotes the complete body pose θ we try to recover, where the unseen parts are initialized as random noise. In the following ablated studies, if not specified, the evaluation of the body pose completion is performed using 10 hypotheses on the AMASS dataset [42] with left leg occlusion.

G.2. Motion Denoising (Noisy Input)

Adhering to Pose-NDF settings [59], we aim to refine noisy joint positions J^t_{obs} over N frames to obtain clean poses θ^t, initialized from mean poses in SMPL with small noise. We formulate the task-specific loss combining an observation fidelity term L_{obs} and a temporal consistency term L_{temp}:

$$L_{\text{obs}} = \sum_{t=0}^{N-1} ||M_J(\theta^t, \beta_0) - J^t_{\text{obs}}||_2^2, \tag{6}$$

$$L_{\text{temp}} = \sum_{t=1}^{N-1} ||M_J(\theta^{t-1}, \beta_0) - M_J(\theta^t, \beta_0)||_2^2, \tag{7}$$

where M_J denotes the 3D joint positions regressed from SMPL [41] and β_0 is the constant mean shape parameters.

G.3. Motion Denoising (Partial Input)

This task focuses on reconstructing clean poses, θ^t, from partially observed joint positions, J^t_{obs}, across N frames, employing a known mask matrix to identify visible joints. The optimization objective mirrors that of motion denoising (Section G.2), but incorporates a mask in Eq. (6) to specifically target visible parts, ensuring that only these segments guide the recovery process.

G.4. Inverse Kinematics

Inverse kinematics (IK) aims to estimate clean poses from noisy or partially observed 3D joint positions, similar to the motion denoising task. The key difference in the implementation is that the inputs are single-frame data, meaning the temporal consistency term L_{temp} is not required.

For inverse kinematics applied to hand poses, we optimize only the hand poses while keeping the hand shape parameters fixed. This simplifies the optimization, focusing solely on pose adjustments. For the face, we optimize both the face expression and shape parameters, as face-related tasks require accurate modeling of both shape and dynamic expressions. In all cases, we employ a similar optimization framework as in the motion denoising task, using only the fidelity loss for observed 3D joints.

G.5. Face Reconstruction

Reconstructing human faces using only 2D keypoints is challenging and typically insufficient for high-quality reconstructions. To address this, we utilize the photometric optimization approach described in [16] to fit a textured FLAME model [36]. The optimization aims to refine the face shape and expression parameters, as well as adjust the appearance and lighting parameters for the face rendering. We use a combination of two key loss functions: the photometric loss (L1-loss between rendered and target images) and the reprojection loss (for 2D face keypoints).

We observe that face shape plays a crucial role in tasks like face reconstruction. To this end, DPoser-face is designed to separately model face shape and expression. Given that these two components (shape and expression) are largely independent, we train the face shape and expression models separately using the WCPA [30] and MICA [69] datasets, respectively. For face-only tasks such as face reconstruction, DPoser regularization is applied to both the face shape and expression models. It is important to note that only the expression component of DPoser-face contributes to the broader DPoser-X framework, with the shape component being reserved for face-specific tasks. For a fair comparison, we implement the same strategy for training the VPoser-face model.

G.6. Hand Mesh Recovery

For hand mesh recovery, we optimize the hand poses using the MANO model [53] instead of the SMPL model. Similar to the body mesh recovery task, we employ a reprojection loss based on 2D hand keypoints. In addition to using our DPoser loss for plausible hand poses, we also employ the L2 prior for hand shape, similar to Eq. (11) in the main text, to maintain natural hand geometry.

G.7. Whole-body Mesh Recovery

Whole-body mesh recovery shares similarities with body mesh recovery (as discussed in Section 2.5) but additionally incorporates the face and hands into the optimization. The goal is to recover the whole-body poses θ (including body, hands, and face) and shape parameters β by optimizing a reprojection loss based on whole-body 2D keypoints. A distinguishing feature is the inclusion of two root-relative reprojection losses, one for the hands and another for the face, to refine local poses. Specifically, the wrists for hands and the mouth for the face are chosen as the root, and the root coordinates are subtracted before calculating the reprojection losses. This ensures that the hand and face poses are localized relative to the body, improving the accuracy of hand and facial mesh recovery.

G.8. Evaluation Metrics

For comprehensive assessment across various tasks, following recent works like NRDF [23] and SMPLer-X [2], we adopt task-specific metrics:

- *Pose Generation*: Diversity and fidelity are evaluated using Average Pairwise Distance (APD) and d_{NN} [23], respectively. d_{NN} measures the distance between the generated pose and its nearest neighbor from the training data. We also report the common metrics for generative models, including FID [24] (distribution similarity), Precision [34] (fidelity), and Recall [34] (diversity).
- *Human Mesh Recovery*: Procrustes-aligned Mean Per-Vertex Position Error (PA-MPJPE) and Procrustes-aligned Mean Per-Joint Position Error (PA-MPVPE) measures the accuracy of recovered human meshes.
- *Multi-hypothesis Pose Completion*: MPVPE and APD on masked parts across multiple hypotheses measure solution accuracy and diversity, respectively.
- *Motion Denoising & Inverse Kinematics*: Both MPJPE and MPVPE are calculated to assess the performance.

All errors are reported in *millimeter* units.

G.9. Implementation of Comparative Methods

In pose generation experiments, we employ standard sampling techniques for generative models, including GMM [1], VPoser (VAE) [48], and GAN-S (GAN) [11]. For Pose-NDF [59] and NRDF [23], we reproduce their projection algorithms using their official repositories. For other

tasks during testing, to ensure a fair comparison, we implement all pose priors within the same optimization framework—using identical task-specific loss functions and optimization iterations—while tuning hyperparameters such as loss weights for each method.

VPoser and GAN-S function as pose priors due to their learned meaningful latent representations. We optimize the pose latents for both methods. Given VPoser's Gaussian assumption, it naturally incorporates L2 regularization on the latent pose [48]. However, we observe that applying spherical loss to the latents of GAN-S [11] degrades human mesh recovery performance. Therefore, we use only GAN-S's generator for decoding without imposing additional constraints on the pose latents. Both NDF [59] and NRDF [23] directly optimize pose rotation representations by minimizing the predicted distance between the current pose and their learned plausible pose fields. We implement these methods using their official code and model weights. Since GAN-S does not provide pre-trained models, we train it from scratch on the same datasets as our DPoser. Additionally, for hand, face, and whole-body models, we train the comparative methods ourselves.

H. Additional Experiments

In this section, we present a series of additional experiments that further demonstrate the efficacy of DPoser-X across various tasks. These experiments cover body mesh recovery, body pose completion, motion denoising, hand/face generation, hand/face inverse kinematics, face reconstruction, and whole-body mesh recovery, with a focus on different input types and datasets.

H.1. Body Mesh recovery

In addition to the priors compared in the main text, we evaluate DPoser against two recent state-of-the-art, generation-based methods: GFPose [8] and HuProSO3 [14]. Unlike optimization-based priors, these methods are designed to produce multiple, diverse hypotheses for a given input.

We report the results for the Human Mesh Recovery (HMR) task on the EHF dataset [48] in Table S-4. The comparison is conducted with both a single hypothesis (hypotheses_num=1) and multiple hypotheses (hypotheses_num=10), reporting the minimum PA-MPJPE and MPJPE. The results clearly show that while GFPose and HuProSO3 can generate diverse potential poses, our DPoser achieves significantly higher accuracy (i.e., lower error) in both evaluation settings. This suggests that DPoser provides a more precise and reliable pose prior for this task.

H.2. Body Pose Completion

In practical scenarios, HMR algorithms often grapple with occlusions leading to incomplete 3D pose estimates. In this context, the task is to recover full 3D poses from partially observed data, initializing the occluded parts with noise. Our DPoser model is employed to refine these initially implausible poses into feasible ones, utilizing an L2 loss on the visible parts to ensure data consistency. In parallel, we employ a comparable optimization strategy for both Pose-NDF [59] and VPoser [48]. As a task-specific baseline, we adapt the original VPoser model into CVPoser by incorporating conditional inputs within its VAE framework [32] for end-to-end training and conditional sampling. The completion experiment is conducted on the AMASS dataset [42] with occlusion of various body parts.

Methods	hypotheses_num=1	hypotheses_num=10
GFPose [8]	68.64/89.88	62.80/83.39
HuProSO3 [14]	72.00/104.52	57.42/84.21
DPoser (ours)	**56.05/79.82**	**53.28/76.53**

Table S-4. Comparison with generation-based methods on the HMR task using the EHF dataset [48]. We report the minimum PA-MPJPE/MPJPE across multiple hypotheses.

Given the uncertainties in this task, we generate multiple hypotheses and evaluate them using minimum, mean, and standard deviation errors against the ground truth. We calculate APD across solutions to assess diversity. As illustrated in Table S-5, DPoser exhibits superior performance across different occlusion scenarios compared to existing pose priors and even the task-specific CVPoser, highlighting its effectiveness in pose completion. The qualitative evaluations are presented in Fig. S-3. Here, we observe that DPoser can generate a multitude of plausible poses, a capability lacking in VPoser [48]. Pose-NDF [59], meanwhile, struggles with generalizing to unseen noisy poses and making plausible adjustments from the mean pose initialization.

H.3. Motion Denoising (Noisy Input)

To further evaluate DPoser's performance in motion denoising, we extend our analysis to scenarios with varying noise levels. In complement to the results presented in Table 3 of our main text, we conduct an in-depth examination that spans a broader range of noise conditions. The extended results, detailed in Table S-6, showcase DPoser's exceptional performance against state-of-the-art (SOTA) pose priors, especially under high noise conditions, manifesting DPoser's resilience to noise.

H.4. Motion Denoising (Partial Input)

We next assess the performance of our model in scenarios involving partial input using the AMASS dataset [42]. Two types of occlusions were considered: legs and left arm. The quantitative results of these experiments are presented in Table S-7, while visual examples can be found in Section L. Errors (in *cm*) are evaluated in terms of MPJPE across visible (Vis.), occluded (Occ.), and all joints, along

Methods	Occ. left leg		Occ. legs		Occ. arms		Occ. torso	
	MPVPE ↓	APD ↑	MPVPE ↓	APD ↑	MPVPE ↓	APD ↑	MPVPE ↓	APD ↑
Pose-NDF [59] ($S = 1$)	168.61	NAN	169.92	NAN	261.11	NAN	115.03	NAN
Pose-NDF ($S = 5$)	157.62/168.49/7.94	1.95	162.30/169.94/5.54	1.96	254.97/261.01/4.38	1.22	108.07/114.98/4.98	0.93
Pose-NDF ($S = 10$)	154.21/168.45/8.66	1.95	159.75/169.86/6.12	1.97	252.90/260.94/4.81	1.20	105.87/114.97/5.43	0.93
VPoser [48] ($S = 1$)	200.23	NAN	221.21	NAN	206.83	NAN	58.66	NAN
VPoser ($S = 5$)	187.38/200.73/10.52	2.38	201.70/221.16/14.57	5.49	191.27/206.55/11.54	4.06	49.88/58.67/6.71	1.59
VPoser ($S = 10$)	182.31/200.51/12.20	2.41	195.76/221.34/16.40	5.44	186.55/206.72/12.91	4.08	47.31/58.71/7.38	1.56
CVPoser† ($S = 10$)	113.48/128.04/10.36	1.91	121.00/134.35/10.17	2.43	153.12/162.82/5.58	1.08	45.16/51.23/4.32	0.57
DPoser ($S = 1$)	78.78	NAN	103.12	NAN	104.59	NAN	44.60	NAN
DPoser ($S = 5$)	46.23/78.13/24.96	**6.58**	72.37/102.73/23.05	7.72	74.32/105.70/24.15	5.67	27.47/44.63/13.26	2.19
DPoser ($S = 10$)	**38.79**/78.31/27.13	6.53	**63.65**/102.46/25.39	**7.75**	**64.72**/104.94/26.44	**5.69**	**22.63**/44.60/14.65	**2.21**

Table S-5. Performance metrics (min/mean/std of MPVPE and APD) for body pose completion on the AMASS dataset [42] under varying occlusion scenarios. S denotes the number of hypotheses. † Task-specific baseline trained with partial poses as conditional input.

