HAND-Ω: SCALING TRAINING DATA FOR 3D HAND POSE ESTIMATION VIA GEOMETRY-PRESERVING IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The performance and generalization of monocular 3D hand pose estimation remain constrained by the limited scale and diversity of existing training data. Although graphics-based synthesis provides accurate geometric annotations, conventional rendering pipelines rely on limited HDR backgrounds and predominantly centered-hand configurations, restricting both visual and perspective diversity. We introduce Hand-Ω, a large-scale synthetic hand dataset with substantial diversity in hand poses, backgrounds, hand–object interactions, hand locations, and camera configurations. We construct the dataset through a geometry-preserving rendering-and-generation pipeline. Starting from hand poses obtained from existing datasets and generative pose models, we render textured meshes under diverse virtual camera settings. Guided by structured prompts, a large image-generation model transforms these renderings into photorealistic images with diverse appearances while preserving the underlying hand geometry. We further preserve geometric fidelity through rendering-based appearance control, geometry-constrained prompting, and geometry-consistency filtering. The resulting dataset comprises approximately 10.5 million images spanning isolated hands, hand–object interactions, third-person views, and egocentric scenarios, with accurate MANO, keypoint, mesh, and camera annotations. We further develop WiLoR† by introducing camera-conditioned inputs and full-image perspective projection with the actual camera intrinsics. Experiments show consistent performance gains across benchmarks as the scale of Hand-Ω increases. Trained with the full Hand-Ω dataset, WiLoR† achieves state-of-the-art performance on both FreiHAND and HO-3D.

1 INTRODUCTION

Recovering 3D hand poses from monocular images has recently attracted considerable attention because of its crucial role in providing useful action signals for pre-training embodied models, such as Vision-Language-Action (VLA) models Zheng et al. (2026); Li et al. (2025) and World Action Models (WAM) Dyna Robotics (2026); Agarwal et al. (2026). Despite remarkable progress in monocular 3D hand pose estimation, existing training datasets Banerjee et al. (2025); Fan et al. (2023); Chao et al. (2021); Zimmermann et al. (2019); Hampali et al. (2020) are still largely collected under controlled or instrumented settings. Consequently, these datasets typically provide limited scene diversity and pose diversity, while their collection processes remain costly and difficult to scale. Such limitations in training data constrain the capabilities of existing hand pose estimation models, hindering their generalization to a broader range of scenes and hand poses.

Several recent studies Si et al. (2026); Zhao et al. (2025); Moon et al. (2023) have sought to scale up hand pose estimation datasets through rendering-based image synthesis. A representative method, AnyHand Si et al. (2026), uses a rendering engine to render textured hand meshes from parametric hand models and composite them with background images, thereby producing synthetic data with accurate hand pose annotations. Such synthesis-based approaches show considerable potential because they enable datasets to be constructed at scale. Nevertheless, they suffer from two major limitations. First, their reliance on conventional rendering pipelines with a limited set of HDR backgrounds restricts the visual diversity of the synthesized images. Second, they typically adopt a

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

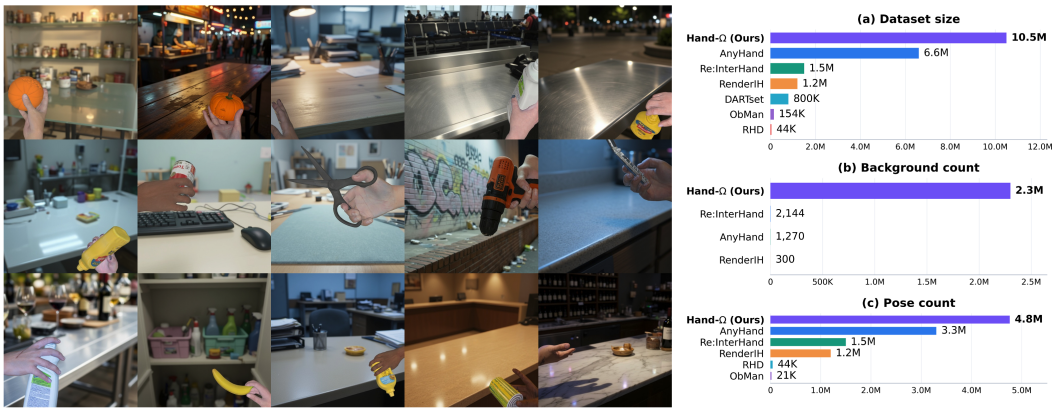

Figure 1: **Overview of Hand-Ω.** Our dataset provides large-scale and geometrically accurate hand images with diverse hand locations, camera perspectives, appearances, and scenes.

fixed rendering configuration in which the hand is centered in the image without varying perspective effects depending on hand positions. Inspired by the remarkable capabilities of recent large-scale image generation models OpenAI (2026); Google (2025); Chen et al. (2026a), we propose a hybrid rendering-and-generation approach that combines the precise pose controllability of graphics-based rendering with the photorealistic synthesis capabilities of large generative models. This approach aims to improve the visual diversity of synthetic training data for enhancing the performance of hand pose estimation models.

In this work, we introduce Hand-Ω, a large-scale synthetic dataset for hand pose estimation that features high visual diversity and substantial diversity in hand locations and perspective effects. First, we develop a scalable hybrid synthesis framework that integrates graphics-based rendering with image generation models, scaling visual diversity while preserving geometric fidelity. Leveraging advanced large-scale image generation models, we scale environmental and photographic diversity through billion-scale compositional text prompting. To preserve geometric fidelity, we introduce multiple geometry-stable strategies, including geometry-stable appearance diversification via rendering, geometry-constrained hand–object prompting, and geometry-consistency filtering. Second, unlike existing methods that predominantly synthesize hands at the image center, we introduce perspective-diverse rendering that explicitly varies hand locations, camera focal lengths, scales, and truncations. Finally, building on the state-of-the-art model of WiLoR Potamias et al. (2025), we develop WiLoR† by introducing camera-conditioned inputs and projection with the actual camera intrinsics. We conduct extensive data ablations under mixed training with Hand-Ω and existing datasets, demonstrating the effectiveness of Hand-Ω for improving hand pose reconstruction.

By combining existing training datasets with Hand-Ω synthetic data, WiLoR† achieves state-of-the-art performance on both FreiHAND Zimmermann et al. (2019) and HO-3D Hampali et al. (2020), outperforming both the original WiLoR and its AnyHand-augmented counterpart. Compared with the original WiLoR, PA-MPJPE is reduced from 5.50 to 5.36 mm on FreiHAND and from 7.50 to 7.27 mm on HO-3D. Ablation studies across different dataset scales further show consistent gains as the amount of synthetic data increases, confirming the scalability and effectiveness of Hand-Ω.

