CDPOSE: COMPOSITIONAL DIFFUSION FOR WHOLE-BODY POSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating 3D whole-body poses that simultaneously capture body articulation, detailed hand configurations, and expressive facial motion remains a challenging task due to joint rotations evolving in an unconstrained space, where small local errors propagate and escalate across parts. To address these limitations, we propose CDPOSE, a diffusion-guided whole-body pose model that represents each articulated joint in 6D rotation and converts inside the generator to an axis–angle before kinematic evaluation. The denoiser is trained under a sub-variance-preserving stochastic differential process with a signal-to-noise–weighted objective. During inference, CDPOSE produces a single-step Tweedie estimate of the clean pose, with an optional short reverse-time refinement for additional sharpening. The whole-body assembly combines part-specific diffusion heads with a fused residual head that learns cross-part coordination, such as mirrored hand motion and face–torso co-articulation. The coverage and diversity are further improved by mixed training on heterogeneous whole-body and part-only data, using selective masking and supervision. Comprehensive experiments demonstrate that CDPOSE is robust and versatile across diverse benchmarks spanning body, hand, face, and whole-body pose modeling. In particular, CDPOSE achieves the lowest errors on the ARCTIC dataset, obtaining a PA-MPJPE of 26.92 mm for body mesh recovery, a PA-MPVPE of 7.10 mm for hands, 2.51 mm for the face, and 23.24 mm for whole-body mesh recovery (PA-MPVPE over all vertices). The code will be publicly released.

1 INTRODUCTION

Accurate reconstruction and generation of whole-body human pose, including body articulation, detailed hand motion, and facial expression, remains a key challenge in vision, graphics, and embodied intelligence. Errors in one body region can propagate through the kinematic chain, while unstable rotation representations can degrade optimization and generalization. This problem is further complicated by the scarcity and heterogeneity of fully annotated whole-body datasets, since body, hand, and face data are often captured using different sensors, protocols, and annotation schemes Moon et al. (2020); Zimmermann et al. (2019).

Existing pose priors include local regularizers, Gaussian mixture priors used in SMPLify-style optimization Bogo et al. (2016); Loper et al. (2023); Cai et al. (2023b), and latent pose priors such as VPoser Pavlakos et al. (2019) in the SMPL-X framework Pavlakos et al. (2019). Although effective in constrained settings, these methods often under-model cross-part dependencies, such as bilateral hand coordination and face–upper-body co-articulation, especially when supervision is incomplete. Parametric models such as SMPL Loper et al. (2023), MANO Romero et al. (2017), FLAME Li et al. (2017), and SMPL-X Pavlakos et al. (2019) provide a common kinematic foundation, but do not offer a robust unified generative prior. Recent diffusion models provide a promising alternative for learning expressive score fields in high-dimensional pose spaces Ho et al. (2020); Song et al. (2020b); however, existing approaches often rely on fragile rotation parameterizations, expensive part-specific regressors, or limited use of heterogeneous whole-body and part-only data.

We propose CDPOSE, a diffusion-based whole-body pose prior designed for stable, expressive, and coordinated body, hand, and face modeling. CDPOSE represents all articulated joints using a *6D* rotation representation Zhou et al. (2019), which is converted inside the generator to axis–

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: CDPOSE is evaluated across a wide range of pose-related tasks. Guided by a diffusion-based framework, CDPOSE serves as a powerful and flexible prior for 3D human body and parts mapping. The figure illustrates examples of body pose recovery, novel body, hand, and facial pose generation, and whole-body pose mesh recovery.

angle form before SMPL, MANO, and FLAME decoding. The score model is trained with a sub-variance-preserving SDE and an SNR-weighted denoising objective Song et al. (2020b). At test time, a single-step Tweedie estimate Chung et al. (2022) provides an efficient regularization signal, with optional short reverse-time refinement when additional sharpening is needed. Compared with previous diffusion-based pose priors such as DPoser Lu et al. (2025), CDPOSE explicitly targets expressive whole-body generation through cross-part coordination and dataset-agnostic supervision.

To support diverse datasets and evaluation protocols, CDPOSE uses a unified sparse vertex-to-keypoint mapping that produces consistent body, hand, and face landmarks while reducing memory and computation overhead Cao et al. (2017). For whole-body modeling, part-specific diffusion heads are combined with a fused residual head that learns cross-part dependencies and corrects inconsistencies between independently generated regions. A mixed training strategy further enables learning from both whole-body and part-only datasets by masking missing components and supervising only observed parts, while random masking on complete samples encourages contextual completion. The resulting prior can be used in inverse tasks such as human mesh recovery, pose completion, and inverse kinematics, improving robustness under sparse or noisy supervision Kanazawa et al. (2018); Kocabas et al. (2020).

This paper makes the following major contributions.

- We introduce a whole-body diffusion model that operates directly in *6D* rotation space for all articulated joints and converts inside the generator to axis-angle immediately before kinematic evaluation, yielding stable optimization and seamless compatibility with standard body, hand, and face models.
- We propose a dual-path architecture with part-specific diffusion heads and a fused residual head to capture cross-part dependencies across the body, hands, and face.
- We develop a mixed training strategy with selective supervision and random masking, allowing CDPOSE to learn from heterogeneous whole-body and part-only datasets.
- Extensive experiments show that CDPOSE achieves strong performance across body, hand, face, and full-body modeling benchmarks.

2 Related Work

2.1 Pose-centric Diffusion

Diffusion models Ho et al. (2020); Song et al. (2020a); Song & Ermon (2019); Song et al. (2020b) have become effective in modeling complex data distributions and generating multiple plausible solutions, which is particularly useful for estimating ambiguous human poses. Recent methods adapt diffusion to pose-centric tasks including DiffPose Holmquist & Wandt (2023), DiffusionPose Qiu et al. (2023), GFPose Ci et al. (2023), and ZeDO Jiang et al. (2024) focus mainly on 2D-to-3D pose lifting or trajectory generation, while Diff-HMR Cho & Kim (2023) and DiffHand Li et al. (2023) estimate SMPL body parameters and hand meshes. Other works, such as EgoWholeBody Wang et al. (2024) and RoHM Zhang et al. (2024), use diffusion to denoize and refine motion. Prior-based methods, including BUDDI Müller et al. (2024) and DPoser Lu et al. (2025), further show the potential of diffusion priors to model human pose distributions and support test-time optimization.

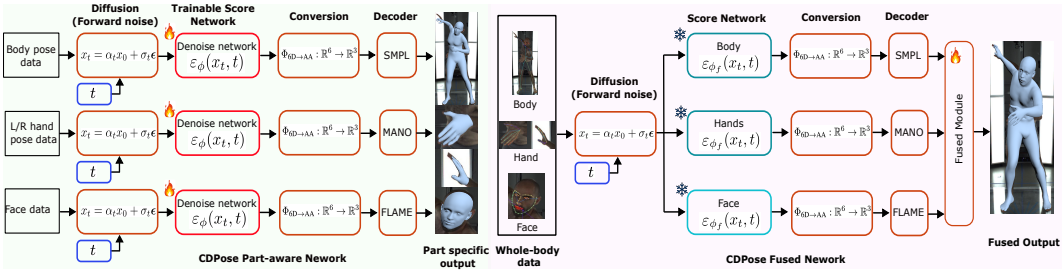

Figure 2: Overview of CDPOSE. The part-aware network first learns independent diffusion priors for the body, hands, and face in *6D* rotation space. Each part-specific prediction is converted to axis–angle using the fixed $\Phi_{6D\to AA}$ operator and decoded by the corresponding SMPL, MANO, or FLAME model. The CDPOSE fused network then freezes the pretrained part-specific priors and trains a fused residual head on whole-body data to combine the part predictions into a coherent whole-body pose.

Our method is closest to BUDDI Müller et al. (2024) and DPoser Lu et al. (2025), but differs in three aspects. First, instead of primarily modeling 3D joint locations, CDPOSE operates in a unified 6D rotation space and converts rotations inside the model to axis–angle for compatibility with SMPL-family models Loper et al. (2023); Pavlakos et al. (2019). Second, CDPOSE jointly models body, hands, and face, and uses a fused cross-part head to improve whole-body coherence beyond body-only or hand-only diffusion priors. Third, it adopts a timestep policy designed for articulated pose refinement, enabling efficient single-step regularization with optional short reverse-time sharpening. These design choices make CDPOSE a flexible plug-and-play diffusion prior to generation and reconstruction of expressive whole-body poses.

2.2 ROTATION REPRESENTATIONS FOR ARTICULATED POSE

Rotation parameterization constitutes a critical design choice for both discriminative and generative models. Axis–angle and quaternions provide minimal or unit-norm constraints but can yield discontinuities or challenging manifolds for optimization. The 6D rotation representation Zhou et al. (2019) improves continuity and gradient flow by representing rotations as two orthonormal columns of a 3D rotation matrix. Previous work has shown benefits for network training stability and downstream fitting. CDPOSE adopts 6D rotations for every articulated joint (body, hands, face) and performs a differentiable in-generator map to axis–angle immediately before kinematic evaluation, harmonizing with SMPL Loper et al. (2023), SMPL-X Pavlakos et al. (2019), MANO Romero et al. (2017), and FLAME Li et al. (2017) while preserving the advantageous optimization characteristics.

3 METHODOLOGY

3.1 PRELIMINARIES

We adopt a score-based diffusion model Ho et al. (2020); Sohl-Dickstein et al. (2015); Song & Ermon (2019); Song et al. (2020b) over a pose vector x. The forward stochastic differential equations (SDE) perturb a clean sample $x_0 \sim p_{\text{data}}$ along $t \in [0, 1]$:

$$x_t \;=\; \alpha_t\, x_0 \;+\; \sigma_t\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

with schedules (α_t, σ_t) monotonically transforming p_{data} to $\mathcal{N}(0, I)$. A denoiser $\varepsilon_\phi(x_t, t)$ is trained using a weighted mean square error (see Sec. 3.4 for the timestep weighting $w(t)$):

$$\mathbb{E}_{x_0,t,\varepsilon}\big[w(t)\,\|\varepsilon - \varepsilon_\phi(x_t, t)\|_2^2\big]. \tag{2}$$

Generation follows the standard reverse trajectory from $x_1 \sim \mathcal{N}(0, I)$. In our method, x is a vectorized 6D whole-body pose (body/hands/face); at inference we convert samples to axis–angle inside the generator via $\Phi_{6D\to AA}$ (Sec. 3.3) before kinematic evaluation.

