

AnyHand: A Large-Scale Synthetic Dataset for RGB(-D) Hand Pose Estimation

Chen Si^{1*} Yulin Liu^{1*} Bo Ai¹ Jianwen Xie²
 Rolandos Alexandros Potamias³ Chuanxia Zheng⁴ Hao Su¹

¹University of California, San Diego ²Lambda, Inc

³Imperial College London ⁴Nanyang Technological University

<https://chen-si-cs.github.io/projects/AnyHand>

Abstract

We present AnyHand, a large-scale synthetic dataset designed to advance the state of the art in 3D hand pose estimation. While recent works with foundation approaches have shown that scaling training data markedly improves hand pose estimation, existing real-world datasets are limited in coverage, and prior synthetic datasets rarely provide occlusions, arm details, and aligned depth together at scale. To address this bottleneck, we propose AnyHand contains 2.5M single-hand and 4.1M hand-object interaction RGB-D images, with rich geometric annotations. We show that extending the original training data recipes of existing RGB baselines with AnyHand yields significant gains on multiple benchmarks (FreiHAND and HO-3D), even when keeping the architectures and training schemes fixed. More impressively, the resulting models generalize better to the out-of-domain HO-Cap dataset at zero-shot testing. Together with extensive ablations on the scale and composition of the training data setups, these results suggest that training data diversity and quality are as critical as scale for advancing hand pose estimation. We further examine the utility of AnyHand’s aligned depth maps in the appendix, showing that scaling RGB-D supervision with AnyHand allows a lightweight depth-fusion variant of existing RGB baselines to outperform prior RGB-D methods.

1 Introduction

Our daily interactions with the physical world are largely mediated by our hands in a 3D space, a capacity widely regarded as *a genesis of human intelligence*. Grasping tools, taking a coffee, or typing on a keyboard are all examples of the remarkable fine-grained dexterity of human hands in manipulating objects of diverse shapes, sizes, and affordances. Equipping machines with similar capabilities is crucial in VR / AR [1, 44] and robotics [11, 15, 31, 43, 49, 60, 75], where accurate hand pose estimation from visual observations is necessary for natural interaction with both the virtual and physical worlds.

In this work, we consider the problem of *3D hand pose estimation*, which aims to build models that can robustly estimate 3D hand pose from RGB inputs across diverse real-world scenarios. Several recent advances have been made in this direction, driven by large-capacity transformer-based pipelines that regress parametric hand representations such as MANO [59] from a single image. Recent contributions such as HaMeR [52], Hamba [16], and WiLoR [55] demonstrate that relatively simple architectures perform well when trained on diverse, large-scale data. However, scaling such data coverage needed for *foundational* training remains difficult in practice. For instance, GigaHands [18] provides sequential hand-object interaction annotations, but as a real-captured dataset, its diversity is constrained by the capture setup/collection scale (e.g. viewing perspectives, subjects, and objects),

*Equal contribution.