In summary, we introduce Hand-Ω, a large-scale synthetic dataset designed to improve the performance of 3D hand pose estimation models. Our main contributions are threefold:

- We propose a scalable hybrid synthesis framework that integrates graphics-based rendering with large-scale image generation models, combining accurate geometric control with diverse visual synthesis. Our framework further introduces perspective-diverse rendering by explicitly varying hand locations and camera focal lengths.
- We build WiLoR† upon the state-of-the-art WiLoR by introducing camera-conditioned inputs and projection with the actual camera intrinsics. Based on WiLoR†, we conduct extensive data ablations to validate the effectiveness of Hand-Ω.
- WiLoR† trained with Hand-Ω establishes new state-of-the-art performance across multiple benchmarks.

Table 1: **Comparison with representative synthetic hand-pose datasets.** "Render+Gen." denotes graphics rendering followed by geometry-preserving image generation. "–" means that it is not emphasized as such in the original dataset.

Dataset	Size	Appearance pipeline	Background	Hand setting	Object	Focal diversity	Position / truncation
RHD Zimmermann & Brox (2017)	44K	Rendering	Static	Single hand	–	–	–
ObMan Hasson et al. (2019)	154K	Rendering	Static	Hand–object	✓	–	–
DARTset Gao et al. (2022b)	800K	Rendering	Static	Single hand	–	–	–
Re:InterHand Moon et al. (2023)	1.5M	Rendering	HDR indoor	Two hands	–	–	–
RenderIH Li et al. (2023)	1.2M	Rendering	HDR indoor/outdoor	Two hands	–	–	–
HandSynthesis Zhao et al. (2025)	460K	Rendering	HDR indoor/outdoor	Single / hand–object	✓	–	–
AnyHand Si et al. (2026)	6.6M	Rendering	HDR indoor/outdoor	Single / hand–object	✓	✓	–
Hand-Ω (Ours)	**10.5M**	**Render+Gen.**	**Generated, diverse**	**Single / hand–object**	✓	✓	✓

2 Related Work

Monocular 3D Hand Reconstruction. Early 3D hand pose estimation evolved from depth-based methods Ge et al. (2016); Kyriazis & Oikonomidis (2011); Tagliasacchi et al. (2015) to monocular RGB reconstruction that directly regresses 3D hand parameters Boukhayma et al. (2019); Romero et al. (2022). Subsequent works explored heatmap supervision and iterative refinement Zhang et al. (2019); Baek et al. (2019), direct mesh regression Kulon et al. (2019; 2020), and robustness to occlusion and motion blur Park et al. (2022); Oh et al. (2023). More recently, HaMeR Pavlakos et al. (2024) shows strong generalization by scaling model capacity and heterogeneous training data, Hamba Dong et al. (2024) models spatial dependencies over the hand mesh, and WiLoR Potamias et al. (2025) combines efficient hand localization with mesh-aligned refinement. Our work complements these model advances by scaling geometrically accurate training data, deliberately diversifying its cameras and layouts, and adapting a WiLoR-style estimator to exploit the benefit of data.

Synthetic Data for Hand Pose Estimation. Synthetic data provides exact 3D labels without costly capture and annotation. RHD Zimmermann & Brox (2017) established rendered supervision for single-hand estimation, while ObMan Hasson et al. (2019) jointly rendered hands and objects to model interaction and occlusion. DARTset Gao et al. (2022b) increased appearance diversity through textures and accessories. For interacting hands, Re:InterHand Moon et al. (2023) relights captured geometry, while RenderIH Li et al. (2023) scales interacting-hand data through pose optimization and rendering. HandSynthesis Zhao et al. (2025) analyzes factors affecting synthetic-to-real transfer, and AnyHand Si et al. (2026) further scales simulation-native data to millions of single-hand and hand–object samples. Despite this progress, these datasets remain primarily based on conventional rendering or compositing, whose image statistics may differ from unconstrained photographs. Moreover, focal length, off-center placement, hand scale, and boundary truncation are rarely modeled jointly as explicit dimensions of dataset diversity (Tab. 1).

Generative Hand Image Synthesis. Generative models offer another route to realism by translating geometric controls into natural images. GANerated Hands Mueller et al. (2018) pioneered image-to-image translation from rendered hands to realistic appearances. More recent diffusion-based methods improve controllability and diversity: HandBooster Xu et al. (2024) synthesizes challenging hand–object examples, HOIDiffusion Zhang et al. (2024b) disentangles 3D interaction structure from appearance, and AttentionHand Park et al. (2024) conditions generation on geometric and textual cues. FoundHand Chen et al. (2025) trains a hand-specific diffusion model at large scale, while SesaHand Zhao et al. (2026) improves semantic and structural alignment for 3D reconstruction. Unlike these works, which mainly focus on generative models themselves, Hand-Ω targets large-scale dataset construction while leveraging off-the-shelf advanced large image generation models.

3 Method

3.1 Overview

Our goal is to construct a large-scale synthetic training set that simultaneously provides broad visual diversity and accurate 3D hand geometry. Hand-Ω is built around three key principles. First, we explicitly model **perspective diversity** by varying hand location, focal length, scale, and truncation during rendering, rather than restricting hands to the image center. Second, we **decouple geometry from appearance**: a graphics renderer determines the hand pose, hand–object configuration, camera geometry, and foreground appearance, while a large image-generation model is used to diversify

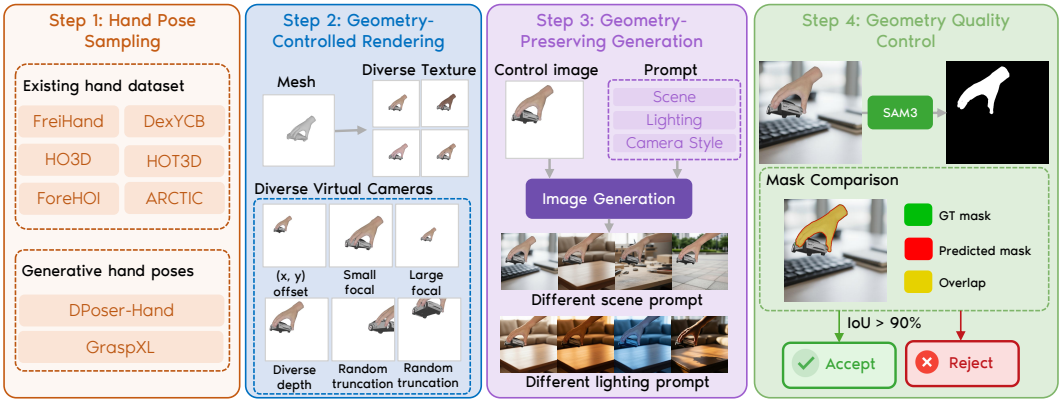

Figure 2: **Overview of the Hand-Ω synthesis pipeline.** Starting from diverse pose sources, we first perform geometry-controlled rendering with diverse textures and cameras. The rendered control image is then combined with structured prompts to generate diverse images. Finally, a geometry quality-control module filters out samples that fail geometric consistency.