3.2 WHOLE-BODY PARAMETERIZATION

We parameterize body, hands, and face in a single representation compatible with SMPL/SMPL-H/SMPL-X, MANO, and FLAME. Let B denote the batch size. The body trunk uses $\boldsymbol{\theta}_{\text{body}} \in \mathbb{R}^{B\times 21\times 3}$ (63 DoF per sample). SMPL-H/X add left/right hands $\boldsymbol{\theta}_{\text{Lh}}, \boldsymbol{\theta}_{\text{Rh}} \in \mathbb{R}^{B\times 15\times 3}$; SMPL-X further includes the jaw $\boldsymbol{\theta}_{\text{jaw}} \in \mathbb{R}^{B\times 3}$ and eye rotations $\boldsymbol{\theta}_{\text{leye}}, \boldsymbol{\theta}_{\text{reye}} \in \mathbb{R}^{B\times 3}$, plus global rigid motion $(\mathbf{R}_g, \mathbf{t}) \in (\mathrm{SO}(3))^B \times \mathbb{R}^{B\times 3}$. We also define a whole-body block

$$\boldsymbol{\theta}_{\text{WB}} = \left[\boldsymbol{\theta}_{\text{body}} \,\|\, \boldsymbol{\theta}_{\text{Lh}} \,\|\, \boldsymbol{\theta}_{\text{Rh}} \,\|\, \boldsymbol{\theta}_{\text{jaw}} \,\|\, \boldsymbol{\psi}\right] \in \mathbb{R}^{B\times(156+E)}, \tag{3}$$

which follows the SMPL-X canonical order and allows constant-time slicing.

Body/hands share shape $\boldsymbol{\beta} \in \mathbb{R}^{B\times N_\beta}$ (typically $N_\beta = 10$); the face uses $\boldsymbol{\beta}_{\text{face}} \in \mathbb{R}^{B\times N_\beta}$ and expressions $\boldsymbol{\psi} \in \mathbb{R}^{B\times E}$ (default $E = 100$).

3.2.1 TRAINING VARIABLE AND CONVERSION.

All joint rotations are represented and diffused in 6d rotation Zhou et al. (2019) (two 3D vectors per joint). The clean variable in Eq. equation 1 is $x_0 = \mathrm{vec}(\boldsymbol{\theta}^{\text{6D}}) \in \mathbb{R}^{B\times J\times 6}$, where J is the number of articulated joints. Before evaluation, 6D blocks are mapped to axis–angle via $\Phi_{\text{6D}\to\text{AA}}$ (Sec. 3.3); meshes and native landmarks are then produced by the SMPL/SMPL-H/SMPL-X, MANO, and FLAME forwards. The standardized and task-specific keypoints follow from the unified sparse projector in Sec. 3.5.

Unless noted otherwise, the global rotation $\mathbf{R}_g$ is included in the 6D diffusion to stabilize orientation, while the global translation $\mathbf{t}$ is estimated from downstream camera/reprojection objectives (not diffused).

3.3 6D ROTATION REPRESENTATION IN DIFFUSION

To ensure stable and consistent optimization across the entire human pose, including *body*, *hands*, and *face*, we adopt a 6D rotation representation for all articulated joints. Rather than using separate rotation formats for each part, we standardize joint rotations to a 6D vector format and perform an internal conversion within the generator to the axis–angle representation required by the SMPL, MANO, and FLAME parametric models. This choice harmonizes the rotation format used by the diffusion-based generative backbone (see Sec. 3.1) with the kinematic parameterizations used for mesh decoding (see Sec. 3.2), including the packed whole-body pose vector defined in Eq. 3. For the handling of the denoising objective and timesteps in this *6D* space, see Sec. 3.4.

3.3.1 CONSTRUCTING 3D ROTATION FROM 6D VECTORS

The rotation of each joint is represented by a 6D vector $\mathbf{r} = [\mathbf{a}\|\mathbf{b}] \in \mathbb{R}^6$, where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$. To convert this to a valid rotation matrix $R \in \mathrm{SO}(3)$, we apply Gram–Schmidt orthonormalization with ε-stabilized norms, where $\varepsilon = 10^{-6}$ used for all experiments in the paper:

$$\begin{aligned}
\mathbf{u}_1 &= \frac{\mathbf{a}}{\max(\|\mathbf{a}\|, \varepsilon)}, \\
\mathbf{u}_2 &= \frac{\mathbf{b} - (\mathbf{u}_1^\top \mathbf{b})\mathbf{u}_1}{\max\left(\left\|\mathbf{b} - (\mathbf{u}_1^\top \mathbf{b})\mathbf{u}_1\right\|, \varepsilon\right)}, \\
\mathbf{u}_3 &= \mathbf{u}_1 \times \mathbf{u}_2, \\
R &= [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \mathbf{u}_3].
\end{aligned} \tag{4}$$

This results in a valid rotation matrix $R \in \mathrm{SO}(3)$, which is then mapped to its axis–angle form $\boldsymbol{\omega} \in \mathbb{R}^3$ using the Rodrigues formula (logarithmic map). We denote this transformation as follows:

$$\Phi_{\text{6D}\to\text{AA}} : \mathbb{R}^6 \to \mathbb{R}^3, \qquad \boldsymbol{\omega} = \Phi_{\text{6D}\to\text{AA}}(\mathbf{r}). \tag{5}$$

The conversion operator $\Phi_{\text{6D}\to\text{AA}}$ is applied uniformly in all body joints, hands, wrists, and finger joints (left and right hands), and face joints, including the neck, jaw, and eyes. The operator is differentiable, so gradients from SMPL/ MANO/ FLAME forwards and task losses back-propagate to the 6D inputs. The ingestion cost is $\mathcal{O}(J)$ and negligible compared to the forwards of the parametric model.

3.4 DENOISING OBJECTIVE

Let $x_0 \in \mathbb{R}^{J\times 6}$ denote the clean whole-body pose represented in the 6D rotation space, where J is the number of modeled rotational joints. For a randomly sampled diffusion time $t \in (0, 1]$, the forward process perturbs x_0 according to the sub-VP SDE marginal distribution Song et al. (2020b):

$$x_t = \alpha_t x_0 + \sigma_t \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, I),\ x_t \quad \sim \mathcal{N}!\left(\alpha_t x_0, , \sigma_t^2 I\right), \tag{6}$$

where α_t and σ_t denote the closed-form signal attenuation and noise standard deviation of the sub-VP SDE. In all experiments, we use the linear noise schedule with $\beta_{\min} = 0.1$ and $\beta_{\max} = 20.0$.

The score network predicts the time-dependent score

$$s_\phi(x_t, t) \approx \nabla_{x_t} \log p_t(x_t), \tag{7}$$

which is used to estimate the clean pose through the continuous-time Tweedie identity Chung et al. (2022):

$$\widehat{x}_0(t) = \frac{x_t + \sigma_t^2 s_\phi(x_t, t)}{\alpha_t}. \tag{8}$$

Here, the correction term $\sigma_t^2 s_\phi(x_t, t)$ moves the noisy sample toward high-density regions of the data distribution, while division by α_t maps the estimate back to the clean pose domain.

During training, $\widehat{x} * 0(t)$ is computed with gradient tracking, and gradients are back-propagated through $s * \phi(x_t, t)$ to update the score network parameters ϕ. Thus, the denoising loss directly trains the diffusion model. This differs from the test-time regularization setting, where the trained score network is frozen and $\widehat{x}_0(t)$ may be used with stop-gradient so that optimization affects only the task variables rather than the score network.

We train the score network by minimizing a clipped signal-to-noise-weighted reconstruction loss between the clean pose and its one-step estimate:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{x_0,t}\left[\frac{1}{2}\sqrt{1 + \min\left(\frac{\alpha_t^2}{\sigma_t^2 + \varepsilon}, \gamma\right)}\ \|x_0 - \widehat{x}_0(t)\|_2^2\right]. \tag{9}$$

We use $\gamma = 5$ to prevent the loss weight from becoming excessively large at very low-noise timesteps, and $\varepsilon = 10^{-6}$ for numerical stability. This clipped weighting keeps the objective focused on informative denoising timesteps while avoiding unstable overemphasis of nearly clean samples.

3.4.1 TIMESTEP POLICY USED IN PRACTICE.

During training, we sample $t \sim \mathcal{U}(0, 1]$; for test-time regularization in inverse problems, we sweep t linearly from $t_{\max}$ down to $t_{\min}$ over S iterations, with defaults $t_{\min} = 0.05$, $t_{\max} = 0.15$.

While the one-step estimator in equation 8 suffices for most scenarios, we optionally refine x_t by integrating a short reverse trajectory over a line segment $t \to t^-$ using N sub-steps. Writing (α_c, σ_c) and (α_b, σ_b) for the signal/noise coefficients at consecutive times $t_c > t_b$, the update is

$$\begin{aligned} \widehat{\varepsilon}(x_{t_c}, t_c) &= -\,\sigma_c \,\mathrm{score}(x_{t_c}, t_c), \\ x_{t_b} &= \frac{\alpha_b}{\alpha_c}\left(x_{t_c} - \sigma_c\,\widehat{\varepsilon}\right) + \sigma_b\,\widehat{\varepsilon}. \end{aligned} \tag{10}$$

After N such steps, with a linear or truncated schedule on t, we obtain a refined state $\tilde{x}_{t^-}$ and its associated $\widehat{x}_0(t^-)$ via equation 8. Empirically, this multi-step sharpening is optional: the one-step estimator already provides a strong signal, whereas multi-step refinement increases computation and is enabled only when a task demonstrably benefits.

When the denoiser regularizes a downstream objective such as completion, inverse kinematics (IK), or human mesh recovery, the timestep t is chosen per iteration either by uniform sampling on $[t_{\min}, t_{\max}]$, by fixing $t = t_{\text{fixed}}$, or by a truncated linear schedule $t_k = t_{\max} - \frac{k}{S-1}(t_{\max} - t_{\min})$ over S iterations. The truncated schedule biases updates toward late times, where pose details are refined, and the one-step target is most informative.

Concretely, the composite objective optimized at test time for a task variable x_0 is

$$\mathcal{L}_{\text{task}}(x_0) = \mathcal{L}_{\text{data}}(x_0) + \lambda_{\text{prior}}\,\mathcal{L}_{\text{denoise}}(x_0; t), \tag{11}$$

where ϕ is frozen, and $\mathcal{L}_{\text{denoise}}$ is evaluated as in Eq. equation 9 using the timestep policy. Only the task variable x_0 receives gradients. The data term $\mathcal{L}_{\text{data}}$ is task-specific; for HMR, it is a loss of 2D keypoint reprojection against detected or ground-truth keypoints through the sparse projector in Sec. 3.5; for completion, it is a loss of masked reconstruction over the observed components of x_0; and for IK, it is a loss of 3D joint-position against observed joints. We use $\lambda_{\text{prior}} = 0.1$ in all experiments.

3.5 Sparse Regression Estimation

After converting the predicted 6D rotations to axis–angle form through $\Phi_{\text{6D}\to\text{AA}}$, we decode the corresponding SMPL/SMPL-H/X, MANO, and FLAME models to obtain part-specific meshes $\mathbf{V}^{(p)} \in \mathbb{R}^{B\times N_v^{(p)}\times 3}$, where $p \in \text{body}, \text{L-hand}, \text{R-hand}, \text{face}$. To support different evaluation protocols and task-specific fitting objectives, we use a unified sparse vertex-to-keypoint projector. Given a sparse regressor $\mathbf{R}^{(p)} \in \mathbb{R}^{N_t^{(p)}\times N_v^{(p)}}$, the target keypoints are obtained channel-wise as

$$\mathbf{K}^{(p)}_{:,:,k} = \mathbf{R}^{(p)}\mathbf{V}^{(p)}_{:,:,k}, \qquad k \in x, y, z, \tag{12}$$

yielding $\mathbf{K}^{(p)} \in \mathbb{R}^{B\times N_t^{(p)}\times 3}$. This sparse projection provides standardized body, hand, and face keypoints for evaluation and downstream fitting losses, including the 2D reprojection term used in Eq. 11. Importantly, the diffusion prior itself is trained only in the 6D rotation parameter space using Eq. 9; the sparse projector is used only for evaluation and task-specific optimization. This design keeps prior training parametric-model-aware at inference time while avoiding expensive dense vertex or joint supervision during diffusion training.