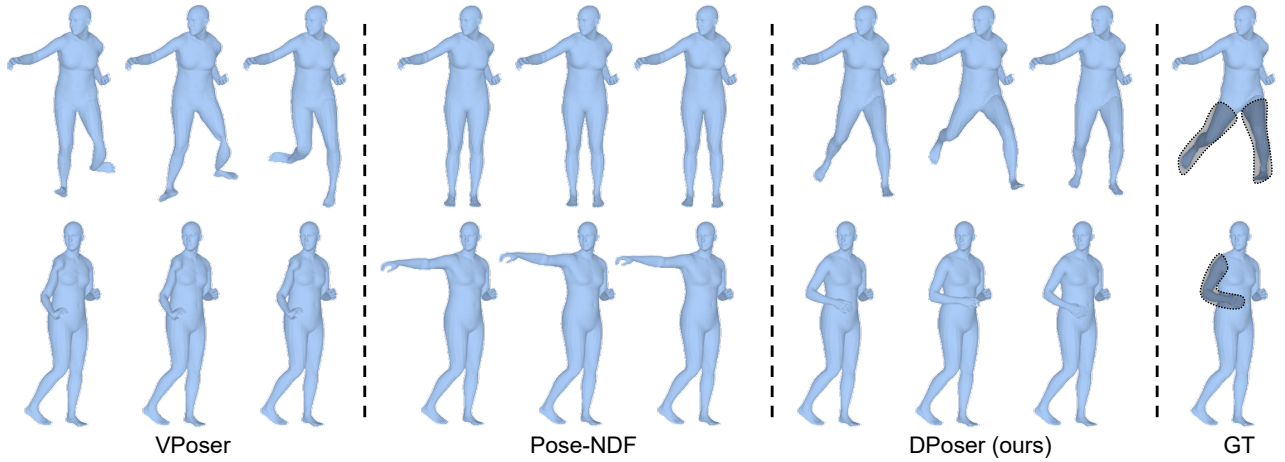


Figure S-3. Visual comparisons of body pose completion. Three hypotheses are drawn for each method. DPoser uniquely offers multiple plausible solutions for partial poses, a scenario where competitors often struggle due to limited generalization.

Methods	AMASS [42]		HPS [21]	
	20mm	100mm	20mm	100mm
w/o prior	15.33	51.48	16.26	50.87
VPoser [48]	15.20	49.10	17.24	46.69
Pose-NDF [59]	13.84	46.10	15.62	47.50
DPoser	**13.64**	**33.18**	**13.45**	**35.32**

Table S-6. Performance comparison of motion denoising under varying noise scales. MPJPE is reported afters denoising.

Methods	Occlusion	MPJPE			MPVPE
		Vis.	Occ.	All	All
w/o prior	legs	0.26	14.72	5.52	5.45
VPoser [48]	legs	1.75	14.29	6.31	7.38
Pose-NDF [59]	legs	**0.25**	15.71	5.87	5.64
DPoser	legs	0.28	**12.24**	**4.63**	**3.65**
w/o prior	left arm	0.26	24.87	4.74	9.91
VPoser [48]	left arm	1.21	13.23	3.40	7.68
Pose-NDF [59]	left arm	**0.25**	17.70	3.42	7.86
DPoser	left arm	0.27	**7.80**	**1.64**	**3.81**

Table S-7. Comparative analysis of methods for motion denoising with different occlusions (legs or left arm) on the AMASS dataset [42].

with MPVPE for all vertices.

In the leg occlusion scenario, where the AMASS dataset primarily consists of straight poses, the lack of diversity allows for reasonable results even without incorporating a pose prior. In this case, the optimization starts from an initial point that closely matches these common poses. However, while VPoser's mean-centered approach struggles to faithfully replicate visible areas, DPoser accurately handles the visible portions and guides the reconstruction of occluded parts, yielding more realistic results. In contrast, Pose-NDF does not effectively enhance the occluded regions. For left arm occlusions, which involve more varied movements, DPoser markedly surpasses other methods, underlining its adaptability and precision in handling diverse

Methods	APD ↑	FID ↓	Prec. ↑	Rec. ↑	d_{NN} ↓
VPoser [48]	1.99	0.21	0.68	0.65	1.85
NRDF [23]	1.76	5.20	0.17	0.65	5.37
DPoser-hand	**2.36**	**0.01**	**0.82**	**0.87**	**1.45**

Table S-8. Quantitative evaluation of hand pose generation.

motion patterns.

H.5. Hand Pose Generation

We evaluate the generated hand poses based on their diversity and realism. As shown in Table S-8, DPoser produces a strong combination of both, outperforming methods like VPoser [48] and NRDF [23]. Specifically, NRDF shows poor realism, reflected in high FID and d_{NN} scores. VPoser, while achieving moderate precision, suffers from limited diversity, as indicated by its low APD. See Fig. S-12 visualization comparison.

H.6. Hand Mesh Recovery

To evaluate DPoser's ability to recover hand meshes, we test its performance on the FreiHAND dataset [70] under two initialization strategies: mean poses and the Hand4Whole [44] prediction poses. The results, detailed in Table S-9 and visually represented in Section L (Fig. S-15 and Fig. S-14), show DPoser's superior performance across various metrics and initialization settings.

DPoser consistently outperforms competing methods, such as VPoser [48] and NRDF [23], achieving the lowest PA-MPJPE and PA-MPVPE values. For example, when using keypoints detected by RTMPose [27], DPoser reduces PA-MPJPE by 20% compared to VPoser. Moreover, the performance is further enhanced when using Hand4Whole [44] initialization, highlighting DPoser's ability to refine results from existing SOTA mesh recovery models. In contrast, methods like the L2 prior and VPoser, which rely on mean-centered priors, fail to match the quality of the initializations, producing poorer results. Additionally, DPoser demonstrates significant advantages over NRDF in modeling hand pose distributions, offering more reliable guidance in mesh recovery. By leveraging ground truth (GT) keypoints, DPoser consistently recovers natural hand meshes that align well with observed 2D keypoints.

H.7. Hand Inverse Kinematics (Noisy Input)

For hand inverse kinematics, we extend our experiments to noisy settings using the ReInterHand dataset [45]. Table S-10 shows that DPoser consistently outperforms alternative methods across different noise levels (2mm, 5mm, 10mm), achieving the lowest MPVPE and MPJPE. While methods like the L2 prior and VPoser [48] perform competitively at lower noise levels, their accuracy deteriorates significantly as noise increases. In contrast, DPoser maintains both stability and precision, showcasing its superior ability to handle noisy input and recover plausible hand poses even under challenging conditions.

H.8. Face Generation

We conduct the face generation experiments for the shape and expression separately since they are uncorrelated attrbutes. As detailed in Table S-11, DPoser outperforms VPoser [48] in terms of FID, achieving values of 5.331 for shape and 0.156 for expression, which highlights DPoser's superior ability to model the distribution of face shapes and expressions. While VPoser achieves higher precision scores, its recall values are considerably lower, indicating a lack of variability in the generated samples. This observation is further corroborated by qualitative results shown in Fig. S-4, which demonstrate DPoser's ability to generate a wide variety of realistic face shapes and expressions. Compared to VPoser, DPoser captures a broader range of subtle variations, especially in expressions, while maintaining fidelity.

H.9. Face Reconstruction

For face reconstruction, along with the NOW beachmark [54], we test on the WCPA [30] dataset, which evaluates both face shape and expression. As shown in Table S-12, DPoser consistently outperforms other methods. It achieves the lowest PA-MPVPE and PA-MPJPE errors across all configurations, with a notable reduction in errors for both overall and side-view cases. When combined with EMOCA [9] initialization, DPoser further refines the reconstruction quality, reducing the mean PA-MPVPE error to 3.10 mmm compared to 3.58 mm for EMOCA alone.

Qualitative visualizations in Fig. S-17 illustrate DPoser's ability to reconstruct detailed and realistic face meshes, even in challenging scenarios involving variations in side-view poses and complex expressions. While other methods often struggle to generalize across such cases, DPoser remains robust and highly accurate, demonstrating its capability to handle the full diversity of facial shapes and expressions in real-world conditions.

H.10. Face Inverse Kinematics

To evaluate DPoser's robustness in face inverse kinematics, we conduct experiments under various noise levels and occlusion scenarios using the WCPA dataset [30]. The results in Table S-13 demonstrate that DPoser consistently achieves the lowest MPVPE and MPJPE errors across all tested conditions. Notably, DPoser retains its strong performance even under extreme noise conditions, whereas VPoser [48] experiences significant degradation as noise levels increase. Qualitative results, visualized in Fig. S-16, further confirm DPoser's ability to reconstruct realistic and aligned facial

Methods	PA-MPJPE ↓	PA-MPVPE ↓	F@5 ↑	F@15 ↑
w/o prior	17.71/16.12	18.40/17.04	0.396/0.446	0.875/0.895
L2 prior	12.87/11.49	12.71/11.59	0.512/0.533	0.924/0.927
VPoser [48]	12.31/10.62	12.23/10.91	0.524/0.609	0.931/0.943
NRDF [23]	13.19/11.04	13.39/11.59	0.469/0.554	0.914/0.937
DPoser-hand	**10.71/8.68**	**10.48/8.70**	**0.574/0.679**	**0.947/0.963**
hand4whole	8.50	7.81	0.651	0.97
+ w/o prior	9.13/6.04	8.97/6.06	0.609/0.749	0.965/0.985
+ L2 prior	9.91/7.16	9.69/7.11	0.568/0.686	0.953/0.974
+ VPoser [48]	9.13/6.42	9.04/6.55	0.605/0.717	0.964/0.981
+ NRDF [23]	9.00/6.15	8.99/6.29	0.595/0.726	0.964/0.983
+ DPoser-hand	**7.96/5.36**	**7.69/5.20**	**0.663/0.793**	**0.973/0.990**

Table S-9. Performance evaluation of hand mesh recovery on the FreiHAND dataset [70]. Results are reported using 2D keypoints detected by RTMPose [27] / ground truth.

Methods	2mm		5mm		10mm	
	MPVPE ↓	MPJPE ↓	MPVPE ↓	MPJPE ↓	MPVPE ↓	MPJPE ↓
No prior	3.95	1.50	5.82	3.46	8.62	5.97
L2 prior	2.10	1.43	4.06	2.92	6.06	4.27
VPoser [48]	2.47	1.36	4.15	2.85	6.32	4.40
NRDF [23]	2.67	1.40	4.57	3.11	7.18	5.06
DPoser-hand	**1.71**	**1.17**	**3.30**	**2.39**	**5.39**	**3.87**

Table S-10. Performance of hand inverse kinematics on the ReInterHand dataset [45] under noisy settings.

Methods	FID ↓	Prec. ↑	Rec. ↑	d_{NN} ↓
VPoser [48] (shape)	31.91	**0.984**	0.105	**6.52**
DPoser-face (shape)	**5.331**	0.689	**0.396**	8.29
VPoser [48] (expression)	0.888	**0.993**	0.019	**0.79**
DPoser-face (expression)	**0.156**	0.818	**0.697**	1.01

Table S-11. Quantitative evaluation for face generation.

Methods	all	side-view
w/o prior	3.67/4.19	3.77/4.46
L2 prior	3.56/3.90	3.58/4.01
VPoser [48]	3.59/4.01	3.62/4.13
DPoser-face	**3.34/3.65**	**3.32/3.61**
EMOCA [9]	3.58/4.07	3.78/4.43
+ w/o prior	3.49/3.88	3.92/4.56
+ L2 prior	3.49/3.82	3.68/4.28
+ VPoser [48]	3.39/3.65	3.56/4.05
+ DPoser-face	**3.10/3.54**	**3.16/3.72**

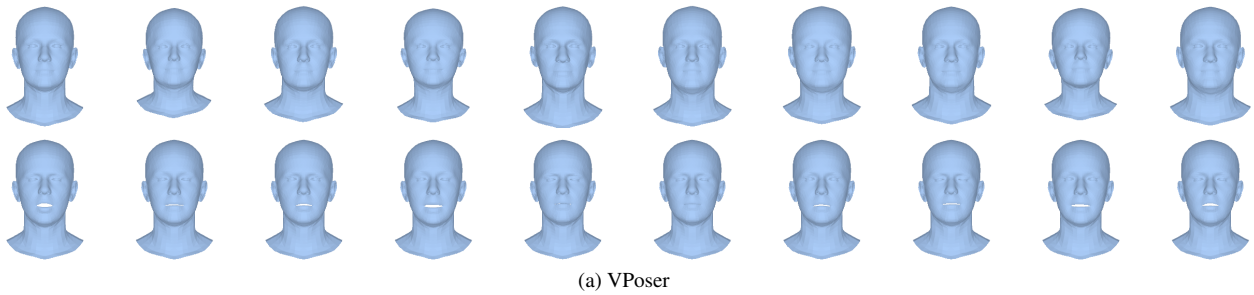
Table S-12. Face reconstruction performance (PA-MPVPE/PA-MPJPE) on the WCPA dataset [30].

details under noisy and occluded conditions.

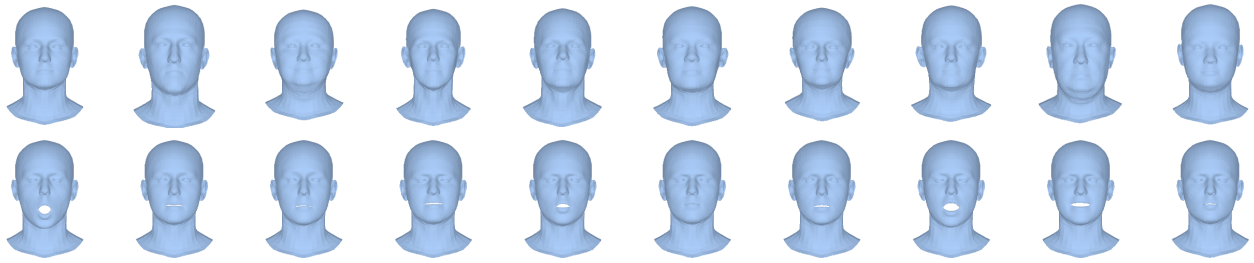
H.11. Whole-body Mesh Recovery

We extend our evaluation of whole-body mesh recovery to include the Fit3D dataset [18], in addition to the comparative results on ARCTIC [15]. For this evaluation, we compare DPoser-X with VPoser-X [48] and the GMM baseline, which utilizes a Gaussian Mixture Model (GMM) [1] for body poses and an L2 prior for hands and face.