the surrounding scene, illumination, and photographic appearance. Third, we enforce **geometry-preserving generation** by constraining the rendered foreground during image synthesis and filtering generated images that deviate from the reference geometry. In essence, Hand-Ω explicitly controls what must remain geometrically correct while aggressively scaling what can vary in appearance.

As illustrated in Fig. 2, Hand-Ω integrates diverse pose sources covering isolated hands and hand–object interactions, drawn from existing datasets including FreiHAND Zimmermann et al. (2019), HO-3D Hampali et al. (2020), DexYCB Chao et al. (2021), ARCTIC Fan et al. (2023), HOT3D Banerjee et al. (2025), and ForeHOI Chen et al. (2026b), as well as generative pose models including GraspXL Zhang et al. (2024a) and DPoser-Hand Lu et al. (2025). The detailed composition and numbers are provided in Appendix A.1. Given MANO pose and shape parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and camera intrinsics $\mathbf{K}$, the rendering stage produces a geometrically accurate control image together with consistent geometric supervision. When an object mesh is available, its camera-space geometry is rendered jointly with the hand to preserve contact and occlusion relationships. The generation stage then transforms the rendered control into a realistic image while preserving the foreground geometry, followed by geometry-aware quality control. This pipeline produces approximately 10.5 million synthetic images in total.

3.2 Perspective-Diverse Geometry-Controlled Rendering

Perspective differences beyond centered-hand rendering. Existing synthetic pipelines often render a hand at the fixed image center, such as AnyHand Si et al. (2026). Although methods such as HaMeR and WiLoR operate on cropped hand images, this does not mean that synthetic training data only need to render hands at the image center. Due to perspective projection, hands located at different positions in the original image can exhibit noticeably different appearances after cropping, even when they share exactly the same underlying 3D hand MANO parameters. This effect becomes particularly pronounced under short focal lengths. As illustrated in Fig. 3, we render the same hand configuration at the image center and near the image boundary using a focal length of 300. Specifically, we keep the MANO pose parameters and global hand rotation identical, and vary only the horizontal hand position in the image. Despite having identical MANO

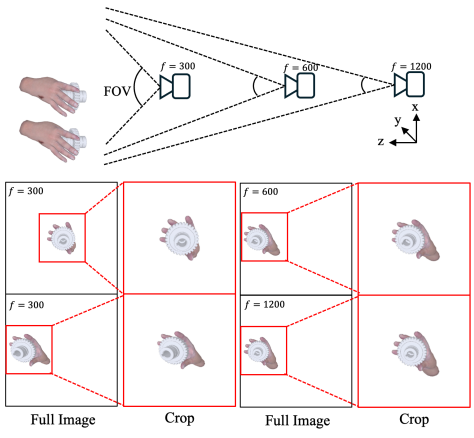

Figure 3: **Perspective differences between centered and off-center hands.**

parameters, the resulting hand crops exhibit clear perspective-induced differences. This observation highlights the importance of explicitly diversifying hand locations during synthetic data generation, rather than restricting rendered hands to the image center.

Perspective-diverse camera sampling. Motivated by the above findings on perspective effects, we render each sampled pose with varying hand positions and camera focal lengths. Let W and H be the image width and height, L the spatial extent of the hand mesh, f the sampled focal length, and r the target hand size as a fraction of $\min(W, H)$. We sample f log-uniformly and r uniformly, and choose the hand depth as

$$z = \mathrm{clip}\left(\frac{fL}{r\min(W,H)}, z_{\min}, z_{\max}\right). \tag{1}$$

This construction approximately preserves the requested image-space hand size while allowing focal length and depth to vary independently. We then sample a target image location and solve for an x–y translation that projects the hand centroid to that location. The admissible range is defined using the projected hand bounding box rather than the image dimensions, which enables both fully visible hands and controlled truncation near image boundaries. Samples with invalid depth or insufficient visible hand vertices are resampled. In our implementation, $f \in [300, 1200]$ pixels, $r \in [0.2, 0.7]$, and $z \in [0.2, 3.0]$ meters; up to half of the projected bounding box may extend beyond an image boundary, while at least 35% of hand vertices must remain visible. The final intrinsic matrix is

$$\mathbf{K} = \begin{bmatrix} f & 0 & W/2 \\ 0 & f & H/2 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

and is stored with the generated sample. This process substantially broadens the distributions of image location, truncation, apparent scale, perspective distortion, and focal length while retaining exact 3D labels.

Geometry-stable appearance diversification via rendering. To diversify hand appearance, we compare two alternatives: relying on text prompts to control hand texture during image generation or introducing texture variation during rendering. We find that the former leads to a higher probability of geometric instability in the experiments, as shown in Fig. 5. We therefore adopt the latter and use the DART Gao et al. (2022a) hand texture library to provide diverse hand textures while preserving geometric consistency. The texture pool contains 284 complete UV atlases: 270 skin textures, covering three skin-tone groups and five appearance categories (plain skin, moles, nail variations, scars, and tattoos), together with 14 glove textures. Following DART, we extend MANO with a short wrist mesh. For each sample, we randomly select a hand texture from this pool and render the textured hand mesh as the RGB control image. For hand-object interaction samples, the object mesh and its texture are rendered jointly with the hand to preserve occlusion and contact cues. We additionally render a binary reference mask of the visible hand region, which is later used for quality-control filtering in the image generation stage.

3.3 Scalable Geometry-Preserving Image Generation

Given the rendered control image $\mathbf{C}$, an image-generation model synthesizes

$$\tilde{\mathbf{I}} = G(\mathbf{C}, \mathbf{P}; \boldsymbol{\epsilon}), \tag{3}$$

where $\mathbf{P}$ is a structured text instruction and $\boldsymbol{\epsilon}$ controls stochastic variation. The instruction explicitly treats the rendered hand geometry as a hard constraint: the generator may change background, illumination, shadows, and image-capture style, but must preserve hand parts.

