AN ANATOMY-INFORMED 20D REPRESENTATION FOR TEXT-CONDITIONED 3D HAND POSE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D hand pose is widely used in sign language, human–computer interaction, and virtual avatars. Its representation determines how finger articulation is measured, controlled, and generated. 3D hand poses are commonly represented using MANO, which drives hand meshes with a 45D vector of local 3D joint rotations. This representation uses more dimensions than the hand's anatomical degrees of freedom, and its components do not directly represent finger flexion and spread angles. In this paper, we introduce AHand20, an anatomy-informed 20D hand pose representation guided by joint structure and degrees of freedom, with angles defined according to how each joint moves. AHand20 supports conversion to and from MANO and enables pose augmentation by varying finger flexion and spread while maintaining anatomical validity. Using the augmented poses, we construct a dataset containing 15,400 MANO poses and 30,800 descriptions across 154 handshapes, with each description linked to multiple pose variations. Trained on this dataset, a text-conditioned generator predicts 20D hand poses from descriptions of unseen handshapes and achieves lower joint position and rotation errors than existing methods. AHand20 also covers the main joint motions and converts to and from MANO with low error. Beyond pose generation, the representation identifies implausible poses in existing datasets, extending its application to hand pose data filtering.

1 INTRODUCTION

3D hand pose supports applications such as virtual reality (Grauman et al., 2022; Han et al., 2020), sign language processing (Yu et al., 2024; Baltatzis et al., 2024), and human–computer interaction (Höll et al., 2018; Yang et al., 2021). Hand poses contain fine-grained patterns of joint articulation. A pose representation determines how these patterns can be measured, controlled, and generated (Spurr et al., 2020; Saleem et al., 2025).

MANO is a widely used parametric model for 3D hands (Romero et al., 2017). It represents hand pose using an axis–angle rotation for each of 15 joints and uses these rotations to drive the hand mesh. Together, the 15 joint rotations form a 45-dimensional pose vector, but the hand has fewer than 45 independent anatomical degrees of freedom (Deshpande et al., 2011). The representation therefore contains rotations not supported by joint anatomy. Learned pose priors compress the MANO pose vector into a lower-dimensional latent space, but the resulting latent variables lack interpretability (Kundu et al., 2026). Moreover, MANO's local joint axes do not directly correspond to anatomical movements. Its pose parameters therefore lack explicit anatomical meanings and cannot be directly interpreted as finger flexion or spread angles.

In this paper, we introduce AHand20, an anatomy-informed 20D hand pose representation. Based on joint degrees of freedom (Yao & Billard, 2020), the representation uses four dimensions per finger: two for the base joint and one flexion angle for each of the other two joints. Joint types with the same degrees of freedom may still move differently. We account for this difference by defining the dimensions for each joint type using the palm reference frame, 3D joint geometry, and local joint rotations, allowing them to be computed consistently across poses.

To generate hand meshes from AHand20, we establish conversion to and from MANO by deriving 20D values from MANO poses and fitting MANO poses to 20D targets. The joint-specific dimensions also provide direct control for pose augmentation, allowing flexion and spread to vary

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.

while maintaining anatomically valid poses. Publicly available datasets pairing 3D hand poses with fine-grained descriptions remain limited. We therefore use AHand20-based pose augmentation to construct a dataset containing 15,400 MANO poses and 30,800 descriptions across 154 selected handshapes. We associate each description with multiple pose variations of the same handshape rather than a single fixed pose. Using this dataset, we train a text-conditioned hand pose generator whose decoder predicts a pose in AHand20 and converts it to a MANO pose.

Experiments show that our text-conditioned generator produces hand poses that follow the input descriptions and achieves the lowest MPJPE, MPVPE, and rotation error among the compared methods. Ablations show that AHand20 and pose and text augmentation improve generation. In addition, AHand20 covers the main motions of hand joints. It can be converted to and from MANO with low errors in joint positions, mesh vertices, and joint rotations. We further use AHand20 to identify anatomically implausible poses in four existing datasets, showing its potential for filtering hand pose annotations. Our contributions are as follows:

- We introduce AHand20, an anatomy-informed 20D hand pose representation that follows joint degrees of freedom and uses joint-specific angle definitions. The representation reduces the 45D MANO pose vector to 20 anatomically defined dimensions that directly measure and control hand articulation.
- We establish conversion between AHand20 and MANO while preserving hand articulation and mesh geometry. The 20D dimensions allow pose augmentation to vary flexion and spread while maintaining anatomical validity. AHand20 can also identify anatomically implausible poses in existing datasets.
- We construct a hand pose–text dataset containing 15,400 MANO poses and 30,800 descriptions across 154 handshapes. Using this dataset, we develop a text-conditioned generator that decodes through AHand20 and achieves lower joint, vertex, and rotation errors than existing text-to-pose methods.

2 Related Work

2.1 Hand Models and Priors

Parametric hand models provide differentiable representations of 3D hand shape and articulation, covering the visible surface, appearance, and internal anatomy (Romero et al., 2017; Qian et al., 2020; Li et al., 2021b; 2022; Wang et al., 2021). Among these models, MANO remains widely adopted across existing datasets and hand-pose pipelines (Ohkawa et al., 2023; Zimmermann et al., 2019; Hampali et al., 2020; Chao et al., 2021).

MANO represents each joint with a general 3D rotation, but in the human hand, joint rotations are constrained by anatomy. Learned pose priors therefore model hand poses in a lower-dimensional space. Both continuous latent models, such as VAEs (Pavlakos et al., 2019; Kundu et al., 2026), and discrete codebook-based models (Saleem et al., 2025; He et al., 2026) have been used for this purpose. However, these learned representations lack interpretability. Biomechanical pose priors impose joint-angle limits (Yu et al., 2024; Ren et al., 2023), constrain finger bending directions (Li et al., 2021a; Spurr et al., 2020; Yang et al., 2021), or model inter-joint dependencies (Zhang et al., 2024b). However, these methods use biomechanical knowledge to constrain hand poses rather than to design the pose representation.

2.2 Hand Pose Data and Text-to-Pose Generation

3D hand pose research is supported by datasets covering a variety of capture and interaction settings. Datasets collected from real hands provide 3D joint or mesh annotations for single hands in PALM (Fan et al., 2026) and FreiHAND (Zimmermann et al., 2019), interacting hands in InterHand2.6M (Moon et al., 2020), and hand–object interactions in HO3D (Hampali et al., 2020) and DexYCB (Chao et al., 2021). But these datasets do not include text annotations for individual hand poses. Other datasets provide language annotations for hand motion and hand–object interaction (Zhang et al., 2024a; Fu et al., 2025; Cha et al., 2024; Zhang et al., 2024c). However, these annotations do not provide fine-grained, pose-level descriptions of hand articulation and structure.

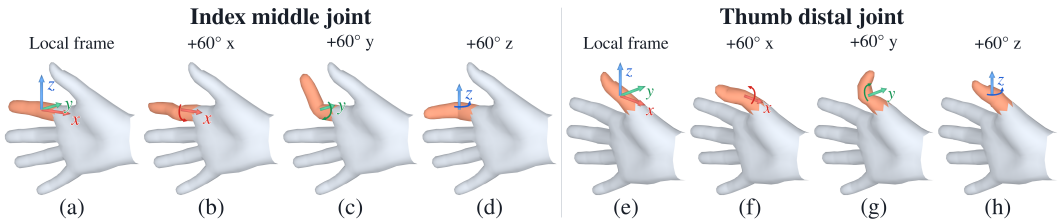


Figure 1: MANO local-axis rotations at the middle joint of the index finger (a–d) and the distal joint of the thumb (e–h). For each joint, the first panel shows the local frame and the remaining panels apply $+60^\circ$ rotations about the local x, y, and z axes.

For text-to-pose generation, existing methods primarily focus on body poses. PoseScript (Delmas et al., 2024) pairs 3D body poses with detailed descriptions and supports text-conditioned pose generation, and PoseFix (Delmas et al., 2023) extends this setting to pose correction through textual instructions. ChatPose (Feng et al., 2024), PoseLLaVA (Feng et al., 2025), and UniPose (Li et al., 2025) integrate SMPL poses into multimodal language models for pose understanding, generation, and editing. However, detailed hand-pose generation remains less explored.

3 AHand20: Anatomy-Informed Hand Pose Representation

MANO (Romero et al., 2017) is a parametric hand model that maps pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$ to a hand mesh $\mathbf{V} = M(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{778\times 3}$ using pose and shape deformations followed by linear blend skinning. We obtain $\mathbf{J} \in \mathbb{R}^{21\times 3}$ by combining the 16 joints returned by MANO with five fingertip vertices. In the non-PCA setting, $\boldsymbol{\theta} \in \mathbb{R}^{15\times 3}$ comprises 15 local joint rotations in axis–angle form.

The three axis–angle components are defined in MANO's joint frames rather than assigned to anatomical movements such as flexion and spread. As illustrated in Fig. 1, equal rotations about the MANO local axes can produce different movements, several of which depart from the natural flexion of the joint. We therefore construct AHand20, an anatomy-informed 20D representation of hand articulation. Each dimension represents one anatomical degree of freedom. The dimensions are defined for each joint type using 3D joint geometry and local joint rotations.

3.1 Hand Articulation and Palm Reference Frame

We model each finger as a three-joint kinematic chain rooted at the palm. From base to tip, the four non-thumb fingers contain the MCP, PIP, and DIP joints, while the thumb contains the CMC, MCP, and IP joints. The non-thumb MCP joint has two degrees of freedom for flexion and spread, whereas the PIP and DIP joints each have one flexion degree of freedom. The thumb also has two degrees of freedom at the CMC joint and one flexion degree of freedom at the MCP and IP joints, but differs from the other fingers in joint orientation and motion pattern (Marzke & Marzke, 2000). We therefore define the thumb dimensions separately.

To measure joint motion relative to the hand, we construct a palm reference frame from four joint positions. Let $\mathbf{j}_\mathrm{W}$ denote the wrist joint position, and let $\mathbf{j}_\mathrm{I}$, $\mathbf{j}_\mathrm{M}$, and $\mathbf{j}_\mathrm{L}$ denote the index-, middle-, and little-finger MCP joint positions, respectively. For any nonzero vector $\mathbf{v}$, we write $\mathrm{unit}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$. Let $\mathbf{d}_\mathrm{P} = \mathbf{j}_\mathrm{M} - \mathbf{j}_\mathrm{W}$ denote the wrist-to-middle-MCP direction. The palm axes are defined as

$$\mathbf{x}_\mathrm{P} = \mathrm{unit}(\mathbf{j}_\mathrm{I} - \mathbf{j}_\mathrm{L}), \quad \mathbf{y}_\mathrm{P} = \mathrm{unit}\big(\mathbf{d}_\mathrm{P} - (\mathbf{d}_\mathrm{P}^\top \mathbf{x}_\mathrm{P})\mathbf{x}_\mathrm{P}\big), \quad \mathbf{z}_\mathrm{P} = \mathbf{x}_\mathrm{P} \times \mathbf{y}_\mathrm{P}. \tag{1}$$

Here, $\mathbf{x}_\mathrm{P}$ spans the palm from the little-finger side to the index-finger side, $\mathbf{y}_\mathrm{P}$ follows the wrist-to-middle-MCP direction within the palm, and $\mathbf{z}_\mathrm{P}$ is the palm normal.

3.2 Joint-Specific Definitions

AHand20 contains four dimensions per finger. To account for differences in joint anatomy, we define these dimensions using the palm frame, adjacent bone directions, and MANO local rotations as appropriate for each joint.

Finger-base joints. The non-thumb MCP joints and the thumb CMC joint are the roots of the finger chains. Their dimensions are defined from the directions of the first bone segments in the palm frame. For a non-thumb finger f, let $\mathbf{j}_{f,\mathrm{MCP}}$ and $\mathbf{j}_{f,\mathrm{PIP}}$ denote its MCP and PIP joint positions. The current MCP-to-PIP direction is $\mathbf{d}_{f,\mathrm{MCP}} = \mathrm{unit}(\mathbf{j}_{f,\mathrm{PIP}} - \mathbf{j}_{f,\mathrm{MCP}})$, and the direction in the MANO flat zero pose is denoted by $\mathbf{d}^0_{f,\mathrm{MCP}}$. For vectors $\mathbf{u}$ and $\mathbf{v}$, we define their signed angle around $\mathbf{a}$ as $\angle_{\mathbf{a}}(\mathbf{u},\mathbf{v}) = \mathrm{atan2}((\mathbf{u}\times\mathbf{v})^\top\mathbf{a}, \mathbf{u}^\top\mathbf{v})$.