3.6 Whole-body assembly.

We construct the whole-body pose by training three part-specific diffusion models aligned with the native kinematics described in Sec. 3.2: a body model (SMPL/SMPL-H/SMPL-X joints), a hand model (MANO, L/R), and a face model (FLAME). As a baseline, the whole-body input is partitioned into body, hand, and face blocks, where each block is processed by its corresponding generator without cross-part interaction, forming the (CDPose-B) model, illustrated as the *Part Specific Network* in Fig. 2.

To capture cross-part dependencies (e.g., mirrored hands or face–upper-body articulation), we introduce a *fused* head trained on whole-body data that consumes the concatenated part outputs and learns residual corrections, referred to as the (CDPose-F) model. To further improve coverage and diversity while preserving coordination, we adopt a *mixed training* strategy resulting in the (CDPose-M) model. Following Lu et al. (2025), training uses a heterogeneous pool of whole-body, body-only, hand-only (one/two hands), and face-only samples. Part-only samples are treated as incomplete whole-body inputs with losses applied only to available parts, while whole-body samples are randomly masked with probability 0.2 and supervised on all parts to predict the masked components (missing parts are initialized with mean poses).

At inference, standardized per-part outputs are produced via the 6D-to-axis–angle conversion and sparse vertex-to-keypoint projection, and then concatenated into a unified whole-body representation, as illustrated by the *Fused Network* in Fig. 2.

4 Experiments

Implementation Details CDPose is trained with sub-VP SDE Song et al. (2020b), where $t \sim \mathcal{U}(0, 1]$ during training and a truncated linear schedule $t \in [0.05, 0.15]$ is used at the test time. All pose parameters are represented as 6D rotations, normalized to zero mean and unit variance, and converted inside the model to axis–angle format for SMPL, MANO, and FLAME decoding.

The architecture follows the fully connected design of DPoser Lu et al. (2025). The part-level diffusion model contains about 7.2M parameters, while the whole-body model adds a fused residual head with about 4.1M parameters. Part models are trained with Adam, a learning rate of 1×10^{-4}, cosine decay, gradient clipping, and no multi-step refinement.

Methods	APD ↑	FID ↓	Prec. ↑	Rec. ↑	dNN ↓
AMASS Mahmood et al. (2019)	15.44	0.005	0.86	0.90	0.001
GMM Bogo et al. (2016)	16.28	1.02	0.13	0.34	4.37
VPoser Pavlakos et al. (2019)	10.75	0.66	0.29	0.42	3.74
GAN-S Davydov et al. (2022)	15.68	0.18	0.61	0.41	2.98
NDF Tiwari et al. (2022)	18.75	5.92	0.02	0.00	9.08
NRDF He et al. (2024)	**22.82**	0.64	0.03	**0.99**	6.69
DPoser Lu et al. (2025)	14.28	**0.07**	**0.72**	0.80	**2.63**
CDPose	**19.03**	**0.13**	**0.77**	**0.82**	**2.58**

Table 1: Comparative evaluation of pose-generation metrics. AMASS denotes the baseline for real-world samples.



Figure 3: Qualitative comparison of generated human poses. CDPose illustrates naturalistic poses aligned with real-world data compared to NRDF He et al. (2024) and DPoser Lu et al. (2025).

Methods	APD ↑	FID ↓	Prec. ↑	Rec. ↑	dNN ↓
VPoser Pavlakos et al. (2019)	1.99	0.21	0.68	0.65	1.85
NRDF He et al. (2024)	1.76	5.20	0.17	0.65	5.37
DPoser Lu et al. (2025)	2.36	0.01	0.82	0.87	1.45
CDPose	**2.90**	**0.02**	**0.85**	**0.88**	**1.05**

Table 2: Quantitative evaluation of hand pose generation.

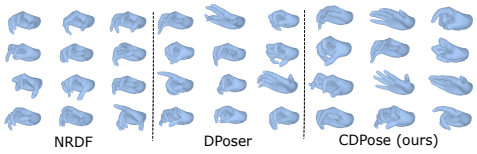

Figure 4: Qualitative comparison of hand pose generation.

For part-level training, we use AMASS Mahmood et al. (2019) for body pose with standard splits Pavlakos et al. (2019); Tiwari et al. (2022), HaMeR-based pseudo-ground-truth annotations Pavlakos et al. (2024) for hands, and MICA-based annotations Zielonka et al. (2022) for face. The body, hand, and face models are trained for 15K, 40K, and 50K iterations with batch sizes of 1280, 1280, and 256, respectively. For final whole-body training, we use BEAT2 Liu et al. (2024), GRAB Taheri et al. (2020), ARCTIC Fan et al. (2023), and EgoBody Zhang et al. (2022), and train for 330K iterations with batch size 1280 using RAdam and a learning rate of 2×10^{-4}.

4.1 Body Pose Generation Task

Body pose generation The unconditional synthesis shows that CDPose learns a strong prior, as shown in Table 1, attaining a low FID (0.13; second to DPoser's 0.07), the best dNN (2.58) and the highest APD (19.03), indicating samples closest to the data manifold with the highest pose diversity (vs. GAN-S 15.68, DPoser 14.28). Precision/Recall reach 0.77/0.82, clearly beating GAN-S (0.61/0.41) and edging DPoser in recall (0.80) while matching its precision (0.72). In contrast, NRDF sacrifices precision (0.03) for extreme recall (0.99) and a high dNN (6.69), and Pose-NDF has poor fidelity (FID 5.92) despite high APD. Overall, CDPose offers the best fidelity–diversity balance, approaching the AMASS oracle; qualitative gains are illustrated in Fig. 3 compared to NRDF He et al. (2024) and DPoser Lu et al. (2025).

4.2 Face and Hand Pose Generation Task

Hand Pose Generation We evaluate the generated hand poses based on their diversity and realism. As shown in Table 2, CDPose produces a strong combination of both, outperforming methods like VPoser Pavlakos et al. (2019), NRDF He et al. (2024), and DPoser Lu et al. (2025). Specifically, NRDF shows poor realism, reflected in high FID and dNN scores. VPoser, while achieving moderate precision, suffers from limited diversity, as indicated by its low APD. See Fig. 4 visualization comparison compared to NRDF He et al. (2024) and DPoser Lu et al. (2025).

Face generation We decouple FLAME shape and expression and sample each with our unified 6D ingestion, converting to axis–angle in-model before mesh decoding, using a late-time score-based schedule to balance fidelity and coverage. As shown in Table 3 CDPose achieves the best distributional match in both sub-tasks: for shape, FID 4.130 with precision/recall 0.989/0.456, surpassing VPoser 0.984/0.105 and DPoser 0.689/0.396; for expression, FID 0.124 with precision/recall 0.882/0.751, improving upon DPoser 0.156, 0.818/0.697 and VPoser 0.888, 0.993/0.019. The dNN metric is competitive (slightly higher for expressions), reflecting the broader coverage of CDPose.

	Methods	FID ↓	Prec. ↑	Rec. ↑	dNN ↓
Shape	VPoser Pavlakos et al. (2019)	31.91	0.984	0.105	6.52
	DPoser Lu et al. (2025)	5.331	0.689	0.396	8.29
	CDPose	**4.130**	**0.989**	**0.456**	**7.72**
Exp.	VPoser Pavlakos et al. (2019)	0.888	0.993	0.019	0.79
	DPoser Lu et al. (2025)	0.156	0.818	0.697	1.01
	CDPose	**0.124**	**0.882**	**0.751**	**1.53**

Table 3: Quantitative comparison of face generation across shape and expression.

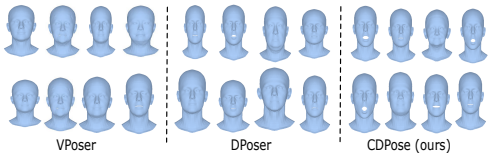

Figure 5: Qualitative comparison of face expression generation.

Methods	PA-MPVPE ↓			PA-MPJPE ↓
	All	Hands	Face	Body
w/o Ingestion	72.94	18.80	11.60	86.40
GMM Bogo et al. (2016)	67.25	17.93	10.92	79.08
VPoser-X Pavlakos et al. (2019)	66.74	17.44	10.99	79.88
DPoser-X Lu et al. (2025)	60.98	15.60	9.75	73.00
CDPose	**57.25**	**13.47**	**9.12**	**70.51**
SMPLer-X Cai et al. (2023a)	26.15	11.21	2.95	29.31
+ w/o Ingestion	26.33	10.34	2.86	28.69
+ GMM Bogo et al. (2016)	25.60	9.99	2.78	28.12
+ VPoser-X Pavlakos et al. (2019)	25.41	9.50	2.83	28.37
+ DPoser-X Lu et al. (2025)	24.65	7.33	2.73	27.87
+CDPose	**23.24**	**7.10**	**2.51**	**26.92**

Table 4: Whole-body mesh recovery results on ARCTIC Fan et al. (2023).

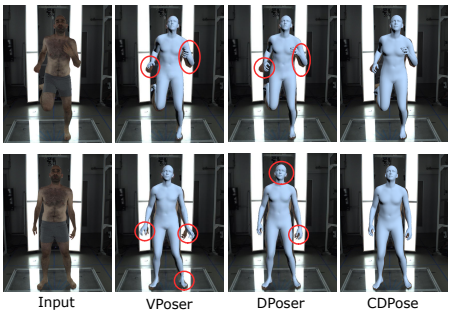

Figure 6: Visual comparison with SOTA VPoser and DPoser.

The qualitative results in Fig. 5 show more realistic facial expressions compared to VPoser Pavlakos et al. (2019) and DPoser Lu et al. (2025).

4.3 Whole-body pose Task

Whole-body mesh recovery We evaluate whole-body HMR on the ARCTIC dataset Fan et al. (2023) in two settings: (i) from scratch with RTMPose Jiang et al. (2023) 2D keypoints, and (ii) SMPLer-X–initialized inference Cai et al. (2023a) with ground-truth 2D keypoints. As shown in Table 4, CDPose achieves state-of-the-art results across all body regions in both configurations. From scratch, it yields the lowest errors on every component (All: 57.25, Hands: 13.47, Face: 9.12, Body PA-MPJPE: 70.51), improving over DPoser-X Lu et al. (2025) by roughly 6–10%, which suggests that our 6D rotation ingestion and diffusion-based denoising capture fine-grained articulation without sacrificing whole-body coherence. As a prior in the SMPLer-X–initialized setting, CDPose further lowers errors (All: 23.24, Hands: 7.10, Face: 2.51, Body: 26.92), surpassing GMM Bogo et al. (2016), VPoser-X Pavlakos et al. (2019), and DPoser-X Lu et al. (2025). Qualitative results in Fig. 6 confirm that CDPose produces coherent meshes with better-aligned body, hand, and face poses, enabling more accurate recovery of expressive articulated motion.