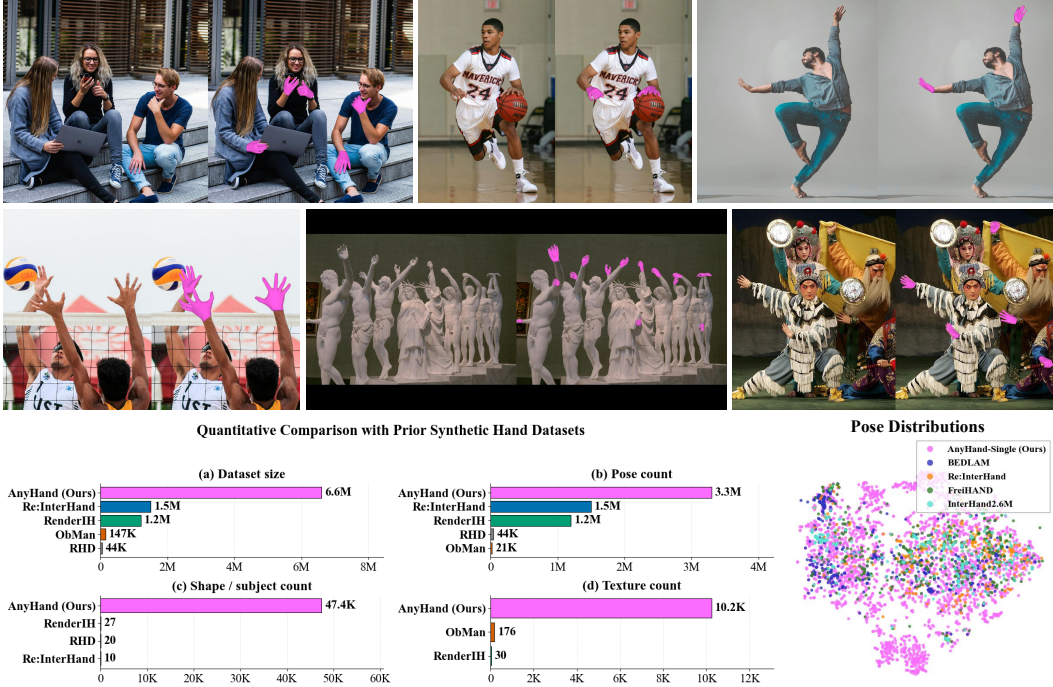


Figure 1: We propose AnyHand as a large-scale synthetic RGB-D dataset that substantially expands coverage of hand poses, hand-object interactions, textures, and viewpoint variations. When added to training data for models such as HaMeR [52] and WiLoR [55], it yields consistent improvements across multiple benchmarks. Predicted hand meshes on in-the-wild images from WiLoR trained with AnyHand in addition are shown in pink. **Bottom-left:** Quantitative comparison with representative prior synthetic hand datasets across four axes: dataset size, unique pose count, shape/subject count, and texture count. AnyHand surpasses all prior datasets on every axis by a large margin. **Bottom-right:** t-SNE visualization of hand pose distributions across datasets. AnyHand-Single densely covers the pose distributions of commonly used real and synthetic hand datasets, while also extending to additional low-density and edge-case regions that are less represented in prior data.

and its 3D annotations might include noisy ones due to limitations of the annotation/reconstruction pipeline, especially under heavy occlusion.

Recent large-scale synthetic 3D corpora such as Objaverse(-XL) [14, 13] demonstrate a practical alternative: scaling synthetic data can measurably improve downstream 3D models on various of tasks [67, 63, 77, 70, 64, 34, 35, 68, 25]. Inspired by this data-scaling paradigm, we introduce AnyHand, a large-scale RGB-D synthetic dataset that consists of hand-only and hand-object interaction scenes with realistic textures and rich annotations (Fig. 1). The resulting dataset provides *guaranteed* ground-truth labels, as the synthetic hands’ poses are “perfectly labeled” by construction, avoiding annotation noise common in real datasets. It substantially expands visual diversity by randomizing camera viewpoints, backgrounds, and illumination, and by increasing subject-level variation in hand texture, skin tone, and hand shape beyond what is feasible with a limited pool of real participants. As a concrete example, we augment the attached-arm context with diverse realistic forearm appearances, including both bare skin and clothing such as short- and long-sleeves, which helps mitigate overfitting to clean capture conditions. **To support reproducibility and future research, we will publicly release the full dataset, generation pipeline, and model checkpoints.**

We then design a mixed-training recipe that combines our created large-scale synthetic dataset AnyHand with existing datasets and validate the state-of-the-art models across multiple standard benchmarks. Without bells and whistles, that is, using the *de facto* architecture and exactly the same training protocol as HaMeR [52] and WiLoR [55], our retrained models surpass all previous state-of-the-art methods, suggesting that scaling training data can be a stronger lever than architecture design. Ablations on data scaling, real-synthetic mixing, and pipeline components further show that both the scale and the quality of the synthetic data contribute to these gains.