As shown in Table S-14, DPoser-X outperforms both VPoser-X [48] and GMM [1] across most metrics, for both hands and the entire body. However, we observe that the L2-prior baseline performs better than DPoser-X in terms of PA-MPVPE on the face. We attribute this result to the low-resolution images in the Fit3D dataset, where the face is depicted with limited pixel density and the 2D keypoints are less expressive. In this case, the neutral face produced by the L2 prior is more likely to yield better results due to the lack of detailed facial features in the input. Nonetheless, DPoser-X still outperforms other methods in handling the full-body mesh recovery, showing its robustness in both body and hand mesh reconstruction.



(a) VPoser



(b) DPoser (ours)

Figure S-4. Visualization of face generation results. Top row shows varying face shapes; bottom row shows varying expressions.

Methods	1mm Noise		2mm Noise		5mm Noise		Half Face Occ.	
	MPVPE ↓	MPJPE ↓	MPVPE ↓	MPJPE ↓	MPVPE ↓	MPJPE ↓	MPVPE ↓	MPJPE ↓
w/o prior	1.460	0.878	2.230	1.702	4.701	4.028	0.752	0.632
L2 prior	1.121	0.865	1.626	1.288	2.570	2.344	0.698	0.512
VPoser [48]	1.153	0.803	1.688	1.480	2.716	2.688	0.671	0.361
DPoser-face	**0.784**	**0.584**	**1.098**	**0.963**	**1.902**	**1.936**	**0.427**	**0.228**

Table S-13. Performance of face inverse kinematics on the WCPA dataset [30] under noisy and occlusion settings.

Methods	PA-MPVPE↓			PA-MPJPE↓
	All	Hands	Face	Body
w/o prior	89.72	23.51	7.26	91.18
GMM [1] & L2 prior	86.95	18.22	5.38	83.58
VPoser-X [48]	81.96	17.59	6.37	86.50
DPoser-X	**70.91**	**15.83**	**5.27**	**74.33**
SMPLerX	25.49	18.89	2.85	28.30
+ w/o prior	24.72	11.92	2.78	22.98
+ GMM [1] & L2 prior	24.28	11.09	**2.58**	22.95
+ VPoser-X [48]	24.41	10.21	2.65	23.03
+ DPoser-X	**23.20**	**8.91**	2.62	**21.22**

Table S-14. Whole-body mesh recovery results on the Fit3d dataset [18].

I. Ablated DPoser's Training

This section dissects the impact of different rotation representations and normalization techniques on DPoser's performance. The ablation of training experiments is conducted for the DPoser-body model trained on AMASS [42].

Initially, we examine axis-angle representation, comparing various normalization strategies: min-max scaling, z-score normalization, and no normalization. Our findings, summarized in Table S-15, indicate that z-score normalization is generally the most effective. Subsequently, using this optimal normalization, we explore 6D rotations [67] as an alternative. As evidenced by Table S-16, axis-angle representation offers superior performance. This preference can be attributed to the effective modeling capabilities of diffusion models, which do not benefit much from a more continuous data representation.

Inspired by HuMoR [52], we experiment with integrating the SMPL body model [41] as a regularization term during training. Alongside the prediction of additive noise, as outlined in Eq. (4) in the main text, we employ a 10-step DDIM sampler [55] to recover a "clean" version of the pose, denoted as $\tilde{\mathbf{x}}_0$, from the diffused $\mathbf{x}_t$. The regularization loss aims to minimize the discrepancy between the original and

Normalization	Body Mesh Recovery	Body Pose Completion	Motion Denoising	
	PA-MPJPE ↓	MPJPE ($S = 10$) ↓	MPVPE ↓	MPJPE ↓
w/o norm	57.88	45.37/102.28/41.08	44.82	24.04
min-max	59.17	47.41/107.00/43.42	42.70	21.29
z-score	**56.49**	**34.37**/72.47/26.32	**38.57**	**20.24**

Table S-15. Comparative performance of normalization methods using axis-angle rotation representation across multiple tasks.

Representation	Body Mesh Recovery	Body Pose Completion	Motion Denoising	
	PA-MPJPE ↓	MPJPE ($S = 10$) ↓	MPVPE ↓	MPJPE ↓
axis-angle	**56.05**	**34.76**/72.41/26.09	**38.21**	**19.87**
6D rotations	57.54	40.89/81.43/27.31	38.44	20.12

Table S-16. Comparative performance of rotation representations using z-score normalization across multiple tasks.

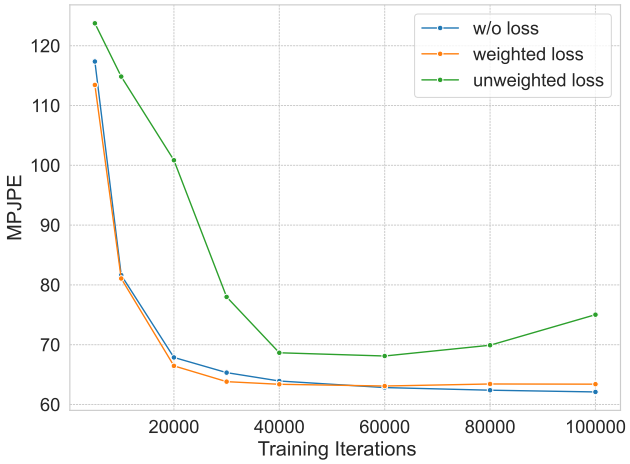


Figure S-5. MPJPE evolution in DPoser training with different regularization loss settings for body pose completion, assessed on AMASS [42] with 10 hypotheses under legs occlusion scenarios.

recovered poses under the SMPL body model M:

$$L_{\text{reg}} = ||M_J(\tilde{\mathbf{x}}_0, \beta_0) - M_J(\mathbf{x}_0, \beta_0)||_2^2 + ||M_V(\tilde{\mathbf{x}}_0, \beta_0) - M_V(\mathbf{x}_0, \beta_0)||_2^2. \quad (8)$$

Here, β_0 represents the mean shape parameters in SMPL. To account for denoising errors, we scale the regularization loss by $\log(1 + \frac{\alpha_t}{\sigma_t})$, thereby increasing the weight for samples with smaller t values (less noise).

Fig. S-5 visualizes the impact of this regularization on MPJPE during the training, specifically for pose completion tasks with occlusion of both legs. We observe that weighted regularization offers slight performance gains in the early training process, while the absence of weighting introduces instability and deterioration in results. Despite these insights, the computational cost of incorporating the SMPL model—especially for our large batch size of 1280—makes the training approximately 8 times slower. Therefore, we opted not to include this regularization in our main experiments.

Steps	Blocks	Hidden Dim	HMR (PA-MPJPE)	Runtime (s)
500	2	1024	56.05	17.34
250	2	1024	56.53	**8.44**
1000	2	1024	**55.74**	34.11
500	4	1024	56.47	18.72
500	2	2048	56.67	19.12

Table S-17. Ablation of DPoser's architecture and optimization steps for the HMR task.

We ablate the architectural hyperparameters of DPoser-body and the number of optimization steps on the HMR task, with results shown in Table S-17. Our findings indicate that a more complex architecture (i.e., 4 blocks or a 2048 hidden dimension) does not improve accuracy. Regarding the optimization, increasing the steps to 1000 offers the best accuracy (55.74 PA-MPJPE) but at a high computational cost (34.11s), while 250 steps are fastest but less accurate. Based on this analysis, we adopt the configuration of *500 steps, 2 blocks, and a 1024 hidden dimension* for our experiments, as it provides a solid trade-off between accuracy and runtime efficiency.

J. Extended DPoser's Optimization

In addressing pose-centric tasks as inverse problems, we propose a versatile optimization framework, which employs variational diffusion sampling as its foundational approach [43]. Our exploration extends to an array of diffusion-based methodologies for solving these complex inverse problems. Among the techniques considered are ScoreSDE [57], MCG [6], and DPS [5]. These methods augment standard generative processes with observa-

Methods	Occ. left leg	Occ. legs	Occ. arms	Occ. torso
ScoreSDE [57]	48.73/106.32/41.30	74.68/128.32/37.27	66.89/127.86/48.15	16.69/34.54/12.21
DPS [5]	40.51/104.32/54.57	64.26/113.46/33.71	60.63/119.85/42.78	15.10/33.90/13.27
MCG [6]	49.04/106.37/41.07	74.90/128.53/37.40	66.17/127.72/48.15	16.69/34.66/12.23
DPoser	**35.37**/74.01/26.47	**59.25**/96.77/24.55	**51.27**/81.76/20.04	**13.95**/28.57/9.85

Table S-18. Comparative evaluation of diffusion-based solvers for body pose completion on the AMASS dataset [42]. The min/mean/std of MPJPE are reported (hypotheses number $S = 10$).

tional data, either by employing gradient-based guidance or back-projection techniques. We compare these methods with our DPoser for body pose completion tasks. Our findings, captured in Table S-18, reveal that DPoser outperforms the competitors under most occlusion conditions. Consequently, DPoser emerges not merely as a universally applicable solution to pose-related tasks, but also as an exceptionally efficient one.

It is worth mentioning that methods rooted in generative frameworks [5, 6, 31, 57] can pose challenges for broader applicability in pose-centric tasks. For instance, in blind inverse problems—certain parameters in $\mathcal{A}$ (e.g., camera models in HMR) are unknown—generative methods are less straightforward to implement. ZeDO [28], a recent study focusing on the 2D-3D lifting task, adopts the ScoreSDE [57] framework and refines camera translations by solving an optimization sub-problem after each generative step. However, directly porting this strategy to HMR is non-trivial, owing to the added complexity of body shape parameter optimization—a feature currently absent in our DPoser model. Although some state-of-the-art techniques [7, 47] offer solutions by jointly modeling operator $\mathcal{A}$ and data distributions, a full-fledged discussion on this subject is beyond this paper's purview and remains an open question for future work.

K. Limitation and future work

A primary limitation of our work is the dependency on the training data's distribution. Our body pose prior is trained on the AMASS dataset [42], which, while diverse in common daily actions, contains limited examples of challenging or extreme poses like those found in yoga. This data imbalance leads to two main issues. First, the learned prior is inherently biased towards common standing poses. Second, when confronted with out-of-distribution inputs, as illustrated in Fig. S-6, the prior may offer limited or even incorrect guidance. This problem is often exacerbated by the failure of off-the-shelf 2D keypoint detectors like ViTPose [64] to produce accurate keypoints for such complex images, which in turn misguides the optimization.

Future work could address these data-driven limitations in several ways. To mitigate the action imbalance, techniques like clustering motions with action labels [50] and performing importance sampling during training could be effective. To improve robustness on challenging poses, incorporating more diverse training data and exploring more robust fitting strategies, such as using predicted dense depth maps for supervision, are promising directions.

Figure S-6. Failure cases of our method on challenging yoga poses. Inaccuracies in the estimated 2D keypoints (middle column), combined with our model's limited exposure to such out-of-distribution poses during training, lead to flawed 3D mesh reconstructions (right column).

Our framework also inherits certain limitations from the variational diffusion sampling [43] process it employs, most notably a tendency towards mode-seeking. For example, minimizing the DPoser regularization loss alone for "generation" results in a high Precision of 0.995 but a low Recall of 0.163. The low recall, compared to standard generative diffusion samplers (see Table 1 in the main text), indicates that the optimization framework captures the primary modes of the data distribution accurately but lacks diversity. To address this, future research could explore techniques like particle-based variational inference [40, 63] to enhance solution diversity. Finally, within the broader context of inverse problems we have framed, a plethora of existing methods [10] could be adapted to leverage our diffusion-based pose prior. Exploring these methods holds great potential for future progress.

L. More Qualitative Results

We show more qualitative results for body pose generation (Fig. S-7), body pose completion (Fig. S-8), body mesh recovery (Fig. S-9), motion denoising (Fig. S-10 and Fig. S-11), hand generation (Fig S-12), hand inverse kinematics (Fig S-13), hand mesh recovery (Fig S-14 and Fig S-15), face inverse kinematics (Fig S-16), face reconstruction (Fig S-17), whole-body pose generation (Fig S-18 and Fig. S-19), whole-body mesh recovery (Fig S-20), whole-body pose completion (Fig S-21 and Fig. S-22).

Figure S-7. Visualization of body pose generation. DPoser can generate diverse and realistic body poses.