5 Ablation Studies

6D Rotation Ingestion. We replace the continuous 6D rotation representation with direct axis–angle optimization and compare across three tasks (Fig. 7a–c). On ARCTIC, 6D lowers PA-MPVPE from 63.92 to 57.25 mm overall, with the largest relative gain on hands (17.43 → 13.47 mm), and reduces body PA-MPJPE from 75.31 to 70.51 mm. On FreiHAND the effect is strongest under detector noise: PA-MPJPE drops from 17.71 to 9.90 mm with detected keypoints, while the gap to the ground-truth setting shrinks from 1.6 to 1.5 mm, showing that 6D mainly improves robustness rather than the clean-input optimum. On ReInterHand IK (σ = 2 mm), MPVPE falls from 3.95 to 1.22 mm. The continuous parameterization therefore stabilizes optimization of articulated parts and reduces whole-body error at no additional cost.

Denoising Objective and Timestep Policy. We evaluate the one-step Tweedie estimator Chung et al. (2022) with the truncated late-time schedule under increasing input noise (Fig. 7d–e). For body motion, MPJPE rises smoothly from 14.07/13.61 mm at σ = 20 mm to 23.65/24.98 mm at σ = 60 mm on AMASS/HPS; for hand IK, MPVPE rises from 1.22 mm at σ = 2 mm to 4.93 mm at

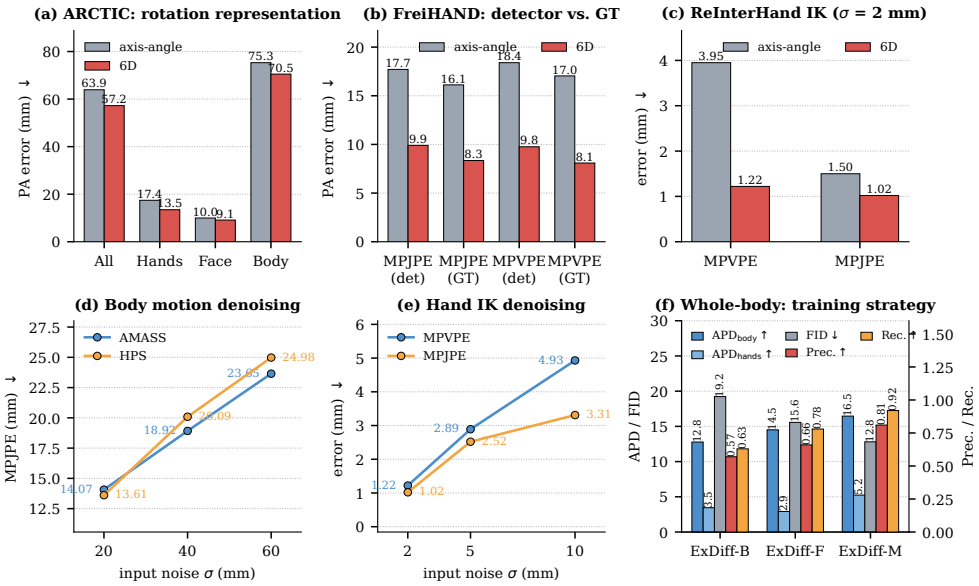

Figure 7: Ablations. (a–c) Replacing axis-angle with the continuous 6D rotation representation lowers error on ARCTIC (whole-body fitting from scratch), FreiHAND (detector and ground-truth keypoints), and ReInterHand hand IK ($\sigma = 2$ mm). (d–e) Denoising degrades gracefully with increasing input noise for body motion (AMASS/HPS, MPJPE) and hand IK (ReInterHand). (f) Mixed training (CDPOSE-M) improves both fidelity (FID, precision) and coverage (APD, recall) over the base and fused variants for whole-body generation.

$\sigma = 10$ mm, and MPJPE from 1.02 to 3.31 mm. Error grows roughly linearly with the noise level rather than collapsing, indicating that the denoising step degrades gracefully and remains useful well beyond the noise range seen at training time.

Part-Specific vs. Fused vs. Mixed Training. Finally, we compare the three training strategies on whole-body generation (Fig. 7f). Adding the cross-part fusion head (CDPOSE-F) to the independent part priors (CDPOSE-B) improves fidelity and coverage (FID 19.23 $\rightarrow$ 15.57, precision 0.57 $\rightarrow$ 0.66, recall 0.63 $\rightarrow$ 0.78) but slightly reduces hand diversity ($\text{APD}_{\text{hands}}$ 3.45 $\rightarrow$ 2.92), as the head pulls hands toward configurations coherent with the body. Mixed training (CDPOSE-M), which samples heterogeneous hand-only, face-only, and joint batches, recovers and surpasses that diversity ($\text{APD}_{\text{hands}}$ 5.24, APD_{body} 16.46) while further improving fidelity (FID 12.81, precision 0.81, recall 0.92). Fusion thus enforces inter-part coherence, and mixed training supplies the coverage that fusion alone sacrifices.

6 Conclusion and Limitations

We presented CDPOSE, a diffusion-guided whole-body pose model operating in a unified 6D rotation space, converting in-model to axis–angle for kinematic evaluation. With a sub-VP SDE, SNR-weighted training, and a one-step Tweedie estimator (plus optional short refinement), it provides a stable, data-consistent prior for body, hands, and face. A unified sparse vertex-to-keypoint projector enables protocol-agnostic evaluation, while a fused residual head and mixed training capture cross-part coordination and increase coverage. Experiments across body, hand, face, and whole-body tasks show consistent improvements over state-of-the-art methods, establishing CDPOSE as a new baseline. Key insights are that 6D-space learning/denoising benefits fine articulations and that cross-part fusion with mixed supervision is crucial for coherent whole-body modeling.

The primary limitations of the current framework arise from its dependence on conventional parametric models and fixed, sparsely defined regressors. Future research will seek to transcend purely parametric representations by investigating non-parametric mesh formulations and dynamically evolving topological structures. In addition, we intend to integrate temporal priors into both uncertainty quantification and self-supervised learning procedures.

REFERENCES

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 *European conference on computer vision*, pp. 561–578. Springer, 2016.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, 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*, 36:11454–11468, 2023a.

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*, 36:11454–11468, 2023b.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

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*, pp. 9044–9053, 2021.

Hanbyel Cho and Junmo Kim. Generative approach for probabilistic human mesh recovery using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4183–4188, 2023.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4800–4810, 2023.

Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial parametric pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10997–11005, 2022.

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*, pp. 12943–12954, 2023.

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*, pp. 9919–9928, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4318–4329, 2021.

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*, pp. 3196–3206, 2020.

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*, pp. 1661–1671, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15977–15987, 2023.

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.

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. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6142–6152, 2024.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

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*, 3(8), 2022.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5253–5263, 2020.

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*, pp. 10138–10148, 2021.

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.

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.

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*, pp. 590–606. Springer, 2022.

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020.

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*, pp. 1144–1154, 2024.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, et al. Dposer-x: Diffusion model as robust 3d whole-body human pose prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9988–9997, 2025.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

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*, pp. 548–564. Springer, 2020.

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*, 36:17689–17701, 2023.

Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9687–9697, 2024.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2024.

Z Qiu, Q Yang, J Wang, X Wang, C Xu, D Fu, K Yao, J Han, E Ding, and J Wang. Learning structure-guided diffusion model for 2d human pose estimation (2023). *arXiv preprint arXiv:2306.17074*, 2023.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11488–11499, 2021.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017.

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*, pp. 7763–7772, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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*, 2020b.

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pp. 581–600. Springer, 2020.

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 *European Conference on Computer Vision*, pp. 572–589. Springer, 2022.

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*, pp. 777–787, 2024.

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

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*, pp. 180–200. Springer, 2022.

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*, pp. 14606–14617, 2024.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753, 2019.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pp. 250–269. Springer, 2022.

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*, pp. 813–822, 2019.

A APPENDIX

This appendix provides comprehensive technical details and supplementary experiments that support the main paper. In Section B, we present additional experiments for the pose generator: Section B.1 analyzes body-pose denoising strategies on multiple noise scales, while Section B.2 evaluates inverse-kinematics–based denoising for hand poses under varying noise magnitudes. Section B.4 reports a comparative study of hand-mesh recovery when initialized from the mean pose.

Face-related experiments are given in two paragraphs. Section B.5 compares the results of face reconstruction between fitting from scratch and initialization with MICA Zielonka et al. (2022), and Section B.6 provides a detailed analysis of face denoising at several noise scales. Section B.8 describes the whole-body pose completion experiments used to fairly evaluate our fused model and the proposed mixed-training regime.

Section C describes the datasets used for training and evaluation, including preprocessing steps and any dataset-specific splits. Section D details the experimental setup, evaluation metrics for each task, and implementation details for the comparative baselines. Section E analyzes runtime performance and the computational overhead introduced by CDPOSE.

We provide additional visual evaluations of CDPOSE from all generators, including body only, hand, face, and whole body pose in the supplementary Section F.

B ADDITIONAL EXPERIMENTS

B.1 BODY POSE DENOISING

We evaluate motion denoising by corrupting AMASS 60-frame sequences with i.i.d. Gaussian noise on multiple scales ($\sigma = 40$ mm, $\sigma = 40$ mm, $\sigma = 40$ mm) on 3D joints and by testing zero-shot on HPS Guzov et al. (2021) without additional training. In line with our denoising methodology (Section 5), CDPOSE operates entirely in the 6D rotation space, applying the one-step Tweedie estimator (Eq. 7 in the main paper) with the SNR-weighted objective (Eq. 8 in the main paper). We sweep a truncated late-time schedule and map the cleaned 6D poses to axis–angle before kinematic decoding. As reported in Table 5, CDPOSE achieve the lowest MPJPE on AMASS (18.92 mm) and HPS (20.09 mm), outperforming VPoser, Pose-NDF, HuMoR, and DPoser, indicating a stronger pose prior and improved generalization under noise.

Methods	**AMASS** Mahmood et al. (2019) ↓	**HPS** Guzov et al. (2021) ↓
w/o Ingestion	23.92	23.02
VPoser Pavlakos et al. (2019)	23.42	22.78
Pose-NDF Tiwari et al. (2022)	22.13	21.60
MVAE Ling et al. (2020)	26.80	N/A
HuMoR Rempe et al. (2021)	22.69	N/A
DPoser Lu et al. (2025)	19.87	20.54
CDPOSE ($\sigma = 20$ mm)	14.07	13.61
CDPOSE ($\sigma = 40$ mm)	18.92	20.09
CDPOSE ($\sigma = 60$ mm)	23.65	24.98

Table 5: Performance metrics (MPJPE) for pose denoising. σ indicates noise scale
.

