DIBIHAND: DIFFUSION GUIDED BILATERAL HAND COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes DIBIHAND, a novel diffusion-based framework for 3D hand mesh recovery. In contrast to conventional discriminative pipelines that directly regress hand parameters from images via multi-stage approaches such as detect, crop, flip, box, and joint analysis, DIBIHAND formulates hand pose estimation as a reverse diffusion process, progressively denoising random samples drawn from a Gaussian prior to produce realistic hand meshes. The method employs the MANO (hand Model with Articulated and Non-rigid defOrmations) parametric model to represent hand poses, supporting our proposed adaptive axis-angle (A-axis) and continuous 6D rotation representations (Rot-6D) for stable and efficient rotation learning. For bilateral hand pose estimation, the model handles both left and right hands simultaneously through a unified pose representation that concatenates rotations for all joints. Training with random masking simulates partial observations and occlusions, enhancing the robustness of the model under real-world conditions. The diffusion-based formulation inherently captures the complex, multi-modal nature of hand articulations and provides a strong generative prior that can serve as a regularizer for conditional refinement tasks. Meanwhile, DIBIHAND achieves state-of-the-art performance on single and bi-hand benchmarks, e.g., 4.6 JPE on FreiHand, 4.5 JPE on DexYCB, and an average 64% PCK@0.05 on HInt. The code will be available online.

1 INTRODUCTION

Accurate 3D hand reconstruction underpins AR/VR Chen et al. (2023); Dong et al. (2024b), human–computer interaction Doosti et al. (2020), robotics Chen et al. (2022b), and behavior understanding Joo et al. (2019). Robust bilateral recovery remains difficult under self-occlusion in palmar, dorsal, and mixed-view configurations, inter-hand interaction, viewpoint variability, and the high articulation of hand kinematics.

Most existing methods Lin et al. (2021b;a); Chen et al. (2022a); Pavlakos et al. (2024); Dong et al. (2024a); Chen et al. (2025) formulate hand pose estimation as deterministic regression; an image encoder maps a hand crop to MANO (hand Model with Articulated and Non-rigid defOrmations) coefficients, as shown in Figure 1 (top). While effective under controlled conditions, regression-based models lack an explicit prior over valid hand configurations, making them sensitive to occlusion, ambiguous evidence, and rotational ambiguity. Pose priors such as VPoser Pavlakos et al. (2019), NRDF He et al. (2024), and DPoser Lu et al. (2025) address this for the body, but a dedicated, diffusion-based prior for hand articulation and its use for bilateral reconstruction remains underexplored.

In this work, we propose DIBIHAND, which separates hand mesh recovery into two stages (Figure 1, bottom). First, we learn an *unconditional* score-based diffusion prior over MANO joint rotations: starting from Gaussian noise, a time-conditioned score network denoises a pose vector into an anatomically coherent articulation, capturing the multi-modal structure of the hand pose space. Second, at inference, the prior regularizes an energy-based fit of articulation, shape, global orientation, and camera to 2D keypoints detected in the image (Sec. 3.5). The image is observed only through these keypoints; the score network itself is image-free.

To make diffusion in rotation space stable, the prior operates in a z-score-normalized parameter space using either a novel Adaptive Axis-angle (A-axis) representation, which applies a smooth bounded

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.

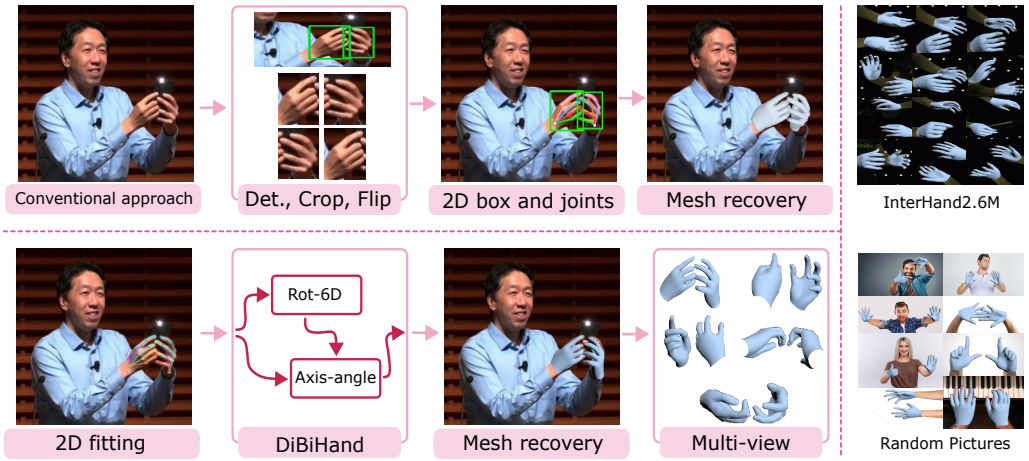

Figure 1: We propose DIBIHAND, a diffusion prior over MANO hand articulations that is used to regularize 2D-keypoint-guided mesh recovery. The framework generalizes to challenging configurations, including palmar and dorsal views, mixed-view settings, and multi-view scenarios where existing methods often fail.

magnitude map before the exponential map, or the proposed continuous 6D rotation representation (Rot-6D) in the diffusion space. The prior is trained with a sub-variance-preserving SDE Song et al. (2020b) and weighted denoising score matching. We further introduce masked diffusion training, in which random pose dimensions are zeroed at the network input so that the prior learns to complete missing articulations. For bilateral recovery, both hands are fitted jointly under a shared shape and camera, and global-orientation ambiguity is resolved by evaluating a small set of flip hypotheses and keeping the one with the lowest fitting energy.

In summary, our contributions are as follows.

- A diffusion prior over MANO hand articulations trained in a normalized parameter space with the proposed A-axis or Rot-6D representations and masked diffusion training, which yields state-of-the-art pose generation and denoising.
- A prior-guided reconstruction procedure that fits single or bilateral hands to 2D keypoints with orientation-hypothesis selection, without any image-conditioned network.
- We incorporate a masked diffusion training strategy to improve robustness under occlusion and noisy initialization.

2 RELATED WORK

Hand Mesh Recovery Monocular 3D hand reconstruction can be grouped into parametric regression, voxel-based representations, and direct vertex regression. Parametric methods Baek et al. (2019); Boukhayma et al. (2019); Cao et al. (2021); Chen et al. (2021c); Hasson et al. (2019); Yang et al. (2020); Zhang et al. (2021); Zhou et al. (2020); Pavlakos et al. (2024); Dong et al. (2024a); Chen et al. (2025) regress MANO Romero et al. (2022) pose and shape from image features and recover the mesh through a differentiable MANO layer. Voxel-based approaches Iqbal et al. (2018); Moon & Lee (2020); Moon et al. (2020); Yang et al. (2021) represent hand structure with volumetric grids or 2.5D heatmaps, offering richer spatial supervision at higher computational cost. Vertex regression methods Lin et al. (2021a;b); Chen et al. (2021b; 2022a); Ge et al. (2019); Kulon et al. (2020) predict mesh vertices directly and typically require additional regularization to preserve plausibility. All of these are feed-forward, image-conditioned regressors. DIBIHAND is complementary; it is a pose prior, and mesh recovery is obtained by fitting to 2D keypoints under that prior.

Pose Priors and Diffusion Models Learned priors over parametric pose have been built with VAEs (VPoser Pavlakos et al. (2019)), neural distance fields (NRDF He et al. (2024)), and diffusion models (DPoser Lu et al. (2025)), primarily for the human body, and are used as regularizers in

optimization-based fitting. Diffusion models Ho et al. (2020); Sohl-Dickstein et al. (2015); Song et al. (2020b); Song & Ermon (2019) model complex high-dimensional distributions by reversing a Gaussian corruption process; applied to pose, they provide a score function that can regularize or refine estimates. Learning such a prior specifically for hand articulation, with rotation representations suited to Gaussian perturbation, and using it for bilateral reconstruction has received little attention.

Difference from Prior Work DIBIHAND differs from previous methods in three aspects. First, rather than image-to-parameter regression, we learn a score-based prior in the MANO articulation space and recover meshes by prior-guided fitting, which makes the reconstruction robust to occlusion and ambiguous evidence. Second, unlike body priors applied to hands, the prior is trained in a normalized rotation space with either the proposed A-axis map or the Rot-6D representation, and with masked training that enables partial-observation completion. Third, both hands are fitted jointly under a shared shape and camera with orientation-hypothesis selection, which provides bilateral consistency and relative root placement without a dedicated two-hand network.

3 METHOD

We divide DIBIHAND into two stages. Stage 1 (Secs. 3.1–3.4) learns an unconditional diffusion prior over MANO hand articulations. Stage 2 (Sec. 3.5) recovers single or bilateral hand meshes from an image by fitting MANO to detected 2D keypoints with the prior as a regularizer. The score network never observes the image.

3.1 HAND POSE PARAMETERIZATION

We parameterize each hand with the MANO model Romero et al. (2022), which maps pose parameters (a global root rotation and $N = 15$ articulated joint rotations) and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ to a mesh $M \in \mathbb{R}^{778\times 3}$ and $K = 21$ joint locations $X \in \mathbb{R}^{K\times 3}$.

The diffusion prior models the N articulated joint rotations of a single hand in a canonical right-hand frame,

$$\boldsymbol{\theta} \in \mathbb{R}^D, \quad D = N \cdot d, \quad d \in \{3, 6\}, \tag{1}$$

where $d = 3$ for the adaptive axis–angle representation (Sec. 3.3) and $d = 6$ for the continuous 6D representation (Sec. 3.4). Left-hand training poses are mirrored into the right-hand frame (Appendix A.4), so a single network serves both hands; the mirroring is inverted when a left hand is decoded. The global root rotation $\mathbf{R} \in SO(3)$, root translation $\boldsymbol{\tau} \in \mathbb{R}^3$, and shape $\boldsymbol{\beta}$ are *not* modelled by the prior; they are free variables of the fitting stage. For bilateral recovery, the per-image state is $\mathbf{x} = \{\boldsymbol{\theta}_h, \mathbf{R}_h, \boldsymbol{\tau}_h\}_{h\in\{L,R\}} \cup \{\boldsymbol{\beta}, \mathbf{c}\}$, with a shape $\boldsymbol{\beta}$ and camera $\mathbf{c}$ shared by both hands (Sec. 3.5).

3.2 LEARNING DIFFUSION PRIOR

Let $\tilde{\boldsymbol{\theta}}_0 \sim p_{\text{data}}$ be a clean, normalized articulation (Eq. equation 6). We follow the sub-variance-preserving (sub-VP) SDE of Song et al. (2020b), whose perturbation kernel at time $t \in [0, 1]$ is

$$\tilde{\boldsymbol{\theta}}_t = \alpha_t \tilde{\boldsymbol{\theta}}_0 + \sigma_t \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

$$\alpha_t = \exp\Big(-\tfrac{1}{2}\int_0^t \beta(s)\,ds\Big), \quad \sigma_t = 1 - \exp\Big(-\int_0^t \beta(s)\,ds\Big), \quad g(t)^2 = \beta(t)\Big(1 - \exp\Big(-2\int_0^t \beta(s)\,ds\Big)\Big), \tag{3}$$

where $\beta(\cdot)$ is a linear variance schedule and $g(t)$ is the diffusion coefficient. We train a time-conditioned noise predictor $\boldsymbol{\varepsilon}_\phi(\tilde{\boldsymbol{\theta}}_t; t)$ with denoising score matching; the score is $\nabla \log p_t(\tilde{\boldsymbol{\theta}}_t) \approx -\boldsymbol{\varepsilon}_\phi/\sigma_t$, and the likelihood weighting $\lambda(t) = g(t)^2$ of Song et al. (2020b) becomes $w(t) = g(t)^2/\sigma_t^2$ in the noise-prediction form:

$$\mathcal{L}_{\text{diff}}(\phi) = \mathbb{E}_{\tilde{\boldsymbol{\theta}}_0, t, \boldsymbol{\varepsilon}}\left[w(t)\left\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\phi(\tilde{\boldsymbol{\theta}}_t; t)\right\|_2^2\right], \qquad t \sim \mathcal{U}(0, 1). \tag{4}$$

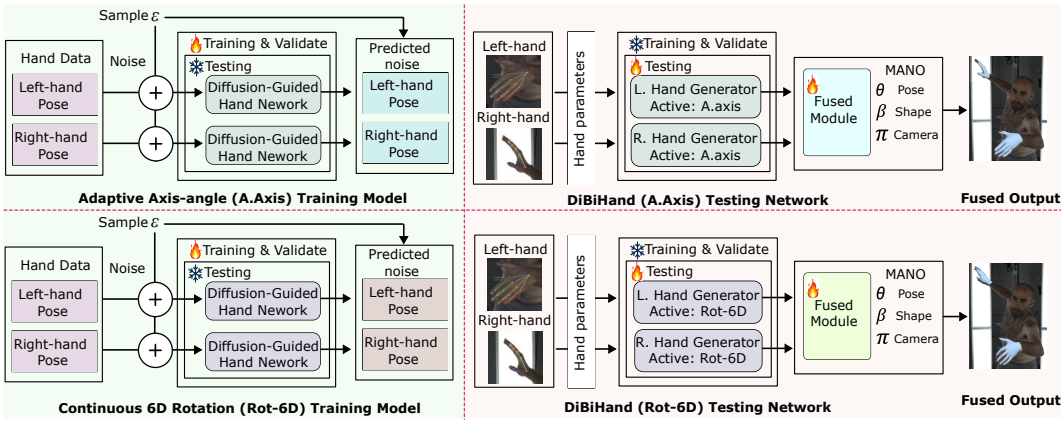

Figure 2: Overview of DIBIHAND. The training part consists of two configurations the adaptive axis-angle and the Rot-6D model. During training, hand pose data (left and right hands) are corrupted with sampled Gaussian noise. A diffusion-guided hand network is trained to predict the injected noise and progressively recover clean hand poses. At test time, the trained diffusion-guided networks act as hand generators. Given noisy or intermediate representations, they iteratively refine and generate plausible left and right-hand poses using a fused module and MANO parameterization.