In summary, our main contributions are as follows:

- We introduce AnyHand, a large-scale RGB-D synthetic hand dataset and a released generation pipeline that includes realistic single-hand and hand-object interaction scenes with aligned depth and arm context.
- We show the benefits of our synthetic data through extensive experiments on standard benchmarks, demonstrating consistent improvements in both in-domain accuracy and out-of-domain generalization across diverse and challenging conditions.
- We provide ablations on data scaling, real-synthetic mixing, and dataset components, offering practical insights into effective synthetic data design for hand pose estimation.

2 Related works

3D hand pose estimation. Early 3D hand pose estimation approaches often relied on depth-based tracking, using geometric cues from depth and articulated alignment for real-time recovery [56, 62]. To move beyond depth sensors, Boukhayma et al. [5] introduced the first fully learnable pipeline that regresses MANO [59] parameters from RGB images. Subsequent RGB-based methods improved reconstruction with stronger 2D supervision, refinement modules, mesh/vertex regression, and mesh convolutions [78, 2, 28, 29]. Other works improve robustness under occlusion and motion blur [50, 48], or incorporate kinematic and biomechanical priors to suppress implausible poses [61, 71].

More recently, the dominant trend has been to scale model capacity and training data with transformer backbones, following their success in body pose and mesh recovery [72, 36]. HandDiff [10] explores a diffusion-based formulation that generates hand poses through iterative denoising. In contrast, HaMeR [52] adopts a minimalist foundation-style design that fine-tunes a ViT backbone [72] to directly regress MANO pose, shape, and camera from a single RGB image, trained on a mixed ~ 2.7 M-image corpus. Hamba [16] replaces attention with a graph-guided Mamba [19] backbone to capture joint spatial relations. WiLoR [55] further scales training to ~ 4.2 M RGB images with a coarse-to-fine refinement module, achieving state-of-the-art RGB performance. These works show that simple architectures can achieve strong performance when trained with large and diverse data.

However, RGB-based pipelines remain ill-posed in depth, leading to global translation and scale errors even when 2D projections seem well aligned. To exploit geometric cues when depth is available, prior work has explored RGB-D formulations that regress from depth maps or multimodal inputs such as image-point-cloud hybrids [39, 58]. As a recent representative, Keypoint-Fusion [40] fuses RGB and depth features around hand keypoints to reduce RGB ambiguity. Nevertheless, the lack of large-scale, aligned RGB-D datasets has limited these methods to benchmark-specific training, hindering cross-dataset generalization. To address this, we include aligned depth in our synthetic data, enabling transformer-based models to use geometric cues while retaining data-driven generalization.

Synthetic data for 3D hand pose estimation and transfer. Collecting large-scale real 3D hand annotations is difficult, especially for hand-object interactions where hands are small, articulated, and frequently occluded. This challenge is even greater for RGB-D learning, since large datasets with well-aligned depth and reliable 3D labels remain scarce. Synthetic data therefore provides a practical way to scale supervision while obtaining paired RGB-D observations with full 3D ground truth.

As summarized in Tab. 1, prior synthetic hand datasets cover different aspects of this problem, including aligned depth [80], hand-object interaction [24], and more realistic rendering with HDR

Table 1: **Comparisons with representative synthetic hand-pose datasets.** AnyHand provides the most comprehensive setting among prior work, combining the largest scale with realistic HDR indoor/outdoor backgrounds, diverse dynamic lighting, aligned depth, and arm/object configurations (AnyHand-Single/Interact).

Dataset	Size	Background	Lighting	Depth	Arm	Object
RHD [80]	44K	static	manual	✓	✗	✗
ObMan [24]	154K	static	manual	✓	✓	✓
Re:InterHand [47]	1.5M	HDR - Indoor	dynamic	✗	✗	✗
RenderIH [32]	1.2M	HDR - Indoor / Outdoor	dynamic	✗	✗	✗
AnyHand-Single (Ours)	2.5M	HDR - Indoor / Outdoor	dynamic	✓	✓	✗
AnyHand-Interact (Ours)	4.1M	HDR - Indoor / Outdoor	dynamic	✓	✓	✓

backgrounds and dynamic lighting [47, 32]. However, existing datasets rarely provide scale, realistic rendering, arm context, object-induced occlusion, and aligned depth together. Recent analysis of the synthetic-to-real gap further suggests that forearm context, interaction realism, occlusion, and rendering variation are important factors for transfer [79]. We provide a more detailed comparison with prior synthetic datasets in Appx. D.