B.2 HAND POSE DENOISING

We evaluate inverse-kinematics denoising on ReInterHand Moon et al. (2023) by corrupting 3D hand joints with i.i.d. Gaussian noise at $\sigma \in \{2, 5, 10\}$ mm and recovering MANO poses via our 6D-rotation prior. Following our denoising scheme, CDPOSE operates in the 6D joint space, applies the one-step Tweedie estimator with SNR-weighted consistency, sweeps a truncated late-time schedule, and maps the cleaned 6D rotations to axis–angle for kinematic decoding. As reported in Table 6, CDPOSE achieve the lowest errors at $\sigma = 2$ mm (MPVPE/MPJPE $= 1.22/1.02$ mm), outperforming

Methods	**MPVPE** ↓	**MPJPE** ↓
w/o Ingestion	3.95	1.50
VPoser Pavlakos et al. (2019)	2.47	1.36
NRDF He et al. (2024)	2.67	1.40
DPoser Lu et al. (2025)	1.71	1.17
CDPose ($\sigma = 2$ mm)	1.22	1.02
CDPose ($\sigma = 5$ mm)	2.89	2.52
CDPose ($\sigma = 10$ mm)	4.93	3.31

Table 6: Performance of hand inverse kinematics on the ReInterHand dataset Moon et al. (2023) under noisy settings.

Initialization	GMM	VPoser	Pose-NDF	NRDF	GAN-S	DPoser	**CDPose**
from scratch	108.57	58.32	58.08	57.87	57.38	56.05	**55.08**
CLIFF	56.62	51.02	49.39	49.50	49.27	49.05	**48.16**

Table 7: Human mesh recovery performance on the EHF Pavlakos et al. (2019) dataset from scratch and CLIFF Li et al. (2022) initialization.

VPoser Pavlakos et al. (2019), NRDF He et al. (2024), and DPoser Lu et al. (2025). Performance degrades gracefully with increased noise ($2.89/2.52$ mm at 5 mm; $4.93/3.31$ mm at 10 mm) and consistently remains well below the no-ingestion baseline, evidencing a robust denoising prior for hand IK.

B.3 Body mesh recovery

We evaluate CDPose on EHF Pavlakos et al. (2019) for monocular human mesh recover (HMR) under two initializations: (i) from scratch (mean pose) and (ii) CLIFF Li et al. (2022) as presented in Table. 7. CDPose achieves the lowest PA-MPJPE across baselines GMM Bogo et al. (2016), VPoser Pavlakos et al. (2019), Pose-NDF Tiwari et al. (2022), NRDF He et al. (2024), GAN-S Davydov et al. (2022), DPoser Lu et al. (2025). Errors drop to 55.08,mm from scratch (vs. ≈ 56–58,mm) and to 48.16,mm with CLIFF (vs. ≈ 49–51,mm). These results show CDPose is an effective 6D-driven prior that both improves strong regressors and enables reliable convergence.

B.4 Hand Mesh Recovery

We evaluate hand mesh recovery on FreiHAND Zimmermann et al. (2019) under mean-pose initialization using MediaPipe Hands Zhang et al. (2020) detected 2D keypoints compared with ground-truth (GT) keypoints. As summarized in Table 8, CDPose attains the best PA-MPJPE/PA-MPVPE with MediaPipe Hands Zhang et al. (2020) detections, demonstrating strong robustness to 2D keypoint noise (Table 8). Compared to VPoser Pavlakos et al. (2019), it yields (19.6%) (PA-MPJPE) and (20.2%) (PA-MPVPE) reductions; vs. NRDF He et al. (2024) the gains are larger (25.0%)/(27.1%); and it also improves over DPoser Lu et al. (2025) under detector noise (7.6%)/(6.9%). Removing 6D Rotation Ingestion causes a large degradation (44.1%), confirming the benefit of operating/-denoising in unified 6D space. With GT keypoints, CDPose matches DPoser at (8.68/8.70) mm (PA-MPJPE/PA-MPVPE), indicating that the method primarily enhances *robustness* while preserving state-of-the-art fidelity.

B.5 Face reconstruction

We assess single–image face reconstruction on NOW Sanyal et al. (2019) under two regimes: (i) fitting from scratch and (ii) initialization with MICA Zielonka et al. (2022). Optimization uses our 6D–to–axis–angle ingestion for FLAME jaw/eye rotations with jointly optimized expressions. From scratch, CDPose achieves the lowest errors in both the full split and the side-view subset, matching DPoser: (med/mean/std=11.68/14.58/12.17) (all) and (11.77/14.67/12.34) (side), outperforming VPoser and the no-ingestion baseline (Table 9). With MICA initialization, CDPose

Methods	PA-MPJPE ↓	PA-MPVPE ↓
w/o Ingestion	17.71/16.12	18.40/17.04
VPoser Pavlakos et al. (2019)	12.31/10.62	12.23/10.91
NRDF He et al. (2024)	13.19/11.04	13.39/11.59
DPoser Lu et al. (2025)	10.71/8.68	10.48/8.70
CDPOSE	9.90/8.68	9.76/8.70

Table 8: Performance evaluation of hand mesh recovery on the FreiHAND dataset Zimmermann et al. (2019). Results are reported using 2D keypoints detected by MediaPipe Hands Zhang et al. (2020) / ground truth.

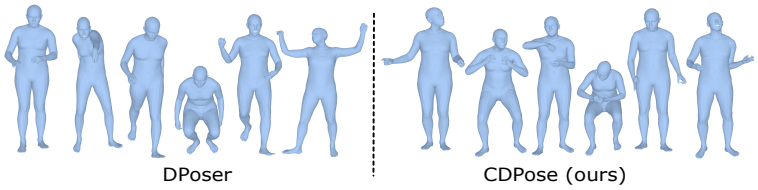

Figure 8: Qualitative comparison of whole-body pose generation with DPoser Lu et al. (2025).

yields further gains—(7.68/9.34/8.72) (all) and (8.21/10.92/9.10) (side), surpassing MICA alone and MICA+VPoser. Adding VPoser to MICA degrades performance relative to MICA, while CDPOSE consistently improves stability and accuracy, especially for side views.

Methods	**all ↓**			**side-view ↓**		
	med.	**mea.**	**std.**	**med.**	**mea.**	**std.**
w/o Ingestion	12.97	16.07	13.15	13.15	16.22	13.26
VPoser Pavlakos et al. (2019)	12.23	15.13	12.71	12.20	15.35	12.98
DPoser Lu et al. (2025)	11.68	14.58	12.17	11.77	14.67	12.34
CDPOSE	10.91	13.96	11.75	11.05	13.11	12.09
MICA Zielonka et al. (2022)	9.03	11.12	9.24	9.29	11.71	10.04
+ w/o Ingestion	9.01	11.09	9.15	9.47	11.80	9.84
+ VPoser Pavlakos et al. (2019)	9.93	12.34	10.43	10.00	12.50	10.65
+ DPoser Lu et al. (2025)	8.76	10.78	9.00	9.18	11.47	9.73
+ CDPOSE	7.68	9.34	8.72	8.21	10.92	9.10

Table 9: Face reconstruction performance on the NOW benchmark Sanyal et al. (2019). Results are reported as median, mean, and standard deviation of MPVPE.

B.6 FACE DENOISING

We evaluate robustness to measurement noise and occlusion on WCPA Kao et al. (2022) using our 6D–space denoising pipeline: one-step Tweedie estimation (Eq. 7 in the main paper) with SNR-weighted consistency loss (Eq. 8 in the main paper), followed by in-model 6D to AA conversion for FLAME decoding. As summarized in Table 10, CDPOSE achieves the lowest error at mild noise ($\sigma = 1$ mm: MPVPE/MPJPE $= 0.715/0.548$), improving over DPoser ($0.784/0.584$) and VPoser ($1.153/0.803$). Error increases with stronger perturbations, yet degrades gracefully ($\sigma = 2, 5$ mm), indicating that denoising in the unified 6D rotation space provides a stable, data-consistent prior under both noise and occlusion.

B.7 WHOLE-BODY POSE GENERATION

As summarized in Table 11, CDPOSE yields the best fidelity–diversity trade-off among all baselines, achieving the lowest FID (12.81), the highest Precision/Recall (0.81/0.92), and the largest diversity $\text{APD}_{\text{body}}/\text{APD}_{\text{hands}}$ (16.46/5.24). In contrast, VPoser-X shows limited variability (APD 7.79/1.16)

Methods	**MPVPE** ↓	**MPJPE** ↓
w/o Ingestion	1.460	0.878
VPoser Pavlakos et al. (2019)	1.153	0.803
DPoser Lu et al. (2025)	0.784	0.584
CDPOSE ($\sigma = 1$ mm)	0.715	0.548
CDPOSE ($\sigma = 2$ mm)	1.102	1.020
CDPOSE ($\sigma = 5$ mm)	1.989	1.971

Table 10: Performance of face inverse kinematics on the WCPA dataset Kao et al. (2022) under noisy and occlusion settings.

Methods	APD_{body}↑	APD_{hands}↑	FID ↓	Prec.↑	Rec.↑
VPoser-X Pavlakos et al. (2019)	7.79	1.16	14.52	0.64	0.07
DPoser-X Lu et al. (2025)	14.08	2.04	13.97	0.02	0.81
CDPOSE	16.46	5.24	12.81	0.81	0.92

Table 11: Quantitative evaluation of whole-body pose generation.

with very low recall (0.07), while DPoser-X achieves a higher recall (0.81), but suffers from extremely low precision (0.02) and a worse FID (13.97). In general, CDPOSE produces more realistic and varied whole-body samples with strong coverage and on-manifold consistency; see Fig. 8 for a visual comparison with DPoser Lu et al. (2025).

B.8 WHOLE-BODY POSE COMPLETION

We evaluate inter-part reasoning by randomly masking one hand and completing the full pose from the remaining body, hand, and face evidence. Completion is performed in the 6D rotation space using our fused models and mixed-training regime, and we report MPVPE as minimum, mean, and standard deviation. As shown in Table 12, CDPOSE achieves the best results on both ARCTIC (15.61/25.58/6.17, mm) and BEAT2 (12.94/21.67/5.39, mm), outperforming VPoser-X and DPoser-X. These gains indicate that the fused technique and mixed training strengthen cross-part dependencies—particularly coordination between the two hands—while preserving diversity for plausible completions.

C DATASETS

This section provides a comprehensive overview of the datasets used to train and evaluate the proposed framework. The datasets are grouped according to the body region or task they focus on, and for each dataset we summarize the intended use, data scale, and relevant preprocessing steps. Together, these datasets cover body, hand, face, and full-body motion domains, enabling a unified and diverse training corpus for the CDPOSE models.

C.1 BODY-POSE DATASETS

AMASS The AMASS dataset Mahmood et al. (2019) aggregates high-quality 3D human body motions from numerous motion capture sources, providing full-body pose sequences represented in the SMPL parameterization. Following the data splits adopted by DPoser Lu et al. (2025) and after removing redundant frames through temporal subsampling, we obtain approximately 55 million SMPL-X Pavlakos et al. (2019) body poses for training the CDPOSE-body model. The evaluation split contains 54,000 poses, which is used to assess model performance on body-pose completion and motion-denoising tasks.