Masked diffusion training. To let the prior complete partially observed articulations, we sample a binary mask $\mathbf{M} \in \{0,1\}^D$ with masking rate $\rho \sim \mathcal{U}(0.05, 0.35)$ and zero the masked dimensions *at the input*, while supervising all dimensions:

$$\mathcal{L}_{\text{masked}}(\phi) = \mathbb{E}_{\tilde{\boldsymbol{\theta}}_0, t, \boldsymbol{\varepsilon}, \mathbf{M}} \left[w(t) \left\| \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\phi(\tilde{\boldsymbol{\theta}}_t \odot \mathbf{M}; t) \right\|_2^2 \right]. \tag{5}$$

Because the network must recover the noise on dimensions it does not see, it learns the conditional structure among joints, which is what the fitting stage exploits when some keypoints are undetected or occluded. At inference $\mathbf{M} = \mathbf{1}$ unless a joint is marked missing by the detector.

Diffusion operates in a z-score-normalized Euclidean space:

$$\tilde{\boldsymbol{\theta}} = \frac{\boldsymbol{\theta} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \qquad \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \odot \boldsymbol{\sigma} + \boldsymbol{\mu}, \tag{6}$$

with per-dimension mean and standard deviation $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^D$ computed on the training set, so that the Gaussian perturbations of Eq. equation 2 are commensurate across dimensions. Whenever a (possibly intermediate) sample is decoded to a mesh, it is denormalized and then mapped to $SO(3)$ per joint by the representation-specific decoder below.

3.3 ADAPTIVE AXIS-ANGLE

As described in Sec. 3.1, with $d = 3$ each articulated MANO joint is parameterized by a 3D axis–angle rotation vector. For a single joint, the rotation is represented as $\boldsymbol{\omega} = \theta \mathbf{v} \in \mathbb{R}^3$, where $\mathbf{v} \in \mathbb{R}^3$ is a unit rotation axis and $\theta \in \mathbb{R}$ denotes the rotation magnitude. The corresponding rotation matrix $\mathbf{R} \in SO(3)$ is obtained via the exponential map, i.e., $\mathbf{R} = \exp(\mathbf{K}(\boldsymbol{\omega}))$, where $\mathbf{K}(\cdot)$ is the skew-symmetric matrix operator. Stacking the $N = 15$ articulated joints of one hand yields the pose vector $\boldsymbol{\theta} \in \mathbb{R}^D$ with $D = 3N = 45$. During diffusion training, the model operates on the normalized pose $\tilde{\boldsymbol{\theta}}$ defined in Eq. equation 6, and noise is injected according to the SDE perturbation kernel in Eq. equation 2 and Eq. equation 3. Although axis–angle is a minimal parameterization of $SO(3)$, it is embedded in the Euclidean space $\mathbb{R}^3$ while valid rotations lie on a curved manifold; the logarithm map is canonical only for $\|\boldsymbol{\omega}\| \leq \pi$, and antipodal vectors at $\|\boldsymbol{\omega}\| = \pi$ represent identical rotations. Gaussian perturbations are applied in $\mathbb{R}^3$ without respecting this constraint, so noisy samples $\tilde{\boldsymbol{\theta}}_t$ at large σ_t, as well as the intermediate iterates of reverse sampling and of the fitting loop (Sec. 3.5), routinely exceed the canonical range. Decoding such vectors naively through the exponential map produces wrapped rotations and discontinuous gradients of the data term, which destabilizes both sampling and optimization.

To improve stability while preserving minimal dimensionality, we introduce an adaptive magnitude reparameterization applied after denormalization and prior to the exponential map, in every decode during sampling and fitting. Given a rotation vector $\hat{\boldsymbol{\omega}} \in \mathbb{R}^3$, we decompose it into magnitude and direction,

$$\hat{\theta} = \|\hat{\boldsymbol{\omega}}\|_2, \qquad \hat{\mathbf{v}} = \frac{\hat{\boldsymbol{\omega}}}{\|\hat{\boldsymbol{\omega}}\|_2 + \epsilon}, \tag{7}$$

where ϵ ensures numerical stability. We then apply a smooth bounded scaling to the magnitude,

$$\tilde{\theta} = \pi \tanh\left(\frac{\hat{\theta}}{\pi}\right), \qquad \tilde{\boldsymbol{\omega}} = \tilde{\theta}\,\hat{\mathbf{v}}, \tag{8}$$

and reconstruct the adapted rotation vector $\tilde{\boldsymbol{\omega}}$. This transformation guarantees $\|\tilde{\boldsymbol{\omega}}\| < \pi$ and softly enforces the canonical axis–angle range without hard clipping. The map is monotone and differentiable and is near-identity for small angles; it compresses large magnitudes, but because the fitting energy of Sec. 3.5 is always evaluated *through* Eq. equation 8, the optimizer acts on the reparameterized variable and no systematic bias enters the decoded rotation. The final rotation matrix is computed as $\exp(\mathbf{K}(\tilde{\boldsymbol{\omega}}))$, ensuring $\mathbf{R} \in SO(3)$.

3.4 Continuous 6D Rotation Representation

In addition to the adaptive axis–angle parameterization, we consider the continuous 6D rotation representation of Zhou et al. (2019), a deliberately redundant embedding of $SO(3)$ in Euclidean space. Unlike axis–angle, which represents rotations by exponential coordinates of $SO(3)$, the 6D representation encodes a rotation by the first two columns of its rotation matrix, providing a continuous and globally smooth parameterization that is well suited to Gaussian perturbation and gradient-based optimization.

Given a rotation matrix $\mathbf{R} \in SO(3)$ with column vectors $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \in \mathbb{R}^3$, the Rot-6D representation stores the first two columns:

$$\mathbf{v} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} \in \mathbb{R}^6. \tag{9}$$

Whenever a sample is decoded, during reverse sampling and at every iteration of the fitting loop (Sec. 3.5), a valid rotation matrix is recovered by projecting $\mathbf{v}$ onto $SO(3)$ by Gram–Schmidt orthogonalization. Specifically, letting $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$ denote the two 3D vectors extracted from $\mathbf{v}$, we compute

$$\mathbf{r}_1 = \frac{\hat{\mathbf{r}}_1}{\|\hat{\mathbf{r}}_1\|_2}, \qquad \mathbf{r}_2 = \frac{\hat{\mathbf{r}}_2 - (\mathbf{r}_1^\top \hat{\mathbf{r}}_2)\mathbf{r}_1}{\|\hat{\mathbf{r}}_2 - (\mathbf{r}_1^\top \hat{\mathbf{r}}_2)\mathbf{r}_1\|_2}, \tag{10}$$

and obtain $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$, yielding $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3] \in SO(3)$. The projection is differentiable almost everywhere, so gradients of the fitting energy pass through it.

For a hand with $N = 15$ articulated joints, the pose is represented as $\boldsymbol{\theta} \in \mathbb{R}^D, \qquad D = 6N = 90,$ where each joint contributes a 6D rotation vector. Consistent with Sec. 3.3, the hand representation is constructed by concatenating the joint parameters and subsequently normalized using Eq. equation 6 prior to diffusion training.

In the SDE-based diffusion formulation Song et al. (2020b), noise is injected directly into the normalized Euclidean space $\mathbb{R}^{6N}$. Unlike axis–angle, the 6D representation imposes no bounded-magnitude constraint and has no discontinuity arising from antipodal equivalence, so perturbed and intermediate samples always decode smoothly. Empirically this eases optimization of the score network $\boldsymbol{\epsilon}_\phi(\cdot, t)$: Rot-6D converges to a lower training loss than A-axis (Appendix A.3) and yields more consistent finger alignment and structural coherence in recovered meshes (Appendix A.2, which also gives the architectural configuration of both variants). The representation itself is standard; our contribution is its use as the diffusion state with the decoder of Eq. equation 10 inside the fitting loop.

3.5 Prior-Guided Hand Recovery from Images

The prior of Sec. 3.2 never observes an image. Mesh recovery is therefore posed as an optimization problem in which MANO is fitted to 2D evidence and the prior acts as a regularizer on the articulation (Fig. 2, right). For each hand $h \in \{L, R\}$ present in the image, ViTPose Xu et al. (2022) provides

$K = 21$ keypoints $\hat{\mathbf{j}}_h \in \mathbb{R}^{K\times 2}$ with confidences $\mathbf{c}_h \in [0,1]^K$.. These keypoints are the only image-derived signal used by DIBIHAND; no image features enter the score network.

Given these observations, we solve for the state $\mathbf{x}$ introduced in Sec. 3.1. Each hand contributes its normalized articulation $\tilde{\boldsymbol{\theta}}_h \in \mathbb{R}^D$, a root rotation $\mathbf{R}_h \in SO(3)$ parameterized in Rot-6D, and a root translation $\boldsymbol{\tau}_h \in \mathbb{R}^3$ expressed in the camera frame, while the shape $\boldsymbol{\beta}$ and a perspective camera $\mathbf{c}$ with fixed focal length are shared by both hands. Sharing $\boldsymbol{\beta}$ and $\mathbf{c}$ couples the two fits, so that the relative root position $\boldsymbol{\tau}_L - \boldsymbol{\tau}_R$, the quantity measured by MRRPE on InterHand2.6M, is determined jointly from both sets of keypoints rather than from two independent single-hand estimates.

Let $J_k(\mathbf{x}_h)$ denote the k-th MANO joint obtained by denormalizing $\tilde{\boldsymbol{\theta}}_h$ through Eq. equation 6, decoding each joint rotation through Eq. equation 8 or Eq. equation 10, and posing the hand with $(\mathbf{R}_h, \boldsymbol{\tau}_h, \boldsymbol{\beta})$, and let $\Pi_{\mathbf{c}}$ denote projection under camera $\mathbf{c}$. We minimize

$$E(\mathbf{x}) = \sum_h \sum_{k=1}^{K} c_{h,k}\, \rho\Big(\big\|\Pi_{\mathbf{c}}\big(J_k(\mathbf{x}_h)\big) - \hat{\mathbf{j}}_{h,k}\big\|_2\Big) + \lambda_{\text{p}} \sum_h E_{\text{prior}}(\tilde{\boldsymbol{\theta}}_h) + \lambda_\beta \|\boldsymbol{\beta}\|_2^2, \tag{11}$$

where ρ is the Geman–McClure robustifier, the first term is the confidence-weighted reprojection error, and the last term keeps the shape close to the MANO mean. The prior term is supplied by the diffusion model following DPoser Lu et al. (2025). At iteration i we draw $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, perturb the current articulation to $\tilde{\boldsymbol{\theta}}_{h,t_i}$ using Eq. equation 2 at a noise level t_i that is annealed from $t_{\max}$ to $t_{\min}$ over the course of the optimization, and use the score-distillation gradient

$$\nabla_{\tilde{\boldsymbol{\theta}}_h} E_{\text{prior}} \;\approx\; w(t_i)\, \Big(\boldsymbol{\varepsilon}_\phi(\tilde{\boldsymbol{\theta}}_{h,t_i}; t_i) - \boldsymbol{\varepsilon}\Big), \tag{12}$$

in which the network output is treated as a constant. This gradient pushes the articulation toward regions of high density under the learned prior, with the annealed noise level allowing coarse corrections early and fine ones late. Because the prior is defined in the canonical right-hand frame, a left hand is mirrored into that frame before Eq. equation 12 is evaluated and mirrored back afterwards. Since the data term is always evaluated through the same decoder as the prior, the A-axis reparameterization of Eq. equation 8 is part of the forward chain and introduces no bias in the recovered rotations.