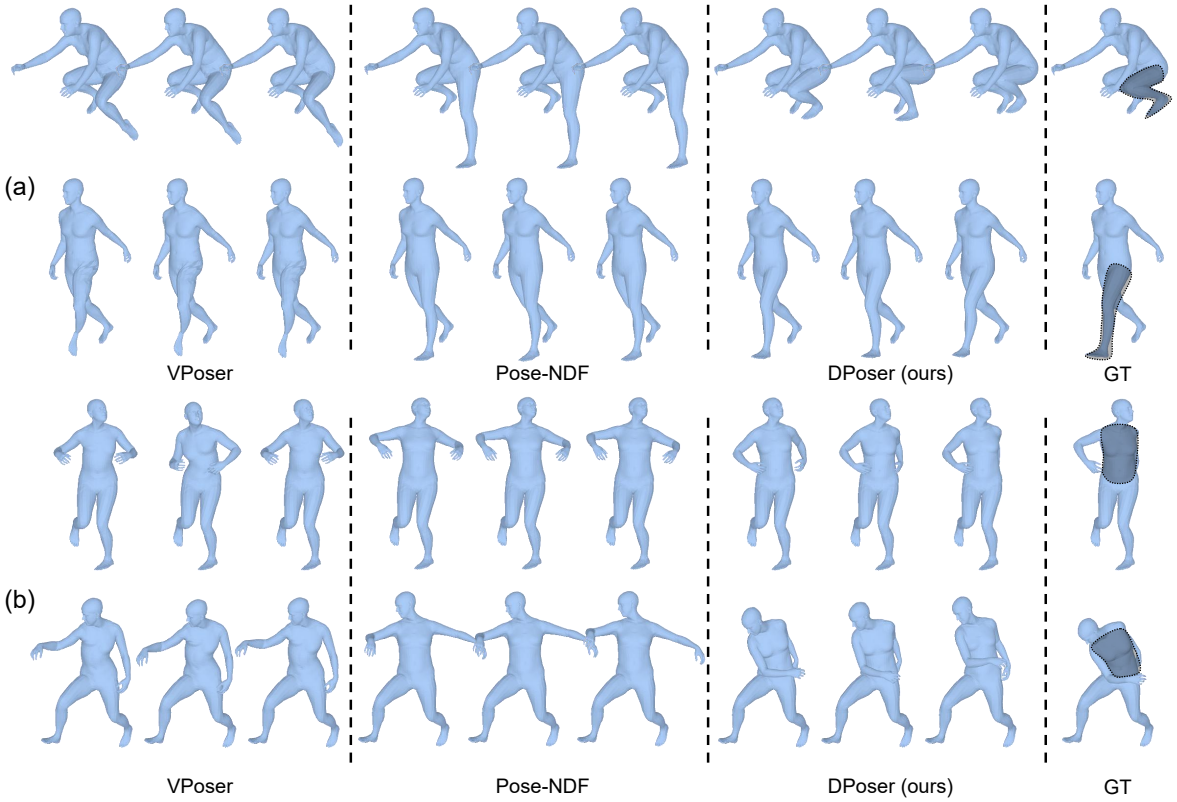


Figure S-8. Visualization of body pose completion. (a) Left leg under occlusion. (b) Torso under occlusion.

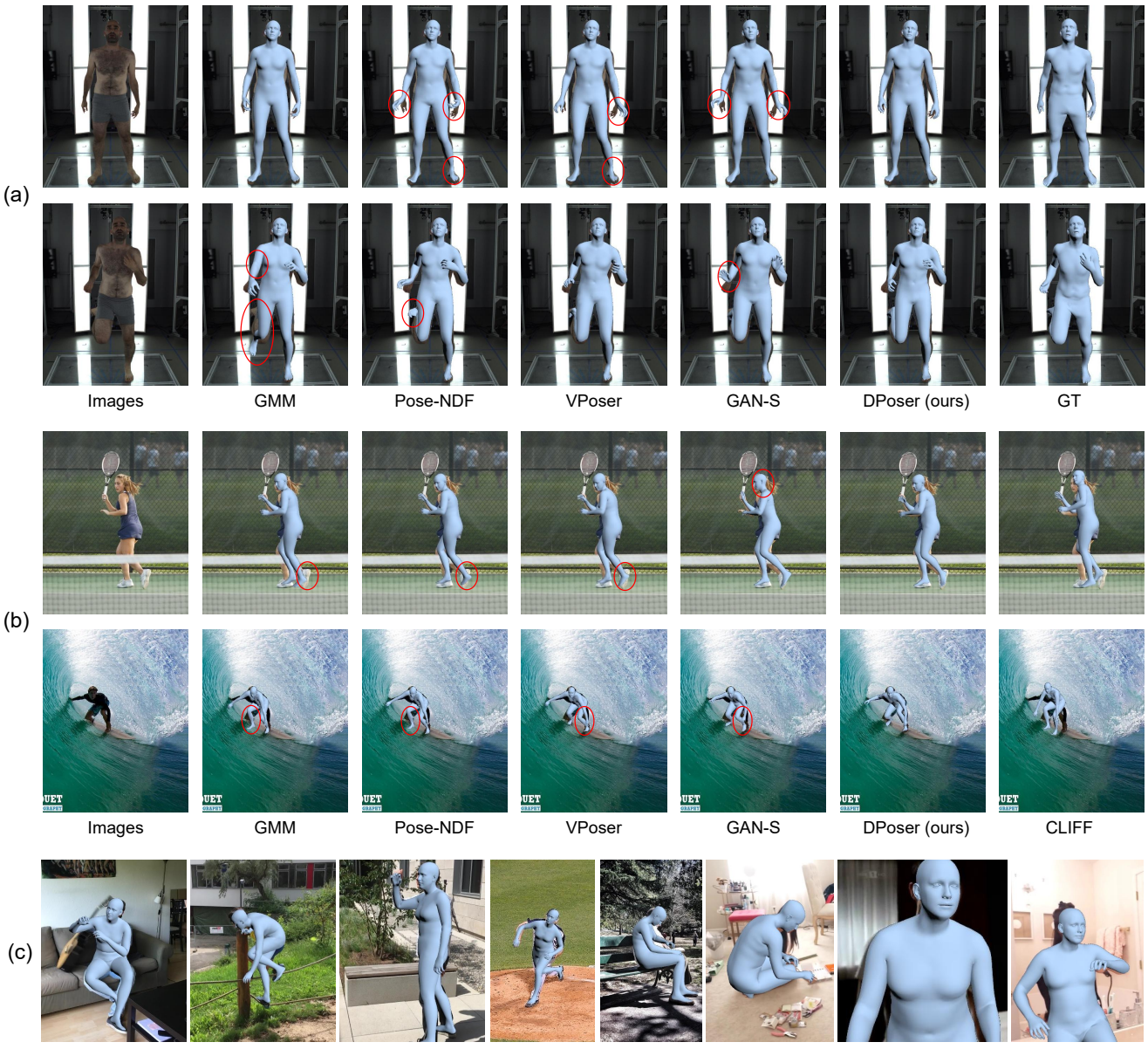
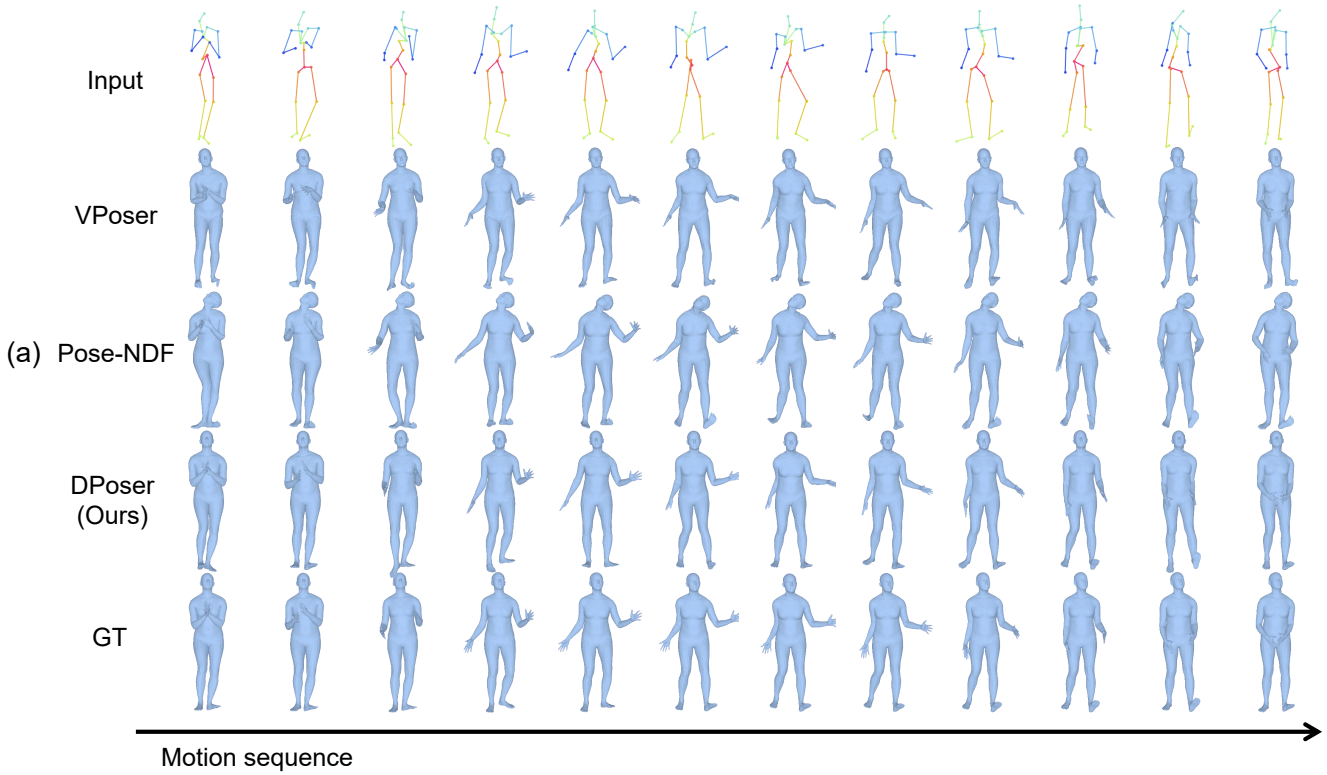
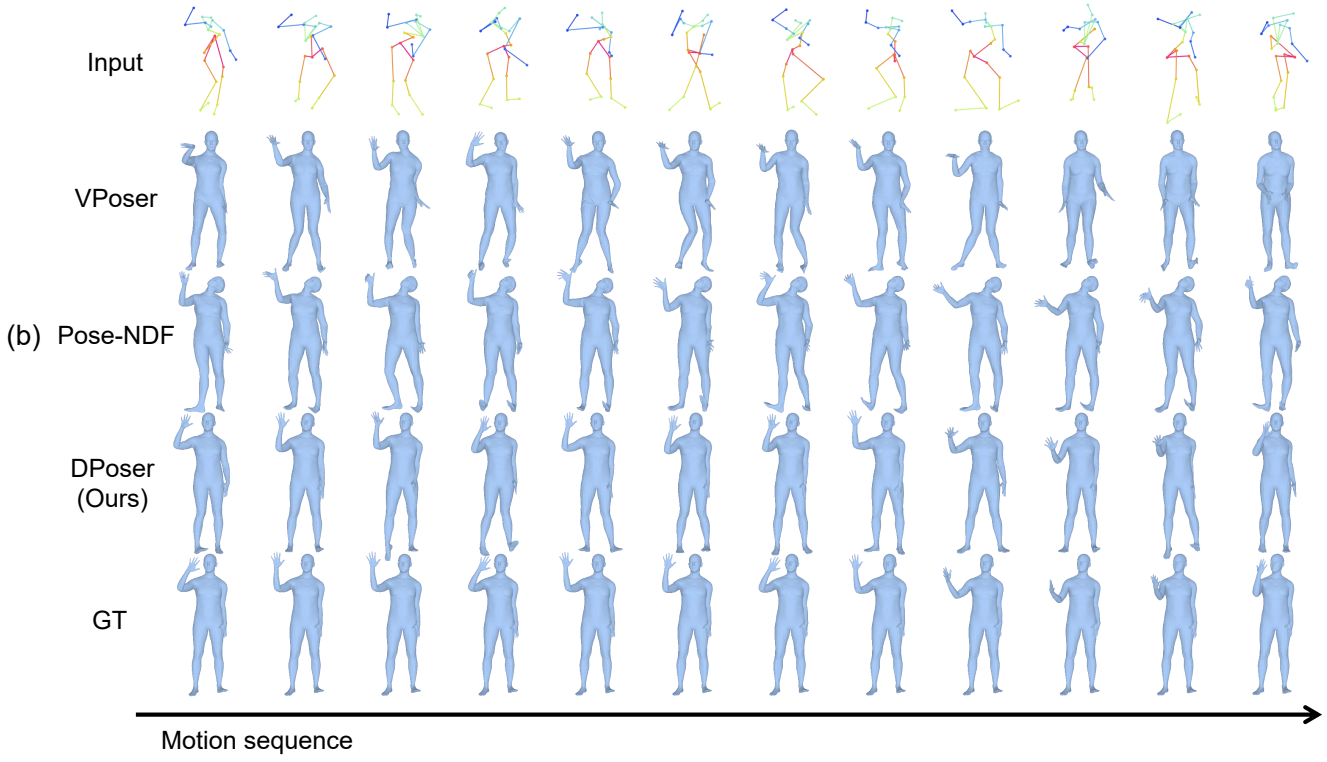


Figure S-9. Visualization of body mesh recovery. (a) Fitting from scratch. (b) Initialization using the CLIFF [37] prediction results. (c) More results of DPoser optimization with CLIFF initialization on in-the-wild images.