Let $\mathbf{d}^{\mathrm{palm}}_{f,\mathrm{MCP}}$ be the normalized projection of $\mathbf{d}_{f,\mathrm{MCP}}$ onto the palm plane spanned by $\mathbf{x}_\mathrm{P}$ and $\mathbf{y}_\mathrm{P}$. The rest direction and $\mathbf{z}_\mathrm{P}$ define a flexion plane with normal $\mathbf{a}^{\mathrm{flx}}_{f,\mathrm{MCP}} = \mathrm{unit}(\mathbf{d}^0_{f,\mathrm{MCP}} \times \mathbf{z}_\mathrm{P})$, and $\mathbf{d}^{\mathrm{flx}}_{f,\mathrm{MCP}}$ denotes the normalized projection of $\mathbf{d}_{f,\mathrm{MCP}}$ onto this plane. As illustrated in Fig. 2(a), the MCP spread and flexion are

$$\alpha^{\mathrm{spr}}_{f,\mathrm{MCP}} = \angle_{\mathbf{z}_\mathrm{P}}(\mathbf{d}^0_{f,\mathrm{MCP}}, \mathbf{d}^{\mathrm{palm}}_{f,\mathrm{MCP}}), \quad \alpha^{\mathrm{flx}}_{f,\mathrm{MCP}} = \angle_{\mathbf{a}^{\mathrm{flx}}_{f,\mathrm{MCP}}}(\mathbf{d}^0_{f,\mathrm{MCP}}, \mathbf{d}^{\mathrm{flx}}_{f,\mathrm{MCP}}). \tag{2}$$

The thumb has a different base orientation and motion pattern from the other fingers. We therefore use the azimuth and elevation of its CMC-to-MCP direction in the palm frame, as illustrated in Fig. 2(b–c). Let $\mathbf{d}_{t,\mathrm{CMC}} = \mathrm{unit}(\mathbf{j}_{t,\mathrm{MCP}} - \mathbf{j}_{t,\mathrm{CMC}})$, with palm-frame components $d_{t,x} = \mathbf{d}^\top_{t,\mathrm{CMC}}\mathbf{x}_\mathrm{P}$, $d_{t,y} = \mathbf{d}^\top_{t,\mathrm{CMC}}\mathbf{y}_\mathrm{P}$, and $d_{t,z} = \mathbf{d}^\top_{t,\mathrm{CMC}}\mathbf{z}_\mathrm{P}$. The two thumb-base dimensions are

$$\alpha^{\mathrm{azi}}_{t,\mathrm{CMC}} = \mathrm{atan2}(d_{t,x}, d_{t,y}), \quad \alpha^{\mathrm{ele}}_{t,\mathrm{CMC}} = \mathrm{atan2}\left(d_{t,z}, \sqrt{d^2_{t,x} + d^2_{t,y}}\right). \tag{3}$$

Non-thumb PIP and DIP joints. Each non-thumb PIP and DIP joint is modeled as a hinge, with flexion determined from its parent and child bone directions. Let $j \in \{\mathrm{PIP}, \mathrm{DIP}\}$ denote a joint of finger f, and let j^- and j^+ denote the preceding and following joints in the finger chain. The parent and child directions are $\mathbf{d}^{\mathrm{par}}_{f,j} = \mathrm{unit}(\mathbf{j}_{f,j} - \mathbf{j}_{f,j^-})$ and $\mathbf{d}^{\mathrm{ch}}_{f,j} = \mathrm{unit}(\mathbf{j}_{f,j^+} - \mathbf{j}_{f,j})$, respectively.

During substantial finger flexion, the parent bone can become nearly parallel to $\mathbf{z}_\mathrm{P}$, making $\mathbf{d}^{\mathrm{par}}_{f,j} \times \mathbf{z}_\mathrm{P}$ unstable. Since $\mathbf{x}_\mathrm{P}$ is orthogonal to $\mathbf{z}_\mathrm{P}$, it provides a stable reference in this configuration. We therefore project $\mathbf{x}_\mathrm{P}$ onto the plane orthogonal to $\mathbf{d}^{\mathrm{par}}_{f,j}$ to define the hinge axis. Let $\mathbf{d}^{\mathrm{ch,pln}}_{f,j}$ denote the normalized projection of the child direction onto the resulting hinge plane. As illustrated in Fig. 2(d), the hinge axis and flexion angle are

$$\mathbf{a}^{\mathrm{hinge}}_{f,j} = \mathrm{unit}\Big(\mathbf{x}_\mathrm{P} - (\mathbf{x}^\top_\mathrm{P}\mathbf{d}^{\mathrm{par}}_{f,j})\mathbf{d}^{\mathrm{par}}_{f,j}\Big), \quad \alpha^{\mathrm{flx}}_{f,j} = \angle_{\mathbf{a}^{\mathrm{hinge}}_{f,j}}(\mathbf{d}^{\mathrm{par}}_{f,j}, \mathbf{d}^{\mathrm{ch,pln}}_{f,j}). \tag{4}$$

Thumb MCP and IP joints. The thumb MCP and IP joints are also modeled as hinges. However, their axes cannot be constructed consistently from the palm frame because CMC motion reorients the remaining thumb chain. We therefore work in the MANO local joint frames, where the MCP and IP hinge axes remain fixed, and estimate these axes from valid MANO poses. We first remove axial twist from each local rotation, fit a fixed axis to the remaining swing rotations, and then use this axis to measure flexion in any MANO pose.

Let $\mathcal{D}_{\mathrm{axis}} = \{\boldsymbol{\theta}_i\}_{i=1}^N$ contain N poses used for axis estimation, and let $\mathbf{R}_{i,j}$ be the local rotation of thumb joint $j \in \{\mathrm{MCP}, \mathrm{IP}\}$. Because this rotation may contain twist around the child bone, we remove the twist using swing–twist decomposition (Ge et al., 2024). Let $\mathbf{q}_{i,j} = (w_{i,j}, \mathbf{v}_{i,j})$ be the unit quaternion of $\mathbf{R}_{i,j}$, and let $\mathbf{d}^0_{j,\mathrm{loc}}$ be the rest child-bone direction in the local joint frame. The twist and swing components are

$$\mathbf{q}^{\mathrm{twist}}_{i,j} = \mathrm{normalize}\big(w_{i,j}, (\mathbf{v}^\top_{i,j}\mathbf{d}^0_{j,\mathrm{loc}})\mathbf{d}^0_{j,\mathrm{loc}}\big), \qquad \mathbf{q}^{\mathrm{swing}}_{i,j} = \mathbf{q}_{i,j} \otimes (\mathbf{q}^{\mathrm{twist}}_{i,j})^{-1}. \tag{5}$$

We convert $\mathbf{q}^{\mathrm{swing}}_{i,j}$ to $\mathbf{R}^{\mathrm{swing}}_{i,j}$, retaining the part of the local rotation that changes the child-bone direction.

Since the thumb hinge axes are difficult to determine from palm geometry, we fit a unit axis whose rotations best approximate the observed swing rotations. For an arbitrary MANO pose, let $\mathbf{R}^{\mathrm{swing}}_j$ denote the swing rotation at joint j. The fitted axis and the resulting thumb flexion angle are

$$\begin{aligned} \mathbf{a}^\star_{t,j} &= \underset{\|\mathbf{a}\|_2=1}{\arg\min} \frac{1}{N}\sum_{i=1}^N \min_{\varphi} \left\| \mathbf{R}^{\mathrm{swing}}_{i,j} - \mathrm{Rot}(\mathbf{a},\varphi) \right\|^2_F, \\ \alpha^{\mathrm{flx}}_{t,j} &= \underset{\varphi}{\arg\min} \left\| \mathbf{R}^{\mathrm{swing}}_{j} - \mathrm{Rot}(\mathbf{a}^\star_{t,j},\varphi) \right\|^2_F. \end{aligned} \tag{6}$$

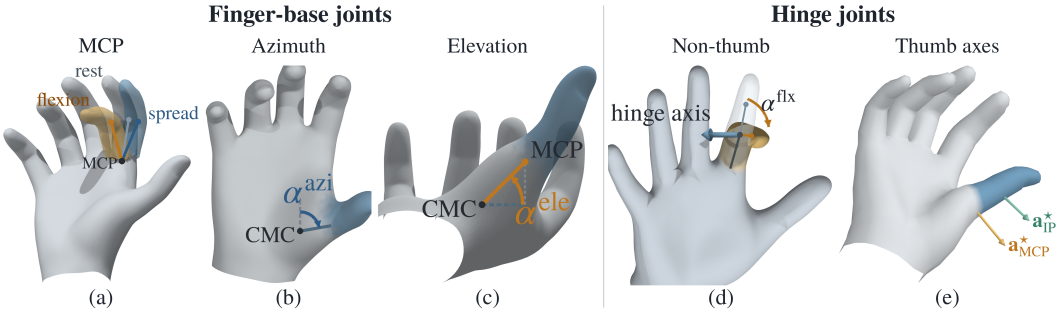


Figure 2: Joint-specific definitions used by AHand20. (a) Non-thumb MCP motion, (b–c) thumb CMC motion, (d) non-thumb hinge flexion, and (e) fitted thumb hinge axes.

The minimization over φ has a closed-form solution derived in Appendix A.2. The MCP and IP axes are estimated independently and then fixed across poses, as shown in Fig. 2(e).

3.3 Fitting Anatomical Targets to MANO Pose

A 20D target describes the desired joint motions, but generating the corresponding 3D hand requires a MANO pose. Since AHand20 and MANO use different spaces to model hand articulation, we optimize a MANO pose that matches the target in the anatomical space. Let $\mathbf{z}^{\mathrm{tar}} \in \mathbb{R}^{20}$ denote a target in the anatomical space. With the shape parameter $\boldsymbol{\beta}$ fixed, the anatomical representation produced by a MANO pose $\boldsymbol{\theta}$ is $\mathbf{z}(\boldsymbol{\theta};\boldsymbol{\beta}) = \Phi(\boldsymbol{\theta};\boldsymbol{\beta}) \in \mathbb{R}^{20}$. Here, Φ evaluates the anatomical quantities defined in Sec. 3.2 from the MANO joints and local rotations.

The fitted pose is $\boldsymbol{\theta}^{\star} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{fit}}$, where

$$\mathcal{L}_{\mathrm{fit}} = \mathcal{L}_{\mathrm{target}} + \lambda_{\mathrm{bio}}\mathcal{L}_{\mathrm{bio}} + \lambda_{\mathrm{ex}}\mathcal{L}_{\mathrm{ex}}. \tag{7}$$

The target term matches the current anatomical representation to $\mathbf{z}^{\mathrm{tar}}$ and is defined as

$$\mathcal{L}_{\mathrm{target}} = \frac{1}{20}\sum_{r=1}^{20} w_r \left(z_r(\boldsymbol{\theta};\boldsymbol{\beta}) - z_r^{\mathrm{tar}}\right)^2, \tag{8}$$

where w_r weights the fitting error of the r-th anatomical dimension.

Matching the target dimensions alone does not prevent a bone from deviating from its intended motion plane. Let $\mathcal{F}_{\mathrm{nt}}$ denote the four non-thumb fingers and let $\mathcal{J}_{\mathrm{nt}} = \{\mathrm{MCP}, \mathrm{PIP}, \mathrm{DIP}\}$. The combined plane loss is

$$\mathcal{L}_{\mathrm{bio}} = \sum_{\substack{f\in\mathcal{F}_{\mathrm{nt}}\\ j\in\mathcal{J}_{\mathrm{nt}}}} \lambda_j \left(\Delta_{f,j}^{\mathrm{plane}}\right)^2, \quad \text{where} \quad \Delta_{f,j}^{\mathrm{plane}} = \begin{cases} \mathbf{d}_{f,\mathrm{MCP}}^{\top}\mathbf{a}_{f,\mathrm{MCP}}^{\mathrm{flx}}, & j = \mathrm{MCP}, \\ (\mathbf{d}_{f,j}^{\mathrm{ch}})^{\top}\mathbf{a}_{f,j}^{\mathrm{hinge}}, & j \in \{\mathrm{PIP}, \mathrm{DIP}\}. \end{cases} \tag{9}$$

Here, λ_j weights the constraint for joint type j. The loss penalizes bone directions outside their motion planes. For MCP joints, this soft plane penalty provides a prior, since flexion is their main motion and spread outside the flexion plane is relatively limited. The thumb uses direction angles and fitted local axes rather than geometric motion planes, so this loss does not apply to the thumb.