Missing evidence is handled through the two mechanisms already present in the model. Keypoints that the detector reports as missing receive $c_{h,k} = 0$ and thus contribute nothing to the data term, and the articulation dimensions that are then unconstrained by 2D evidence are completed by the prior, which was trained with input masking (Eq. equation 5) precisely for this situation. All variables are optimized jointly with Adam in three phases: first the camera and root parameters, then root and articulation, and finally all variables together, each for a fixed number of iterations.

The remaining difficulty is global orientation, which is the most ambiguous variable under a monocular 2D data term because palmar and dorsal configurations project almost identically. Articulation is initialized at the mean pose ($\tilde{\boldsymbol{\theta}}_h = \mathbf{0}$) and shape at $\boldsymbol{\beta} = \mathbf{0}$, and the fit is run from a small set $\mathcal{H}$ of root-rotation initializations. We consider $\mathcal{H}_{\text{none}} = \{\mathbf{I}\}$, a single initialization; $\mathcal{H}_{\text{both}} = \{\mathbf{I}, \mathbf{R}_z(\pi)\}$, which adds a 180° flip about the viewing axis; and $\mathcal{H}_{\text{all}} = \{\mathbf{I}, \mathbf{R}_x(\pi), \mathbf{R}_y(\pi), \mathbf{R}_z(\pi)\}$, which adds flips about all three axes. Each hypothesis is optimized independently and the one with the lowest final energy $E(\mathbf{x})$ is retained. Selection relies solely on Eq. equation 11 and involves no ground truth, so the procedure is identical at evaluation and deployment time. Table 5 ablates the choice of $\mathcal{H}$.

The same components serve the two auxiliary tasks reported in Sec. 4. Unconditional samples for pose generation are obtained by integrating the reverse sub-VP SDE from $\tilde{\boldsymbol{\theta}}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with 10 Euler–Maruyama steps and decoding the final sample once. For inverse-kinematics denoising (Sec. 4.3), the 2D data term in Eq. equation 11 is replaced by a 3D joint term $\sum_k \|J_k(\mathbf{x}_h) - \hat{\mathbf{X}}_k\|_2^2$ on the noisy input joints, while the prior term and optimizer are unchanged.

4 EXPERIMENTAL SETUP

Training the prior. The prior is trained on MANO articulations pooled from FreiHAND Zimmermann et al. (2019), DexYCB Chao et al. (2021), HO3D Hampali et al. (2020), H2O Kwon et al. (2021), and ReInterHand Moon et al. (2023), with left hands mirrored to the right-hand frame

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
DIBIHAND (A-axis)	2.90	0.02	0.85	0.88	1.05
DIBIHAND (Rot-6D)	3.24	0.01	0.87	0.91	0.93

Table 1: Hand pose generation evaluated against the FreiHAND Zimmermann et al. (2019) training poses.

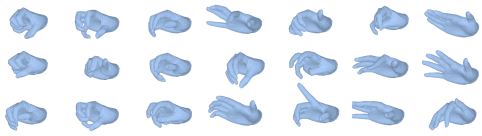

Figure 3: Unconditional samples from the DIBIHAND prior.

(Appendix A.4); no images are used at this stage. Both variants are trained on two RTX 3090 GPUs for 116,534 iterations with AdamW (learning rate 2×10^{-4}, weight decay 0.05, cosine schedule with 3,000 warmup steps, minimum learning rate 2×10^{-6}), EMA with decay 0.999, gradient clipping at 1.0, and input masking with rate 0.05–0.35 (Eq. equation 5). Training curves are in Appendix A.3.

Inference. Mesh recovery follows Sec. 3.5. 2D keypoints from ViTPose Xu et al. (2022), joint fitting of both hands when two are detected, and orientation-hypothesis selection with $\mathcal{H}_{\text{all}}$ by fitting energy. Unconditional generation uses 10 reverse-SDE steps. The same detector, energy, and schedule are used on every benchmark; only the prior variant (A-axis or Rot-6D) changes.

Evaluation metrics. For mesh recovery we report Procrustes-aligned mean per-joint and per-vertex position error (PA-MPJPE / PA-MPVPE, abbreviated JPE/VPE, in mm), F-scores at 5 and 15 mm, and AUC. On InterHand2.6M we follow the protocol of Han et al. (2026) and report root-relative MPJPE/MPVPE and MRRPE (relative root position error between the hands). On HInt we report 2D PCK of the projected joints at thresholds 0.05/0.1/0.15 of the box size, for all, visible, and occluded joints. For generation we report APD (diversity), dNN (distance to the nearest training pose), FID Heusel et al. (2017), and Precision/Recall Kynkäänniemi et al. (2019). Throughout, differences in PCK are stated in percentage points and differences in error metrics as relative changes. Results in color indicate the **best**, second, and third best.

4.1 EXPERIMENTAL RESULTS

Hand Pose Generation Table 1 compares the learned prior with VPoser, NRDF, and DPoser. Relative to DPoser, the A-axis prior raises APD by 22.9% and lowers dNN by 27.6% at comparable FID and recall. The Rot-6D prior is best on every metric, raising APD by 37.3%, precision by 6.1%, and recall by 4.6%, and lowering dNN by 35.9% while matching the lowest FID. Samples are shown in Fig. 3 and Appendix A.7.

	Method	NewDays			VISOR			Ego4D		
		@0.05	@0.1	@0.15	@0.05	@0.1	@0.15	@0.05	@0.1	@0.15
All Joints	METRO Lin et al. (2021a)	14.7	38.8	57.3	16.8	45.4	65.7	13.2	35.7	54.3
	FrankMocap Rong et al. (2021)	16.1	41.4	60.2	16.8	45.6	66.2	13.1	36.9	55.8
	MeshGraphormer Lin et al. (2021b)	16.8	42.0	59.7	19.1	48.5	67.4	14.6	38.2	56.0
	HandOccNet Park et al. (2022)	9.1	28.4	47.8	8.1	27.7	49.3	7.7	26.5	47.7
	HaMeR Pavlakos et al. (2024)	48.0	78.0	88.8	43.0	76.9	89.3	38.9	71.3	84.4
	Hamba Dong et al. (2024a)	48.7	79.2	90.0	47.2	80.2	91.2	41.7	72.9	85.5
	HandOS Chen et al. (2025)	53.7	75.9	85.1	64.8	85.4	92.0	63.4	85.3	92.9
	DIBIHAND	55.6	81.4	92.1	66.6	87.5	94.8	66.8	84.1	94.4
Visible	HaMeR Pavlakos et al. (2024)	60.8	87.9	94.4	56.6	88.0	94.7	52.0	83.2	91.3
	Hamba Dong et al. (2024a)	61.2	88.4	94.9	61.4	89.6	95.6	56.0	84.3	91.9
	HandOS Chen et al. (2025)	65.7	84.2	90.5	78.1	92.1	95.6	77.4	92.8	96.6
	DIBIHAND	68.5	89.8	95.2	78.6	93.2	96.2	76.5	94.1	96.9
Occluded	HaMeR Pavlakos et al. (2024)	27.2	60.8	78.9	25.9	60.8	80.7	23.0	56.9	76.3
	Hamba Dong et al. (2024a)	28.2	62.8	81.1	29.9	66.6	84.3	25.2	59.2	77.6
	HandOS Chen et al. (2025)	35.8	64.4	77.5	50.6	78.5	88.0	46.3	76.1	87.3
	DIBIHAND	37.3	68.2	83.6	50.1	80.1	89.2	50.7	78.1	89.2

Table 2: 2D PCK on the in-the-wild HInt Pavlakos et al. (2024) benchmark (Rot-6D prior). Baselines are feed-forward image regressors; DIBIHAND fits to 2D keypoints from the detector of Sec. 3.5.

Hand Mesh Recovery in the Wild Table 2 reports PCK on HInt (NewDays, VISOR, Ego4D). DIBIHAND obtains the best score in 24 of the 27 cells; the strongest prior method, HandOS Chen et al. (2025), remains ahead on Ego4D all-joints @0.1 (85.3 vs. 84.1), Ego4D visible @0.05 (77.4 vs. 76.5), and VISOR occluded @0.05 (50.6 vs. 50.1). Over all joints, the gains over HandOS at @0.05

Method	MPJPE↓	MPVPE↓	MRRPE↓
InterWild Moon (2023)	11.12	13.01	29.29
Ren et al. Ren et al. (2023)	10.49	10.26	28.98
IntagHand Li et al. (2022)	8.79	9.03	–
MEMAHand Wang et al. (2023a)	8.65	8.89	–
ACR Yu et al. (2023)	8.41	8.53	–
Zuo et al. Zuo et al. (2023)	8.34	8.51	–
4DHands Lin et al. (2024)	7.49	7.72	24.58
Han et al. Han et al. (2026)	5.36	5.58	21.60
DIBIHAND (A-axis)	5.15	5.36	20.75
DIBIHAND (Rot-6D)	4.64	4.91	18.86

Table 3: Performance comparison on the InterHand2.6M 5fps dataset. Errors are measured in mm.

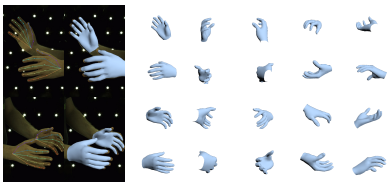

Figure 4: Bilateral reconstructions on InterHand2.6M Moon et al. (2020) shown from multiple views.

are +1.9, +1.8, and +3.4 points on the three subsets. The largest and most consistent gains are on occluded joints, where the prior contributes most +1.5/+3.8/+6.1 points at @0.05/@0.1/@0.15 on NewDays and +4.4/+2.0/+1.9 on Ego4D. Because DIBIHAND projects a 3D fit onto 2D evidence, PCK on visible joints is bounded above by the detector's own 2D accuracy; the occluded-joint results, where no 2D evidence exists, isolate what the prior adds.

Bi-Hand Mesh Recovery. Table 3 reports InterHand2.6M results. Both hands are fitted jointly under a shared shape and camera, so the relative root position measured by MRRPE is a direct output of the fit. Relative to Han et al. Han et al. (2026), the A-axis variant reduces MPJPE, MPVPE, and MRRPE by 3.9% each, and the Rot-6D variant by 13.4%, 12.0%, and 12.7%. Fig. 4 shows multi-view renderings of the recovered bilateral meshes.

4.2 COMPUTATION ANALYSIS

Table 4 breaks down compute by stage. The only learnable component of DIBIHAND is the 7.2M-parameter TimeFC score network, which costs ≈0.05 GFLOPs per evaluation and occupies 28.8 MB; unconditional generation with 10 reverse-SDE steps therefore costs ≈0.5 GFLOPs per sample. Mesh recovery is more expensive because the network is evaluated inside the fitting loop; with T=300 iterations, one network evaluation and one MANO decode per iteration amount to ≈15 GFLOPs per hand, and evaluating the four orientation hypotheses of $\mathcal{H}_{\text{all}}$ raises this to ≈60 GFLOPs. The 2D detector, ViTPose-H, adds ≈123 GFLOPs and 632M parameters per image, giving totals of ≈183 GFLOPs for a single-hand image and ≈243 GFLOPs for a two-hand image, since the detector runs once and each hand is fitted separately. In compute, the detector thus dominates (≈2/3 of a single-hand image), whereas in wall-clock time the fitting loop dominates because its iterations are sequential and cannot be batched across steps.