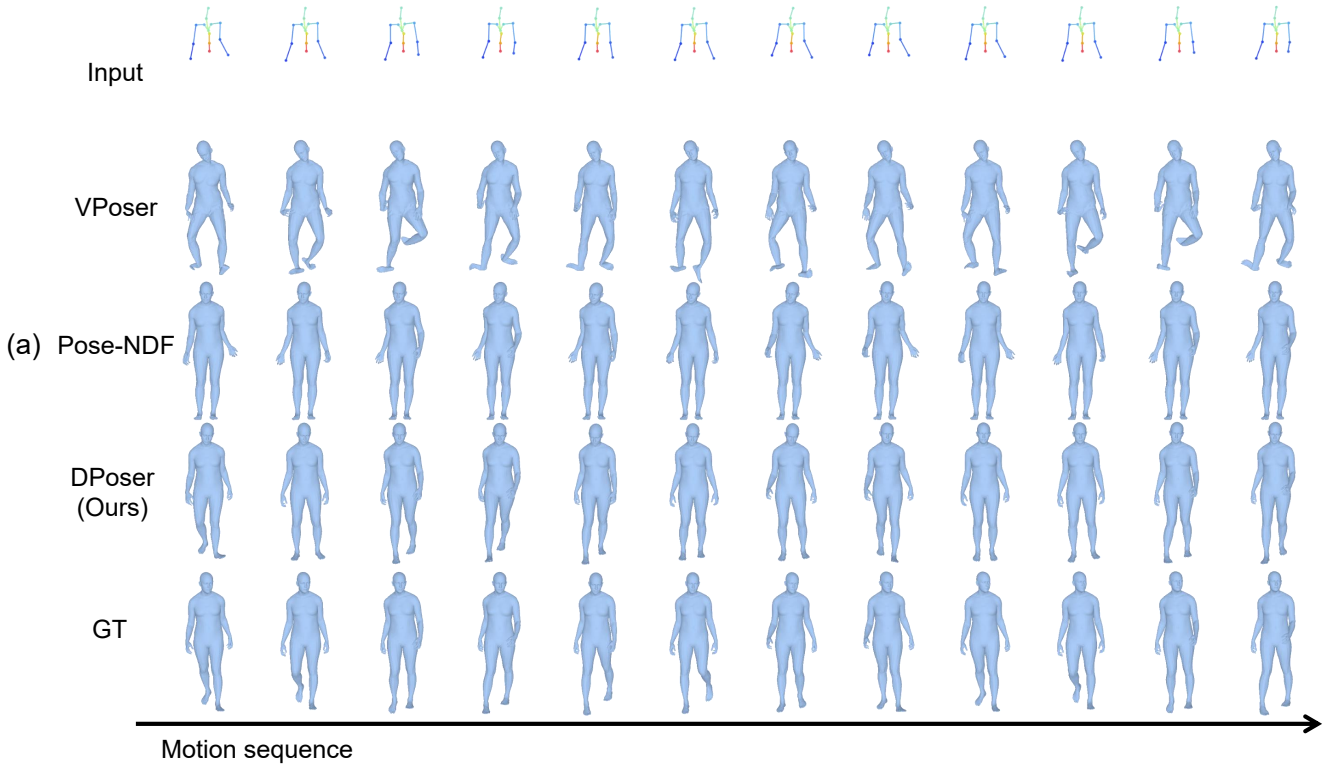


(a) Gaussian noise with 40 mm standard deviation.

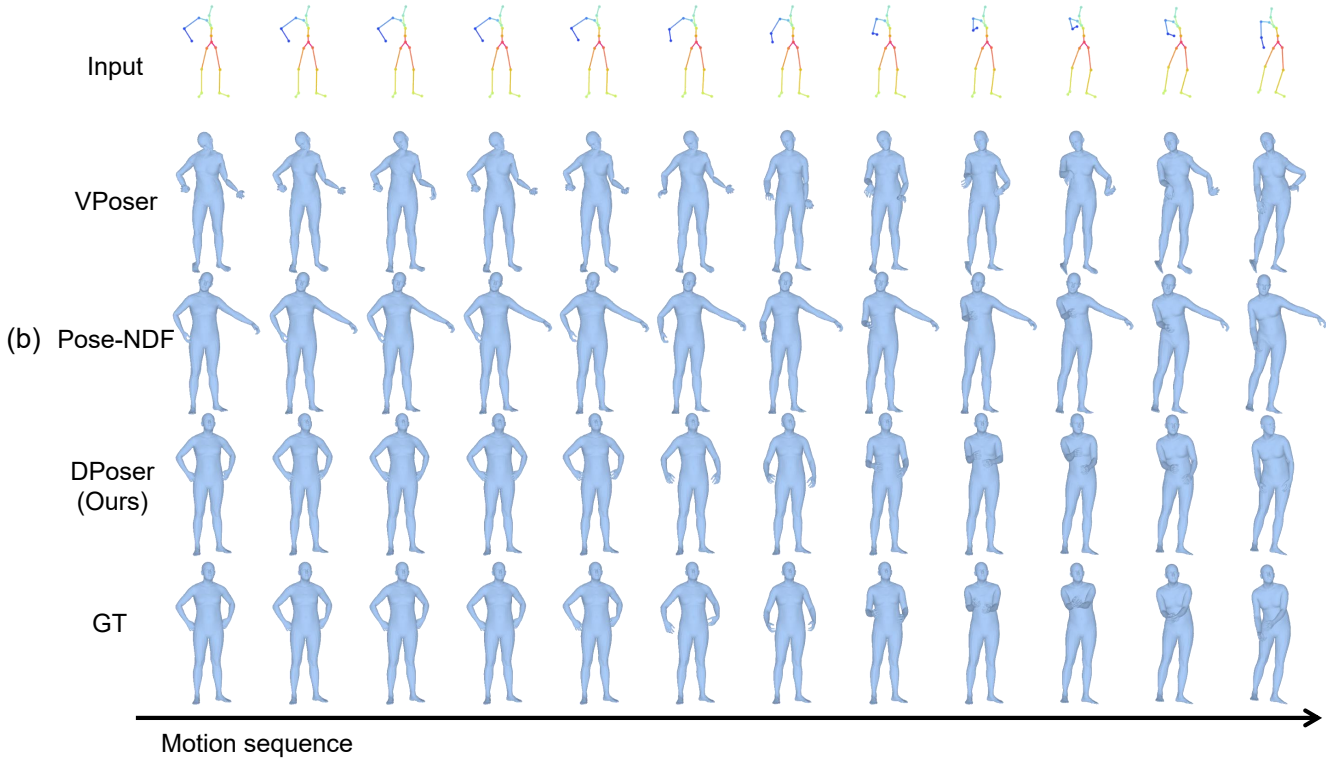


(b) Gaussian noise with 100 mm standard deviation.

Figure S-10. Visualization of motion denoising with noisy observations. We visualize every 20^{th} of the sequence.

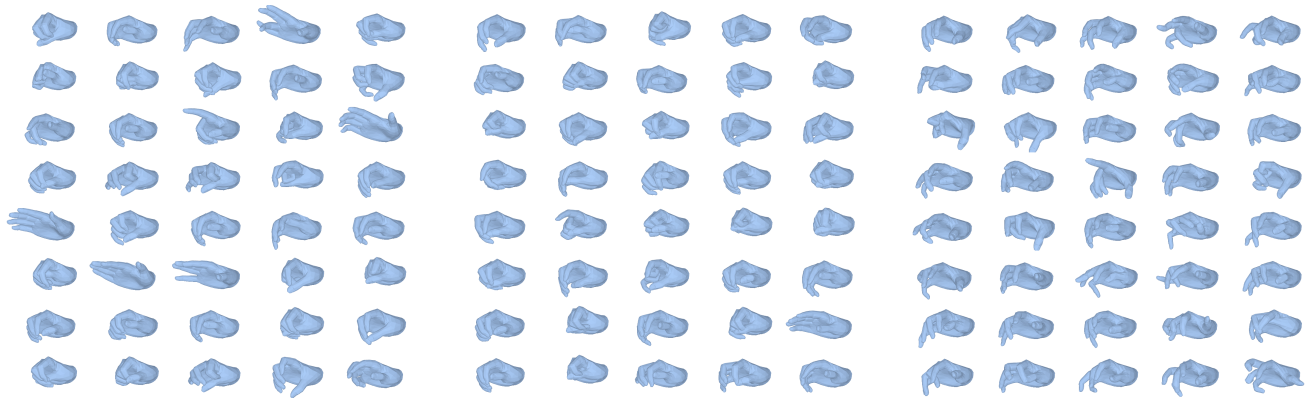


(a) Legs under occlusion.



(b) Left arm under occlusion.

Figure S-11. Visualization of motion denoising with partial observations. We visualize every 20^{th} of the sequence.



(a) DPoser (ours) (b) VPoser (c) NRDF

Figure S-12. Visualization of hand pose generation. DPoser produces more diverse and realistic hand poses compared to VPoser [48] and NRDF [23].

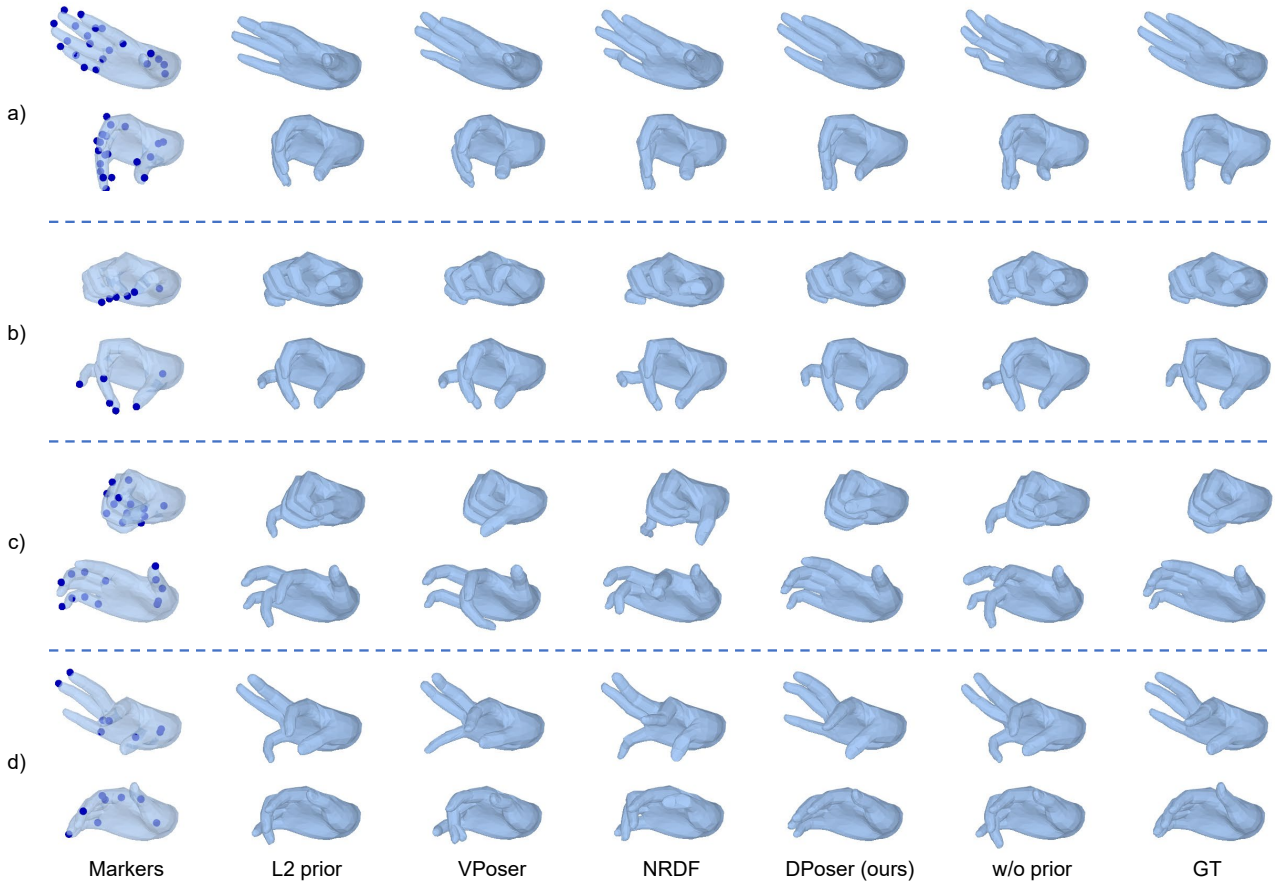


Figure S-13. Visualization of hand inverse kinematics under multiple challenging settings. Comparison across (a) noisy keypoints, (b) fingertip keypoints, (c) partial finger keypoints, and (d) sparse keypoints settings.