Overall, existing datasets rarely combine large scale, hand-object occlusion, arm context, realistic rendering variation, and RGB-aligned depth cues with rich 3D annotations. This gap motivates a synthetic data source that can complement real RGB training while supporting depth-aware extensions.

3 AnyHand dataset

Motivated by the need for scalable RGB-D supervision, we propose AnyHand, a large-scale synthetic dataset designed to complement existing real-world data and support both RGB-only and RGB-D model training, guided by two principles. First, the data should be *diverse* in pose, shape, appearance, viewpoint, and interaction patterns to support large-capacity models. Second, the data should be *geometrically grounded*: we provide depth cues and precise hand annotations obtained from the rendering and simulation pipeline. Concretely, AnyHand comprises two complementary branches: AnyHand-Single with 2.5M RGB-D samples for diverse single-hand settings, and AnyHand-Interact with 4.1M RGB-D samples for hand-object interactions with object-induced occlusions. Detailed dataset statistics are provided in Appx. C.

3.1 Dataset creation

Hand shapes. To cover a broad range of hand shapes, but also to avoid unrealistic geometry, we sample 47,438 MANO shape parameters β from the empirical distribution of FreiHAND [81] and InterHand2.6M [46], which are real datasets with large subject-level diversity and provide a good proxy for the true distribution of hand shapes in the population.

Hand poses. Hand realism requires plausible articulation. Rather than sampling poses from simple heuristics, we leverage DPoser-Hand [42], a hand pose model trained on a mixture of large real datasets (e.g. FreiHAND [81], HO-3D [20], DexYCB [7], H2O [30], and Re:InterHand [47]). The *key* advantage of such a diffusion prior is that it captures multiple pose distributions observed in real data, enabling us to generate a broad range of poses while avoiding unnatural articulations that often arise from naive sampling. During dataset generation, hand poses are sampled on the fly from this prior.

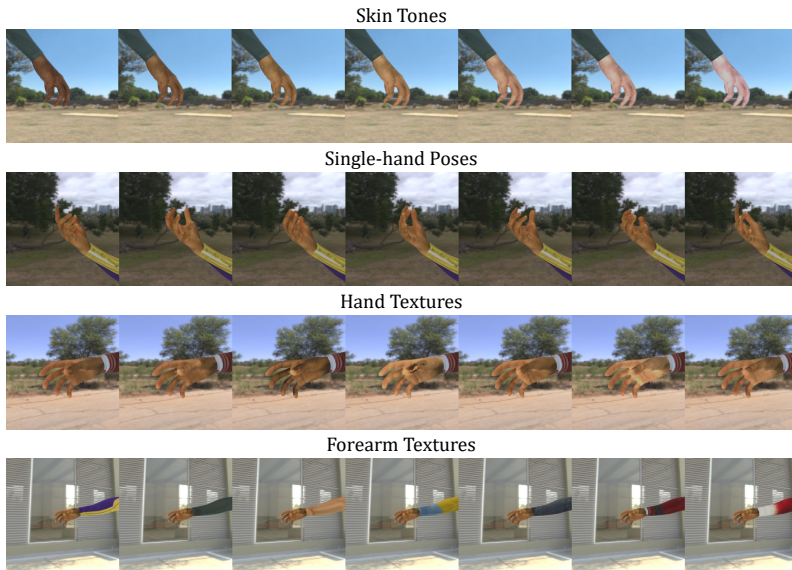


Figure 2: **Qualitative visualizations of controllable variations.** We show representative samples from our generator by varying one factor at a time: *skin tones*, *single-hand poses* from DPoser-Hand [42], *hand textures* from Handy [54], and *forearm appearance* from SMPLitex [6]. These examples demonstrate the diversity of appearance and context that we leverage in AnyHand.