Scalable generation with low cost and latency. We compare GPT-Image-2 OpenAI (2026), Nano Banana Pro Google (2025), and the open-source SOTA model Boogu-Image Chen et al. (2026a) as candidate generators. While GPT-Image-2 and Nano Banana Pro deliver strong visual quality, their high inference cost and generation latency make million-scale synthesis impractical. We therefore adopt Boogu-Image, which provides substantially higher throughput at much lower cost. Importantly, our renderer-guided pipeline explicitly constrains hand geometry, hand-object interactions, and foreground appearance, thereby reducing the burden on the image generator. Under this structured synthesis framework, Boogu-Image achieves generation quality close to GPT-Image-2 while enabling efficient large-scale data generation.

Billion-scale structured prompt combinations. Previous rendering-only synthetic pipelines are limited by the size and diversity of manually prepared scene assets. In contrast, our rendering-generation framework can leverage diverse text prompts to synthesize a substantially broader range of scenes, illumination conditions, and photographic effects, enabling scalable appearance diversification beyond a fixed asset library. We construct each generation instruction by sampling from a structured, factorized prompt space rather than maintaining a fixed list of complete prompts:

- **Scene and surface.** We define 300 places from 21 semantic categories, including home, office, retail, transportation, outdoor, industrial, medical, sports, and entertainment environments. Each place is paired only with category-compatible surfaces. The 55 material or ground types produce approximately 3.0K valid place–surface pairs.
- **Background attributes.** We sample four levels of background complexity, four depth-of-field settings, seven color palettes, and seven times of day. Combined with the place–surface pairs, these dimensions provide approximately 2.3M background configurations.
- **Imaging and challenge conditions.** Each background is combined with six lighting conditions, four camera styles, and 23 observation conditions. The effects include defocus and motion blur, JPEG artifacts, low-light noise, low contrast, hard shadows, reflections, distractors, small hand scale, edge composition, and partial shadow occlusion. Together, these dimensions provide approximately 1.3B structured prompt combinations.

Geometry-constrained hand-object prompting. To preserve the hand geometry of the rendered control image, we adopt a strict prompting strategy that explicitly instructs the generation model to preserve the rendered hand and object regions and prohibits regenerating either component. It further enumerates the geometric attributes that must remain unchanged: hand pose and handedness, finger order and fingertip locations, wrist position, object shape, scale, position and orientation, and, critically, the hand–object contact and front–back occlusion relationships. Only parts outside the rendered hand–object foreground may be edited to synthesize the scene, lighting, and camera appearance. This asymmetric prompting design decouples open-ended environmental synthesis from geometry-preserved rendering, preserving accurate hand-object geometry while retaining the scalability and visual diversity of a large image-generation model.

Geometry-consistency filtering. Image generation can occasionally alter finger silhouettes, move the hand, or create additional hands. We therefore validate each generated image against its rendered geometry before adding it to Hand-Ω. We use the SOTA promptable segmentation model SAM3 Carion et al. (2026) to extract a hand mask $\hat{\mathbf{M}}$ from $\tilde{\mathbf{I}}$. We compare it with the rendered reference mask $\mathbf{M}$ using

$$\mathrm{IoU}(\mathbf{M}, \hat{\mathbf{M}}) = \frac{|\mathbf{M} \cap \hat{\mathbf{M}}|}{|\mathbf{M} \cup \hat{\mathbf{M}}|}. \tag{4}$$

We accept a sample only if its mask overlap exceeds a specified threshold. Specifically, we set the IoU threshold to 0.9. Finally, the accepted images and their inherited annotations are included into training datasets.

3.4 Training with Hand-Ω

To validate the effectiveness of our synthetic Hand-Ω dataset, we train and evaluate models based on WiLoR, a state-of-the-art method for monocular 3D hand pose estimation. WiLoR takes cropped hand images as input, adopts a ViT-H backbone, and appends additional MANO tokens to the ViT input sequence to regress the MANO pose $\hat{\boldsymbol{\theta}}$, shape $\hat{\boldsymbol{\beta}}$, and hand position parameters $\hat{\mathbf{c}} = (\hat{s}, \hat{t}_x, \hat{t}_y)$.

Camera conditioning. A key distinction of Hand-Ω from previous synthetic hand datasets is that we explicitly diversify both hand locations and camera focal lengths during synthesis. As shown in Fig. 3, identical MANO poses rendered at different image locations can produce substantially different cropped appearances due to perspective effects. Such variations are largely absent from previous synthetic datasets, where hands are predominantly rendered near the image center.

To exploit this perspective diversity, the reconstruction model must retain the camera context lost during hand cropping. Following prior camera-aware reconstruction methods Patel & Black (2025); Li et al. (2022), we augment WiLoR with a box token encoding the hand location and scale relative

Table 2: **Comparison with the state of the art on FreiHAND.** PA-MPJPE and PA-MPVPE are reported in mm. Best and second-best results are highlighted in bold and underlined, respectively.

Method	PA-MPJPE ↓	PA-MPVPE ↓	F@5 ↑	F@15 ↑
AMVUR Jiang et al. (2023)	6.20	6.10	0.767	0.987
HaMeR Pavlakos et al. (2024)	6.00	5.70	0.785	0.990
Hamba Dong et al. (2024)	5.70	5.30	0.806	0.992
WiLoR Potamias et al. (2025)	5.50	5.10	0.825	<u>0.993</u>
AnyHand-WiLoR Si et al. (2026)	<u>5.39</u>	<u>5.05</u>	<u>0.827</u>	**0.994**
WiLoR† w/ Hand-Ω (Ours)	**5.36**	**5.01**	**0.828**	**0.994**

to the camera intrinsics. Given the full-image box center (c_x, c_y), box size b, image dimensions (W, H), and focal length f, the token is defined as

$$\mathbf{b} = \left[\frac{c_x - W/2}{f}, \frac{c_y - H/2}{f}, \frac{b}{f}\right]. \tag{5}$$

After linear embedding, the box token is prepended to the MANO regression and image tokens, providing camera context that cannot be reliably recovered from a normalized crop alone.

Perspective-consistent projection. The original WiLoR uses a fixed, large virtual focal length for all samples, and thus does not explicitly model sample-specific perspective variations. In contrast, we project the predicted 3D joints in the full image using the per-sample focal length provided by Hand-Ω, enabling perspective-consistent supervision and hand localization under the corresponding camera geometry. We refer to WiLoR equipped with the camera-aware box token and perspective-consistent projection as WiLoR†.