Stage	Component	Params	GFLOPs
Prior	score network (TimeFC), per eval	7.2M	≈0.05
Prior	10-step reverse SDE (generation)	–	≈0.5
MANO	decode + eval layer	19.2K	≈0.0003
Fitting	per hand, T=300 iterations	–	≈15
Fitting	$\times\,\lvert\mathcal{H}_{\text{all}}\rvert = 4$ hypotheses	–	≈60
2D detector	ViTPose-H Xu et al. (2022)	632M	≈123
Total, single-hand image		639M	≈183
Total, two-hand image		639M	≈243

Table 4: Parameters and per-sample compute. Prior and MANO rows are measured; fitting rows count one network evaluation and one MANO decode per hand per iteration over the three phases of Sec. 3.5; detector FLOPs are as reported by Xu et al. (2022) at 256×192 input.

4.3 ABLATION STUDIES

Qualitative comparison. Fig. 5 compares the Rot-6D variant with HaMeR and Hamba on FreiHAND and on HInt-EpicKitchens cooking scenes with severe occlusion, hand-to-hand contact, and truncation at the image border. Additional results and failure cases are in Appendices A.8 and A.6.

Hand Pose Denoising We corrupt ReInterHand Moon et al. (2023) 3D joints with Gaussian noise $\sigma \in \{1, \ldots, 10\}$ mm and recover MANO poses with the 3D variant of Eq. equation 11. Fig. 6 reports MPJPE and MPVPE against σ; the Rot-6D prior is consistently more accurate than A-axis at every noise level. Fig. 7 shows examples at 2 and 3 mm.

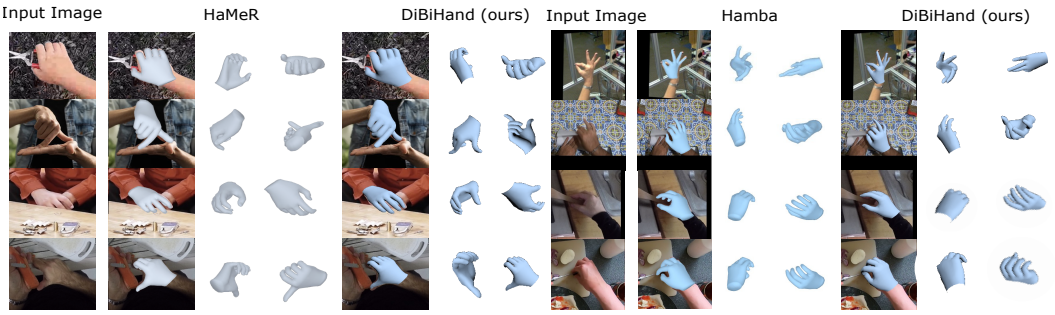

Figure 5: Qualitative comparison of DIBIHAND (Rot-6D) with HaMeR Pavlakos et al. (2024) and Hamba Dong et al. (2024a) on FreiHAND and HInt-EpicKitchens.

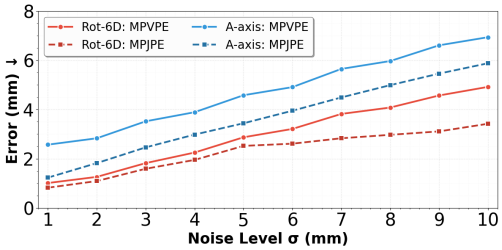

Figure 6: Inverse-kinematics denoising on ReInterHand: MPJPE/MPVPE vs. input joint noise σ.

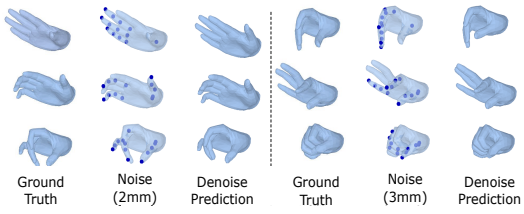

Figure 7: Denoised meshes at 2 mm and 3 mm joint noise.

Initialization and Orientation Hypotheses Table 5 ablates the fitting initialization and the hypothesis set $\mathcal{H}$ of Sec. 3.5. For the A-axis prior, mean-pose initialization reduces MPVPE by 21.3% and MPJPE by 7.2% over a random articulation initialization; adding the 180° hypothesis ($\mathcal{H}_{\text{both}}$) reduces them by 35.8% and 14.4%, and the full set $\mathcal{H}_{\text{all}}$ by 53.3% and 18.3%. The same ordering holds for Rot-6D, which is better than A-axis in every row (66.2% lower MPVPE and 33.3% lower MPJPE already without hypotheses) and reaches 0.83 mm MPVPE and 0.74 mm MPJPE with $\mathcal{H}_{\text{all}}$. Since hypotheses are selected by energy rather than by ground truth, these gains transfer to the image-based setting, where $\mathcal{H}_{\text{all}}$ is used throughout.

ReInterHand Moon et al. (2023)	**MPVPE ↓**	**MPJPE ↓**
DIBIHAND (A-axis), random init	3.94	1.53
$\mathcal{H}_{\text{none}}$ (mean-pose init)	3.10	1.42
$\mathcal{H}_{\text{both}}$	2.53	1.31
$\mathcal{H}_{\text{all}}$	1.84	1.25
DIBIHAND (Rot-6D), random init	1.33	1.02
$\mathcal{H}_{\text{none}}$ (mean-pose init)	1.02	0.93
$\mathcal{H}_{\text{both}}$	0.95	0.81
$\mathcal{H}_{\text{all}}$	**0.83**	**0.74**

Table 5: Initialization and orientation-hypothesis sets (Sec. 3.5) on ReInterHand denoising at $\sigma = [-]$ mm; the winning hypothesis is chosen by fitting energy.

5 CONCLUSION AND LIMITATIONS

In this work, we presented DIBIHAND, a diffusion-based framework for 3D single- and bi-hand mesh recovery that departs from conventional multi-stage regression pipelines. Leveraging the MANO parametric model with adaptive axis-angle and continuous Rot-6D rotation representations, our approach facilitates stable rotation learning and improved global orientation reasoning. The unified bilateral formulation and the random masking strategy further enhance robustness under occlusions and real-world interactions. Extensive experiments on both controlled and in-the-wild benchmarks demonstrate that DIBIHAND consistently achieves state-of-the-art performance.

Although DIBIHAND leverages the strong bilateral hand representation with masking strategies, its ability to handle highly diverse in-the-wild scenarios remains limited. The iterative denoising process of diffusion models incurs higher inference latency compared to conventional single-step regression methods. Improving computational efficiency remains an important direction for future research.

REFERENCES

Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, pp. 1067–1076, 2019.

Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pp. 10843–10852, 2019.

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–supplementary material–.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, pp. 12417–12426, 2021.

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 *CVPR*, pp. 9044–9053, 2021.

Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12929–12938, 2021a.

Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *CVPR*, pp. 13274–13283, 2021b.

Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20544–20554, 2022a.

Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8683–8693, 2023.

Xingyu Chen, Zhuheng Song, Xiaoke Jiang, Yaoqing Hu, Junzhi Yu, and Lei Zhang. Handos: 3d hand reconstruction in one stage. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17304–17314, 2025.

Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5150–5163, 2022b.

Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10451–10460, 2021c.

Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37:2127–2160, 2024a.

Haoye Dong, Tiange Xiang, Sravan Chittupalli, Jun Liu, and Dong Huang. Physical-space multi-body mesh detection achieved by local alignment and global dense learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1267–1276, 2024b.

Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6608–6617, 2020.

Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6353–6363, 2024.

Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10833–10842, 2019.

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.

Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11090–11100, 2022.

Gaoge Han, Yongkang Cheng, Zhe Chen, Shaoli Huang, and Tongliang Liu. From 2d alignment to 3d plausibility: Unifying heterogeneous 2d priors and penetration-free diffusion for occlusion-robust two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42800–42809, 2026.

Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pp. 11807–11816, 2019.

Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *CVPR*, pp. 1661–1671, 2024.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 118–134, 2018.

Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 758–767, 2023.

Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *CVPR*, pp. 10873–10883, 2019.

Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pp. 4990–5000, 2020.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *CVPR*, pp. 10138–10148, 2021.

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.

Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2761–2770, 2022.

Dixuan Lin, Yuxiang Zhang, Mengcheng Li, Yebin Liu, Wei Jing, Qi Yan, Qianying Wang, and Hongwen Zhang. 4dhands: Reconstructing interactive hands in 4d with transformers. *arXiv preprint arXiv:2405.20330*, 3, 2024.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pp. 1954–1963, 2021a.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *CVPR*, pp. 12939–12948, 2021b.

Haofan Lu, Shuiping Gou, and Ruimin Li. Spmhand: Segmentation-guided progressive multi-path 3d hand pose and shape estimation. *IEEE Transactions on Multimedia*, 26:6822–6833, 2024.

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 *ICCV*, pp. 9988–9997, 2025.

Gyeongsik Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *CVPR*, pp. 17028–17037, 2023.

Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, pp. 752–768. Springer, 2020.

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 *ECCV*, 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. *NeurIPS*, 36:17689–17701, 2023.

JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *CVPR*, pp. 1496–1505, 2022.

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*, pp. 10975–10985, 2019.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, pp. 9826–9836, 2024.

Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *ICCV*, pp. 8014–8025, 2023.

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

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV*, pp. 1749–1759, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 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. *NeurIPS*, 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.

Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pp. 211–228. Springer, 2020.

Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11698–11707, 2021.

Congyi Wang, Feida Zhu, and Shilei Wen. Memahand: Exploiting mesh-mano interaction for single image two-hand reconstruction. In *CVPR*, pp. 564–573, 2023a.

Shuaibing Wang, Shunli Wang, Dingkang Yang, Mingcheng Li, Ziyun Qian, Liuzhen Su, and Lihua Zhang. Handgcat: Occlusion-robust 3d hand mesh reconstruction from monocular images. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2495–2500. IEEE, 2023b.

Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17048–17058, 2023.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Linlin Yang, Shicheng Chen, and Angela Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *CVPR*, pp. 11364–11373, 2021.

Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. *arXiv preprint arXiv:2008.05079*, 2020.

Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *CVPR*, pp. 2750–2760, 2022.

Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *CVPR*, pp. 12955–12964, 2023.

Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *CVPR*, pp. 11354–11363, 2021.

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.

Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5346–5355, 2020.

Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *CVPR*, pp. 1367–1376, 2024.

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 *ICCV*, pp. 813–822, 2019.

Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *ICCV*, pp. 9054–9064, 2023.

A APPENDIX

This appendix provides additional details on the proposed framework. We first give a preliminary description of diffusion models (Section A.1) and the architectural configuration of the adaptive axis-angle and continuous 6D rotation variants (Section A.2). We then present the training and validation procedure of the prior (Section A.3), the dataset selection (Section A.4), and the evaluation metrics (Section A.5). We also discuss failure cases and limitations together with additional quantitative results on hand mesh recovery (Section A.6), examples of novel hand poses and denoising (Section A.7), and qualitative results on FreiHAND Zimmermann et al. (2019), HO3Dv3 Hampali et al. (2020), InterHand2.6M Moon et al. (2020), DexYCB Chao et al. (2021), and in-the-wild HInt Pavlakos et al. (2024) (Section A.8). Finally, Section **??** lists the hyperparameters of the prior-guided fitting stage of Sec. 3.5.

A.1 PRELIMINARIES

Diffusion models Ho et al. (2020); Sohl-Dickstein et al. (2015); Song et al. (2020b); Song & Ermon (2019) are a class of generative models that operate by reversing a predefined forward noising process. This forward process systematically corrupts the data $x_0 \sim p_{\text{data}}$ by adding Gaussian noise indexed by a continuous or discrete time variable $t \in [0, 1]$. A noisy sample x_t at any given time t is produced by

$$x_t = \alpha_t \, x_0 + \sigma_t \, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{13}$$

where α_t and σ_t are predefined schedule functions such that as t increases from 0 to 1, the data distribution is gradually transformed into a tractable prior, typically an isotropic Gaussian distribution $\mathcal{N}(0, I)$. For the sub-VP SDE used in this paper, α_t and σ_t are given in Eq. equation 3.

The generative aspect of the model lies in learning to reverse this process. This is achieved by training a neural network, typically parameterized as a noise predictor $\varepsilon_\phi(x_t, t)$, to estimate the noise component ε from a given noisy sample x_t. The network is optimized using the following weighted L2-loss objective, where $w(t)$ is a positive weighting function:

$$\mathbb{E}_{x_0 \sim p_{\text{data}}, \varepsilon \sim \mathcal{N}(0,I), t \sim U[0,1]} \, w(t) \, \|\varepsilon - \varepsilon_\phi(x_t, t)\|_2^2 \, . \tag{14}$$

The noise prediction is related to the score of the perturbed distribution by $\nabla_{x_t} \log p_t(x_t) \approx -\varepsilon_\phi(x_t, t)/\sigma_t$, which is the quantity used as a prior gradient in Sec. 3.5.