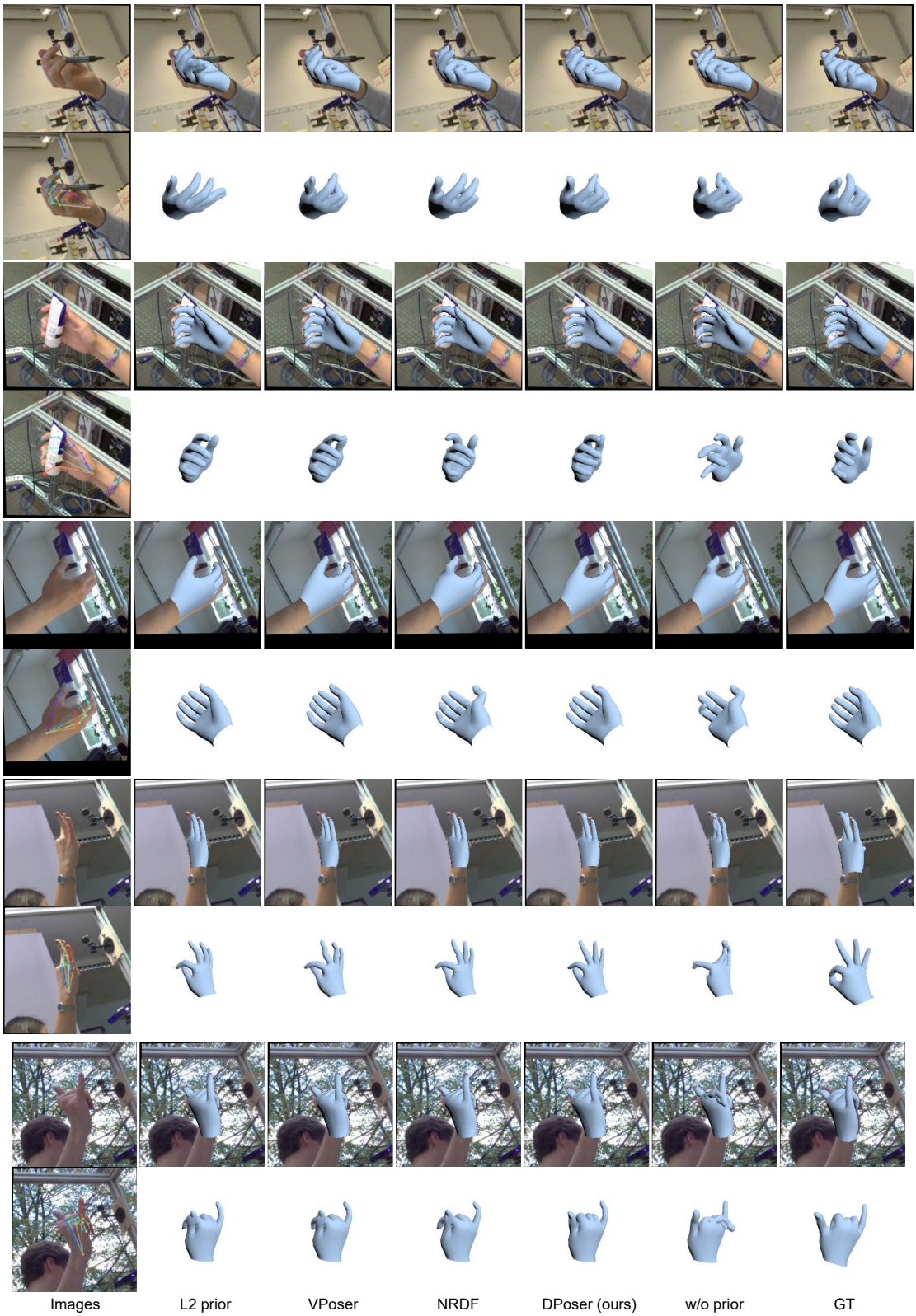


Figure S-14. Visualization of hand mesh recovery with mean pose initialization.

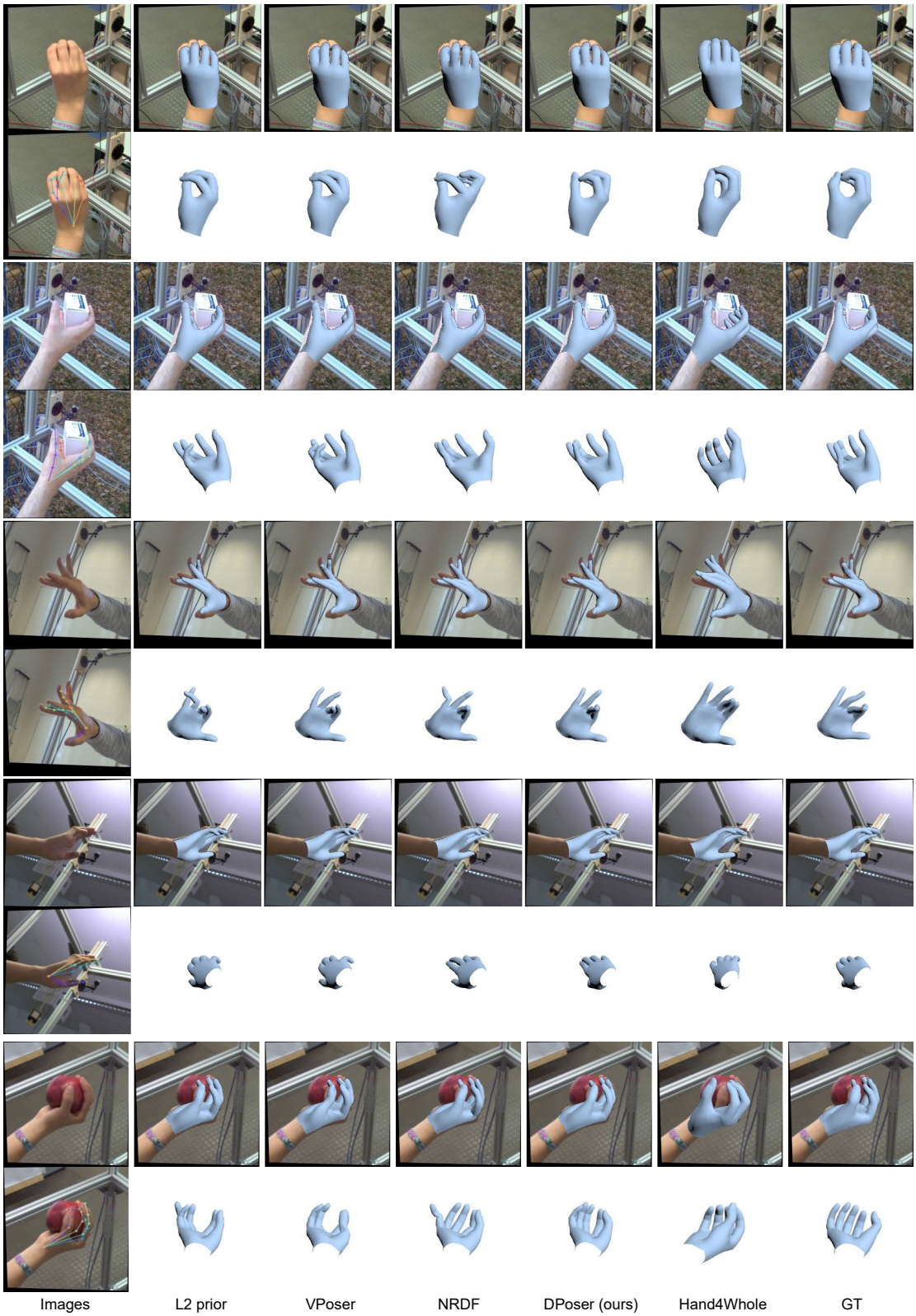


Figure S-15. Visualization of hand mesh recovery with Hand4Whole [44] initialization.

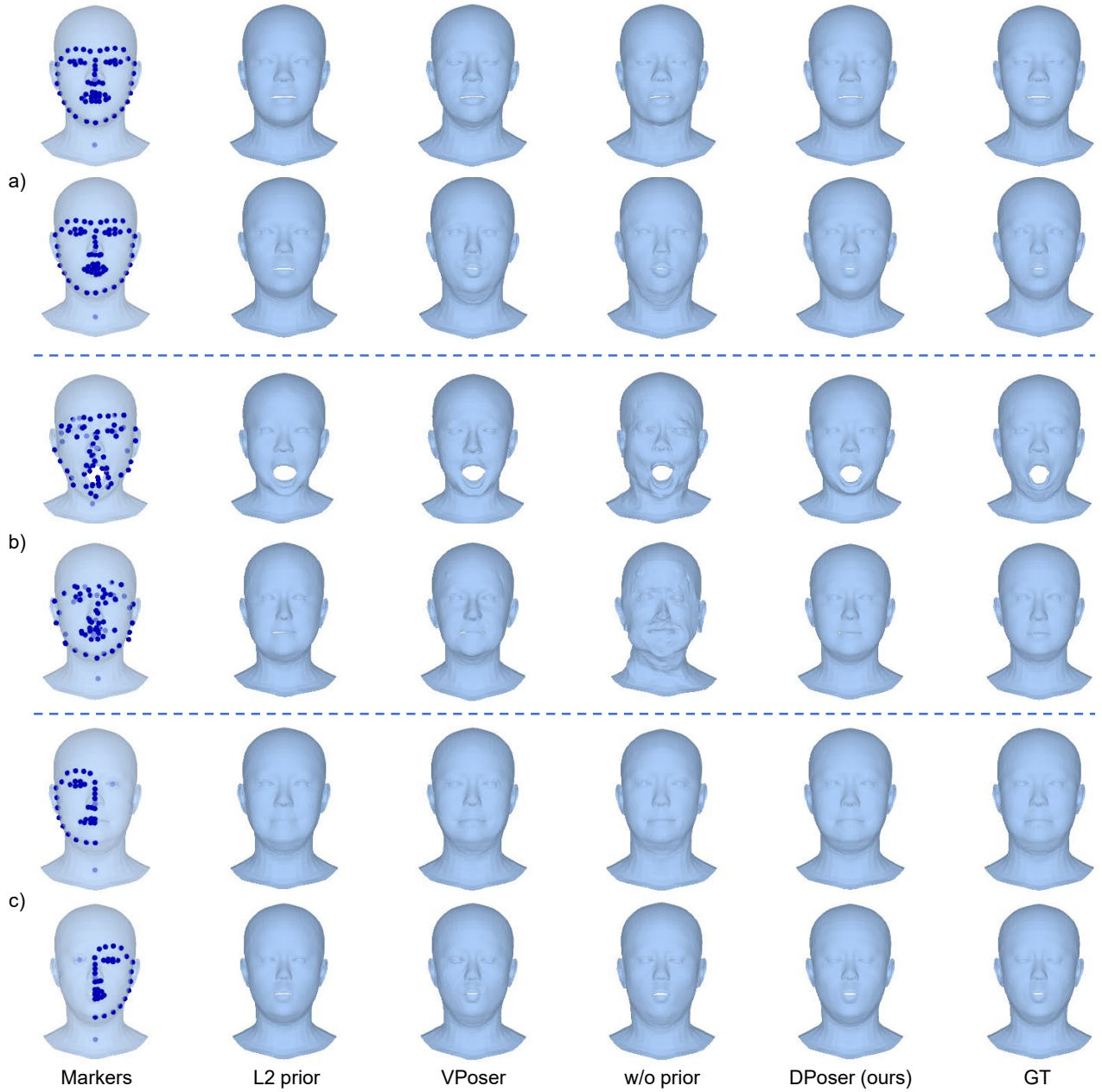


Figure S-16. Qualitative results of face inverse kinematics on the WCPA dataset [30]. Comparison across (a) 1 mm noise, (b) 5 mm noise, and (c) half-face occlusion. Better zoom in and compare the human eyes and chin.