HPS The HPS dataset Guzov et al. (2021) contains more than 300K synchronized RGB images with corresponding 3D pose and location annotations from seven subjects interacting with large-scale real-world scenes. The dataset covers a wide range of activities, including exercising, reading, computer use, lecturing, and dancing. In line with prior work such as Pose-NDF Tiwari et al. (2022)

Methods	**ARCTIC** Fan et al. (2023)			**BEAT2** Liu et al. (2024)		
	min.	**mea.**	**std.**	**min.**	**mea.**	**std.**
VPoser-X Pavlakos et al. (2019)	37.34	43.24	4.60	27.49	35.46	5.06
DPoser-X Lu et al. (2025)	21.81	30.99	6.10	15.92	25.89	6.04
CDPOSE	15.61	25.58	6.17	12.94	21.67	5.39

Table 12: Performance metrics for whole-body pose completion. MPVPE is reported as *minimum/mean/standard deviation* (mm); lower is better for MPVPE.

and DPoser Lu et al. (2025), the HPS dataset is used exclusively for evaluation of motion denoising, without using it for training. After uniform sequence sampling, we obtain 350 sequences, each consisting of 60 frames, for testing.

C.2 HAND-POSE DATASETS

FreiHAND The FreiHAND dataset Zimmermann et al. (2019) provides large-scale 3D single-hand pose annotations paired with multi-view RGB images. It includes 130,240 training samples (four rendered views per pose) and 3,960 evaluation samples. From this dataset, we use 32,560 MANO Romero et al. (2017) hand poses to train the CDPOSE-hand model, while the 3,960 evaluation samples are used for hand-mesh recovery benchmarks.

DexYCB DexYCB Chao et al. (2021) captures 3D hand-object interaction sequences. For the purposes of this work, only the hand-pose annotations are used. The training set contains 407,000 single-hand poses, which contribute to training the CDPOSE-hand model.

HO3D HO3D Hampali et al. (2020) provides paired 3D annotations for hand poses and interacting objects. Similar to DexYCB, we use only the hand-pose annotations, yielding 83,000 single-hand poses for training CDPOSE-hand.

H2O The H2O dataset Kwon et al. (2021) includes 3D annotations for two-hand poses and hand–object interactions. For training CDPOSE-hand, only the right-hand poses are used, resulting in 58,000 poses.

C.2.1 REINTERHAND

ReInterHand Moon et al. (2023) is a synthetic photorealistic dataset designed for interacting-hand pose estimation. It provides 3D annotations for both hands, which we convert to a unified right-hand format by horizontally flipping left-hand poses. With an 8:1:1 train/val/test split, we use approximately 186,000 poses for training and 23,000 poses for testing, with the test set supporting hand inverse-kinematics evaluation.

C.3 FACE DATASETS

MICA The MICA dataset Zielonka et al. (2022) merges eight smaller datasets into a unified collection of 2,315 subjects, represented using the FLAME model Li et al. (2017). As it provides only shape geometry (no expressions), it is used to train the shape component of CDPOSE-face, focusing on high-fidelity 3D identity reconstruction.

WCPA The WCPA dataset Kao et al. (2022) contains 200 subjects and 356,640 training samples, with detailed 3D expression annotations under perspective projection. We use this dataset to train the expression component of CDPOSE-face, allocating 10

NOW The NOW benchmark Sanyal et al. (2019) provides 352 validation images for evaluating 3D face shape reconstruction under challenging conditions such as varying viewpoint, illumination, and occlusions. Since the ground-truth contains only shape, evaluation follows prior work such as

DECA **?**, where expression parameters are set to zero in the FLAME model to produce a neutral face mesh for non-metrical comparison.

C.4 WHOLE-BODY DATASETS

BEAT2 BEAT2 Liu et al. (2024) is a co-speech motion dataset combining MoSh'ed SMPL-X body parameters with FLAME facial parameters, enabling coherent modeling of body, head, and finger motions. After de-duplication, we obtain 1.48 million whole-body poses for training CDPOSE-X. The test split contains 172,000 poses, used for whole-body pose completion.

GRAB The GRAB dataset Taheri et al. (2020) captures full-body shape and motion sequences of subjects interacting with 51 everyday objects. We use 391,000 whole-body poses derived from this dataset to train CDPOSE-X, informing whole-body mesh recovery during grasping and manipulation tasks.

ARCTIC ARCTIC Fan et al. (2023) provides 2.1 million video frames of two-hand object manipulation with full 3D annotations for hands and objects. After de-duplication, we use 77,000 whole-body poses to train CDPOSE-X. The validation set contains 10,000 poses, used for whole-body mesh recovery and pose completion. As facial expressions are not annotated, we set FLAME expression parameters to zero during evaluation; thus, face-related metrics depend only on shape.

EgoBody EgoBody Zhang et al. (2022) captures full 3D human motions during dyadic social interactions from both an egocentric and a third-person viewpoint. It includes SMPL-X annotations for body shape, pose, and motion. The training set comprises 38,896 whole-body pose instances (duplicated for both subjects), and the test set includes 24,665 instances, used to evaluate whole-body pose completion.

Fit3D Fit3D Fieraru et al. (2021) contains more than 3 million images with the corresponding 3D human motion and shape annotations in 37 types of exercises. To evaluate generalization, we use subject s04 exclusively, yielding 612 sampled images, without incorporating Fit3D into training.

EHF The EHF dataset Pavlakos et al. (2019) consists of 100 curated images with pseudo ground-truth whole-body pose annotations. Following Pose-NDF Tiwari et al. (2022) and DPoser Lu et al. (2025), we use EHF to evaluate total body mesh recovery, reporting PA-MPJPE for body joints.

D EXPERIMENTAL DETAILS

Scope of the diffused variable. Unless stated otherwise, we diffuse only joint rotations in 6D; facial expressions $\boldsymbol{\psi}$ and shapes $\boldsymbol{\beta}$ are kept fixed during diffusion and can be optimized as auxiliary variables at test time. We denote the number of articulated joints by J so that $x_0 \in \mathbb{R}^{J\times 6}$.

Training/test-time timestep policy and SDE. We adopt the sub-VP SDE; during training, t is drawn uniformly from $(0, 1]$. For test-time regularization, we use a truncated linear schedule $t \in [t_{\min}, t_{\max}]$ with a late-time emphasis (default $[0.05, 0.15]$), detailed in Section 5 of the main paper.

Optimization details. Unless specified, we train with AdamW, batch size B, learning rate 10^{-4}, cosine decay, and gradient clipping. Multi-step refinement is disabled by default.

Body Training. We train the body model on AMASS Mahmood et al. (2019), adopting the standard data splits used by previous work Pavlakos et al. (2019); Tiwari et al. (2022). Joint rotations are represented in a unified 6D format and ingested in-model to axis–angle before kinematic evaluation; per-joint features are standardized to zero mean and unit variance. The backbone is a fully connected network with approximately 7.2,M trainable parameters, following the unconditional pose prior design of Lu et al. (2025). Optimization uses RAdam for 15,000 iterations with a learning rate of $2 \cdot 10^{-4}$ and a batch size of 1,280.

Hand Training. We train the hand model on HAMER Pavlakos et al. (2024). Joint rotations are represented in a unified 6D format and ingested in-model to axis–angle before kinematic evaluation; per-joint features are standardized to zero mean and unit variance. The backbone is a fully connected network with approximately 7.2,M trainable parameters, following the unconditional pose prior design of Lu et al. (2025). Optimization uses RAdam for 40,000 iterations with a learning rate of $2 \cdot 10^{-4}$ and a batch size of 1,280.

Face Training. We train the face model on MICA datasets Zielonka et al. (2022). Joint rotations are represented in a unified 6D format and ingested in-model to axis–angle before kinematic evaluation; per-joint features are standardized to zero mean and unit variance. The backbone is a fully connected network with approximately 7.2,M trainable parameters, following the unconditional pose prior design of Lu et al. (2025). Optimization uses RAdam for 50,000 iterations with a learning rate of $2 \cdot 10^{-4}$ and a batch size of 256.

Whole-body Pose Training. We train the whole-body model on BEAT2 Liu et al. (2024), GRAB Taheri et al. (2020), ARCTIC Fan et al. (2023), and EgoBody Zhang et al. (2022) datasets. Joint rotations are represented in a unified 6D format and ingested in-model to axis–angle before kinematic evaluation; per-joint features are standardized to zero mean and unit variance. The backbone is a fully connected network with approximately 4.1,M trainable parameters, following the unconditional pose prior design of Lu et al. (2025). Optimization uses RAdam for 330,000 iterations with a learning rate of $2 \cdot 10^{-4}$ and a batch size of 1280.

D.1 Evaluation Metrics

For a complete assessment of various tasks, following recent work such as NRDF He et al. (2024) and SMPLer-X Cai et al. (2023a), we adopt task-specific metrics:

Pose Generation: Diversity and fidelity are evaluated using the average pairwise distance (APD) and dNN He et al. (2024), respectively. dNN measures the distance between the generated pose and its nearest neighbor from the training data. We also report common metrics for generative models, including FID (distribution similarity), Precision (fidelity), and Recall (diversity).

Human Mesh Recovery: The Procrustes-aligned Mean PerVertex Position Error (PA-MPJPE) and the Procrustes-aligned Mean Per-Joint Position Error (PA-MPVPE) measure 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 performance. All errors are reported in millimeter units.

Method	w/o Ingestion	GMM	VPoser	Pose-NDF	GAN-S	DPoser	CDPose
Recovery	14.39	16.14	16.83	21.88	74.60	17.34	15.05
Generation	10.78	14.56	14.90	19.56	71.89	15.12	12.78
Denoising	11.39	13.02	13.89	17.37	72.55	16.03	13.41

Table 13: Runtime comparison (seconds) of different prior models for human mesh recovery, mesh-generation, and mesh-denoising evaluated on 100 images using an NVIDIA RTX 3090 Ti GPU.

E Runtime Comparison

Diffusion-based methods typically rely on iterative denoising procedures, which are computationally more demanding than feedforward architectures such as VAEs and GANs. Despite this general drawback, the integration of CDPose into downstream optimization pipelines introduces only minimal additional overhead. This efficiency is primarily attributed to two design choices: (1) CDPose

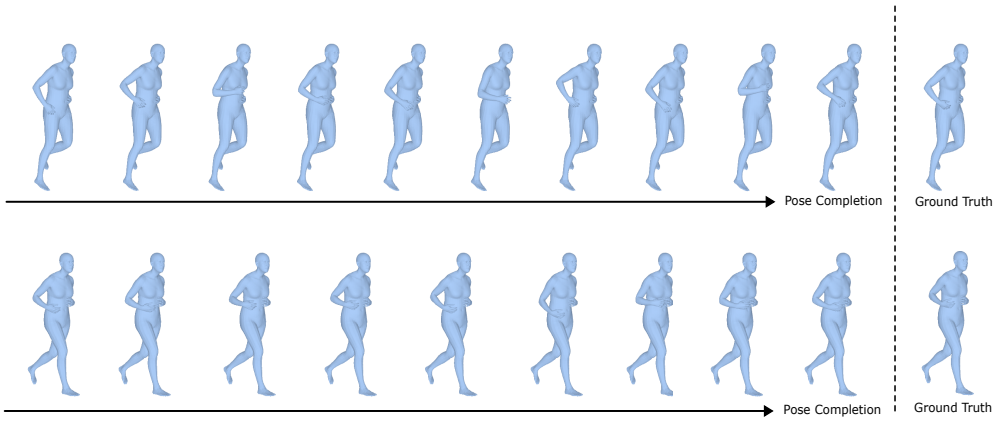

Figure 9: Pose completion for the right hand. CDPOSE offers multiple plausible solutions for partial poses, a scenario where competitors often struggle due to limited generalization.

employs a single-step denoising operation at each iteration of the optimization process, rather than a full multi-step diffusion schedule; and (2) the use of a stop-gradient operator, which ensures that the regularization term does not require gradient backpropagation through the pretrained diffusion network.