Training objective. The total reconstruction objective is

$$\mathcal{L} = \lambda_{2D}\mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_\theta\mathcal{L}_\theta + \lambda_\beta\mathcal{L}_\beta, \tag{6}$$

where $\mathcal{L}_{2D}$ is a confidence-weighted ℓ_1 loss on projected joints, $\mathcal{L}_{3D}$ is a wrist-aligned ℓ_1 loss on 3D joints, and $\mathcal{L}_\theta$ and $\mathcal{L}_\beta$ supervise the MANO pose and shape, respectively.

4 Experiments

Training configuration. We use the original WiLoR training data as our base training set, which contains approximately 4.2M images. Each training batch is constructed by sampling from different datasets according to predefined sampling weights. When jointly training with Hand-Ω, we follow the sampling strategy of Si et al. (2026) and set the sampling probabilities of the base training set and Hand-Ω to 0.6 and 0.4, respectively.

4.1 Comparison with the State of the Art Models

We evaluate our model on the FreiHAND Zimmermann et al. (2019) and HO-3D v2 Hampali et al. (2020) benchmarks. Tables 2 and 3 compare WiLoR† trained with Hand-Ω against recent state-of-the-art methods. On FreiHAND, WiLoR† w/ Hand-Ω achieves a PA-MPJPE of 5.36 mm, outperforming both the original WiLoR and its AnyHand-trained version. Similar improvements are observed on HO-3D v2, where our model obtains the best AUC_J, PA-MPJPE, AUC_V, PA-MPVPE, and F@5 scores among the compared methods. In particular, compared with AnyHand-WiLoR, Hand-Ω reduces PA-MPJPE from 7.36 mm to 7.27 mm, while improving F@5 from 0.649 to 0.663. These consistent gains across both benchmarks indicate that Hand-Ω provides consistent benefits for monocular 3D hand reconstruction across different evaluation benchmarks.

4.2 Qualitative Comparison

We qualitatively compare our model with the original WiLoR and WiLoR trained with AnyHand on the HO-3D test set, as shown in Fig. 4, with all reconstructions projected using the real camera focal length. WiLoR and AnyHand-WiLoR are inferred using their respective official virtual focal

Table 3: **Comparison with the state of the art on HO-3D v2.** PA-MPJPE and PA-MPVPE are reported in mm. Best and second-best results are highlighted in bold and underlined, respectively.

Method	$AUC_J \uparrow$	PA-MPJPE ↓	$AUC_V \uparrow$	PA-MPVPE ↓	F@5 ↑	F@15 ↑
AMVUR Jiang et al. (2023)	0.835	8.30	0.836	8.20	0.608	0.965
HaMeR Pavlakos et al. (2024)	0.846	7.70	0.841	7.90	0.635	0.980
Hamba Dong et al. (2024)	0.850	7.50	0.843	7.70	0.648	0.982
WiLoR Potamias et al. (2025)	0.851	7.50	0.846	7.70	0.646	<u>0.983</u>
AnyHand-WiLoR Si et al. (2026)	<u>0.853</u>	<u>7.36</u>	<u>0.848</u>	<u>7.62</u>	<u>0.649</u>	**0.984**
WiLoR† w/ Hand-Ω (Ours)	**0.855**	**7.27**	**0.849**	**7.53**	**0.663**	<u>0.983</u>

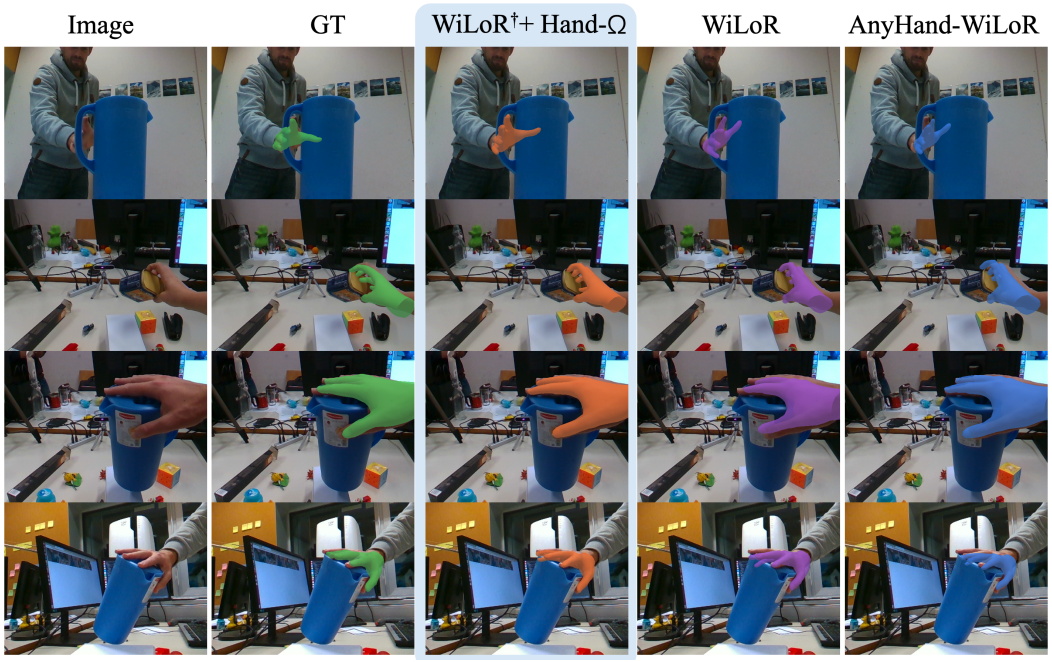

Figure 4: **Qualitative comparison.** Visual comparison with WiLoR and WiLoR trained with AnyHand on the HO3D testset.

lengths. We then use PnP to transform their predictions into the 3D coordinates corresponding to the real camera focal length. In the first row, our model reconstructs the hand more accurately under severe object occlusion than both WiLoR and AnyHand-WiLoR. The second row shows that our method achieves more accurate global hand rotation estimation. Both WiLoR and AnyHand-WiLoR exhibit noticeable errors in wrist orientation. In the third example, our reconstruction closely matches the scale of the observed hand. Finally, the last row shows that our method also achieves more accurate finger alignment. Overall, these qualitative results show that our model performs better in challenging cases.