Hand textures. A key limitation of prior synthetic datasets is the lack of high-fidelity hand textures, which are crucial for closing the sim-to-real gap. To address this, we leverage a hand texture generator, Handy [54], to produce a large variety of realistic high-frequency skin patterns, which improves the visual fidelity of rendered creases and shading transitions compared to the canonical MANO texture space. In particular, we adopt Handy as our primary source of hand textures and then augment it by applying controlled color transformations, such as hue and saturation perturbations, that broaden skin-tone distribution while preserving the high-frequency texture structure. This yields the diverse textures of 10,240 unique hand appearances, which is significantly more than the limited hand-crafted texture libraries used in prior works [80, 24, 47, 32].

Forearm textures. To maintain realistic appearance continuity across the hand-forearm junction, we attach a full arm mesh from SMPL-H [41, 51, 59] during dataset generation. We texture the arm using SMPLiteX [6], which provides 254 high-quality human-body textures for realistic rendering. This introduces diverse and natural skin appearances for the visible forearm regions.

Backgrounds. To diversify environments, we use two sources of high-resolution backgrounds: the MIT Indoor Scenes dataset [57] (536 images) and a pool of 734 HDRI environment maps from Poly Haven, formerly HDRI Haven [53]. For each rendered hand sample, a background is randomly selected, rotated, zoomed in, and then cropped with a random size and location to a patch before being used as the sample-specific background. This ensures the effective background appearance vary across samples, even when the same source image or HDRI map is selected multiple times.

Lighting. For foreground-background consistency, we randomize a small set of scene lights and correlate their color statistics with the background patch, so that the rendered hand inherits the dominant illumination tone of the scene (e.g. warm indoor tungsten vs. cooler daylight). For HDRI environment maps, we directly use the environment illumination to obtain coherent scene lighting.

Cameras. To ensure viewpoint diversity, we randomize camera intrinsics and extrinsics within realistic ranges based on calibration statistics from real datasets such as HO-3D [20] and DexYCB [7]. The intrinsics, including focal length and field of view, are sampled to mimic those of the real capture setups. For the extrinsics, we keep the camera-to-hand distance within realistic bounds (within 0.6-0.8m) and randomize the camera viewpoint over the full 3D viewing sphere around the hand, enabling observations from all-possible directions while constraining the hand to remain well framed, avoiding severe truncation artifacts.

Rendering AnyHand-Single dataset. After preparing the above-described components, we instantiate them into a unified rendering-and-compositing pipeline. We render all scenes in SAPIEN [69], using its ray-tracing renderer to better capture realistic shading, cast shadows, and specular effects. For each sample, we first draw a plausible MANO shape and pose, and apply diffused hand textures