A MANO pose can match the anatomical target while still containing unwanted axial twist, leading to unnecessarily large local rotations. For $j \in \{\mathrm{MCP}, \mathrm{PIP}, \mathrm{DIP}\}$, let $\boldsymbol{\theta}_{f,j} \in \mathbb{R}^3$ be the MANO local axis–angle rotation. Its magnitude, $\rho_{f,j}^{\mathrm{MANO}} = \|\boldsymbol{\theta}_{f,j}\|_2$, gives the local rotation angle. We then estimate the rotation magnitude required to realize the anatomical target as

$$\rho_{f,j}^{\mathrm{anat}} = \begin{cases} \sqrt{(\alpha_{f,\mathrm{MCP}}^{\mathrm{spr,tar}})^2 + (\alpha_{f,\mathrm{MCP}}^{\mathrm{flx,tar}})^2}, & j = \mathrm{MCP}, \\ \left|\alpha_{f,j}^{\mathrm{flx,tar}} - \alpha_{f,j}^{\mathrm{flx},0}\right|, & j \in \{\mathrm{PIP}, \mathrm{DIP}\}. \end{cases} \tag{10}$$

where $\alpha_{f,j}^{\mathrm{flx},0}$ is the anatomical flexion angle of the MANO initial pose. The excess loss is

$$\mathcal{L}_{\mathrm{ex}} = \sum_{\substack{f\in\mathcal{F}_{\mathrm{nt}}\\ j\in\mathcal{J}_{\mathrm{nt}}}} \eta_j \left(\Delta_{f,j}^{\mathrm{ex}}\right)^2, \quad \text{where} \quad \Delta_{f,j}^{\mathrm{ex}} = \left[\rho_{f,j}^{\mathrm{MANO}} - \rho_{f,j}^{\mathrm{anat}}\right]_+. \tag{11}$$

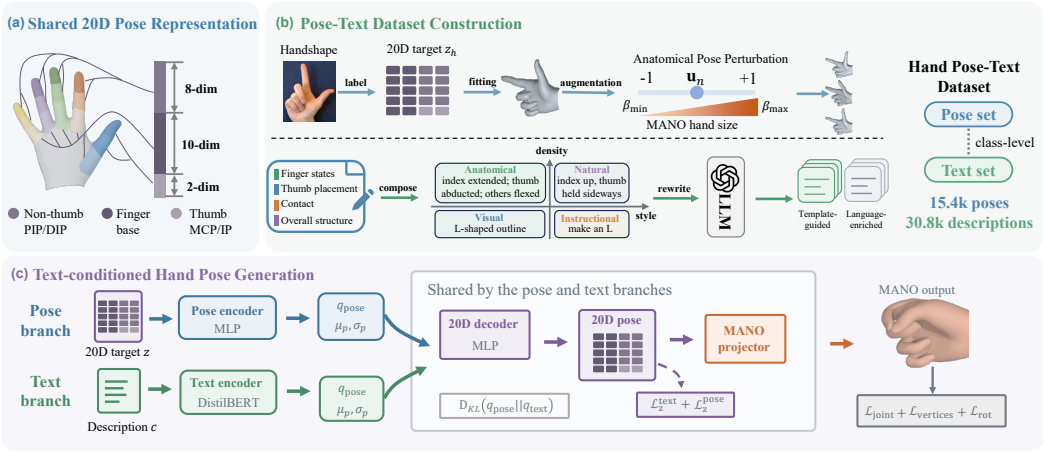


Figure 3: Overview of pose–text dataset construction and text-conditioned hand pose generation. (a) AHand20 provides a shared space for pose construction and generation. (b) Pose and shape augmentation and text construction establish many-to-many associations between poses and descriptions of the same handshape. (c) Pose and text encoders map inputs into a shared latent space, and a shared decoder predicts 20D poses that are converted to MANO.

Here, η_j weights the excess term for joint type j. Since $[x]_+ = \max(x, 0)$, the loss penalizes local rotation beyond the target. Appendix A.4 provides the optimization settings and loss weights.

4 Pose–Text Data and Generation

To assess how well AHand20 represents common handshapes, we use the MANO conversion established in Sec. 3 to construct a hand pose–text dataset and train a text-conditioned generator. We select handshapes, augment their pose and shape while preserving handshape identity, and construct descriptions from structured handshape facts. A variational model maps poses and text into a shared latent space and decodes through AHand20 to generate MANO poses. Figure 3 illustrates how AHand20 supports pose–text dataset construction and text-conditioned hand pose generation.

4.1 Handshape Selection and Pose–Shape Augmentation

AHand20, shown in Fig. 3(a), can generate anatomically valid hand poses, but many applications concern recognizable handshapes rather than arbitrary poses. We therefore construct the pose collection around handshapes selected from Global Signbank (glo) and Asian SignBank (Tang et al., 2023; asi). Since these inventories provide images rather than 3D hand parameters, we manually specify a 20D target for each selected handshape and fit a corresponding MANO pose.

A selected handshape allows limited variation in finger articulation while retaining its defining structure. As shown in the upper branch of Fig. 3(b), we generate these variations by perturbing the 20D target using a bounded vector $\mathbf{u}_n \in [-1, 1]^{20}$. We also vary hand size and finger length using MANO shape parameters. A shape vector sampled from a pool built from 3D hand datasets is paired with a perturbed 20D target. Appendix B.2 details the perturbation sampling, shape selection, and pairing procedure.

Augmentation is intended to preserve the defining structure of each selected handshape. We therefore check each fitted pose using handshape-specific conditions on articulation, contact, and separation. Articulation is evaluated in the 20D space, while contact and separation are evaluated from the MANO joints and mesh. Candidates with severe mesh self-penetration are also discarded (Forte et al., 2023). Appendix B.3 gives the preservation criteria, threshold definitions, and examples.

4.2 Pose–Text Dataset Construction

For each selected handshape, we first record a structured fact set containing finger states, thumb placement, spatial relations, and overall hand structure. As shown in the lower branch of Fig. 3(b),

Table 1: Text-conditioned generation on unseen handshapes. MPJPE and MPVPE are in millimeters, rotation error is in degrees, and R@K is in percent.

Method	Text encoder	Geometric error ↓			Retrieval ↑		FID ↓
		MPJPE	MPVPE	Rot.	R@1	R@5	
PoseScript (Delmas et al., 2024)	DistilBERT	12.04	11.67	<u>15.45</u>	32.76	70.91	<u>73.42</u>
PoseFix (Delmas et al., 2023)	DistilBERT	22.96	22.33	28.85	3.63	15.92	149.56
UniPose (Li et al., 2025)	DistilBERT	23.15	22.38	30.85	10.09	18.97	189.68
ChatPose (Feng et al., 2024)	LLaVA-v1.5-13B	<u>10.55</u>	<u>10.58</u>	29.96	**47.76**	**85.54**	98.83
Ours	DistilBERT	**10.19**	**9.90**	**12.90**	<u>42.03</u>	<u>81.54</u>	**71.34**

we then construct two types of descriptions from these facts. Template-guided descriptions vary their writing style and information scope. Language-enriched descriptions are produced by rewriting the template-guided descriptions with GPT-5.6 (OpenAI, 2026) while preserving their content. All rewritten descriptions are checked against the structured facts. Appendix B.4 describes the five writing styles and four information scopes, with examples.

This process produces 100 template-guided and 100 language-enriched descriptions for each handshape. The dataset contains 15,400 pose variations and 30,800 descriptions across 154 handshapes. Each description specifies the defining structure of a handshape rather than one pose variation, so it can be associated with multiple variations of the same handshape. The dataset therefore forms handshape-level many-to-many pose–text associations rather than fixed pose–text pairs. Appendix B presents representative source images, handshape statistics, and an analysis of the descriptions.

4.3 Text-Conditioned Hand Pose Generation

We implement text-conditioned hand pose generation with a variational autoencoder containing pose and text encoders and a shared decoder, as shown in Fig. 3(c). Given a paired 20D pose $\mathbf{z}$ and description c, the encoders predict distributions $q_{\text{pose}}(\boldsymbol{\ell} \mid \mathbf{z})$ and $q_{\text{text}}(\boldsymbol{\ell} \mid c)$ over a shared latent code $\boldsymbol{\ell}$. For either branch $b \in \{\text{pose}, \text{text}\}$, the shared decoder D_z converts a sampled code $\boldsymbol{\ell}_b$ into a 20D pose $\widehat{\mathbf{z}}_b = D_z(\boldsymbol{\ell}_b)$. A learned mapping G_R then produces 15 MANO local rotation matrices, $\widehat{\mathbf{R}}_b = G_R(\widehat{\mathbf{z}}_b)$. For each branch $b \in \{\text{pose}, \text{text}\}$, the reconstruction loss is

$$\mathcal{L}^b_{\text{rec}} = \lambda_z \mathcal{L}^b_z + \lambda_R \mathcal{L}^b_{\text{rot}} + \lambda_J \mathcal{L}^b_{\text{joint}} + \lambda_V \mathcal{L}^b_{\text{vertex}}. \tag{12}$$

Following PoseScript (Delmas et al., 2024), we use Laplace negative log-likelihood losses for the 20D angles, joint positions, and mesh vertices, and a Gaussian negative log-likelihood loss for the MANO rotation matrices. Applying the same reconstruction loss to the text branch directly trains its latent sample to reproduce the target hand pose. The training objective is

$$\mathcal{L} = \mathcal{L}^{\text{pose}}_{\text{rec}} + \lambda_{\text{text}} \mathcal{L}^{\text{text}}_{\text{rec}} + \lambda_{\text{align}} D_{\text{KL}}(q_{\text{pose}} \| q_{\text{text}}) + \sum\nolimits_b \lambda^{\text{KL}}_b D_{\text{KL}}(q_b \| \mathcal{N}(\mathbf{0}, \mathbf{I})) \,. \tag{13}$$

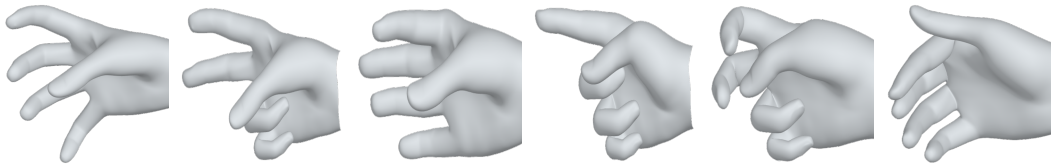
The first KL term aligns the pose and text distributions, while the summed terms regularize both distributions toward a standard normal prior. Appendix C specifies the network architecture, loss weights, and training protocol.

5 Experiments

5.1 Text-Conditioned Hand Pose Generation

Experimental setup. We divide the 154 handshapes, together with their poses and descriptions, into 124 for training, 15 for validation, and 15 for testing. We report MPJPE over 21 joints, MPVPE over 778 vertices, and the mean geodesic error over 15 MANO local rotations. We also report retrieval recall at ranks 1 and 5 (R@1 and R@5) and Fréchet inception distance (FID). Appendix D.1 specifies reference matching, sample aggregation, and the semantic evaluation protocol.

Overall performance. Table 1 compares our method with four baselines on unseen handshapes. Our method achieves the lowest MPJPE, MPVPE, rotation error, and FID, and the highest retrieval



The four fingers slope palmward in an even spread row as the thumb points forward under the index across visible space.

The extended index and middle diverge as a moderate V-shaped upper jaw opposing a curved lower thumb across a wide gap while the ring-little pair stays tucked.

The middle, ring, and little fingers form a close straight row with only slight middle-MCP flexion; a flat MCP-folded index lies above thumb2 and thumb3 across a clear noncontacting gap.

The index remains upright and unbent while the other three fingers form a fist and the thumb aligns along its side.

Separated index-middle hooks sit over a closed ring-little base crossed by a thumb contacting the distal ring side.

Keep the four fingers from index through little together and bend them forward at their bases without curling them, then extend the thumb vertically at the side.

Figure 4: Text-conditioned generation on unseen handshapes. Each example pairs an input description with a generated pose, illustrating finger articulation, separation, and thumb placement.

Table 2: Ablation and data-scale results on unseen handshapes. (a) Effects of AHand20 and language-enriched text. (b) Effects of pose and description counts. MPJPE is in millimeters and rotation error in degrees; lower is better.

(a) Ablation

AHand20	Enriched text	MPJPE	Rot.
–	–	14.08	17.39
–	✓	12.04	15.45
✓	–	<u>11.35</u>	<u>13.86</u>
✓	✓	**10.19**	**12.90**

(b) Data scale

Pose variants		Descriptions	
Count	MPJPE	Count	MPJPE
1	17.95	1	20.77
10	12.81	10	15.68
25	<u>10.74</u>	40	<u>11.84</u>
100	**9.87**	200	**9.87**

recall among methods using DistilBERT. ChatPose achieves higher retrieval recall with its larger language backbone, which may contribute to stronger text–pose alignment.

Figure 4 shows generated poses for unseen handshapes. The outputs generally capture the finger configurations described in the text, although details such as finger spacing and the degree of bending are not always reproduced accurately. Appendix D.2 provides additional comparisons with the baselines and failure cases.