4.3 Ablation Study

The effect of random hand positions. To evaluate the performance gains brought by increased diversity in hand locations, we conduct an ablation study using only the centered-hand samples from our synthetic dataset while keeping all other training settings unchanged. Specifically, we exclude samples generated with positional offsets, resulting in a 2.5M-sample subset in which the hands are predominantly centered in the image. As shown in the results, training exclusively on these centered-hand samples leads to a substantial performance degradation on the HO3D benchmark, where hand locations exhibit considerably greater spatial variation. In contrast, the impact on FreiHAND is relatively limited, as hands in its test set are largely centered. These results highlight the importance of diversifying hand locations during synthetic data generation. Existing synthetic hand datasets, such as AnyHand, predominantly place hands at the image center and therefore do not explicitly account for this source of spatial diversity.

Table 4: **Ablation study on data configurations, including hand-position sampling, data scale and pose source.** PA-MPJPE and PA-MPVPE are reported in millimeters.

		FreiHAND				HO-3D v2			
Ablation	**Data config**	**PA-MPJPE ↓**	**PA-MPVPE ↓**	**F@5 ↑**	**F@15 ↑**	**PA-MPJPE ↓**	**PA-MPVPE ↓**	**F@5 ↑**	**F@15 ↑**
Hand position	Centered only	5.44	5.10	0.824	**0.994**	7.66	7.93	0.642	0.980
Pose source	w/o Gen. poses	5.47	5.21	0.816	0.993	7.36	7.62	0.658	**0.983**
Data scale	Hand-Ω 3M	5.55	5.25	0.814	0.992	7.73	8.01	0.636	0.980
	Hand-Ω 6M	5.48	5.19	0.817	0.993	7.33	7.58	0.660	**0.983**
	Hand-Ω 10.5M	**5.36**	**5.01**	**0.828**	**0.994**	**7.27**	**7.53**	**0.663**	**0.983**

Impact of the data scale. To further investigate the effect of synthetic data scale on hand reconstruction accuracy, we compare models trained on randomly sampled subsets of 3M and 6M examples from the full synthetic dataset, while keeping all other training settings unchanged. Increasing the training data from 3M to 6M samples, and further from 6M to the full 10.5M dataset, consistently improves performance across all benchmarks. These results demonstrate the effectiveness of scaling up synthetic training data. A larger dataset provides broader coverage of hand poses, hand appearances and textures, backgrounds, and illumination conditions, thereby improving the diversity of the training distribution and the generalization capability of the resulting model. Notably, Hand-Ω contains 10.5M synthetic samples, substantially exceeding the scale of prior synthetic hand datasets.

Ablation of the pose source. To investigate the impact of different hand pose sources on the effectiveness of the synthetic dataset, we remove the samples generated from generative pose sources (i.e., DPoser-Hand and GraspXL), retaining only the samples constructed from poses sampled from existing datasets. We then train the model under the same training settings and evaluate its performance. The results show a clear performance degradation when the generative pose sources are excluded. This suggests that generative pose sources can complement the pose distributions covered by existing datasets by introducing additional hand configurations, thereby improving the diversity of the training data and the overall performance of the model.

Ablation of hand texture control method. We compare two strategies for introducing hand-appearance diversity. The first directly specifies skin tone and local appearance attributes through the generation prompt. The second first applies the desired appearance using a DART texture atlas during mesh rendering and then instructs the image generator to preserve the rendered hand while synthesizing only the surrounding scene. As illustrated in Fig. 5, prompt-based appearance control is more likely to disrupt the underlying hand geometry. In contrast, rendering the texture into the control image provides a direct pixel-level appearance constraint and yields substantially more consistent hand geometry across texture variations. We quantify this effect on 320 images for each method. Rendering-based texture control improves the mean mask IoU from 76.0 to 94.4.

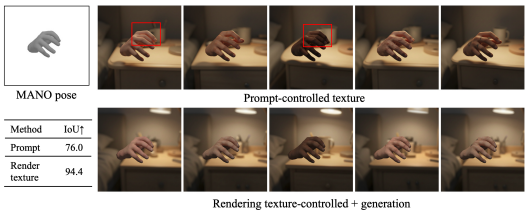

Figure 5: **Comparison of hand texture control methods.**

5 Conclusions

We introduce Hand-Ω, a large-scale synthetic dataset designed for monocular 3D hand reconstruction. Hand-Ω combines geometry-controlled rendering with image generation to preserve accurate hand geometry while increasing the diversity and realism of training data. By explicitly covering variations in hand location, camera focal length, object occlusion, and appearance, our pipeline addresses key limitations of existing synthetic datasets. Building on WiLoR, we further develop WiLoR† with camera-conditioned inputs and projection using the actual camera intrinsics, and use it to systematically evaluate the contribution of Hand-Ω through extensive data ablations. When jointly trained on existing datasets and Hand-Ω, WiLoR† achieves state-of-the-art performance on multiple benchmarks.

AI USE STATEMENT

Generative AI tools were used in this work for several purposes. First, large language models were used to assist with language polishing and drafting parts of the manuscript. Second, they were used to support literature retrieval and discovery of potentially relevant prior work; all cited references were subsequently checked by the authors. Third, generative AI was used as part of our data construction pipeline to generate synthetic training images, as described in Sec. 3.3 and the appendix. All AI-assisted text, retrieved literature, and generated data used in this work were reviewed and verified by the authors. The authors take full responsibility for the final content, claims, experiments, and artifacts presented in this paper.

ETHICS STATEMENT

This work focuses on constructing synthetic training data for monocular 3D hand pose estimation and does not involve the collection of new personally identifiable information or human-subject experiments. Hand-Ω is built from existing hand pose and texture resources together with generative image synthesis; we follow the applicable licenses and usage terms of the corresponding datasets, models, and assets. As with other large-scale generative models, the image generator may inherit biases in appearance, scene composition, or demographic representation from its training data.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the Hand-Ω data construction and training procedures to facilitate reproducibility. The appendix specifies the geometry and annotation processing, virtual-camera sampling procedure, structured image-generation protocol, geometry quality-control criteria, and complete WiLoR† training configuration. We also report the dataset mixing strategy and evaluation protocols used in all experiments.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1067–1076. IEEE, 2019.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071. IEEE, 2025.

Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10843–10852, 2019.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9040–9049. IEEE, 2021.

Guoxuan Chen, Chufeng Xiao, Haoran Yang, Siyue Xie, Binxiao Huang, Ming Zhang, Cheuk Him Chau, Xinyu Fu, Yingzhao Lian, Tom SY Li, et al. Boogu-image-0.1: Boosting open-source unified multimodal understanding and generation. *arXiv preprint arXiv:2607.13125*, 2026a.