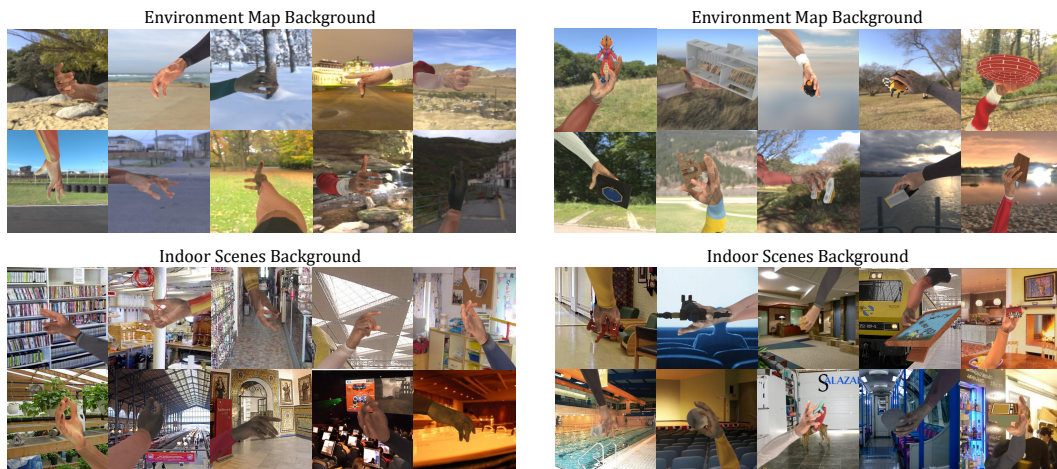


Figure 3: **Qualitative visualizations.** Examples of AnyHand-Single (left) and AnyHand-Interact (right), with both HDR environment-map backgrounds (top) and real indoor scenes (bottom). In addition to diverse hand/arm appearance and poses, we have additional diversity on the interacted objects and grasp configurations, producing a wide range of object-induced hand occlusions and self-occlusions under varying perspectives.

with optional color perturbations to get the textured 3D hand mesh. Then, we attach a textured forearm segment to the hand in 3D, which provides consistent geometry and appearance across the hand–forearm junction and avoids boundary artifacts common in 2D compositing. In particular, we extract the arm mesh from a parametric body model (SMPL/SMPL+H family [41, 51, 59]), align it to the MANO wrist frame, and texture it using SMPLitex [6]. Finally, we render the foreground hand-arm with background-aware lighting by using the above-described camera randomization. For each scene, we render two images from two independently sampled camera viewpoints, improving simulator efficiency while increasing perspective diversity. To further diversify the rendered data, we composite the rendered foreground onto a randomly cropped high-resolution background patch and pair it with HDR environment illumination, while keeping foreground-background consistency by correlating scene-light color statistics with the sampled background patch.

Aligned depth maps. We also provide aligned depth maps for all synthetic samples. To this end, we first render accurate metric depths for the hand and forearm from SAPIEN. For the background patch, we estimate a dense metric depth map using MoGe-2 [66]. We then directly fuse foreground and background depth in camera space to obtain a dense depth image for the final composite. While this is not a “perfect” ground truth depth map due to differences in camera intrinsics between the rendered foreground and background, as well as noise in the estimated background depth, it provides a useful approximation for training and evaluation purposes. Besides, we also store a foreground mask so models can optionally restrict losses or depth usage to valid hand/arm regions.

Rendering AnyHand-Interact dataset. To model occluded hands in real-world scenarios at scale, we render a second branch of the dataset, AnyHand-Interact, by rendering the grasping behaviors from GraspXL [76], which provides over 10M physics-simulation-based hand-object interaction sequences on more than 500k realistically textured objects spanning diverse categories and surface appearances from Objaverse [14], with contact-consistent grasps and natural occlusion patterns. Therefore, we directly use the full GraspXL corpus and inherit its associated object set. The rendering pipeline follows the same strategy as the single-hand branch, but now includes realistic mutual occlusions between the hand and the manipulated object.

3.2 Dataset statistics

Overall, AnyHand consists of AnyHand-Single (with 1.25M scenes and 2.5M images) and AnyHand-Interact (with 2.05M scenes and 4.1M images), which are rendered with a combination of 47,438 hand shapes, 10,240 hand textures, 254 arm textures, 1,270 backgrounds, and more than 500k objects from Objaverse [14]. Note that, for all rendered samples, we store RGB, depth, foreground mask, 2D bounding boxes, together with camera intrinsics and extrinsics. We also provide precise 3D hand pose and shape parameters directly from the simulation, which can be used for supervised training or evaluation. More details on the implementation can be found in Appx. C.1.

4 Assessing AnyHand dataset on RGB-only settings

4.1 Experiment

Evaluation setups. To evaluate the quality of AnyHand and its effectiveness for improving foundation-style hand mesh reconstruction, we study two representative frameworks: HaMeR [52], trained on 2.7M images, and WiLoR [55], trained on 4.2M images. For both frameworks, we augment their original training corpora with AnyHand while keeping the model architectures and training hyperparameters unchanged. To isolate dataset design from simply adding synthetic data, we also train HaMeR with RenderIH [32], a representative large-scale synthetic hand dataset containing 1.2M images with HDR backgrounds and dynamic lighting, using the same mixed-training protocol.