5.2 Ablation and Data Scale

Ablation. We examine the effects of AHand20 and language-enriched descriptions on pose generation. Without AHand20, the decoder directly predicts MANO joint rotations using a 6D representation per joint. Without enriched text, training uses only template-guided descriptions. Table 2(a) shows that AHand20 alone reduces MPJPE and rotation error more than language enrichment alone, with their combination achieving the lowest errors. These results suggest that modeling joint anatomy benefits text-conditioned hand pose generation.

Data scale. We examine how the numbers of pose variations and text descriptions per handshape affect generation accuracy. Table 2(b) shows that increasing pose variants from 1 to 100 reduces MPJPE from 17.95 to 9.87 mm, while increasing descriptions from 1 to 200 reduces it from 20.77 to 9.87 mm. These results support the effectiveness of our AHand20-based pose augmentation and text augmentation in improving generation accuracy.

5.3 AHand20 Pose Coverage and MANO Conversion

Pose coverage. We examine whether AHand20 captures the joint motions observed in hand pose data despite reducing the rotational degrees of freedom provided by MANO. We select FreiHAND and DexYCB for this evaluation. For each pose, we measure the angular residual between the observed joint directions and those reconstructed from AHand20. Table 3(a) reports the results. The thumb CMC and non-thumb MCP joints have near-zero residuals because their two angles are sufficient to fully represent the direction of the first finger segment. For the remaining joints, the

Table 3: Evaluation of AHand20. (a) Mean and P90 angular residuals for finger-base direction reconstruction and hinge approximation, in degrees. (b) Mean and P95 errors for 20D target fitting and MANO round-trip conversion on the 154 reference handshapes used to construct our dataset.

(a) AHand20 pose coverage

	FreiHAND		DexYCB	
Component	Mean	P90	Mean	P90
Thumb CMC	< 0.001	< 0.001	< 0.001	< 0.001
Non-thumb MCP	< 0.001	< 0.001	< 0.001	< 0.001
Non-thumb PIP	7.747	14.889	7.520	14.064
Non-thumb DIP	5.659	11.000	6.608	12.833
Thumb MCP	8.915	17.670	6.455	13.202
Thumb IP	9.511	19.776	2.564	5.468

(b) MANO conversion

Metric	Mean	P95
AHand20 $\rightarrow$ MANO $\rightarrow$ AHand20		
20D error (°)	0.380	0.827
MANO $\rightarrow$ AHand20 $\rightarrow$ MANO		
20D error (°)	0.295	0.638
MPJPE (mm)	0.518	1.084
MPVPE (mm)	0.494	1.029
Rot. (°)	0.617	1.442

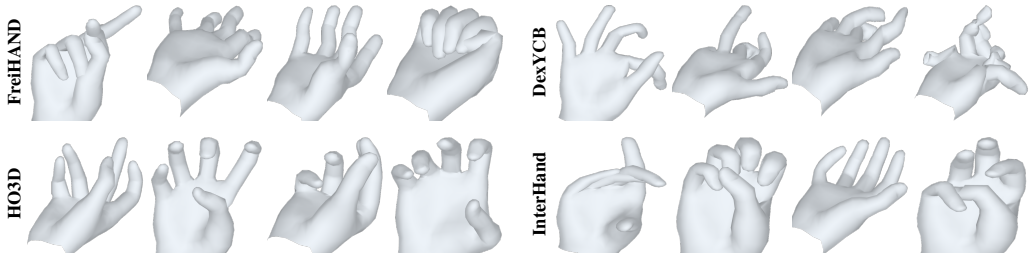


Figure 5: Examples of anatomically implausible poses identified using AHand20 in four hand pose datasets.

relatively small residuals support the hinge approximations used by AHand20, suggesting that they capture the main joint variations in both datasets.

MANO conversion. Accurate conversion between AHand20 and MANO is important for preserving hand articulation. We evaluate this conversion on the 154 reference handshapes used to construct our dataset by fitting a MANO pose to a 20D target and then recomputing its 20D values. We also perform a MANO–AHand20–MANO round trip. As shown in Table 3(b), the fitted poses have a mean 20D error of $0.380°$. The MANO round trip yields a mean MPJPE of 0.518 mm. These results demonstrate that AHand20 preserves hand pose through conversion to and from MANO.

5.4 Identifying Anatomically Implausible Poses

MANO rotation parameters alone make it difficult to identify implausible poses. In contrast, AHand20 allows us to identify such poses by detecting deviations from expected joint bending directions. We use this approach to identify implausible MANO poses in FreiHAND, DexYCB, HO3D, and InterHand2.6M.

Figure 5 shows representative identified poses. The examples from FreiHAND and DexYCB contain visible abnormalities such as excessive flexion and lateral finger bending, while those from HO3D and InterHand2.6M are less severe but still contain joint motions outside the expected ranges or directions. These results support using AHand20 to inspect and filter hand poses and construct cleaner datasets. Appendix D.3 details the diagnostic scores, reference intervals, and selection of the displayed cases.

6 Conclusion

This work shows that joint anatomy provides a practical basis for representing hand poses in a compact and interpretable form. AHand20 connects joint-level control with language-conditioned generation while remaining compatible with MANO. Its use in both pose synthesis and annotation inspection highlights the value of explicit anatomical definitions for modeling fine-grained hand articulation.

AI USE STATEMENT

As part of our dataset construction, we used an LLM to augment textual descriptions, as detailed in Sec. 4.2 and Appendix B.4. Generative AI tools also assisted with language polishing, LaTeX formatting, and code implementation and debugging. The authors determined the research design and the manuscript's content, organization, and final wording, and reviewed the data and the complete manuscript. The authors take full responsibility for the final paper, code, and experimental conclusions.

REPRODUCIBILITY STATEMENT

Appendix A provides the geometric definitions, thumb-axis calibration procedure, and MANO conversion settings. Appendix B describes dataset construction and augmentation, while Appendix C specifies the model architecture and training settings. Appendix D details the evaluation protocol and metrics. The supplementary material includes code for MANO–AHand20 conversion and text-conditioned generator training.

REFERENCES

Asian SignBank: Explanation: Handshape. `http://www.cslds.org/asiansignbank/search/explanation/handshape.jsp`. Accessed: 2026-07-24.

Global Signbank: Show all handshapes. `https://signbank.cls.ru.nl/handshapes/show_all/`. Accessed: 2026-07-24.

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3D sign language production from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1985–1995, 2024.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3D motion generation for hand-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00156.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. 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.

Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. PoseFix: Correcting 3D human poses with natural language. In *IEEE/CVF International Conference on Computer Vision*, pp. 14972–14982. IEEE, 2023.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. PoseScript: Linking 3D human poses and natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5146–5159, 2024.

Ashish D. Deshpande, Zhe Xu, Michael J. Vande Weghe, Benjamin H. Brown, Jonathan Ko, Lillian Y. Chang, David D. Wilkinson, Sean M. Bidic, and Yoky Matsuoka. Mechanisms of the anatomically correct testbed hand. *IEEE/ASME Transactions on Mechatronics*, 18(1):238–250, 2011.

Zicong Fan, Edoardo Remelli, David Dimond, Fadime Sener, Liuhao Ge, Bugra Tekin, Cem Keskin, and Shreyas Hampali. PALM: A dataset and baseline for learning multi-subject hand prior. In *International Conference on 3D Vision*, pp. 693–703. IEEE, 2026.

Dong Feng, Ping Guo, Encheng Peng, Mingmin Zhu, Wenhao Yu, and Peng Wang. PoseLLaVA: Pose-centric multimodal LLM for fine-grained 3D pose manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2951–2959, 2025.

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. ChatPose: Chatting about 3D human pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2093–2103. IEEE, 2024.

Maria-Paola Forte, Peter Kulits, Chun-Hao P. Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. Reconstructing signing avatars from video using linguistic priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12791–12801, 2023.

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. GigaHands: A massive annotated dataset of bimanual hand activities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17461–17474. IEEE, 2025.

Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. DART: Articulated hand model with diverse accessories and rich textures. *Advances in Neural Information Processing Systems*, 35:37055–37067, 2022.

Linkang Ge, Qing Zhu, and Wenzhe Zhu. 3D hand mesh reconstruction based on a monocular RGB camera. In *5th International Conference on Computer Engineering and Intelligent Control*, pp. 238–245. IEEE, 2024.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

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.

Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. MegaTrack: Monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics*, 39(4): 87:1–87:17, 2020.

Xinguo He, Yixin Shen, and Rahul Chaudhari. TokenHand: Discrete token representation for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8921–8931, 2026.

Markus Höll, Markus Oberweger, Clemens Arth, and Vincent Lepetit. Efficient physics-based implementation for realistic hand–object interaction in virtual reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces*, pp. 175–182, 2018.

Stephen Joe and Frances Y. Kuo. Constructing Sobol sequences with better two-dimensional projections. *SIAM Journal on Scientific Computing*, 30(5):2635–2654, 2008. doi: 10.1137/070709359.

Kaustubh Kundu, Hrishav Bakul Barua, Lucy Robertson-Bell, Zhixi Cai, and Kalin Stefanov. DexAvatar: 3D sign language reconstruction with hand and body pose priors. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5842–5852. IEEE, 2026.

Guiqing Li, Zihui Wu, Yuxin Liu, Huiqian Zhang, Yongwei Nie, and Aihua Mao. 3D hand reconstruction from a single image based on biomechanical constraints. *The Visual Computer*, 37(9): 2699–2711, 2021a.

Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. UniPose: A unified multimodal framework for human pose comprehension, generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27805–27815. IEEE, 2025.

Yuwei Li, Minye Wu, Yuyao Zhang, Lan Xu, and Jingyi Yu. PIANO: A parametric hand bone model from magnetic resonance imaging. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 816–822. International Joint Conferences on Artificial Intelligence Organization, 2021b. doi: 10.24963/ijcai.2021/113.

Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. NIMBLE: A non-rigid hand model with bones and muscles. *ACM Transactions on Graphics*, 41(4), 2022. doi: 10.1145/3528223.3530079.

Vicente Coelho Lobo-Neto and Helio Pedrini. LSWH100: A handshape dataset for brazilian sign language (Libras) using SignWriting. *Data in Brief*, 56:110780, 2024.

Mary W. Marzke and Robert F. Marzke. Evolution of the human hand: Approaches to acquiring, analysing and interpreting the anatomical evidence. *The Journal of Anatomy*, 197(1):121–140, 2000.

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*. Springer, 2020.

Jimson G Ngeo, Tomoya Tamei, and Tomohiro Shibata. Continuous and simultaneous estimation of finger kinematics using inputs from an EMG-to-muscle activation model. *Journal of NeuroEngineering and Rehabilitation*, 11(1):122, 2014. doi: 10.1186/1743-0003-11-122.

Takehiko Ohkawa, Ryosuke Furuta, and Yoichi Sato. Efficient annotation and learning for 3D hand pose estimation: A survey. *International Journal of Computer Vision*, 131(12):3193–3206, 2023. doi: 10.1007/s11263-023-01856-0.

OpenAI. GPT-5.6 model guidance. `https://developers.openai.com/api/docs/guides/latest-model`, 2026. Accessed September 2, 2026.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *IEEE/CVF International Conference on Computer Vision*, pp. 10985–10995. IEEE, 2021.

Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. HTML: A parametric hand texture model for 3D hand reconstruction and personalization. In *European Conference on Computer Vision*, pp. 54–71. Springer, 2020.

Jinwei Ren, Jianke Zhu, and Jialiang Zhang. End-to-end weakly-supervised single-stage multiple 3D hand mesh reconstruction from a single RGB image. *Computer Vision and Image Understanding*, 232:103706, 2023.

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. doi: 10.1145/3130800.3130883.

Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Mayur Jagdishbhai Patel, Hongfei Xue, Ahmed Helmy, Srijan Das, and Pu Wang. MaskHand: Generative masked modeling for robust hand mesh reconstruction in the wild. In *IEEE/CVF International Conference on Computer Vision*, pp. 8372–8383. IEEE, 2025.

Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3D hand pose estimation via biomechanical constraints. In *Computer Vision – ECCV 2020*, pp. 211–228. Springer, 2020. doi: 10.1007/978-3-030-58520-4_13.

Gladys Tang, Joe Mak, Ka Yiu Cheng, and Felix Y. B. Sze. Developing an inventory of handshapes, locations, and movements in Hong Kong Sign Language. In *Advances in Sign Language Corpus Linguistics*, pp. 309–352. John Benjamins Publishing Company, 2023. doi: 10.1075/scl.108.12tan.

Bohan Wang, George Matcuk, and Jernej Barbič. Modeling of personalized anatomy using plastic strains. *ACM Transactions on Graphics*, 40(2):1–21, 2021. doi: 10.1145/3443703.

Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, and Chi-Wing Fu. HandBooster: Boosting 3D hand-mesh reconstruction by conditional synthesis and sampling of hand-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10159–10169. IEEE, 2024.

Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, and Chi-Wing Fu. HandBooster+: Boosting 3D hand-mesh reconstruction from data synthesis to progressive multi-hypothesis aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11201–11219, 2025. doi: 10.1109/TPAMI.2025.3596986.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand–object interaction. In *IEEE/CVF International Conference on Computer Vision*, pp. 11077–11086, 2021.

Kunpeng Yao and Aude Billard. An inverse optimization approach to understand human acquisition of kinematic coordination in bimanual fine manipulation tasks. *Biological Cybernetics*, 114:63–82, 2020. doi: 10.1007/s00422-019-00814-9.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. SignAvatars: A large-scale 3D sign language holistic motion dataset and benchmark. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

Menghe Zhang, Haley M. So, Mohammad Asadi, Dongfang Zhao, Yangwen Liang, Shuangquan Wang, Gordon Wetzstein, Kee-Bong Song, and Donghoon Kim. Synthetic hands meet legacy data: A synthetic dataset for structured, controllable, and multimodal evaluation. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp. 466–477. IEEE, 2025.

Wenqian Zhang, Molin Huang, Yuxuan Zhou, Juze Zhang, Jingyi Yu, Jingya Wang, and Lan Xu. BOTH2Hands: Inferring 3D hands from both text prompts and body dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2393–2404. IEEE, 2024a.

Yufei Zhang, Jeffrey O. Kephart, and Qiang Ji. Weakly-supervised 3D hand reconstruction with knowledge prior and uncertainty guidance. In *European Conference on Computer Vision*, pp. 106–125. Springer, 2024b.

Zhongqun Zhang, Hengfei Wang, Ziwei Yu, Yihua Cheng, Angela Yao, and Hyung Jin Chang. NL2Contact: Natural language guided 3D hand-object contact modeling with diffusion model. In *European Conference on Computer Vision*, pp. 284–300. Springer, 2024c.

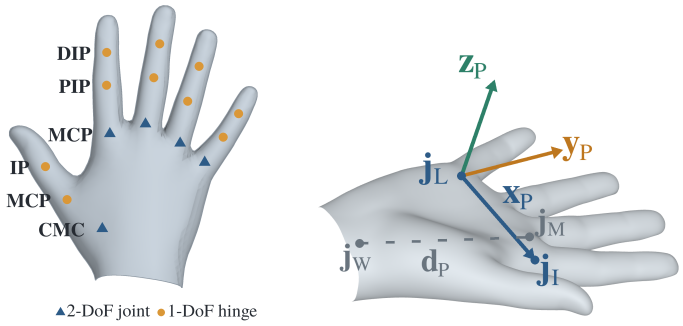
Zimu Zhang, Yucheng Zhang, Xiyan Xu, Ziyin Wang, Sirui Xu, Kai Zhou, Bing Zhou, Chuan Guo, Jian Wang, Yu-Xiong Wang, et al. HandX: Scaling bimanual motion and interaction generation. *arXiv preprint arXiv:2603.28766*, 2026.

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. doi: 10.1109/ICCV.2019.00090.

A JOINT GEOMETRY AND THUMB-AXIS CALIBRATION

A.1 JOINT MOTION AND PALM GEOMETRY

Hand joints differ in both their degrees of freedom and their motion patterns. These differences guide the choice of joint angles and the references used to measure them. This section expands on Sec. 3.1 by explaining the joint motion constraints and the geometric construction of the palm reference frame.



(a) (b)

Figure 6: (a) Joint locations and degrees of freedom in AHand20. Triangles mark joints with two degrees of freedom; circles mark joints with one degree of freedom. (b) Palm reference frame and the four joint positions used to construct it.

Figure 6(a) locates the joints used in AHand20. Fingertip positions are taken from right-hand MANO mesh vertices 745, 317, 444, 556, and 673 for the thumb, index, middle, ring, and little fingers, respectively (0-indexed). For the non-thumb fingers, the metacarpophalangeal (MCP) joint supports flexion–extension, which controls finger bending and straightening, and abduction–adduction, which controls lateral spreading between neighboring fingers. The proximal and distal interphalangeal (PIP and DIP) joints have flexion–extension as their dominant motion and are modeled as hinges. Axial twist is not permitted at any joint in this model. Measuring joint motion as a signed angle requires a zero state and a positive direction for each dimension.

Figure 6(b) illustrates the palm axes, which follow the palm's width and length. The direction from the little-finger MCP to the index-finger MCP runs across the palm and defines $\mathbf{x}_P$. The wrist-to-middle-MCP direction $\mathbf{d}_P$ runs along the palm. We remove its component along $\mathbf{x}_P$ and normalize it to obtain $\mathbf{y}_P$. Their cross product gives the palm normal $\mathbf{z}_P$. This frame supplies the reference directions for the geometric angle definitions, while the zero states and positive directions are specified for each joint type in Sec. 3.2.

Left-hand poses can be mirrored to the right-hand convention by transforming both joint coordinates and local rotations, allowing the same AHand20 definitions to be used. Generated right-hand meshes can also be mirrored to obtain left-hand meshes.

A.2 THUMB MCP AND IP HINGE-AXIS FITTING

This section derives the minimization over φ and details the optimization of the thumb hinge axes used for flexion measurement in Sec. 3.2. We start from the swing rotation $\mathbf{R}_{i,j}^{\text{swing}}$ obtained after removing axial twist, where $j \in \{\text{MCP}, \text{IP}\}$.

Closed-form solution for the flexion angle. We approximate the swing rotations across poses using a shared hinge axis and a separate flexion angle for each pose. Once the unit axis $\mathbf{a}$ is fixed, the closest rotation is determined by a single angle φ. The Frobenius fitting error depends on this

angle only through the alignment between the observed and candidate rotations:

$$E_{i,j}(\mathbf{a},\varphi) = \left\| \mathbf{R}_{i,j}^{\text{swing}} - \text{Rot}(\mathbf{a},\varphi) \right\|_F^2 = 6 - 2\,\text{tr}\Big((\mathbf{R}_{i,j}^{\text{swing}})^\top \text{Rot}(\mathbf{a},\varphi) \Big). \tag{14}$$

Minimizing the error therefore maximizes the trace. Rodrigues' formula separates the candidate rotation into a part fixed by the axis and two angle-dependent terms:

$$\text{Rot}(\mathbf{a},\varphi) = \mathbf{a}\mathbf{a}^\top + (\mathbf{I} - \mathbf{a}\mathbf{a}^\top)\cos\varphi + [\mathbf{a}]_\times \sin\varphi. \tag{15}$$

Here, $[\mathbf{a}]_\times \mathbf{v} = \mathbf{a} \times \mathbf{v}$. Because the trace is linear, its angle-dependent part has the form $c_{i,j}(\mathbf{a})\cos\varphi + s_{i,j}(\mathbf{a})\sin\varphi$, with coefficients

$$\begin{aligned} c_{i,j}(\mathbf{a}) &= \text{tr}(\mathbf{R}_{i,j}^{\text{swing}}) - \mathbf{a}^\top \mathbf{R}_{i,j}^{\text{swing}} \mathbf{a}, \\ s_{i,j}(\mathbf{a}) &= \mathbf{a}^\top \text{vee}\Big(\mathbf{R}_{i,j}^{\text{swing}} - (\mathbf{R}_{i,j}^{\text{swing}})^\top \Big). \end{aligned} \tag{16}$$

Here, vee is the inverse of the skew-symmetric matrix operator: for a skew-symmetric matrix $\mathbf{S}$, $\text{vee}(\mathbf{S}) = (S_{32}, S_{13}, S_{21})^\top$.

For coefficients that are not both zero, the angle-dependent terms combine into a single cosine:

$$\begin{aligned} c_{i,j}(\mathbf{a})\cos\varphi + s_{i,j}(\mathbf{a})\sin\varphi &= \sqrt{c_{i,j}(\mathbf{a})^2 + s_{i,j}(\mathbf{a})^2}\,\cos\big(\varphi - \varphi_{i,j}^\star(\mathbf{a})\big), \\ \varphi_{i,j}^\star(\mathbf{a}) &= \text{atan2}(s_{i,j}(\mathbf{a}), c_{i,j}(\mathbf{a})). \end{aligned} \tag{17}$$

The cosine is maximized at $\varphi = \varphi_{i,j}^\star(\mathbf{a})$ modulo 2π, giving the global minimum of Eq. 14 for the fixed axis. If both coefficients vanish, every angle gives the same fitting error.

Axis-fitting objective. The maximum angle-dependent contribution is $\sqrt{c_{i,j}(\mathbf{a})^2 + s_{i,j}(\mathbf{a})^2}$. Combining it with the constant trace term $\mathbf{a}^\top \mathbf{R}_{i,j}^{\text{swing}} \mathbf{a}$ gives an objective that depends only on the axis:

$$\mathcal{L}_{\text{axis},j}(\mathbf{a}) = \frac{1}{N}\sum_{i=1}^{N} \left\{ 6 - 2\left[\mathbf{a}^\top \mathbf{R}_{i,j}^{\text{swing}} \mathbf{a} + \sqrt{c_{i,j}(\mathbf{a})^2 + s_{i,j}(\mathbf{a})^2} \right] \right\}. \tag{18}$$

Minimizing $\mathcal{L}_{\text{axis},j}$ over unit axes gives $\mathbf{a}_{t,j}^\star$. Each candidate axis is evaluated with its optimal per-pose angles, so the numerical optimization searches only for the shared hinge direction. The MCP and IP joints are fitted independently.

Axis fitting and flexion measurement. We calibrate the right-hand MANO thumb MCP and IP axes from $N = 32{,}560$ FreiHAND pose annotations. Candidate axes are initialized along the positive and negative local coordinate directions and normalized during optimization. We optimize the axes using Adam with a learning rate of 0.05 for 160 iterations and retain the axis with the lowest observed loss.

Since reversing both the axis and angle represents the same rotation, we orient each fitted axis so that the median projected angle over the calibration poses is non-negative. The resulting axes in the right-hand MANO local joint frames are $\mathbf{a}_{t,\text{MCP}}^\star \approx (-0.57, 0.14, -0.81)^\top$ and $\mathbf{a}_{t,\text{IP}}^\star \approx (-0.59, 0.38, -0.71)^\top$. Neither fitted axis aligns with any of the MANO local x, y, or z axes. These FreiHAND-calibrated axes are used by default. For the DexYCB results in Table 3(a), we instead use axes fitted on DexYCB. Given a swing rotation $\mathbf{R}_j^{\text{swing}}$, its flexion angle $\alpha_{t,j}^{\text{flx}}$ is obtained by evaluating Eq. 17 with the corresponding fitted axis.

A.3 Thumb-Axis Generalization Across Datasets

The thumb MCP and IP dimensions use hinge axes estimated from pose data. We evaluate how these axes transfer beyond their calibration dataset by fitting each axis on one dataset and measuring its residual on three other datasets. This experiment examines the fitted axes used to measure joint motion.

A.3.1 Experimental Setup

We use right-hand MANO poses from FreiHAND (Zimmermann et al., 2019), DexYCB (Chao et al., 2021), HO3D v3 (Hampali et al., 2020), and InterHand2.6M (Moon et al., 2020). Table 4 gives the fitting and evaluation counts. HO3D poses come from its training annotations, and InterHand2.6M poses are collected from the train, validation, and test partitions of its 5-fps MANO annotations.

We unify the MANO local rotation conventions across datasets before swing–twist decomposition, excluding global wrist rotation. The thumb MCP and IP axes are fitted separately on each source dataset using the method in Appendix A.2 and kept fixed when evaluating the other datasets. We measure the geodesic distance between each target swing rotation and its closest rotation about the transferred axis. Mean and P90 residuals are reported in degrees for swing rotations with magnitude at least 1°.

Table 4: Right-hand poses used for thumb-axis generalization. Fitting counts apply to each joint; evaluation counts include only swing rotations of at least 1°.

Dataset	Fitted	MCP evaluated	IP evaluated
FreiHAND	32,560	32,498	32,517
DexYCB	31,750	31,651	31,372
HO3D v3	80,000	77,202	78,620
InterHand2.6M	18,449	18,435	18,449

A.3.2 Results and Analysis

Table 5: Thumb-axis generalization across datasets. Rows specify the axis-fitting dataset and columns specify the evaluation dataset. Entries report mean / P90 geodesic residuals in degrees. Diagonal entries are same-dataset fitting references.