Kefan Chen, Chaerin Min, Linguang Zhang, Shreyas Hampali, Cem Keskin, and Srinath Sridhar. Foundhand: Large-scale domain-specific learning for controllable hand image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17448–17460. IEEE, 2025.

Yuantao Chen, Jiahao Chang, Chongjie Ye, Chaoran Zhang, Zhaojie Fang, Chenghong Li, and Xiaoguang Han. Forehoi: Feed-forward 3d object reconstruction from daily hand-object interaction videos. *arXiv preprint arXiv:2602.06226*, 2026b.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

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

Dyna Robotics. Dyna-2: A 1-million-hour scaling law for world-action models, August 2026. URL https://dyna.co/dyna-2.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954. IEEE, 2023.

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, 2022a.

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. *NeurIPS*, 35:37055–37067, 2022b.

Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3593–3601, 2016.

Google. Introducing nano banana pro, 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pp. 3196–3206, 2020.

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

Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *CVPR*, pp. 758–767, 2023.

Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. *arXiv preprint arXiv:1905.01326*, 2019.

Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4990–5000, 2020.

Nicholas Kyriazis and Iason Oikonomidis. Efficient model-based 3d tracking of hand articulations using kinect. In *Procedings of the British Machine Vision Conference 2011*, 2011.

Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *ICCV*, pp. 20395–20405, 2023.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.

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

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

Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *NeurIPS*, 36:17689–17701, 2023.

Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 49–59. IEEE, 2018.

Yeonguk Oh, JoonKyu Park, Jaeha Kim, Gyeongsik Moon, and Kyoung Mu Lee. Recovering 3d hand mesh sequence from a single blurry image: A new dataset and temporal unfolding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 554–563. IEEE, 2023.

OpenAI. Introducing chatgpt images 2.0, 2026. URL `https://openai.com/index/introducing-chatgpt-images-2-0/`.

JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1486–1495. IEEE, 2022.

Junho Park, Kyeongbo Kong, and Suk-Ju Kang. Attentionhand: Text-driven controllable hand image generation for 3d hand reconstruction in the wild. In *European Conference on Computer Vision*, pp. 329–345. Springer, 2024.

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

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

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *CVPR*, pp. 12242–12254, 2025.

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

Chen Si, Yulin Liu, Bo Ai, Jianwen Xie, Rolandos Alexandros Potamias, Chuanxia Zheng, and Hao Su. Anyhand: A large-scale synthetic dataset for rgb (-d) hand pose estimation. *arXiv preprint arXiv:2603.25726*, 2026.

Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. In *Computer graphics forum*, volume 34, pp. 101–114. Wiley Online Library, 2015.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10159–10169. IEEE, 2024.

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024a.

Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8521–8531. IEEE, 2024b.

Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2354–2364. IEEE, 2019.

Zhuoran Zhao, Linlin Yang, Pengzhan Sun, Pan Hui, and Angela Yao. Analyzing the synthetic-to-real domain gap in 3d hand pose estimation. In *CVPR*, pp. 12255–12265, 2025.

Zhuoran Zhao, Xianghao Kong, Linlin Yang, Zheng Wei, Pan Hui, and Anyi Rao. Sesahand: Enhancing 3d hand reconstruction via controllable generation with semantic and structural alignment. *arXiv preprint arXiv:2603.00443*, 2026.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pp. 4903–4911, 2017.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pp. 813–822, 2019.

A Implementation Details of Data Synthesis

This appendix provides implementation details for the Hand-Ω data construction and WiLoR† training pipelines.

A.1 Pose Sources and Dataset Composition

Hand-Ω combines annotated poses from existing hand datasets with poses sampled from generative hand-pose models. Table 5 summarizes the composition of the final dataset used for training. The reported numbers are the generated images retained after quality control. GraspXL is a generative model for synthesizing hand grasp motions. GraspXL team already used the model with objects from Objaverse Deitke et al. (2023) to construct a large-scale grasp-pose database. In Hand-Ω, we directly use the pose database released by them rather than resampling poses with the model.

A.2 Geometry and Annotation Processing

We use PyTorch3D as the rendering engine. Control images are rendered at 512×512 pixels with $2\times$ supersampling and then downsampled using area interpolation. The same PyTorch3D rasterization pipeline produces both the textured RGB foreground and the binary hand reference mask. Hand–object samples additionally retain an object mask when it is available. We attach a short wrist extension to the MANO mesh and cap the open wrist boundary. This avoids presenting the image generator with an unnatural open mesh while leaving the annotated MANO hand unchanged. The reference mask is rendered from the same camera and geometry as the RGB control.

Table 5: **Pose sources and composition of Hand-Ω.** "Dataset" denotes poses obtained from an existing annotated dataset, while "generative" denotes poses sampled from a learned pose model. The image counts refer to the final quality-controlled training images and sum to 10,485,768 (approximately 10.5M).

Pose source	Source type	Hand setting	Images	Fraction
FreiHAND Zimmermann et al. (2019)	Dataset	Single hand	434,697	4.15%
HO-3D Hampali et al. (2020)	Dataset	Hand–object	616,576	5.88%
ForeHOI Chen et al. (2026b)	Dataset	Hand–object	1,940,750	18.51%
DexYCB Chao et al. (2021)	Dataset	Hand–object	155,633	1.48%
ARCTIC Fan et al. (2023)	Dataset	Hand–object	1,129,981	10.78%
HOT3D Banerjee et al. (2025)	Dataset	Egocentric hand–object	765,400	7.30%
DPoser-Hand Lu et al. (2025)	Generative	Single hand	3,818,350	36.41%
GraspXL Zhang et al. (2024a)	Generative	Hand–object	1,624,381	15.49%
Total	–	–	**10,485,768**	**100.00%**

A.3 Virtual Camera Sampling

We provide the complete sampling procedure corresponding to Eq. 1. Let $\mathbf{C} \in \mathbb{R}^3$ be the mean of the camera-space hand vertices and let $L = \|\mathbf{v}_{\max} - \mathbf{v}_{\min}\|_2$ denote the diagonal extent of the 3D hand. We independently sample

$$\log f \sim \mathcal{U}(\log 300, \log 1200), \qquad r \sim \mathcal{U}(0.2, 0.7), \tag{7}$$

and compute the target centroid depth

$$Z = \mathrm{clip}\left(\frac{fL}{r \min(W, H)}, 0.2, 3.0\right). \tag{8}$$

We first translate the hand centroid to $(0, 0, Z)$ and project the hand to measure its image-space bounding-box width w_b and height h_b. With the maximum off-screen fraction denoted by $\delta = 0.5$, the top-left corner of the target box is sampled as

$$x_b \sim \mathcal{U}(-\delta w_b,\, W - (1-\delta)w_b), \qquad y_b \sim \mathcal{U}(-\delta h_b,\, H - (1-\delta)h_b). \tag{9}$$