Once trained, the model can generate novel data by simulating the reverse diffusion trajectory. This procedure starts with a random sample from the prior, $x_1 \sim \mathcal{N}(0, I)$, and iteratively applies the learned denoiser ε_ϕ to produce a clean sample x_0. In practice, this reverse simulation is implemented by numerical solvers Sohl-Dickstein et al. (2015); Song et al. (2020b), such as the discrete samplers introduced in DDPM Ho et al. (2020) and DDIM Song et al. (2020a), or, as in this paper, Euler–Maruyama integration of the reverse SDE.

In DIBIHAND, the diffused variable x is the normalized articulation $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^D$ of one hand in the canonical right-hand frame (Sec. 3.1); shape, global orientation, root translation, and camera are not part of the diffusion state. The trained network is used in two ways. For pose generation, the reverse process is run from Gaussian noise to a clean articulation, which is then decoded to a MANO mesh. For mesh recovery from an image, the network is not run as a generator; instead its noise prediction supplies the prior gradient of Eq. equation 12 inside the energy minimization of Sec. 3.5, which fits articulation, shape, orientation, and camera to 2D keypoints detected in the image.

A.2 ARCHITECTURAL CONFIGURATION

Both variants share the same TimeFC score network, which receives the normalized pose $\tilde{\boldsymbol{\theta}}_t$ (with masked dimensions zeroed during training, Eq. equation 5) together with a sinusoidal embedding of the diffusion time t, and outputs a noise estimate of the same dimension as its input. The two variants differ only in the dimension D of the diffusion state and in the decoder that maps a denormalized pose to per-joint rotation matrices. The decoder is applied whenever a sample is converted to a mesh: at the end of reverse sampling for generation, and at every iteration of the fitting loop, where the data term of Eq. equation 11 is evaluated through it.

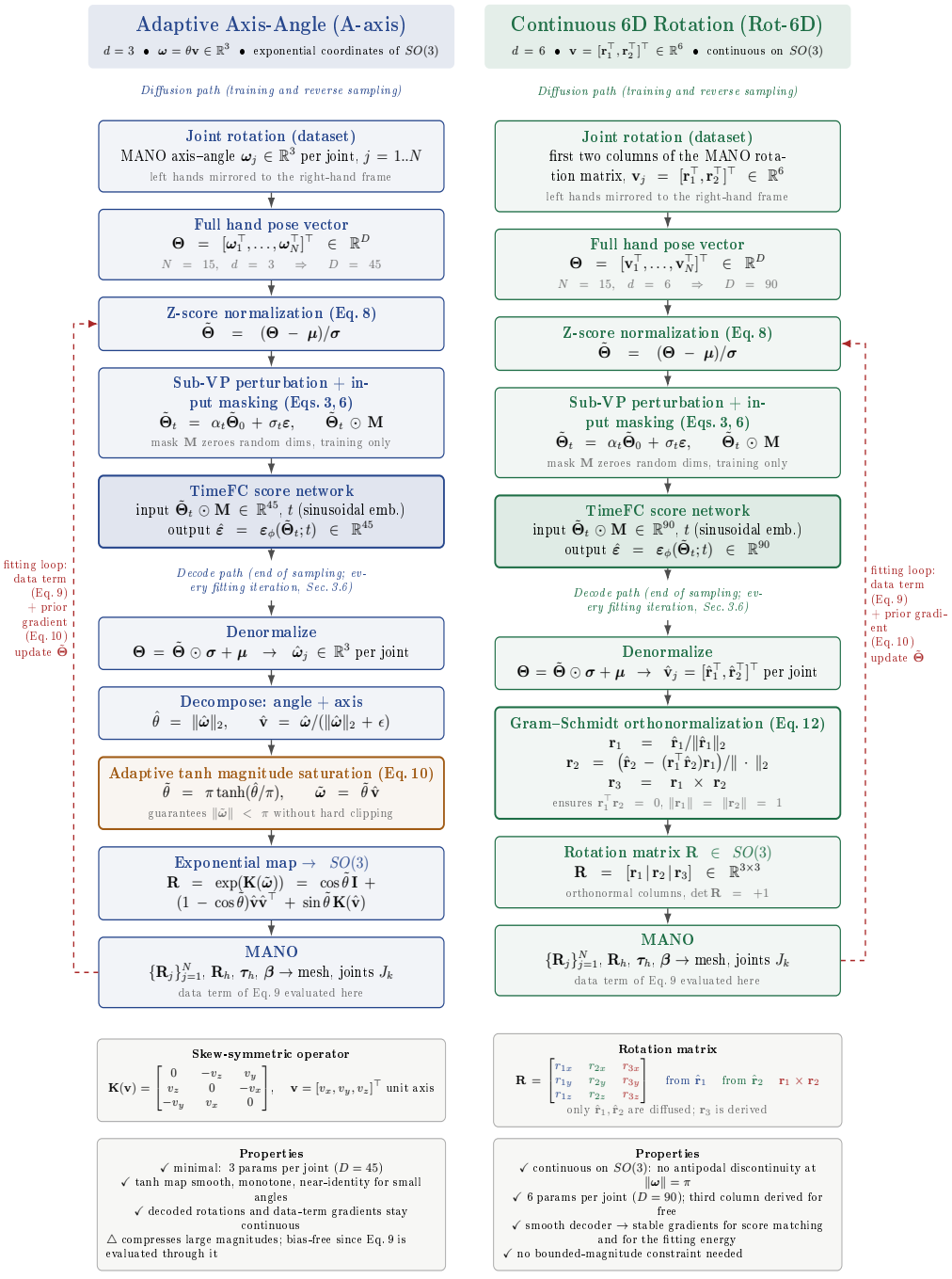

Figure 8: Configuration of the adaptive axis-angle (left) and continuous 6D rotation (right) variants.

Adaptive Axis–Angle (A-axis): The architectural configuration of A-axis is shown in Figure 8 (left), where each joint rotation is represented as a compact 3D vector in the Lie algebra $\mathfrak{so}(3)$, so that the full pose is $\Theta \in \mathbb{R}^{45}$. A denormalized vector $\hat{\boldsymbol{\omega}}$ is decomposed into magnitude $\hat{\theta} = \|\hat{\boldsymbol{\omega}}\|_2$ and unit axis $\hat{\mathbf{v}}$. The smooth saturation $\tilde{\theta} = \pi \tanh(\hat{\theta}/\pi)$ of Eq. equation 8 softly enforces $\|\tilde{\boldsymbol{\omega}}\| < \pi$ without hard clipping, and the adapted vector $\tilde{\boldsymbol{\omega}} = \tilde{\theta}\hat{\mathbf{v}}$ is mapped to a rotation matrix through the exponential map $\mathbf{R} = \exp(\mathbf{K}(\tilde{\boldsymbol{\omega}})) \in SO(3)$. Because noisy samples at large σ_t and intermediate fitting iterates routinely exceed the canonical axis–angle range, applying this map in every decode

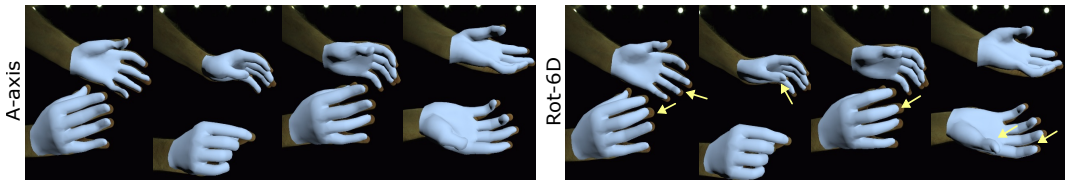

Figure 9: Hand mesh recovery with the A-axis vs. Rot-6D prior under the identical fitting pipeline of Sec. 3.5. Improvements of Rot-6D are marked in yellow.

keeps the decoded rotations and the gradients of the data term continuous. The pose Θ is z-score normalized (Eq. equation 6) for diffusion training.

Continuous 6D Rotation (Rot-6D): The architectural configuration of Rot-6D is shown in Figure 8 (right). Each joint rotation stores the first two columns $\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2$ of its rotation matrix as a 6D vector $\mathbf{v} = [\hat{\mathbf{r}}_1^\top, \hat{\mathbf{r}}_2^\top]^\top$, so that the full pose is $\Theta \in \mathbb{R}^{90}$. In every decode, Gram–Schmidt orthonormalization (Eq. equation 10) reconstructs $\mathbf{r}_1 = \hat{\mathbf{r}}_1/\|\hat{\mathbf{r}}_1\|$, obtains $\mathbf{r}_2$ as the normalized component of $\hat{\mathbf{r}}_2$ orthogonal to $\mathbf{r}_1$, and computes $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$, resulting in $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3] \in SO(3)$. This representation is continuous on $SO(3)$, avoids antipodal ambiguity, and provides stable gradients both for score matching and for the fitting energy. The pose Θ is normalized and provided to the TimeFC score network as $\tilde{\boldsymbol{\theta}}_t$. Figure 9 compares reconstructions from the two variants under the identical fitting pipeline.

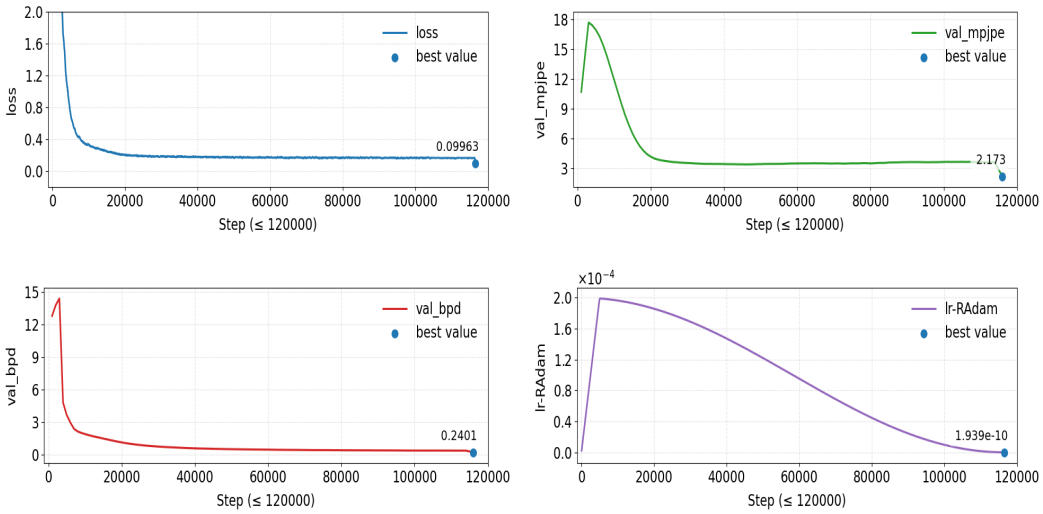

Figure 10: Training curves of the adaptive axis-angle prior.

A.3 Training and Validation

Both priors are trained with the setting of Sec. 4: AdamW with a peak learning rate of 2×10^{-4}, weight decay 0.05, a cosine schedule with 3,000 warm-up steps decaying to a minimum of 2×10^{-6}, exponential moving average with decay 0.999, gradient clipping at 1.0, and input masking with rate 0.05–0.35, for 116,534 iterations on two RTX 3090 GPUs. Three quantities are tracked during training. The training loss is the masked objective of Eq. equation 5. The validation bits-per-dimension (BPD) is the negative log-likelihood of held-out normalized poses (the ReInterHand validation split, Section A.4) computed with the probability-flow ODE. The validation MPJPE probes the prior in isolation: held-out articulations are perturbed at a fixed noise level, denoised by the reverse process without any data term, decoded through the variant's decoder, and compared to the clean MANO joints. It therefore excludes the fitting loop, the data term, and orientation-hypothesis selection, all of which are used in the results of Sec. 4.