To quantify the computational impact, runtime measurements were conducted in 100 images for three tasks—human mesh recovery, mesh generation, and mesh denoising—under a consistent hardware and software environment. Table 13 summarizes the results and compares CDPOSE to several alternative pose priors, including a baseline without any prior. Incorporating CDPOSE yields only a modest increase in the overall optimization time (approximately 10% relative to the baseline), demonstrating its practical efficiency. In contrast, GAN-S Davydov et al. (2022) exhibits substantially higher runtime due to its reliance on a computationally expensive GAN-inversion stage, which maps initial poses into the model's latent space before optimization proceeds.



Figure 10: Visualization of body-pose generation. The generator produces diverse, anatomically plausible, and visually realistic configurations across a broad range of articulations and viewpoints.

F COMPARATIVE VISUAL RESULTS

A comprehensive set of additional qualitative results is provided to further illustrate the capabilities of the proposed CDPOSE framework across all tasks. Body-pose completion for the right hand examples is presented in Fig. 9, followed by qualitative results for body-pose generation in Fig. 10. Figure 11 showcases body-mesh recovery outputs. Hand-related tasks are visualized in Fig. 12 for

hand-pose generation, Fig. 13 for inverse-kinematics–based hand-pose denoise prediction, and Figs. 14 for hand-mesh recovery. Face-specific visualizations include face generation in Fig. 15. Fig. 16 shows inverse-kinematics denoising and face-reconstruction results in Fig. 17. The whole-body results are presented in Figs. 18 for the generation of whole-body poses, Fig. 19 for whole-body whole-body pose completion, and Figs. 20 for whole-body mesh recovery. Together, these results provide a comprehensive view of the performance and generalizability of the proposed model across the full spectrum of body, hand, face, and whole-body tasks.

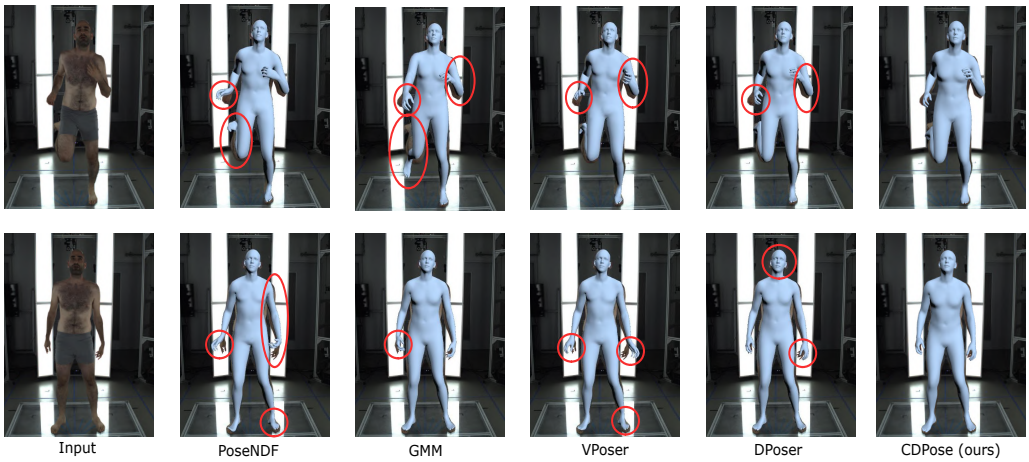

Figure 11: Visualization of body-mesh recovery results. For a fair comparison, the visual outputs of PoseNDF Tiwari et al. (2022), GMM Bogo et al. (2016), VPoser Pavlakos et al. (2019), and DPoser Lu et al. (2025) were obtained directly from the published DPoser resources, and the corresponding predictions from CDPOSE are shown alongside for qualitative comparison.

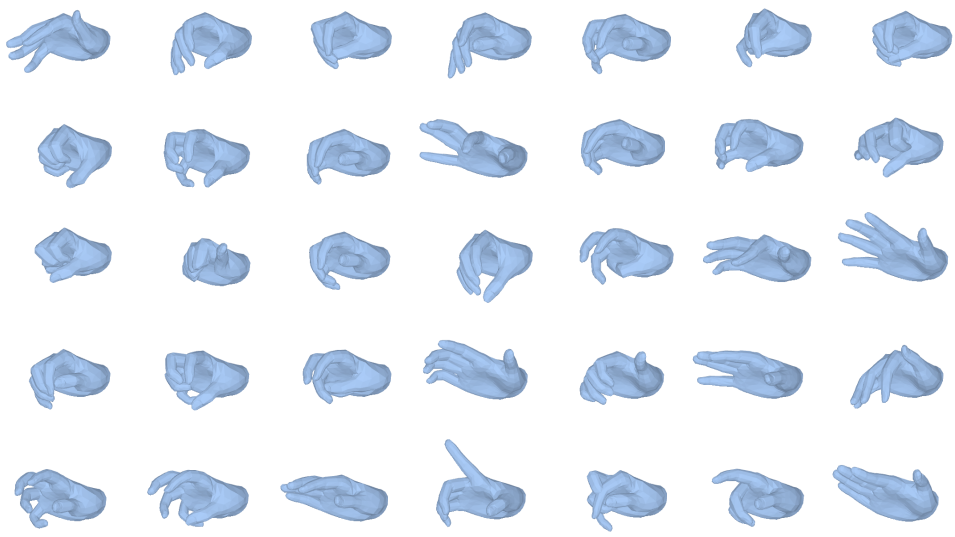

Figure 12: Visualization of hand-pose generation. CDPOSE synthesizes plausible and realistic hand poses.

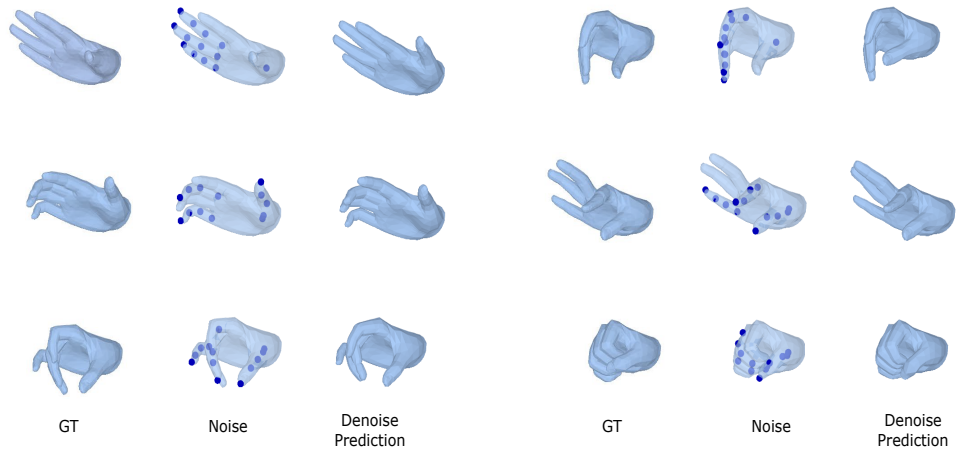

Figure 13: Visualization of hand-denoising prediction using via inverse kinematics. The ground truth (GT) is obtained from the ReInterHand Moon et al. (2023), with all samples corrupted by 2 mm noise.

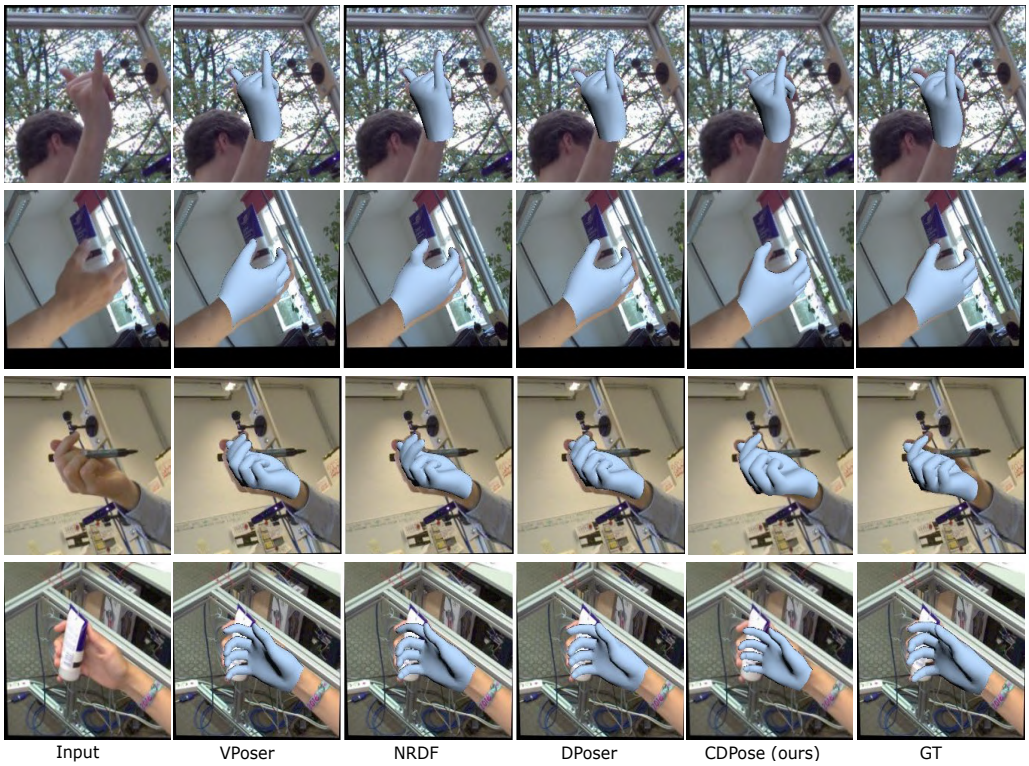

Figure 14: Visualization of hand mesh recovery with mean pose initialization. For a fair comparison, the visual outputs of VPoser Pavlakos et al. (2019), NRDF He et al. (2024), and DPoser Lu et al. (2025) were obtained directly from the published DPoser resources, and the corresponding predictions from CDPOSE are shown alongside for qualitative comparison.