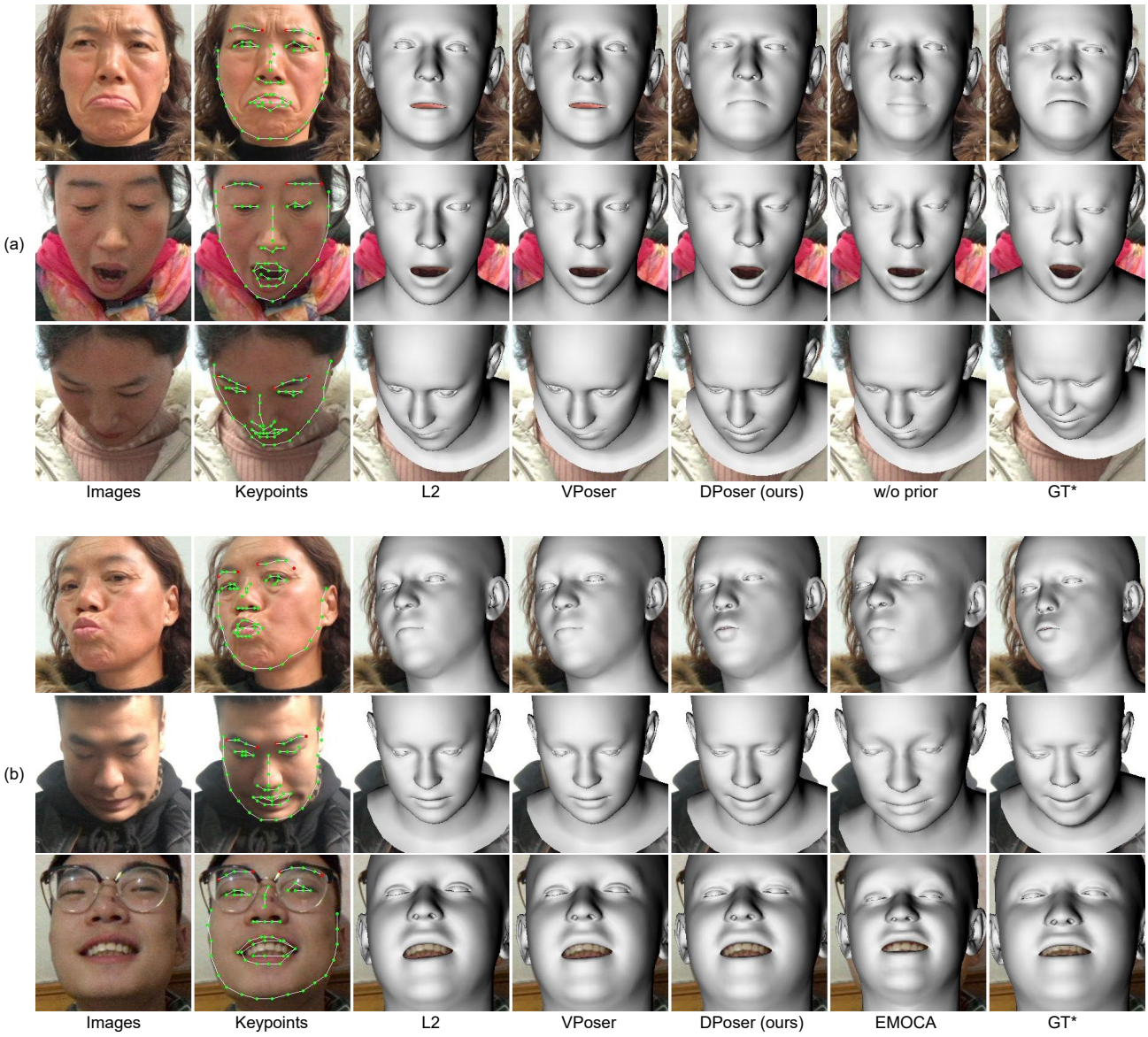
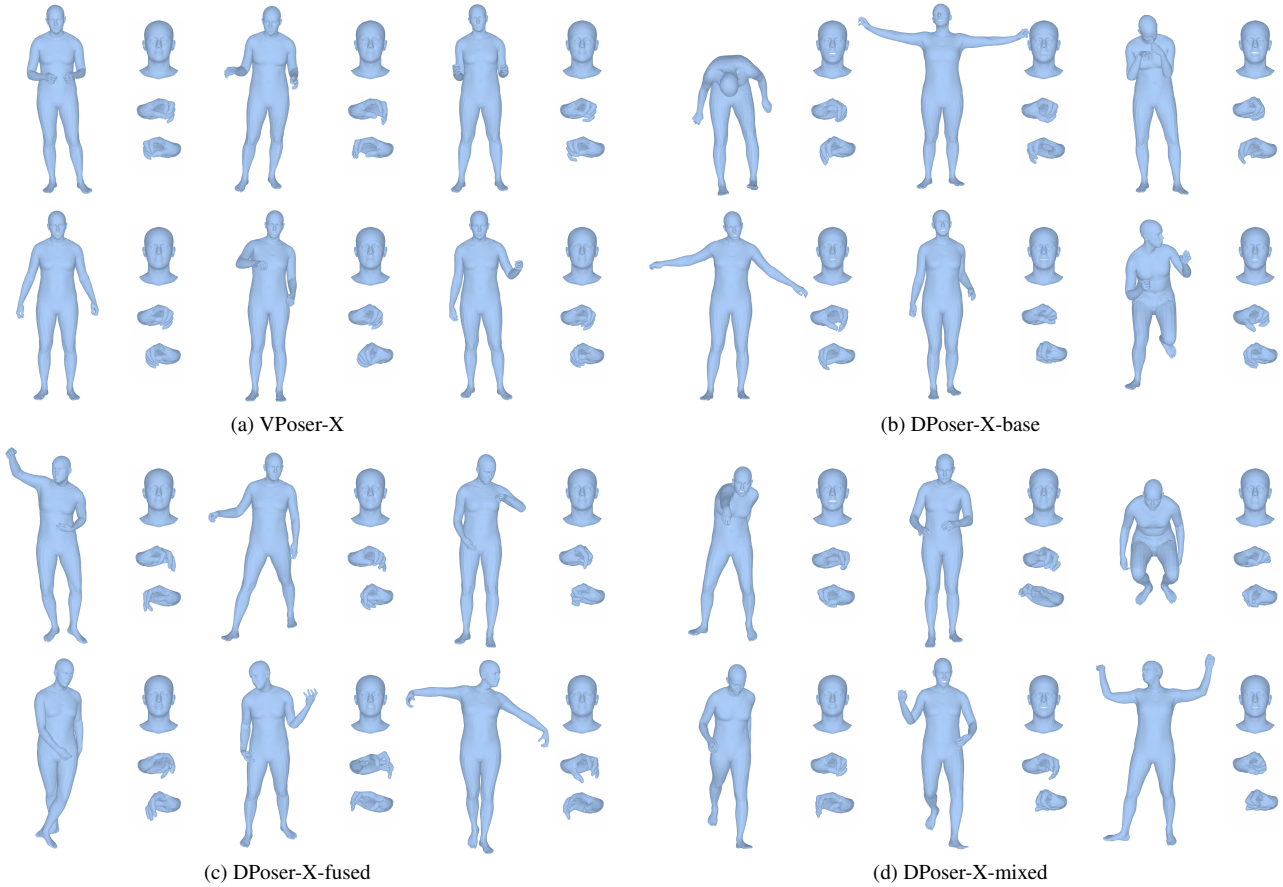


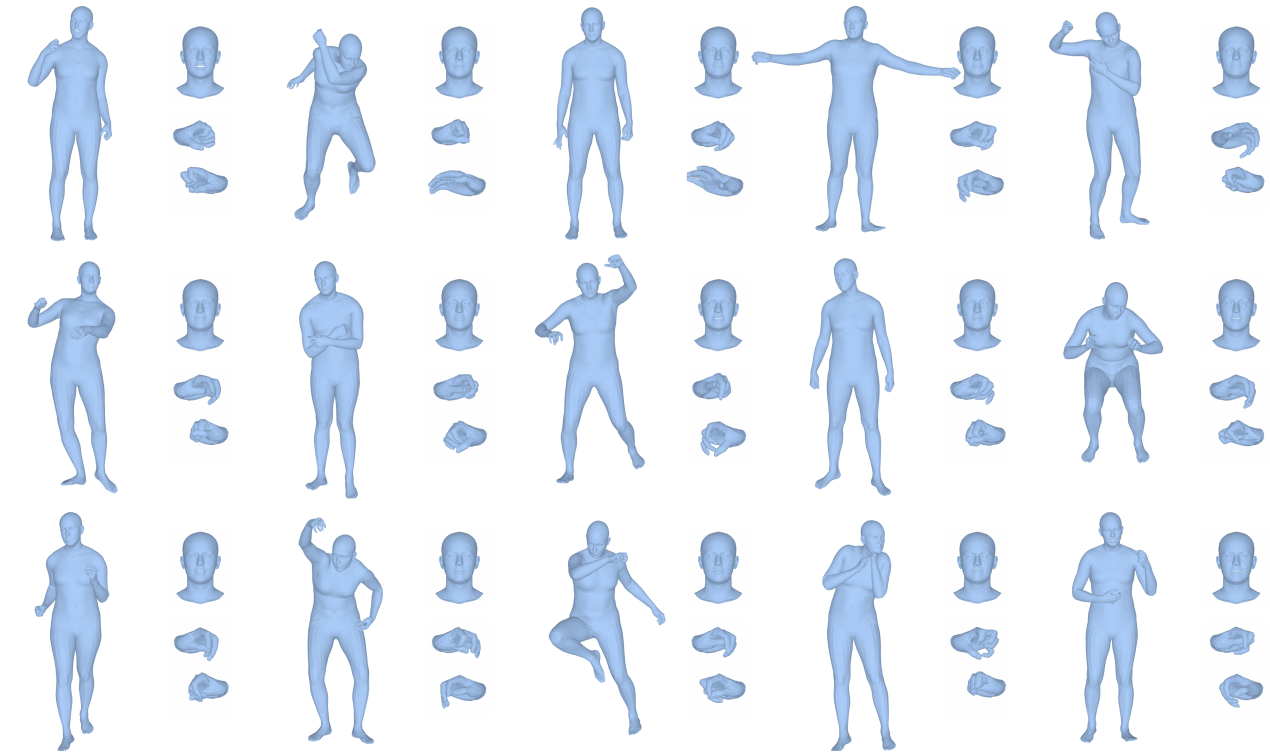
Figure S-17. Visualization of face reconstruction results on the WCPA dataset [30]. Comparisons include (a) fitting from scratch and (b) initialization using EMOCA [9] results. *Ground truth lacks global orientation and translational data; these are fitted for visualization.



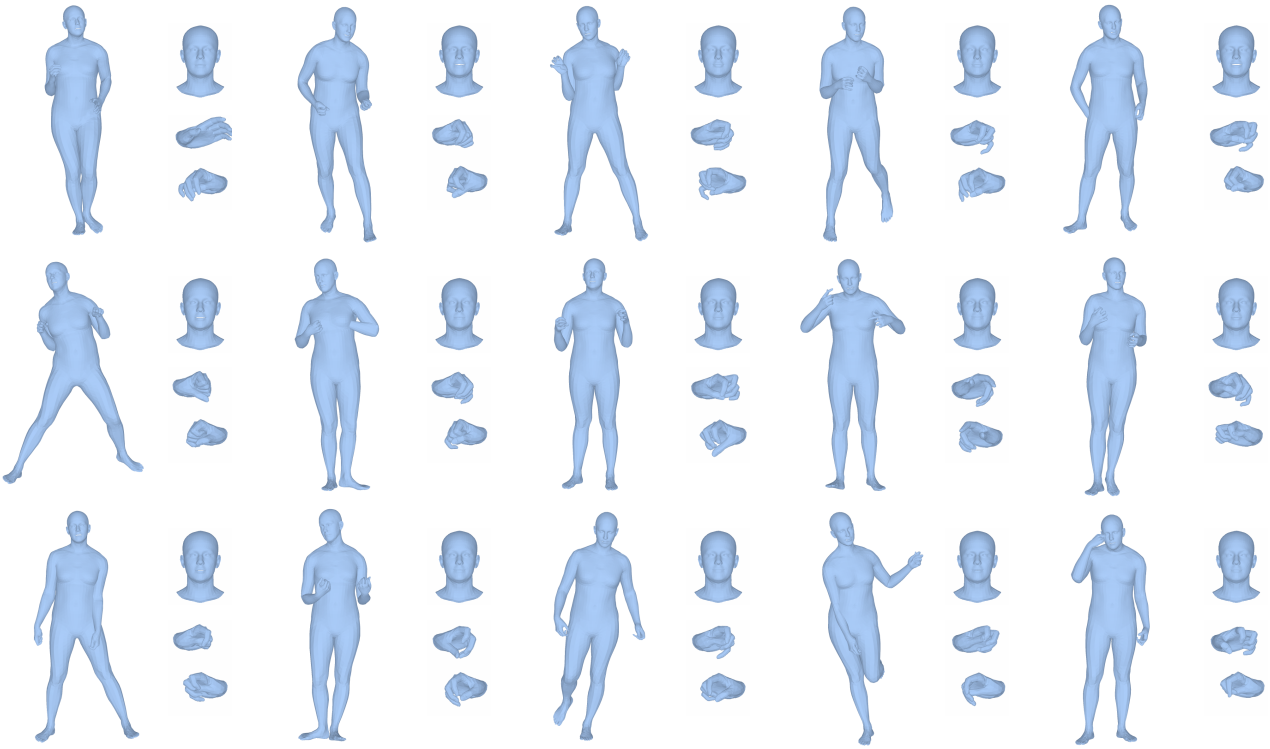
(a) VPoser-X (b) DPoser-X-base

(c) DPoser-X-fused (d) DPoser-X-mixed

Figure S-18. Visualization of whole-body pose generation. (a) VPoser-X primarily generates standing poses with limited diversity. (b) DPoser-X-base generates diverse samples but lacks realism in hand interactions and facial expressions. (c) DPoser-X-fused produces less diverse samples while maintaining plausible whole-body poses. (d) DPoser-X-mixed achieves a well-balanced trade-off between diversity and realism.



(a) DPoser-X-mixed



(b) DPoser-X-fused

Figure S-19. Extended visualization of whole-body pose generation. DPoser-X-mixed generates a diverse range of whole-body poses while maintaining realistic hand interactions and facial expressions. In contrast, DPoser-X-fused retains high realism but produces less diverse results.