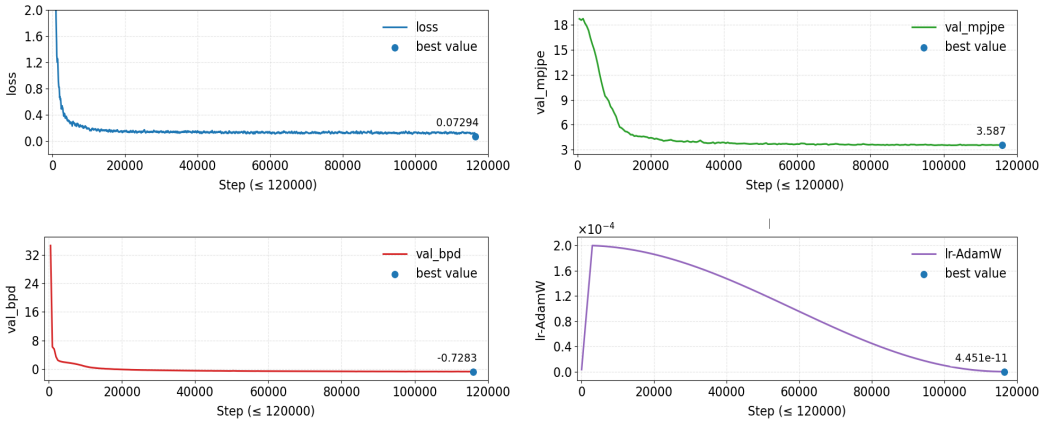

Figure 11: Training curves of the Rot-6D prior.

Adaptive Axis–Angle Representation. The training dynamics of the adaptive axis–angle representation are illustrated in Fig. 10. The training loss decreases rapidly during the early optimization stage, starting near 2.0 and dropping sharply within the first 10k–20k iterations as the network begins to capture the dominant structure of the pose distribution. After this phase, the loss gradually stabilizes and converges to approximately 0.0996 by the end of training, indicating stable convergence of the diffusion score network.

The validation MPJPE follows a similar trend. Initially, the validation error is relatively high (around 15 mm), reflecting the limited predictive capacity at the beginning of training. As optimization progresses, the error decreases rapidly and stabilizes after approximately 20k iterations, reaching a best value of 2.17 mm. This steady reduction indicates that the adaptive axis–angle parameterization supports reliable optimization of the prior.

The validation BPD also exhibits a sharp decrease during the early stages of training, dropping from values above 14 to approximately 0.24 as the training converges. This behavior indicates improved likelihood estimation of the pose distribution throughout the diffusion training process.

Rot-6D Representation. The training and validation curves for the continuous 6D rotation representation are shown in Fig. 11. The training loss begins near 2.0 and rapidly decreases during the initial optimization phase. After roughly 10k–15k iterations, the loss stabilizes and gradually converges to approximately 0.0729, which is lower than the loss obtained with the axis–angle formulation.

The validation MPJPE initially starts around 18 mm and decreases significantly during the early training phase. The curve stabilizes after approximately 20k iterations and converges to a final value of 3.59 mm.

The validation BPD decreases sharply from values above 30 in the early training stage and converges to approximately -0.73, indicating strong likelihood modeling of the pose distribution; negative values are possible because BPD is a differential entropy on continuous, unit-variance data.

Isolated vs. downstream comparison. The isolated validation MPJPE favors A-axis (2.17 vs. 3.59 mm), whereas every downstream result that uses the full fitting procedure (Tables 1, 3, and 5 in the main text, and Tables 6, 7, and 8 below) favors Rot-6D. The two validation numbers are not directly comparable across variants, since they are computed in spaces of different dimension ($D = 45$ vs. 90) and through different decoders, and the isolated metric measures single-pass denoising of the prior alone. The downstream metrics additionally depend on the smoothness of the decoder under gradient-based optimization, which is where the continuity of Rot-6D on $SO(3)$ matters.

A.4 HAND-POSE DATASETS

The prior is trained on MANO articulations only; no images are used in Stage 1. Left-hand articulations are mirrored into the canonical right-hand frame (Sec. 3.1), so all training poses are right-hand poses. In total, the pooled training set contains approximately 766,000 articulations from the five datasets below. Image-based evaluation (Sec. 4) uses only the official evaluation splits, whose poses are never included in prior training; InterHand2.6M and HInt are used for evaluation only and contribute no training poses.

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 the 32,560 unique MANO Romero et al. (2022) articulations of the training split to train the prior, while the 3,960 evaluation samples are used for the hand-mesh-recovery benchmark.

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 prior; the official test split is used for the mesh-recovery benchmark.

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 the prior; the HO3Dv3 evaluation split is used for the mesh-recovery benchmark.

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

ReInterHand ReInterHand Moon et al. (2023) is a photorealistic synthetic dataset designed for interacting-hand pose estimation. It provides 3D annotations for both hands, which we convert to the canonical right-hand frame by mirroring the 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 validation; the 23,000-pose test split supports the inverse-kinematics denoising evaluation of Sec. 4.3 and Table 5.

A.5 EVALUATION METRICS

For comprehensive assessment across various tasks, following recent works like NRDF He et al. (2024) and SMPLer-X Cai et al., we adopt task-specific metrics:

Hand Generation: Diversity and fidelity are evaluated using Average Pairwise Distance (APD) and dNN He et al. (2024), respectively. APD is the mean pairwise joint distance among generated poses, and dNN measures the distance between a generated pose and its nearest neighbor in the FreiHAND training articulations. We also report the common metrics for generative models, including FID Heusel et al. (2017) (distribution similarity), Precision (fidelity), and Recall (diversity) Kynkäänniemi et al. (2019), all computed between generated poses and the same reference set.

Mesh Recovery: Procrustes-aligned Mean Per-Joint Position Error (PA-MPJPE, abbreviated JPE) and Procrustes-aligned Mean Per-Vertex Position Error (PA-MPVPE, abbreviated VPE or PVE) measure the accuracy of recovered hand meshes after rigid alignment to the ground truth. F@5 and F@15 are the F-scores of vertex correspondences at 5 mm and 15 mm, and AUC (AUC_J, AUC_V) is the area under the PCK curve of joints or vertices over thresholds from 0 to 50 mm. On InterHand2.6M we follow the protocol of Han et al. (2026) and report MPJPE and MPVPE together with the Mean Relative-Root Position Error (MRRPE), the error of the left-hand root position relative to the right-hand root, which is a direct output of the joint two-hand fit of Sec. 3.5. On HInt we report 2D PCK of the projected joints at thresholds of 0.05, 0.1, and 0.15 of the hand-box size, separately for all, visible, and occluded joints.

Denoising & Inverse Kinematics: Both MPJPE and MPVPE are calculated between the recovered and ground-truth MANO joints and vertices after fitting to the noisy 3D joints with the 3D variant of Eq. equation 11. All errors are reported in millimeter units.



Figure 12: Illustration of representative failure cases of DIBIHAND on challenging test images, including incorrect finger alignment, ambiguous fingers due to occlusion, incorrect bimanual interaction, missing viewpoint information, and incorrect bi-hand palm orientation.

A.6 FAILURE CASES

Figure 12 shows several representative failure cases of DIBIHAND in challenging scenarios. The model mainly fails under severe ambiguity, including incorrect finger alignment, finger ambiguity caused by occlusion, errors in bimanual interaction estimation, missing viewpoint information, and incorrect palm orientation between hands. Given the two-stage design, these failures have three identifiable sources. First, errors of the 2D detector propagate directly into the data term: missing or swapped keypoints under heavy occlusion leave finger articulations constrained only by the prior, which then returns a plausible but not necessarily correct configuration. Second, when two hands are in contact, the shared camera and shape couple the fits but no contact or interpenetration term is used, so bimanual interactions can be misestimated. Third, the palmar/dorsal ambiguity is resolved by selecting among the flip hypotheses of $\mathcal{H}_{\text{all}}$ by fitting energy; when the 2D evidence is nearly symmetric, the wrong hypothesis can attain the lower energy, which produces the incorrect palm orientations in the last column. These cases highlight the difficulty of hand mesh recovery in situations that involve heavy occlusion, complex hand interactions, and ambiguous visual cues.

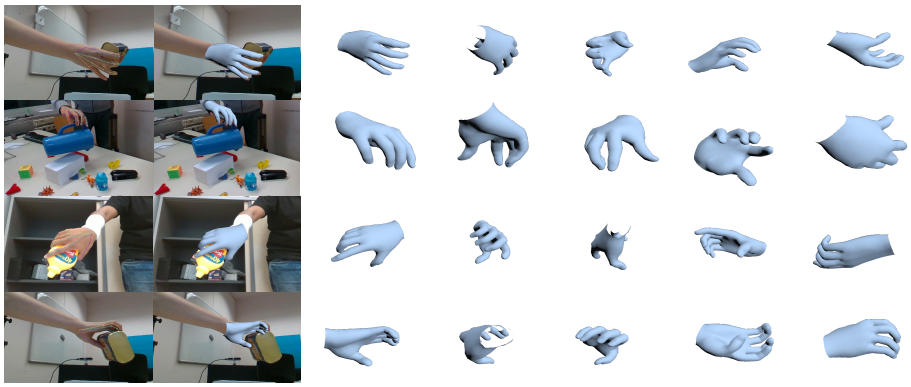

Figure 13: Visuals from HO3Dv3 Hampali et al. (2020). Multi-view renderings of the recovered mesh provide detailed views of occluded parts.

A.6.1 ADDITIONAL HAND MESH RECOVERY RESULTS

All results in this section are obtained with the prior-guided fitting pipeline of Sec. 3.5 using the same detector, energy, and schedule as the main-text experiments; only the prior variant changes between the two DIBIHAND rows of each table. Baseline numbers are taken from the respective papers.

HO3Dv3. Table 6 reports comparisons on the HO3Dv3 Hampali et al. (2020) benchmark. Compared to the strongest prior method Hamba Dong et al. (2024a), DIBIHAND A-axis reduces JPE by 4.3% and VPE by 2.9%, while increasing AUC_J by 2.6% and AUC_V by 3.1%. The Rot-6D variant achieves

Method	JPE ↓	VPE ↓	AUCJ ↑	AUCV ↑
S2HAND Chen et al. (2021c)	11.5	11.1	0.769	0.778
KPT-Transf. Hampali et al. (2022)	10.9	-	0.785	-
ArtiBoost Yang et al. (2022)	10.8	10.4	0.785	0.792
HandGCAT Wang et al. (2023b)	9.3	9.1	0.814	0.818
AMVUR Jiang et al. (2023)	8.7	8.3	0.826	0.834
HMP Duran et al. (2024)	10.1	-	-	-
SPMHand Lu et al. (2024)	8.8	8.6	-	-
HandOS Chen et al. (2025)	8.4	8.4	-	-
Hamba Dong et al. (2024a)	6.9	6.8	0.861	0.864
DIBIHAND(A-axis)	6.6	6.6	0.883	0.891
DIBIHAND(Rot-6D)	5.5	6.0	0.901	0.927

Table 6: Evaluation on the HO3Dv3 Hampali et al. (2020) benchmark. Only state-of-the-art methods are reported; baseline numbers are taken from the respective papers, and DIBIHAND rows use the fitting pipeline of Sec. 3.5.

the best overall performance, reducing JPE to 5.5 mm (a 20.3% reduction over Hamba) and VPE to 6.0 mm (11.8% reduction), while improving AUC_J by 4.6% and AUC_V by 7.3%. These quantitative improvements are reflected in Fig. 13, where our method produces more precise geometry and better recovers occluded regions under challenging hand-object interactions.

Method	JPE ↓	VPE ↓	F@5 ↑	F@15 ↑
I2UV-HandNet Chen et al. (2021a)	6.7	6.9	0.707	0.977
METRO Lin et al. (2021a)	6.5	6.3	0.731	0.984
Tang et al. Tang et al. (2021)	6.7	6.7	0.724	0.981
Mesh Graph Lin et al. (2021b)	5.9	6.0	0.764	0.986
MobRecon Chen et al. (2022a)	5.7	5.8	0.784	0.986
AMVUR Jiang et al. (2023)	6.2	6.1	0.767	0.987
HaMeR Pavlakos et al. (2024)	6.0	5.7	0.785	0.990
Hamba Dong et al. (2024a)	5.8	5.5	0.798	0.991
HandOS Chen et al. (2025)	5.0	5.3	0.812	0.991
DIBIHAND (A-axis)	4.8	5.3	0.827	0.994
DIBIHAND (Rot-6D)	4.6	5.0	0.848	0.995

Table 7: Comparison with the SOTA methods on the FreiHAND dataset Zimmermann et al. (2019). Errors are measured in mm.