Thumb MCP				
Axis source	FreiHAND	DexYCB	HO3D v3	InterHand2.6M
FreiHAND	9.15 / 18.13	10.06 / 17.48	15.07 / 30.04	7.94 / 14.58
DexYCB	14.59 / 29.24	6.47 / 13.21	6.20 / 13.16	8.69 / 15.96
HO3D v3	18.27 / 34.31	6.96 / 15.15	0.64 / 1.48	14.78 / 24.53
InterHand2.6M	10.41 / 22.42	8.12 / 14.51	11.70 / 23.74	4.33 / 9.08
Thumb IP				
Axis source	FreiHAND	DexYCB	HO3D v3	InterHand2.6M
FreiHAND	10.23 / 21.26	5.93 / 11.00	7.39 / 16.04	8.93 / 21.63
DexYCB	16.93 / 33.02	2.59 / 5.49	0.33 / 0.88	9.34 / 15.84
HO3D v3	16.83 / 32.84	2.59 / 5.49	0.36 / 0.76	9.27 / 15.76
InterHand2.6M	13.07 / 26.89	3.47 / 6.84	3.19 / 6.69	7.72 / 14.41

Overall generalization. The fitted thumb axes show reasonable generalization across datasets. For HO3D-to-DexYCB, the mean MCP and IP errors are 6.96° and 2.59°, compared with 6.47° and 2.59° for axes fitted on DexYCB itself. In the reverse direction, the transferred IP axis gives 0.33° mean error on HO3D, compared with 0.36° for its own fitted axis. The FreiHAND-fitted IP axis also gives 8.93° on InterHand2.6M, compared with 7.72° for the InterHand2.6M-fitted axis. These cases demonstrate that axes estimated from one dataset can describe thumb motion in another without being refitted. The fitted axes are sensitive to the calibration dataset, but mean residuals across all transfer settings remain below 20°, ranging from 0.33° to 18.27°. Thus, axes fitted on one dataset provide an approximation of thumb motion across the other datasets.

MCP and IP comparison. With axes fitted and evaluated on the same dataset, IP has lower mean error on DexYCB and HO3D v3, while MCP has lower error on FreiHAND and InterHand2.6M.

Across datasets, IP has lower mean error in eight of the twelve transfer directions. This suggests that the IP axis generally transfers more readily than the MCP axis, possibly because the MCP joint has a larger flexion range than the IP joint.

Dataset comparison. Low error on the fitting dataset does not necessarily imply low transfer error. HO3D v3 has the lowest same-dataset errors for both MCP and IP (0.64° and 0.36°), but its axes give 18.27° and 16.83° on FreiHAND. Conversely, FreiHAND has higher same-dataset errors (9.15° and 10.23°), yet its axes give lower errors on InterHand2.6M (7.94° and 8.93°) than the HO3D-fitted axes (14.78° and 9.27°). These differences may arise from variations in data quality and pose distribution across datasets.

A.4 Fitting AHand20 Targets to MANO

For the 20D-to-MANO fitting in Sec. 3.3, we use Adam with a learning rate of 0.035 for 1,100 steps. The MANO input pose parameters are initialized to zero, while the shape vector $\boldsymbol{\beta}$ and the calibrated thumb axes remain fixed. Table 6 lists the loss weights used for the conversion.

Table 6: Implementation settings for fitting 20D targets to MANO. Non-thumb target weights apply to each of the four fingers.

Parameter	Value
Target weight: non-thumb MCP spread	0.8
Target weight: non-thumb MCP flexion	1.1
Target weight: non-thumb PIP flexion	1.25
Target weight: non-thumb DIP flexion	1.0
Target weight: each thumb dimension	1.0
Overall plane-loss weight	0.15
Plane weights: MCP / PIP / DIP	0.2 / 0.7 / 0.7
Excess weight: non-thumb MCP	0.25
Excess weight: non-thumb PIP and DIP	0.5

Squared plane residuals are averaged over the four fingers for each joint type. Squared excess residuals are averaged over the four MCP joints and the eight PIP/DIP joints separately before applying the corresponding weights. Angular differences are wrapped to $[-\pi, \pi]$.

B Handshape Inventory, Augmentation, and Descriptions

Datasets pairing individual 3D hand poses with fine-grained textual descriptions remain relatively scarce or require access approval (Zhang et al., 2024a). GigaHands (Fu et al., 2025) and HandX (Zhang et al., 2026) pair text with hand motion sequences rather than individual poses. LSWH100 (Lobo-Neto & Pedrini, 2024) contains static handshapes but lacks 3D mesh annotations. We therefore construct a hand pose–text dataset using AHand20.

B.1 Reference Images and Inventory Selection

The selected inventory contains 87 handshapes from Global Signbank (glo) and 67 from Asian SignBank (Tang et al., 2023; asi). Repeated or visually indistinguishable handshapes across the sources are retained once, while variants with different articulation or contact relations are preserved. Figure 7 shows six representative source images from Global Signbank and their fitted MANO poses. They include open finger arrangements, crossed fingers, fingertip contact, and compact configurations. The source images guide manual specification of the 20D targets, and the resulting fitted poses are visually reviewed.

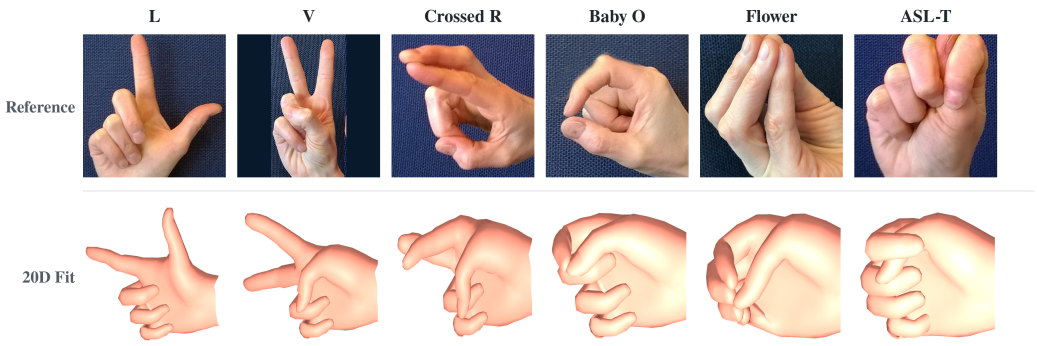


Figure 7: Representative source images from Global Signbank (top) and MANO poses fitted from manually specified 20D anatomical targets (bottom). The examples are L, V, Crossed R, Baby O, Flower, and ASL-T.

We summarize finger articulation and required contacts in the 154 selected handshapes. Non-thumb fingers are grouped into four states based on MCP and PIP flexion. We evaluate the two flexion angles separately, treating a joint as flexed when its angle is at least 30°. Their combination determines the finger state: *extended* when both angles are below the threshold, *MCP-flexed* when only the MCP angle reaches the threshold, *hook-like* when only the PIP angle reaches it, and *curled* when both do. Table 7 summarizes these four combinations.

Table 7: Coarse articulation states used for inventory statistics. Both angles are the non-thumb flexion dimensions defined in the main text.

State	MCP flexion	PIP flexion
Extended	$< 30^\circ$	$< 30^\circ$
MCP-flexed	$\geq 30^\circ$	$< 30^\circ$
Hook-like	$< 30^\circ$	$\geq 30^\circ$
Curled	$\geq 30^\circ$	$\geq 30^\circ$

Figure 8(a) shows all four states for each non-thumb finger. The index finger is extended in 62 handshapes and hook-like in 40, whereas the ring and little fingers are curled in 93 and 89 handshapes, respectively. The index finger is more evenly distributed across the four states, whereas the ring and little fingers are predominantly curled, suggesting that index-finger articulation is more diverse and may play a greater role in distinguishing handshapes.

Of the 154 handshapes, 77 have at least one required surface-contact condition: 50 from Global Signbank and 27 from Asian SignBank. Among these, 57 have one contact condition and 20 have multiple conditions. The thumb participates in a required contact in 72 handshapes. As shown in Fig. 8(b), thumb–index and thumb–middle are the most frequent pairs, occurring in 32 and 26 handshapes. These counts show that contact is common among the selected handshapes and describe the contacts explicitly required to preserve handshape identity.

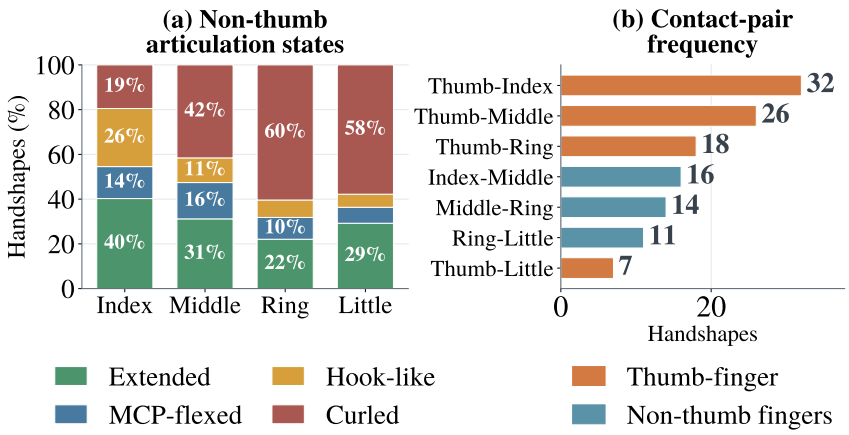


Figure 8: Statistics of the 154 reference handshapes. (a) Non-thumb articulation states classified by MCP and PIP flexion. (b) Number of handshapes with each required contact pair in their preservation configuration. A handshape may contribute to multiple contact pairs.

B.2 Pose Perturbation and MANO Shape Selection

This section details how the 20D targets are perturbed and paired with MANO shape parameters for the augmentation described in Sec. 4.

Related work on pose augmentation and data synthesis. Data augmentation increases the diversity of hand pose datasets without requiring additional data collection. Existing approaches include image-level and pose-level augmentation. Image-level augmentation varies hand appearance, viewpoints, and rendering conditions (Xu et al., 2024; 2025; Gao et al., 2022), and pose-level augmentation generates poses through latent sampling or gesture composition (Petrovich et al., 2021; Zhang et al., 2025). These methods have limitations: image-level augmentation does not expand the pose distribution, whereas pose-level augmentation provides limited control over joint-level articulation.

Sampling anatomical perturbations. For handshape h, let $\mathbf{z}_h \in \mathbb{R}^{20}$ denote the target used as the sampling center. We specify a perturbation magnitude $\boldsymbol{\delta}_h \in \mathbb{R}^{20}$ separately for each handshape and anatomical dimension. We sample $\mathbf{u}_n \in [-1, 1]^{20}$ using scrambled Sobol sequences (Joe & Kuo, 2008) and scale it element-wise by $\boldsymbol{\delta}_h$ to obtain the perturbed target:

$$\widetilde{\mathbf{z}}_{h,n} = \mathbf{z}_h + \boldsymbol{\delta}_h \odot \mathbf{u}_n, \tag{19}$$

where $\odot$ denotes element-wise multiplication.

The configured magnitudes vary with the handshape and joint, as summarized in Table 8. A magnitude of 5° allows perturbations of up to $\pm 5^\circ$. For example, the open-five configuration uses MCP spread magnitudes of $(8, 5, 5, 8)^\circ$ for the index, middle, ring, and little fingers, respectively, and 8° for their MCP flexion. Its thumb CMC azimuth and elevation magnitudes are 10° and 8°, and its thumb MCP and IP flexion magnitudes are both 8°.

Shape pool construction and selection. We select 20 ten-dimensional MANO shape vectors from FreiHAND, DexYCB, HO3D, and InterHand2.6M using farthest-point sampling in the component-wise median/IQR-normalized shape space. The final pool contains eight FreiHAND, eleven HO3D, and one InterHand2.6M shape vectors.

Pairing shape vectors with perturbed targets. For each handshape, we pair the selected shape vectors with pose variants. Most handshapes use all 20 shapes; two configurations restrict the allowed subset to four and six shapes, respectively. Each candidate keeps its assigned $\boldsymbol{\beta}$ fixed while its MANO pose is fitted to the 20D target. Candidate counts range from 400 to 8,000 per handshape.

Table 8: Minimum and maximum configured perturbation magnitudes across the 154 handshapes. Non-thumb rows aggregate the four non-thumb fingers. Values are in degrees.

Dimension	Minimum	Maximum
Non-thumb MCP spread	2.0	8.0
Non-thumb MCP flexion	2.5	8.0
Non-thumb PIP flexion	2.5	9.0
Non-thumb DIP flexion	2.5	7.0
Thumb CMC azimuth	2.0	20.0
Thumb CMC elevation	2.0	15.0
Thumb MCP flexion	2.5	35.0
Thumb IP flexion	2.5	27.0

Figure 9 illustrates the combined effect of changing the 20D target and $\boldsymbol{\beta}$ for an open-five handshape. Within each row, the shape vector is fixed while non-thumb MCP spread varies; within each column, the 20D target is fixed while the shape vector varies.