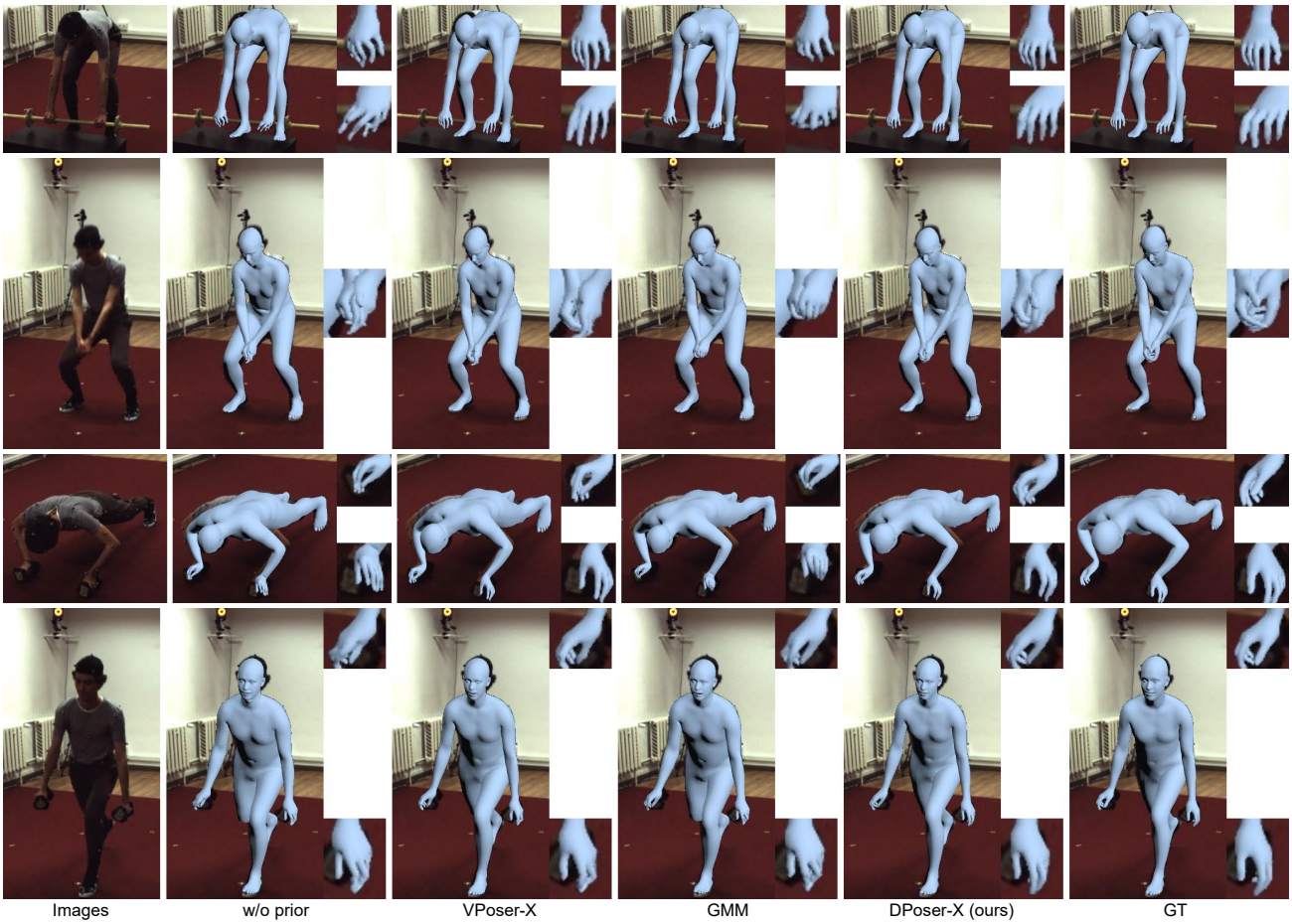


Figure S-20. Visualization of whole-body mesh recovery on the Fit3d dataset [18].

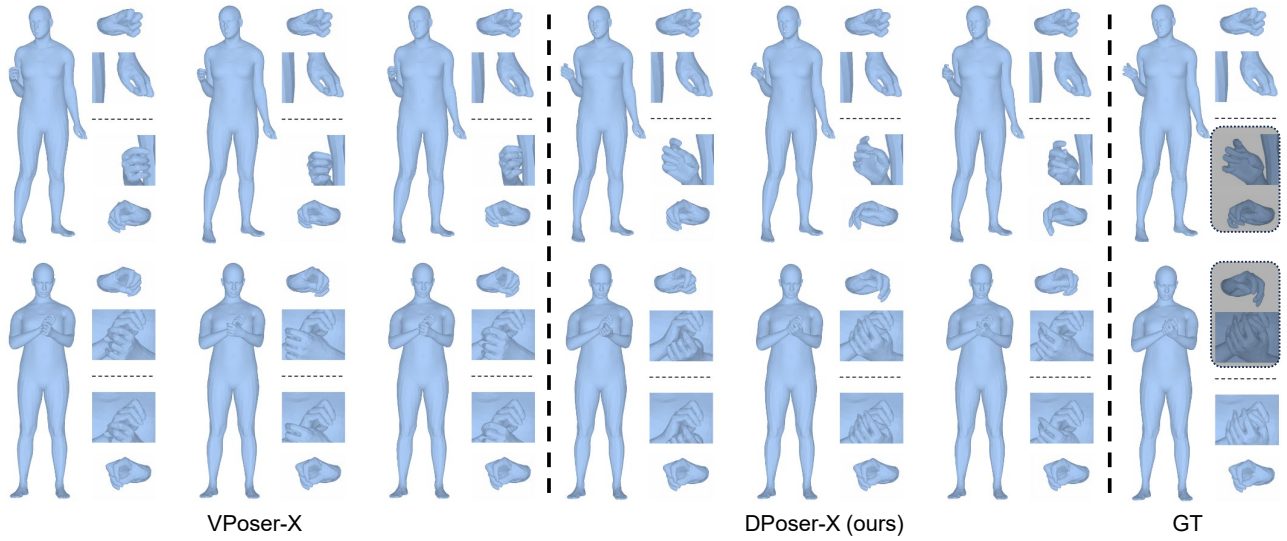


Figure S-21. Qualitative comparison of whole-body pose completion. One hand is masked randomly.

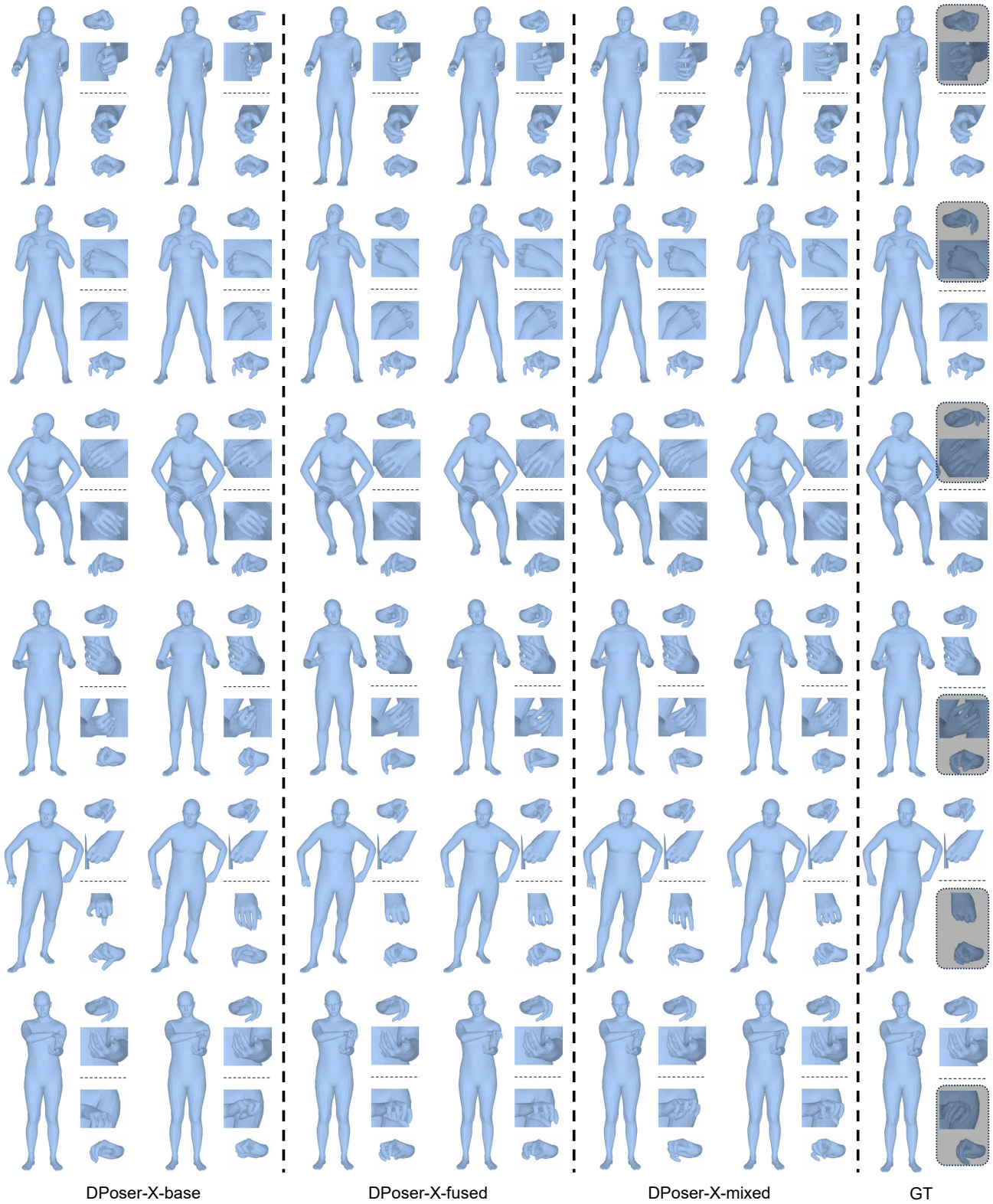


Figure S-22. Visualization of whole-body pose completion for three DPoser-X variants. One hand is masked randomly.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 1, 3, 6, 10, 11

[2] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[3] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 4

[4] Hanbyel Cho and Junmo Kim. Generative approach for probabilistic human mesh recovery using diffusion models. In *ICCV*, 2023. 1

[5] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 12, 13

[6] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *NeurIPS*, 2022. 12, 13

[7] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *CVPR*, 2023. 13

[8] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *CVPR*, 2023. 1, 7

[9] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 9, 10, 23

[10] Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024. 2, 13

[11] Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial parametric pose prior. In *CVPR*, 2022. 1, 2, 3, 6, 7

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 3

[14] Olaf Dünkel, Tim Salzmann, and Florian Pfaff. Normalizing flows on the product space of so (3) manifolds for probabilistic human pose modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2024. 7

[15] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 5, 10

[16] Haven Feng. Photometric optimization. `https://github.com/HavenFeng/photometric_optimization`, 2020. 6

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 5

[18] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9919–9928, 2021. 5, 10, 11, 26

[19] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 1

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 1

[21] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR*, 2021. 4, 8

[22] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 4

[23] Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1661–1671, 2024. 1, 6, 7, 9, 10, 19

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1, 2

[26] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *ICCV*, 2023. 1

[27] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 9, 10

[28] Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, and Jenq-Neng Hwang. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. *arXiv preprint arXiv:2307.03833*, 2023. 1, 13

[29] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1

[30] Yueying Kao, Bowen Pan, Miao Xu, Jiangjing Lyu, Xiangyu Zhu, Yuanzhang Chang, Xiaobo Li, Zhen Lei, and Zixiong Qin. Single-image 3d face reconstruction under perspective projection. *arXiv preprint arXiv:2205.04126*, 2022. 5, 6, 9, 10, 11, 22, 23

[31] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *NeurIPS*, 2022. 13

[32] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 7

[33] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 4

[34] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 6

[35] Lijun Li, Li'an Zhuo, Bang Zhang, Liefeng Bo, and Chen Chen. Diffhand: End-to-end hand mesh reconstruction via diffusion models. *arXiv preprint arXiv:2305.13705*, 2023. 1

[36] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 5, 6

[37] 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 *ECCV*, 2022. 16

[38] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *TOG*, 2020. 1

[39] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024. 5

[40] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *NeurIPS*, 2016. 13

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 1, 6, 11

[42] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 3, 5, 7, 8, 11, 12, 13

[43] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023. 3, 12, 13

[44] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2308–2317, 2022. 9, 21

[45] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *Advances in Neural Information Processing Systems*, 36, 2024. 4, 9, 10

[46] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *arXiv preprint arXiv:2306.09337*, 2023. 1

[47] Naoki Murata, Koichi Saito, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Gibbsddrm: A partially collapsed gibbs sampler for solving blind inverse problems with denoising diffusion restoration. *arXiv preprint arXiv:2301.12686*, 2023. 13

[48] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1, 3, 5, 6, 7, 8, 9, 10, 11, 19

[49] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2

[50] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. 13

[51] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Xiyu Wang, Chang Xu, Dongmei Fu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Learning structure-guided diffusion model for 2d human pose estimation. *arXiv preprint arXiv:2306.17074*, 2023. 1

[52] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 1, 11

[53] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 4, 6

[54] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 5, 9

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 11

[56] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019. 2

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based

generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2, 12, 13

[58] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 5

[59] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *ECCV*, 2022. 1, 3, 4, 5, 6, 7, 8

[60] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 2011. 2

[61] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 1, 2

[62] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 777–787, 2024. 1

[63] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 13

[64] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 13

[65] 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 *European conference on computer vision*, pages 180–200. Springer, 2022. 5

[66] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14606–14617, 2024. 1

[67] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 11

[68] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 2

[69] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 5, 6

[70] 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. 4, 9, 10