If either interval is inverted because the projected hand is larger than the frame, the box is centered along that axis. The desired projected hand-centroid location is then $u_c = x_b + (W/2 - x_b^0)$ and $v_c = y_b + (H/2 - y_b^0)$, where (x_b^0, y_b^0) is the top-left corner obtained when the centroid is at the principal point. The camera-space translation is

$$\mathbf{t} = \left[\frac{(u_c - W/2)Z}{f} - C_x,\ \frac{(v_c - H/2)Z}{f} - C_y,\ Z - C_z\right]^\top. \tag{10}$$

This construction samples truncation relative to the projected hand size, rather than relative to the full image size. We reject a proposal if any hand vertex has depth below 10^{-3} meters or if fewer than 35% of the hand vertices project inside the image. We make at most 12 attempts; the original camera is retained if no valid proposal is found.

After camera sampling, we project all joints and vertices with the stored intrinsics. The training box is computed from the projected hand vertices and expanded by a factor of 2.5 before conversion to the square crop convention used by WiLoR.

A.4 Structured Image Generation

Prompt construction. Prompts are assembled compositionally and deterministically from the sample identifier. A prompt contains: (i) a geometry-preservation instruction; (ii) scene and support-surface descriptions; (iii) background complexity, depth of field, color palette, and time of day; (iv) illumination and camera style; and (v) an optional observation challenge.

The preservation instruction explicitly fixes the hand count, handedness, finger order, fingertip positions, finger gaps and overlap, palm orientation, knuckles, silhouette, scale, and wrist location. For isolated-hand samples, the generator is prohibited from introducing a second hand or an interacting object. For hand–object samples, it is instructed to preserve object geometry, relative pose, contact, and occlusion ordering, and to avoid interpenetration.

Generation protocol. We specifically use Boogu-Image-0.1-Edit-Turbo Chen et al. (2026a), the 10B text-and-image-to-image variant distilled with Decoupled Distribution Matching (DMD). Its image-generation transformer contains 40 transformer layers with a hidden width of 3360 and 28 attention heads. The pipeline uses a separate Qwen3-VL-8B instruction encoder and an image encoder/decoder; the complete checkpoint therefore contains approximately 19.1B parameters across the image transformer, instruction encoder, and VAE. We run all components in bfloat16 without parameter quantization. For each sample, the model receives one rendered control image and the structured instruction described above. We use the distilled student path with four denoising steps, text and image guidance scales both set to 1.0, empty-instruction guidance set to 0, and DMD conditioning noise set to 0. The output resolution follows the control resolution.

A.5 Geometry Quality Control and Packaging

For a generated image, we obtain a hand mask using text-prompted detection followed by segmentation. The predicted mask is resized to the resolution of the rendered reference mask and their intersection-over-union is computed as in Eq. 4. We retain samples with $\mathrm{IoU} \geq 0.9$. We also reject empty or degenerate detections and require the predicted-to-reference mask-area ratio to lie in $[0.65, 1.6]$. For hand–object images, the rendered hand mask is the visible-hand region used for comparison, preventing correct object occlusion from being counted as a hand-geometry error.

B WiLoR† Training Details

Camera-aware forward pass. The three-dimensional vector in Eq. 5 is mapped to the ViT embedding dimension by a learned linear layer and prepended before the MANO pose, shape, camera, and image tokens. For samples with valid intrinsics, we use $f = K_{11}$. The network still predicts the weak-perspective crop camera (s, t_x, t_y). We convert it to a full-image perspective translation using

$$T_z = \frac{2f}{bs}, \qquad T_x = t_x + \frac{2(c_x - W/2)}{bs}, \qquad T_y = t_y + \frac{2(c_y - H/2)}{bs}. \tag{11}$$

The predicted 3D joints are projected with focal length (f, f) and principal point $(W/2, H/2)$ in the full image. The resulting points are then transformed by the same affine mapping used to produce the 256×256 network crop and normalized for the 2D keypoint loss.

Optimization and augmentation. We initialize the ViT-H backbone with the pretrained ViTPose weights used by the WiLoR training setup and optimize the trainable parameters with AdamW, a learning rate of 10^{-5}, weight decay of 10^{-4}, and a per-GPU batch size of 64. Training runs for 10^6 optimization steps on 16 GPUs using bfloat16 mixed precision. The loss weights for 3D keypoints, 2D keypoints, global orientation, articulated hand pose, and shape are 0.05, 0.01, 0.001, 0.001, and 0.0005, respectively. Scale, translation, and color augmentation follow the WiLoR settings, with factors 0.3, 0.02, and 0.2, respectively. In-plane rotation augmentation is disabled because rotating the crop without applying the identical transformation to the full-image intrinsics and box token would break camera consistency. Horizontal flipping is also disabled. Each batch is sampled from the original WiLoR training mixture and Hand-Ω with probabilities 0.6 and 0.4, respectively.

C Limitations

Limited forearm coverage. The hands in Hand-Ω contain only a short wrist segment and do not include the full forearm. During pipeline development, we observed that allowing the image-generation model to synthesize an extended forearm frequently produced severely bent wrists, inconsistent limb directions, or anatomically implausible connections between the hand and arm. Because these failures can invalidate the inherited geometric annotations, our prompts explicitly prohibit the generator from adding a forearm. This restriction has limited impact when training crop-based hand reconstruction models such as HaMeR and WiLoR, whose normalized hand crops typically contain only a small region beyond the wrist. However, Hand-Ω is not directly suitable for training models that take the entire image as input and are expected to reason jointly about the hand, wrist, and full arm. Extending the pipeline to generate anatomically consistent forearms while retaining accurate full-body camera geometry is an important direction for future work.

Coverage of long-tail hand poses. The pose distribution of Hand-Ω is derived from annotated poses in existing datasets and samples produced by generative hand-pose models. Although we aggregate as many complementary pose sources as practical, their combined support cannot cover every articulation encountered in unconstrained images. Rare gestures, unusual viewpoints, extreme self-occlusion, and task-specific hand configurations may therefore remain underrepresented. A promising future direction is to expand the pose space using long-tail examples mined from in-the-wild images and videos. Such examples could be reconstructed, verified, and incorporated into the rendering-and-generation pipeline while preserving the same geometry-aware quality-control requirements.