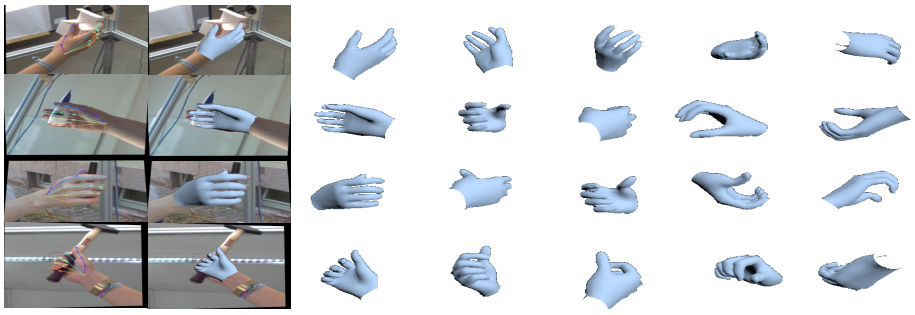

Figure 14: Visuals on FreiHAND Zimmermann et al. (2019). Recovered meshes rendered from multiple views, left to right: front, left, right, top, bottom.

FreiHAND. Table 7 shows performance of hand mesh recovery on FreiHAND Zimmermann et al. (2019). Compared to the strongest baseline HandOS Chen et al. (2025), the A-axis variant reduces JPE by 4.0% and improves F@5 by 1.8%, while matching its VPE. The Rot-6D variant achieves the best overall results, reducing JPE by 8.0% and VPE by 5.7% relative to HandOS (9.1% relative to Hamba Dong et al. (2024a)), and increasing F@5 by 4.4% and F@15 by 0.4%. These quantitative improvements are supported by the qualitative results in Fig. 14, where our reconstructions exhibit more accurate shapes and finer details across multiple viewpoints.

DexYCB. Table 8 presents comparisons on the DexYCB Chao et al. (2021) dataset. Compared to the strongest prior method HandOS Chen et al. (2025), the DIBIHAND A-axis variant reduces JPE by 3.8% and PVE by 4.0%, while increasing AUC by 2.3%. The Rot-6D variant achieves the best overall performance, reducing JPE by 13.5% and PVE by 12.0% relative to HandOS, while improving AUC by 3.9%. The qualitative results in Fig. 15 further illustrate that our method recovers more accurate and realistic hand meshes across multiple views, highlighting its robustness compared to previous methods.

Method	JPE ↓	PVE ↓	AUC ↑
Spurr et al. Spurr et al. (2020)	6.8	–	0.864
MobRecon Chen et al. (2022a)	6.4	5.6	–
HandOccNet Park et al. (2022)	5.8	5.5	–
H2ONet Xu et al. (2023)	5.7	5.5	–
Zhou et al. Zhou et al. (2024)	5.5	5.5	–
HandOS Chen et al. (2025)	5.2	5.0	0.896
DIBIHAND (A-axis)	5.0	4.8	0.917
DIBIHAND (Rot-6D)	4.5	4.4	0.931

Table 8: Results on DexYCB Chao et al. (2021). Errors are measured in mm. Best results are highlighted in color.

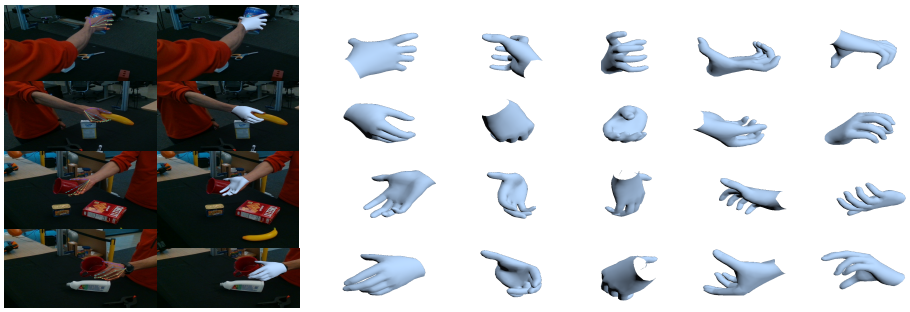

Figure 15: Visuals from DexYCB Chao et al. (2021). Multi-view renderings of the recovered mesh provide detailed views of occluded parts.

A.7 NOVEL HANDS AND DENOISING

Figure 16 presents examples of novel hand poses generated by the diffusion prior. Starting from Gaussian noise in the normalized rotation space, the reverse diffusion process (10 Euler–Maruyama steps, Sec. 3.5) progressively refines the pose to produce anatomically plausible hand configurations, which are decoded to MANO meshes with a mean shape. The generated samples exhibit diverse finger articulations and realistic inter-finger relationships, demonstrating that the learned pose prior captures the underlying distribution of natural hand poses.

Figure 17 illustrates the denoising capability of the prior under varying noise levels, following the inverse-kinematics protocol of Sec. 4.3. Gaussian noise with standard deviation σ (in mm) is added to the ground-truth 3D joints of ReInterHand test poses, and the clean MANO pose is recovered by minimizing the 3D variant of Eq. equation 11, in which the 2D reprojection term is replaced by a 3D joint term on the noisy joints and the prior term is unchanged. Even at higher noise levels, the reconstructed poses closely align with the ground truths, preserving the correct finger orientations and joint relationships.

Overall, the results demonstrate that the proposed diffusion prior can both synthesize novel, plausible hand poses and, when used as a regularizer, robustly recover accurate poses from noisy inputs, highlighting its effectiveness as a strong prior for hand pose modeling.

A.8 ADDITIONAL QUALITATIVE RESULTS

Following the main results, additional quantitative and qualitative results on the HO3Dv3 dataset Hampali et al. (2020) are provided in Section A.6.1. The visual performance of DIBIHAND is further

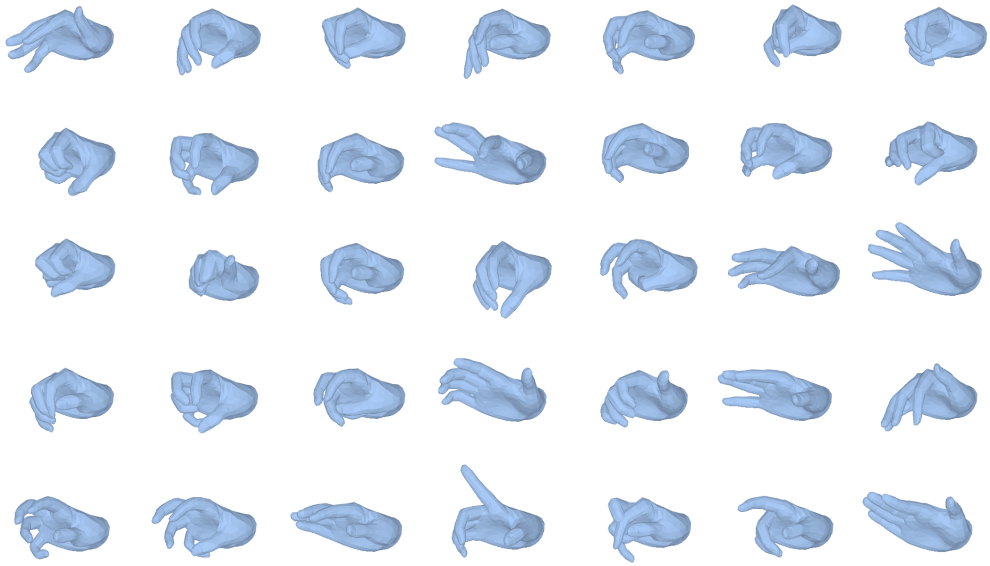

Figure 16: Visualization of hand-pose generation. Unconditional samples from the DIBIHAND prior are plausible and realistic.

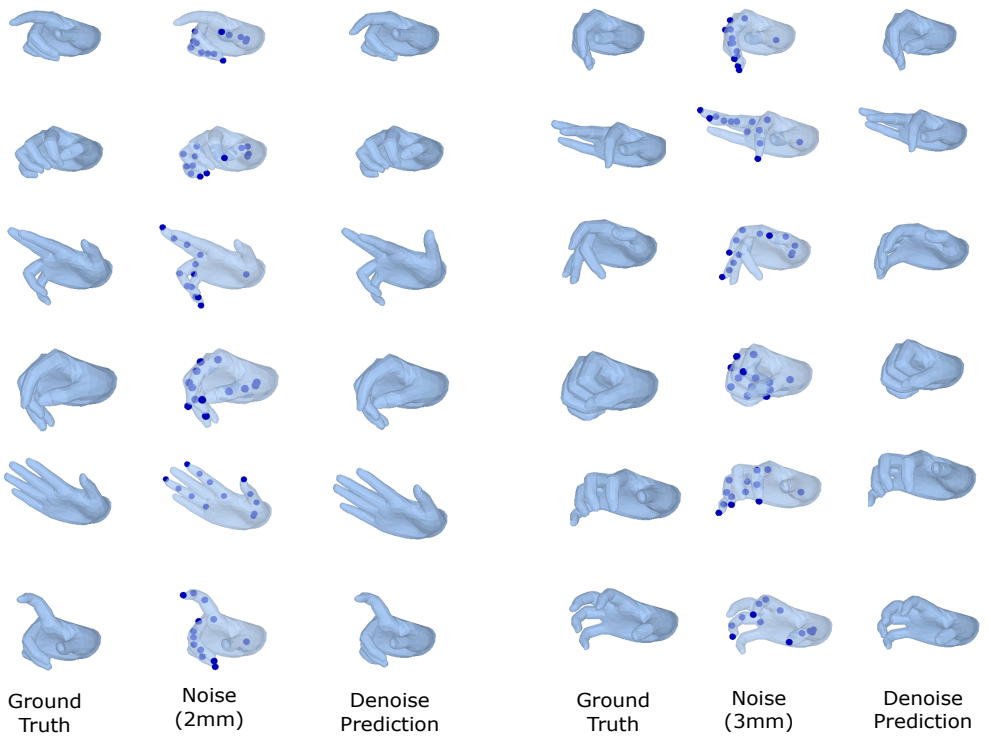

Figure 17: Visualization of hand-denoising prediction via inverse kinematics at increasing joint-noise levels. The ground truth is obtained from ReInterHand Moon et al. (2023).

presented in Figures 18, 19, 20, 21, 22, and 23. In Figures 19–23, the columns labelled *2D fitting* show the input image with the detected 2D keypoints and the projection of the fitted mesh overlaid, i.e., the observations and the data term of Sec. 3.5; the columns labelled *Mesh recovery* and the

multi-view columns show the recovered MANO mesh. These figures show reconstructed hand meshes across multiple datasets, including bi-hand scenarios from InterHand2.6M Moon et al. (2020) and the in-the-wild HInt dataset Pavlakos et al. (2024), as well as single-hand datasets such as FreiHAND Zimmermann et al. (2019), HO3Dv3 Hampali et al. (2020), and DexYCB Chao et al. (2021). Overall, the visual results demonstrate the robustness of DIBIHAND across diverse capture conditions, object interactions, and complex hand articulations.