Figure 15: Visualization of face generation results. Rows a and b show varying face shapes; rows c, d, and e show varying expressions.

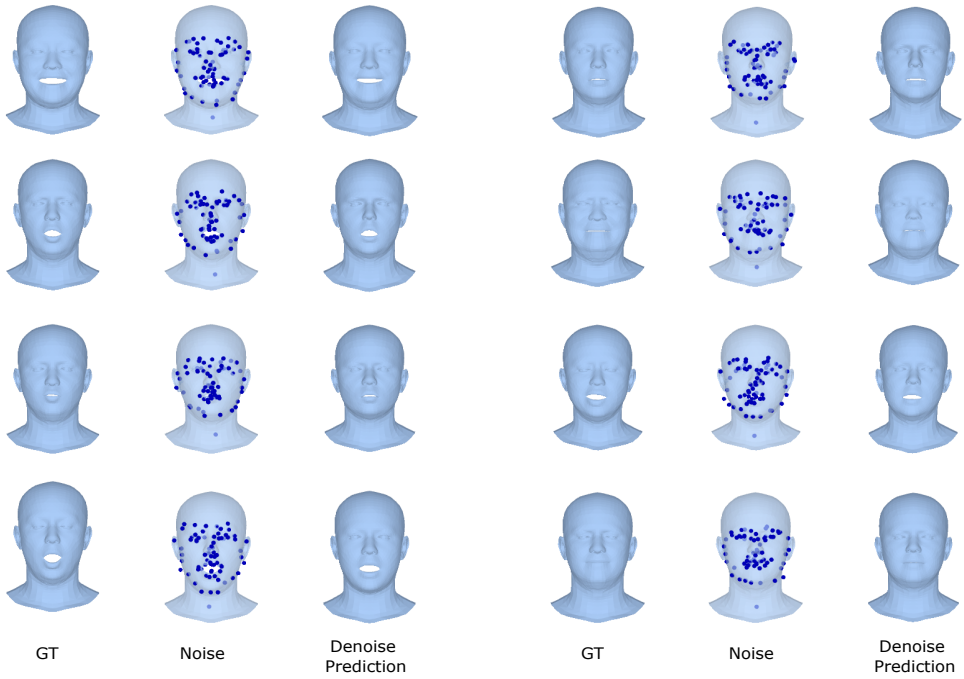

Figure 16: Visualization of face-denoising results obtained via inverse kinematics. The ground truth (GT) is obtained from the WCPA dataset Kao et al. (2022), with all samples corrupted by 5 mm noise.

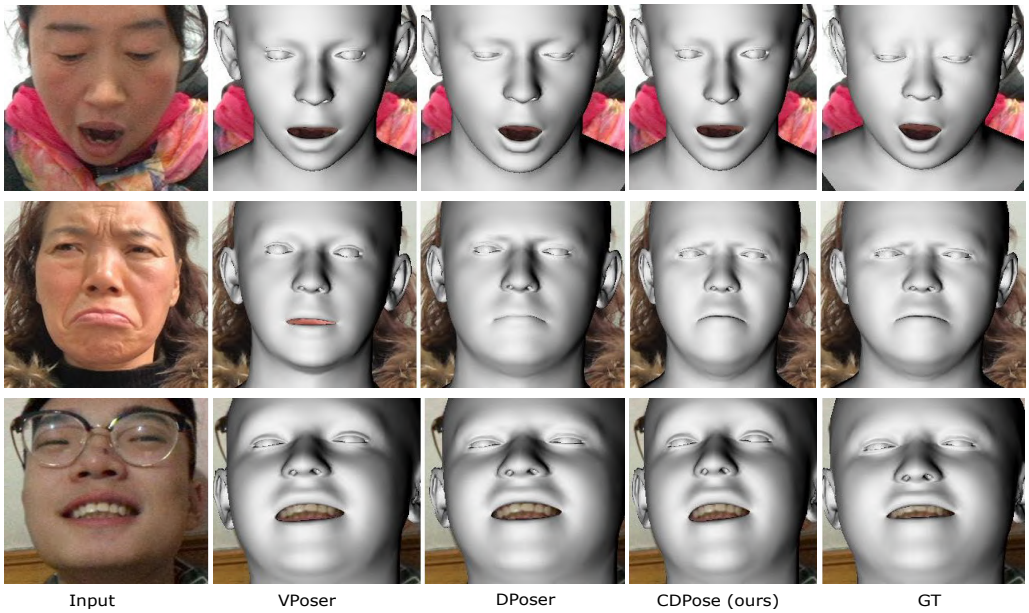

Figure 17: Face-reconstruction visualizations for representative samples from the WCPA dataset Kao et al. (2022). For a fair qualitative comparison, the visual outputs of VPoser Pavlakos et al. (2019) and DPoser Lu et al. (2025) were obtained from the published DPoser resources; the corresponding reconstructions produced by CDPOSE are shown alongside.

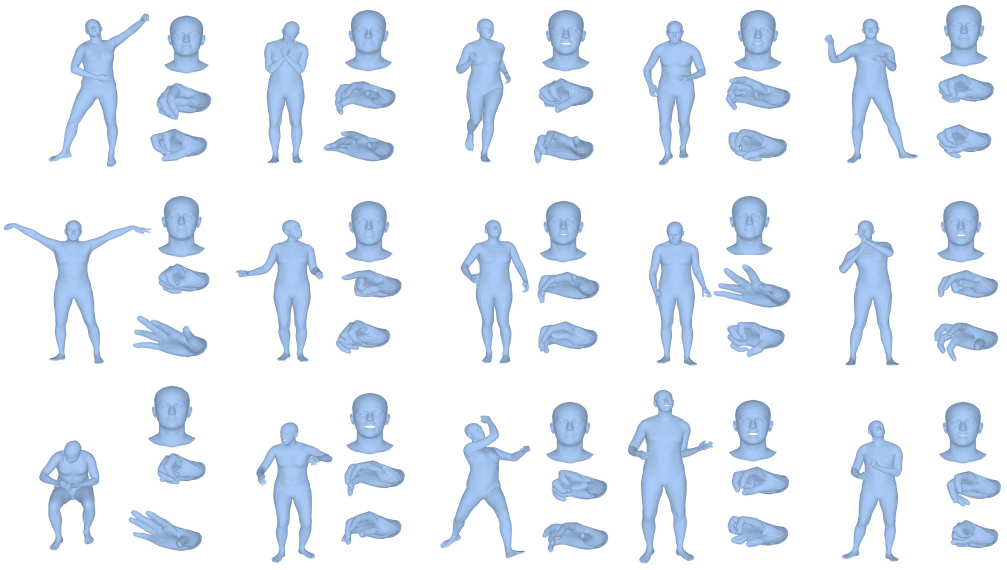

Figure 18: Visualization of whole-body pose generation. CDPOSE produces samples that achieve a favorable trade-off between diversity and anatomical/visual realism, yielding both varied and physically plausible whole-body configurations suitable for downstream tasks.

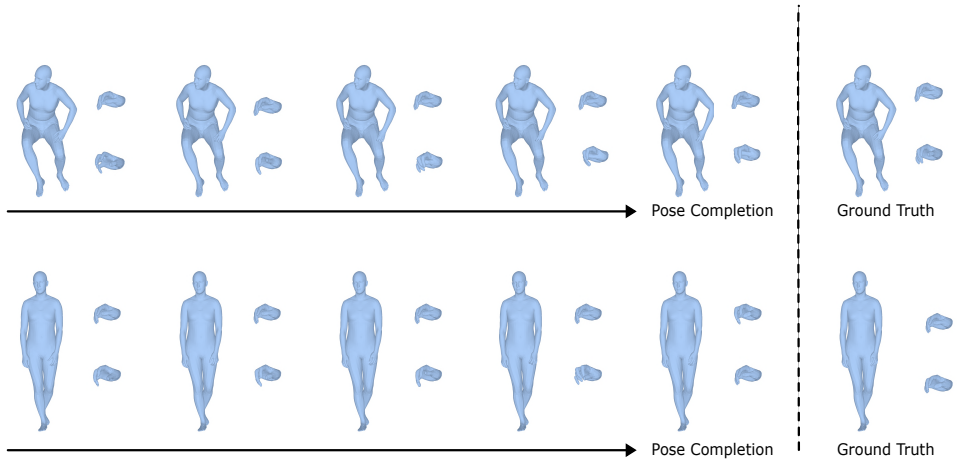

Figure 19: Whole-body Pose-completion results for the left and right hands. CDPOSE generates multiple plausible completions for partially observed hand poses, a scenario in which competing methods often struggle due to limited generalization. Best viewed zoomed in.

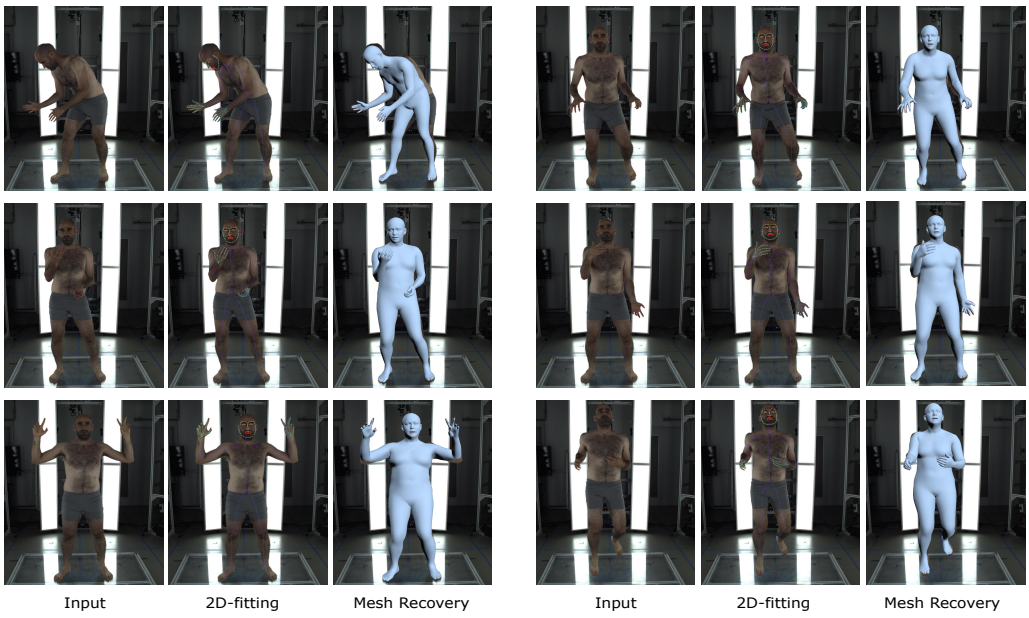

Figure 20: Visualization of whole-body mesh recovery results on the EHF dataset Pavlakos et al. (2019). The proposed CDPOSE model reconstructs plausible and coherent whole-body poses even when provided with imperfect 2D keypoints detected by RTMPose Jiang et al. (2023), whereas alternative methods exhibit reduced robustness and struggle to recover accurate meshes under noisy input conditions.