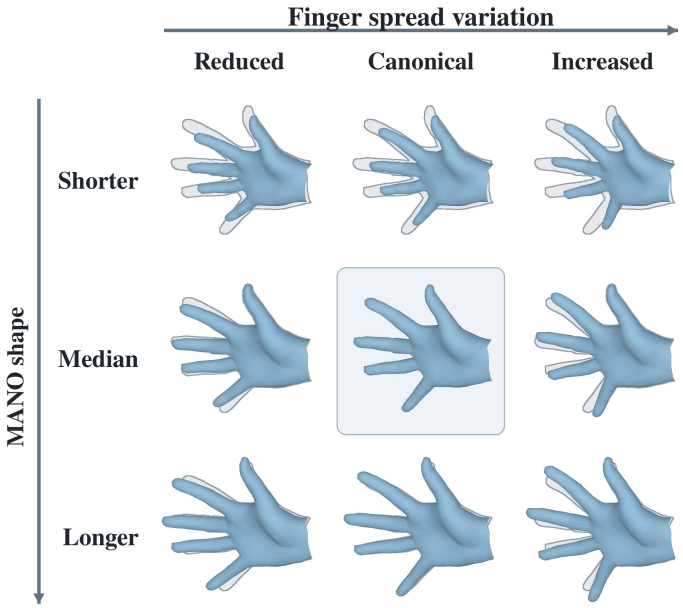


Figure 9: Pose and MANO shape augmentation around a selected pose. Columns vary non-thumb MCP spread through the 20D target, and rows vary hand size and finger length through $\boldsymbol{\beta}$. The highlighted central example is the reference pose and shape. Gray outlines show its silhouette for comparison.

B.3 Hand-Pose Preservation Criteria

Changing articulation and MANO shape can alter the finger states and spatial relations that define a handshape. We therefore evaluate each fitted candidate using handshape-specific articulation, contact, and separation conditions. Figure 10 illustrates these three types of conditions.

Articulation conditions. Articulation conditions preserve finger states through bounds on individual anatomical angles and differences between selected pairs of angles. For candidate $\boldsymbol{\theta}^{\star}_{h,n}$ fitted with shape $\boldsymbol{\beta}_n$, let $\mathbf{z}_{h,n} = \Phi(\boldsymbol{\theta}^{\star}_{h,n}; \boldsymbol{\beta}_n)$ denote its measured AHand20 representation. Each defining dimension r satisfies

$$L^{\text{abs}}_{h,r} \leq z_{h,n,r} \leq U^{\text{abs}}_{h,r}. \tag{20}$$

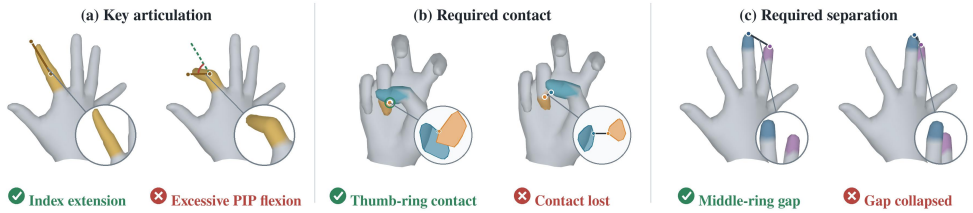


Figure 10: Examples of hand-pose preservation criteria. Each pair shows a preserved condition (left) and a violated condition (right): (a) index extension, (b) thumb–ring contact, and (c) middle–ring separation.

These bounds allow variation while retaining the required finger state. For dimensions whose relation defines the handshape, we additionally impose

$$L^{\mathrm{rel}}_{h,rs} \le z_{h,n,r} - z_{h,n,s} \le U^{\mathrm{rel}}_{h,rs}. \tag{21}$$

An interval around zero keeps two articulations similar, while an asymmetric interval preserves their ordering or difference.

Contact conditions. Required contacts are evaluated on the fitted MANO mesh. For two relevant vertex regions $\mathcal{V}_A$ and $\mathcal{V}_B$, their normalized surface distance is

$$d_{\mathrm{surf}}(\mathcal{V}_A, \mathcal{V}_B) = \frac{1}{w_{\mathrm{P}}} \min_{\substack{\mathbf{v}_A \in \mathcal{V}_A \\ \mathbf{v}_B \in \mathcal{V}_B}} \|\mathbf{v}_A - \mathbf{v}_B\|_2, \tag{22}$$

where $w_{\mathrm{P}} = \|\mathbf{j}_{\mathrm{I}} - \mathbf{j}_{\mathrm{L}}\|_2$ is the palm width. The contact is preserved when

$$d_{\mathrm{surf}}(\mathcal{V}_A, \mathcal{V}_B) \le \tau^{\mathrm{contact}}_{h,A,B}. \tag{23}$$

Normalizing by palm width makes this condition comparable across MANO shapes. The handshape-specific threshold allows a small gap relative to the reference contact distance, capped by a fraction of palm width.

Separation conditions. Required gaps are preserved through a minimum surface or fingertip distance. For two surface regions, the condition is

$$d_{\mathrm{surf}}(\mathcal{V}_A, \mathcal{V}_B) \ge \tau^{\mathrm{sep,surf}}_{h,A,B}. \tag{24}$$

When the gap is defined by two fingertips, we use

$$\frac{\|\mathbf{j}_{\mathrm{tip},A} - \mathbf{j}_{\mathrm{tip},B}\|_2}{w_{\mathrm{P}}} \ge \tau^{\mathrm{sep,tip}}_{h,A,B}. \tag{25}$$

The thresholds retain a specified fraction of the corresponding normalized reference gap. A candidate is retained only if all articulation, contact, and separation conditions assigned to its handshape are satisfied.

Mesh self-penetration. Candidates must additionally pass a check for severe intersections between hand regions (Forte et al., 2023). We allow slight mesh penetration where finger contact is concerned. We use the predefined MANO vertex masks from KNOWN-Hand (Zhang et al., 2024b) to define six regions: the palm and the five fingers. These masks are fixed across poses and shapes. For each region, we retain faces whose vertices all belong to its vertex set and close each boundary loop with triangles connecting its edges to the boundary centroid. For penetration queries, vertices shared by the palm and a finger are assigned to the finger. For each region pair, we identify vertices inside the other region's closed surface and compute their distances to its nearest triangle, in both directions. The implementation uses a winding-number threshold of 0.9 for the inside test. All penetration depths and thresholds below are in millimeters (mm).

For the thumb against the palm and non-thumb fingers, let $\mathcal{D}_{\mathrm{thumb}}$ contain the penetrating vertex depths. A candidate is rejected when both the total and maximum depths exceed their thresholds:

$$\sum_{d \in \mathcal{D}_{\mathrm{thumb}}} d > 25\,\mathrm{mm} \quad \text{and} \quad \max_{d \in \mathcal{D}_{\mathrm{thumb}}} d > 5\,\mathrm{mm}. \tag{26}$$

For pairs of non-thumb fingers, $\mathcal{D}_{\text{inter}}$ contains the penetration depths across all pairs. We sum the excess depth beyond 6 mm and reject candidates when

$$\sum_{d \in \mathcal{D}_{\text{inter}}} [d - 6\,\text{mm}]_+ > 5.5\,\text{mm}, \tag{27}$$

where $[x]_+ = \max(x, 0)$.

For non-thumb fingers against the palm, we apply the same excess-depth measure to their penetration depths $\mathcal{D}_{\text{palm}}$, rejecting candidates when

$$\sum_{d \in \mathcal{D}_{\text{palm}}} [d - 6\,\text{mm}]_+ > 10\,\text{mm}. \tag{28}$$

Empty collections contribute zero to both the sum and the maximum. These checks use summed vertex penetration depths and the maximum depth. Candidates passing all assigned preservation conditions and penetration checks form the accepted set for the final diversity selection.

Selecting the final pose variations. After the articulation, contact, separation, and self-penetration checks, we retain 100 candidates per handshape using farthest-point sampling in a joint pose–shape feature space to encourage diversity in both shape and articulation.

B.4 Description Styles and Language Enrichment

The template-guided descriptions in Sec. 4 vary along two axes: writing style and information scope. Writing style determines how the recorded handshape facts are presented. Figure 11 shows examples of all five styles for the same V handshape. Anatomical descriptions emphasize finger and joint states; visual descriptions emphasize the visible arrangement; instructional descriptions state how to form the pose. Natural descriptions use fluent, continuous language, while detailed-neutral descriptions present the configuration without emphasizing a particular perspective. These styles share the same underlying facts and may overlap in wording.

Information scope determines the coverage and grouping of those facts. *Full* provides the most comprehensive description; *grouped* combines fingers sharing a state or relation; *selective* emphasizes the defining facts; and *minimal* retains the essential configuration. Each of the five styles is combined with all four scopes, with five formulations per combination, giving $5 \times 4 \times 5 = 100$ template-guided descriptions per handshape. Each has a corresponding language-enriched rewrite that preserves its handshape facts, style, and scope.

Anatomical — Finger and joint states

The index and middle fingers are extended and spread into a clear V, while the ring and little fingers are fully curled into the palm. The thumb bends inward and rests over the ring finger. Together, these states produce a clear forked upper outline above a compact base.

Visual — Visible structure and spatial relations

The index and middle create two separated straight projections, giving the hand a clear forked upper outline above a compact base. The ring and little fingers disappear into a tight curl, and the thumb bends inward and rests over the ring finger.

Instructional — Actions used to form the pose

Extend the index and middle fingers and spread them into a V. Next, curl the ring and little fingers fully into the palm. Finally, bend the thumb inward over the ring finger.

Natural — Fluent, continuous language

The index and middle stay straight and open apart like a V. The ring and little fingers are fully curled into the palm. The thumb bends inward and rests over the ring finger. This leaves a clear forked upper outline above a compact base.

Detailed-neutral — Neutral description of the pose

The index and middle fingers are extended and spread into a clear V. The ring and little fingers are fully curled into the palm. The thumb bends inward and rests over the ring finger, leaving a clear forked upper outline above a compact base.

Figure 11: Five writing styles for the same V handshape. The gray box contains unedited template-guided descriptions from the dataset; all five examples use the Full information scope. The index and middle fingers form a V, the ring and little fingers are curled, and the thumb rests over the ring finger.

Language-enriched descriptions. Figure 12 pairs three template-guided descriptions with their corresponding language-enriched rewrites from the dataset. The examples cover the V, baby-beak, and L handshapes and three style–scope combinations. In the V example, the rewrite uses anatomical terms such as "abducted" and "flex" to express finger spread and bending. The baby-beak example reformulates the fingertip contact and compact hand structure, while the L example expands the formation instruction into a sequence of actions. The examples illustrate changes in wording and sentence organization while retaining the defining finger states and spatial relations.

V handshape — Anatomical / Full

Template-guided

The index and middle fingers are extended and spread into a clear V, while the ring and little fingers are fully curled into the palm. The thumb bends inward and rests over the ring finger. Together, these states produce a clear forked upper outline above a compact base.

Language-enriched

The index and middle fingers remain extended and are abducted from one another to form a V. In contrast, the ring and little fingers flex completely into the palm. The thumb bends inward and lies over the curled ring finger.

Baby-beak handshape — Visual / Selective

Template-guided

A straight upper index and a curved lower thumb touch at their tips, forming a small closed beak above a tucked-finger base.

Language-enriched

A straight upper index and a curved lower thumb meet at the tips, drawing a small closed beak above three tucked fingers.

L handshape — Instructional / Minimal

Template-guided

Form an L with an extended upright index and an abducted sideways thumb; keep three fingers curled into the palm.

Language-enriched

Point the straight index upward and extend the thumb sideways to make an L, then curl the other three fingers fully into the palm.

Figure 12: Examples of language-enriched descriptions paired with their template-guided sources. Each pair describes the same handshape and retains the same style and information-scope labels.

Description lengths. Table 9 reports mean word counts for the two description sets, each with 15,400 descriptions and 3,850 per scope. Consistent with their information coverage, Full descriptions are the longest and Minimal descriptions are the shortest in both sets. Language-enriched descriptions are shorter on average in every scope.

Table 9: Mean number of words per description, by information scope.

Information scope	Template-guided	Language-enriched
Full	54.5	43.5
Grouped	45.8	31.7
Selective	35.9	25.9
Minimal	29.4	20.7
All scopes	41.4	30.5

Vocabulary. Figure 13 summarizes the vocabulary of all 30,800 descriptions. Tokens are lowercased, selected inflected forms are grouped (e.g., *curled* and *curling* under *curl*), and stopwords and generic terms such as *hand*, *finger*, and *pose* are removed. Finger names and terms describing finger states are the most common words in the word cloud, reflecting the descriptions' focus on finger configurations.