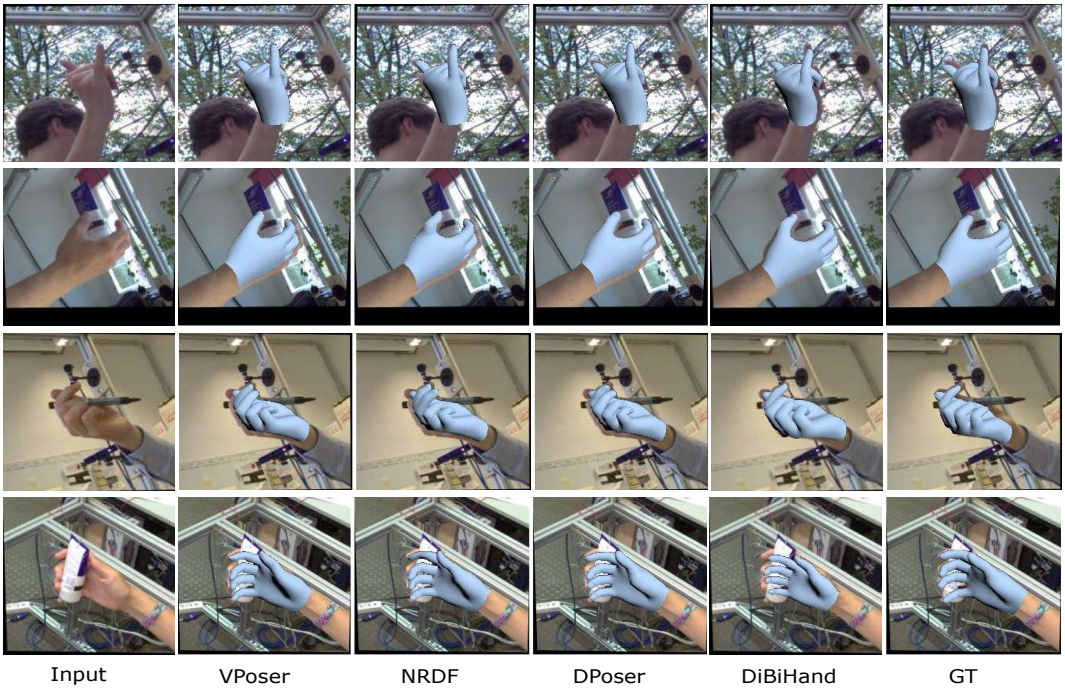

Figure 18: Comparison of prior-based hand mesh recovery. The visual outputs of VPoser Pavlakos et al. (2019), NRDF He et al. (2024), and DPoser Lu et al. (2025) are taken directly from the published DPoser resources, i.e., they were produced with DPoser's fitting pipeline rather than ours; the DIBIHAND predictions on the same images are shown alongside for qualitative comparison. GT denotes the ground-truth mesh provided with the dataset.

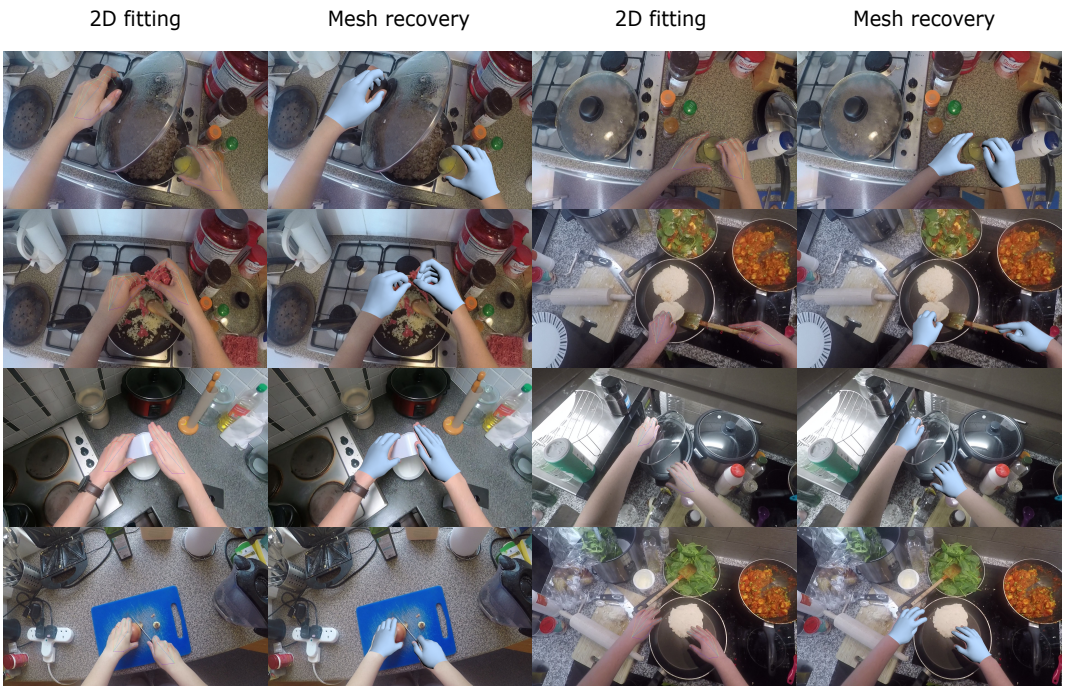

Figure 19: Visuals from the HInt-EpicKitchens-VISOR Pavlakos et al. (2024) subset. Left of each pair: detected 2D keypoints and the projected fit; right: recovered mesh.

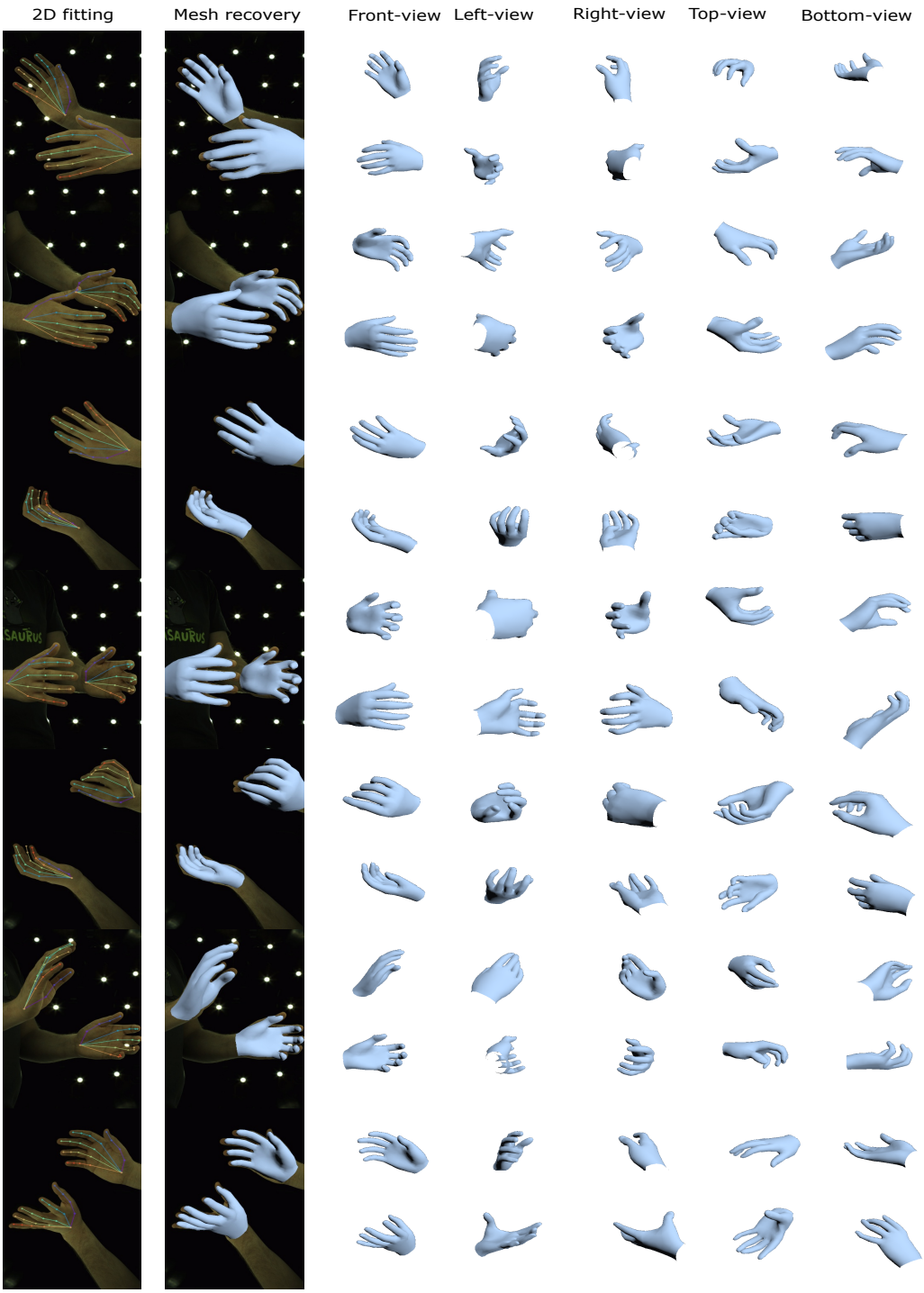

Figure 20: Visuals from the InterHand2.6M Moon et al. (2020) dataset. Both hands are fitted jointly under a shared shape and camera; multi-view renderings provide a detailed view of occluded parts that are hard to assess from the input view alone.

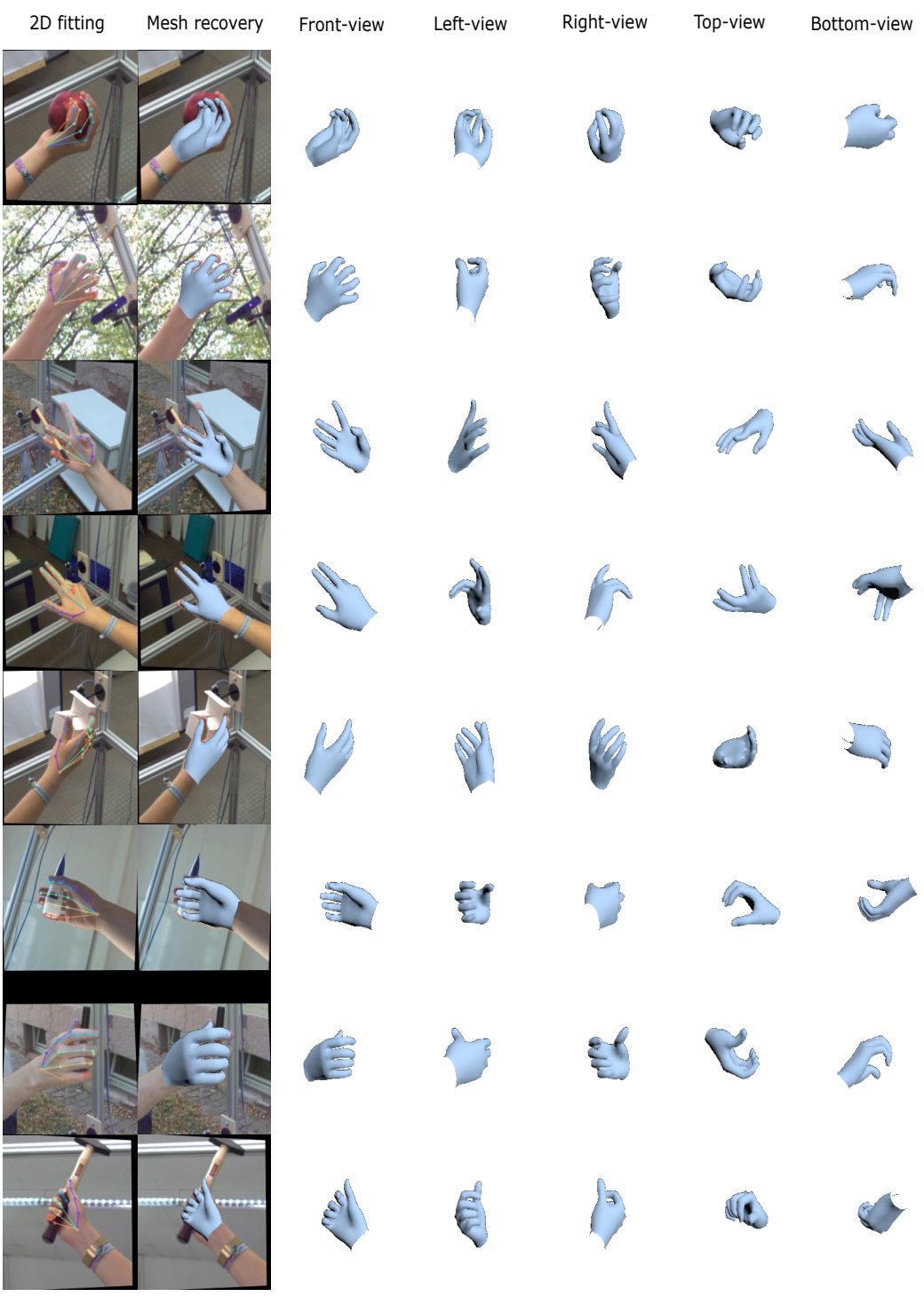

Figure 21: Visuals from the FreiHAND Zimmermann et al. (2019) dataset. Multi-view renderings provide a detailed view of occluded parts that are hard to assess from the input view alone.



Figure 22: Visuals from the HO3Dv3 Hampali et al. (2020) dataset. Multi-view renderings provide a detailed view of occluded parts that are hard to assess from the input view alone.



Figure 23: Visuals from the DexYCB Chao et al. (2021) dataset. Multi-view renderings provide a detailed view of occluded parts that are hard to assess from the input view alone.