Figure 13: Frequent content words after normalization and stopword removal.

C Generator Implementation

This appendix describes the architecture and training settings of the AHand20-based generator in Sec. 4.

C.1 Network Architecture

Pose encoder. A multilayer perceptron (MLP) maps the 20D anatomical pose to a 32-dimensional latent Gaussian distribution. It uses three hidden linear layers of width 512 and separate output heads for the mean and scale.

Text encoder. A frozen DistilBERT backbone extracts 768-dimensional token features, which are projected to dimension 512 and processed by a trainable Transformer encoder. The Transformer has four layers, four attention heads, a feed-forward dimension of 1,024, and dropout 0.1. Mean pooling over non-padding tokens and two output heads produce a Gaussian distribution in the same 32-dimensional latent space.

Shared decoder. The MLP decoder D_z maps a latent sample to a 20D pose through two 512-dimensional hidden layers. A second MLP, G_R, uses two hidden layers of the same width to predict a 6D rotation representation for each of the 15 MANO joints, which is converted to a rotation matrix. Both MLPs use LeakyReLU activations and are shared by the pose and text branches. The encoders and decoder are trained jointly, with the DistilBERT backbone frozen.

The predicted rotations are passed to a fixed right-hand MANO layer. During training, we use each sample's recorded shape vector β to generate MANO joints and vertices from both the predicted and reference rotations. The joint and vertex losses therefore compare poses with the same hand shape, so differences in hand size and finger length do not affect the comparison.

C.2 Loss Implementation

For the 20D, joint, and vertex terms, the per-component Laplace negative log-likelihood is

$$\mathcal{L}_{\mathrm{Lap}}(e; s) = \log 2 + s + |e| \exp(-s), \tag{29}$$

where s is a learned log-scale. For joint and vertex coordinates, e is the coordinate difference in meters. For 20D angles, it is the wrapped difference $e = \mathrm{atan2}(\sin(\widehat{z} - z), \cos(\widehat{z} - z))$. The rotation term applies a Gaussian negative log-likelihood to each entry of the predicted and target rotation matrices:

$$\mathcal{L}_{\mathrm{Gauss}}(e; s_R) = \tfrac{1}{2} \left[\log(2\pi) + s_R + e^2 \exp(-s_R)\right], \tag{30}$$

where s_R is a learned log-variance. There are four learned scalar uncertainty parameters, one per reconstruction quantity, shared by the pose and text branches and initialized to zero. Alignment uses $D_{\mathrm{KL}}(q_{\mathrm{pose}} \| q_{\mathrm{text}})$. Table 10 gives the effective coefficients in the training objective.

Table 10: Effective weights in the generator training objective. The same reconstruction weights apply to both branches.

Term	Weight
20D reconstruction λ_z	2
Rotation reconstruction λ_R	2
Joint reconstruction λ_J	2
Vertex reconstruction λ_V	4
Text-branch reconstruction λ_{text}	1
Pose–text KL alignment λ_{align}	0.1
Pose standard-normal KL $\lambda_{\text{pose}}^{\text{KL}}$	0.001
Text standard-normal KL $\lambda_{\text{text}}^{\text{KL}}$	0.001

C.3 Training Protocol and Generation

Pose–text pairing. For the unseen-handshape experiment, the training, validation, and test sets contain 124, 15, and 15 handshapes, respectively, corresponding to 12,400, 1,500, and 1,500 pose variations. Each handshape has 200 descriptions from the template-guided and language-enriched sets. When a training pose is loaded, one description is sampled uniformly from its handshape's pool. This forms different pose–text pairs across epochs, while validation uses fixed pose–text pairs.

Optimization. We train on NVIDIA L40S GPUs using Adam with learning rate 10^{-5}, weight decay 10^{-4}, moment coefficients $(0.9, 0.999)$, and stability constant $\epsilon = 10^{-8}$. All trainable components use the same learning rate. We train for 5,000 epochs with a batch size of 128 and a constant learning rate. DistilBERT loads pretrained weights and MANO uses its fixed model parameters.

Text-conditioned generation. At test time, the text encoder predicts $q_{\text{text}}(\boldsymbol{\ell} \mid c)$, and samples from this distribution are passed through D_z and G_R. We use one sampled pose per description for geometric evaluation. The predicted rotations can be applied to a supplied MANO shape vector.

D Evaluation Protocols and Supplementary Results

D.1 Evaluation Protocol and Metrics

This section explains the reference matching and sample aggregation used for text-conditioned generation, together with the retrieval and distribution metrics reported in Sec. 5.

Existing hand-focused text-conditioned methods primarily address motion sequences or hand–object interactions (Fu et al., 2025; Zhang et al., 2026; Cha et al., 2024). Our task instead generates an individual hand pose from a description of finger states and spatial relations. We therefore adapt established text-conditioned pose generation and editing methods to MANO, training and evaluating them on the same hand pose–text dataset.

For PoseScript (Delmas et al., 2024), we resize the pose encoder and decoder to 15 hand joints and use MANO for geometric supervision. For ChatPose (Feng et al., 2024), we replace the body-pose output head with one predicting 15 hand-joint rotations. For PoseFix (Delmas et al., 2023), we adapt its pose encoder and decoder to MANO, fix the source pose, and replace editing instructions with handshape descriptions. For UniPose (Li et al., 2025), we retrain the pose tokenizer on MANO rotations and replace its language backbone with frozen DistilBERT, retaining parallel pose-token prediction.

Each description specifies a handshape rather than a unique 3D pose. The 100 reference poses represent variations generated under handshape-preservation constraints, allowing changes in articulation while retaining the described handshape. Matching a prediction to this reference set accommodates such valid variation instead of penalizing differences from a single fixed pose. We compute MPJPE, MPVPE, and rotation error against all 100 references and retain the minimum for each metric independently. The reported scores are the averages of these minimum errors over all test descriptions. MPJPE measures the mean position error over 21 joints, MPVPE over 778 mesh vertices, and rotation error the mean geodesic distance over 15 local joint rotations. Position errors are reported in

millimeters and rotation errors in degrees. Table 1 uses 1,500 fixed test descriptions from the 15 unseen handshapes, with 750 template-guided and 750 language-enriched descriptions.

Retrieval recall and FID use a separate pose–text evaluator that is fixed and shared across all generation methods. Retrieval recall measures whether a generated pose matches its input text. Each generated pose is matched against 32 candidate descriptions: its input text and one test description from each of 31 distinct distractor handshapes sampled from the full inventory. Candidates are ranked by cosine similarity in the shared embedding space. R@1 and R@5 report the percentage of generated poses for which the input text appears among the top one or five candidates, respectively.

FID measures how closely the generated pose distribution matches the reference distribution. We compute the Fréchet distance between the feature distributions of the 1,500 generated poses and their 1,500 paired reference poses. Lower FID indicates closer generated and reference distributions in this hand-pose feature space.

D.2 Generation Comparisons and Failure Cases

Description	GT	PoseScript	PoseFix	UniPose	ChatPose	Ours
C1 Three joined finger lines accompany a noncontacting two-level index and thumb outline.						
C2 Use the spread index-middle pair as the top and the extended thumb as the lower beak edge.						
C3 One raised thumb stands next to four straight joined fingers bent forward as a unit.						
C4 Fully extended index-little horns and a narrow little-ring side surround the complete middle-ring curl with an obliquely bent thumb between both central fingers.						
C5 Below three splayed extended shafts, an inward thumb meets the distal DIP region of the little finger fully curled into the palm.						
C6 Four evenly spaced fingers in a rounded palmward surface and a gently extended thumb below the index outline a distinct unclosed opening.						
C7 The index, middle, and thumb fingertips gather tightly in front of ring and little digits curled evenly into the palm.						
C8 The oblique thumb tip comes up from below directly against the descending middle fingertip, with the other three fingers straight and separated.						

Figure 14: Selected comparisons on eight unseen handshapes. Columns show the input description, its paired reference pose (GT), PoseScript, PoseFix, UniPose, ChatPose, and Ours. Every method uses its first saved output for the exact displayed description.

Figure 15 shows difficult descriptions for five unseen handshapes. F1–F2 contain incorrect finger configurations; F3–F4 fail to reproduce the specified thumb–middle arrangement; F5 does not recover the middle-over-index crossing. These cases show that accurately generating finger states and spatial relations for unseen handshapes remains challenging. Expanding training to cover more handshapes may improve the accuracy of text-conditioned generation.

Description	GT	Ours
F1 The near-neutral index root supports a distal hook apart from the thumb's contact with little. The thumb contact belongs to the curled little finger and remains separate from the index hook. The thumb contacts the distal part of the curled little finger.		
F2 The index and nearby little finger form two straight horns, the middle and ring fingers curl together, and an oblique thumb hooks between the middle and ring fingers. The little-finger horn remains closer to the center than the index horn.		
F3 The obliquely elevated thumb ends in contact over the nail-side tip region of the middle finger. A straight extended index remains separate from the contact. The middle finger curves down toward the hand. Deep palmward flexion closes the ring and little fingers. Strong distal curvature exposes the middle-finger nail region to the thumb.		
F4 Deep palmward bending fully closes both the ring and little fingers. The middle finger rotates palmward from its base. The index provides the single extended finger projection. Limited distal flexion guides the middle fingertip toward the thumb. A rising thumb closes the remaining distance to the middle fingertip. The index remains separate from the middle-thumb contact.		
F5 The hand combines a tucked thumb-ring-little group with a middle-over-index cross in which the middle palmar surface meets the back of the straight index. The middle finger remains above the index throughout the crossing.		

Figure 15: Failure cases, arranged as description, paired GT, and Ours.

D.3 SCREENING MANO POSE ANNOTATIONS

This section details the measurements and case selection used for Fig. 5. The experiment uses anatomical angles and rotation residuals to rank existing MANO annotations for inspection. Each original pose is decoded with its recorded shape parameters, and joint measurements and visualizations use the same decoded mesh.

Reference angle intervals. Table 11 lists reference intervals for angular screening. Flexion–extension intervals follow the theoretical ranges reported by Ngeo et al. (2014). Since that study does not report MCP spread ranges, we estimate these intervals from the 1st–99th percentiles in FreiHAND and DexYCB. These intervals identify unusual joint angles for further inspection.

Table 11: Reference angle intervals for annotation screening, in degrees.

Finger	MCP spread	MCP flexion	PIP flexion	DIP flexion
Index	$[-20, 40]$	$[-45, 90]$	$[0, 100]$	$[0, 80]$
Middle	$[-15, 35]$	$[-45, 90]$	$[0, 100]$	$[0, 80]$
Ring	$[-20, 25]$	$[-45, 90]$	$[0, 100]$	$[0, 80]$
Little	$[-55, 20]$	$[-45, 90]$	$[0, 100]$	$[0, 80]$

MCP direction and flexion-range violations. For an anatomical angle α and its reference interval $[L, U]$, the angular violation is

$$v(\alpha; L, U) = [L - \alpha]_+ + [\alpha - U]_+, \qquad [x]_+ = \max(x, 0). \tag{31}$$

The score is zero within the interval and equals the distance to its nearest endpoint outside it. We apply this score to the non-thumb MCP spread and flexion dimensions, and to PIP and DIP flexion. Angles are measured using the definitions in Sec. 3.2.

Deviation from a hinge plane. For a non-thumb PIP or DIP joint j, let $\mathbf{a}_{f,j}^{\text{hinge}}$ be the unit hinge axis and $\mathbf{d}_{f,j}^{\text{ch}}$ the unit child-bone direction defined in Sec. 3.2. The angular deviation from the hinge plane is

$$e_{f,j}^{\text{plane}} = \arcsin \left| (\mathbf{d}_{f,j}^{\text{ch}})^\top \mathbf{a}_{f,j}^{\text{hinge}} \right|. \tag{32}$$

We report this angle in degrees. It measures lateral departure from the expected bending plane, which can remain large even when the projected flexion angle falls inside its reference interval. The plotted cases use PIP deviations; the extended HO3D and InterHand2.6M audit also computes the corresponding DIP scores.

Thumb-axis residual. We remove twist from the effective local thumb MCP and IP rotations and project each swing rotation onto the closest rotation about its fixed FreiHAND-calibrated axis from Appendix A.2. The diagnostic score is the geodesic residual between the swing rotation and this projection, computed with the same geodesic distance as in Appendix A.3. The two thumb joints are scored separately.

Case selection. We rank poses by their joint-wise angular violations and residuals, then select high-scoring cases with visually clear deviations for Fig. 5. Since the datasets do not provide anomaly labels, these examples illustrate the diagnostic value of the measurements and do not estimate the prevalence of annotation errors.