Each training batch is constructed by sampling from different datasets according to predefined weights. We assign a weight of 0.4 to the added synthetic dataset and scale the weights of the original datasets by 0.6, preserving the original data distribution while introducing substantial synthetic supervision. Details on the mixed-training scheme are provided in Appx. A.1. All experiments run on 8 NVIDIA H100 GPUs and are trained for 5–7 days until convergence. Compute details are provided in Appx. E.

Metrics. Following protocols in the original papers of HaMeR[52] and WiLoR[55], we report Procrustes-aligned Mean per Joint and Vertex Error (PA-MPJPE, PA-MPVPE) and the F-score of vertices at 5mm and 15mm (F@5, F@15) [81, 27]. For the HO-3D[20] dataset, we additionally report AUC_j and AUC_v , defined as area under the PCK curve over joint and vertex error thresholds.

Table 2: **Comparison with the state-of-the-art on the FreiHAND benchmark [81].** Top results are emphasized in **top1**, **top2**, and **top3**. Notably, augmented training with AnyHand yields a **7.6%** PA-MPJPE improvement for HaMeR and a **1.9%** improvement for WiLoR. A full comparison with more prior works can be found in Appx. in Tab. 6.

Method	PA-MPJPE ↓	PA-MPVPE ↓	F@5mm ↑	F@15mm ↑
AMVUR [26] _{CVPR23}	6.2	6.1	0.767	0.987
HaMeR [52] _{CVPR24}	6.0	5.7	0.785	0.990
Hamba [16] _{NeurIPS24}	5.7	5.3	0.806	0.992
WiLoR [55] _{CVPR25}	5.5	5.1	0.825	0.993
HaMeR w/ RenderIH	5.8 ₅₇	5.4 ₈₆	0.797	0.992
HaMeR w/ AnyHand (Ours)	5.5 ₄₅	5.2 ₄₆	0.811	0.993
WiLoR w/ AnyHand (Ours)	5.3 ₉₄	5.0 ₄₆	0.827	0.994

Table 3: **Comparison with the state-of-the-art on the HO-3D v2 benchmark [20].** Top results are emphasized in **top1**, **top2**, and **top3**. Using AnyHand reduces PA-MPJPE by **3.0%** for HaMeR and by **1.9%** for WiLoR. A full comparison of more prior works can be found in Appx. Tab. 7.

Method	AUC _j ↑	PA-MPJPE ↓	AUC _v ↑	PA-MPVPE ↓	F@5 ↑	F@15 ↑
AMVUR [26]	0.835	8.3	0.836	8.2	0.608	0.965
HaMeR [52]	0.846	7.7	0.841	7.9	0.635	0.980
Hamba [16]	0.850	7.5	0.843	7.7	0.648	0.982
WiLoR [55]	0.851	7.5	0.846	7.7	0.646	0.983
HaMeR w/ RenderIH	0.846	7.6 ₈₅	0.842	7.9 ₀₂	0.634	0.982
HaMeR w/ AnyHand (Ours)	0.851	7.4 ₇₁	0.847	7.6 ₇₆	0.645	0.984
WiLoR w/ AnyHand (Ours)	0.853	7.3 ₅₅	0.848	7.6 ₂₄	0.649	0.984

4.2 Results

In-domain results. We perform an in-domain evaluation on the popular FreiHAND [81] and HO-3D v2 [20] benchmarks. As shown in Tab. 2 and 3, adding AnyHand consistently improves performance across all metrics for both HaMeR and WiLoR, demonstrating the effectiveness of our synthetic data for enhancing RGB-based hand pose estimation. On FreiHAND, WiLoR w/ AnyHand achieves the best results, while the effect of synthetic augmentation is even more pronounced for HaMeR. The PA-MPJPE drops from 6.0 mm to 5.54 mm, and PA-MPVPE decreases from 5.7 mm to 5.24 mm, lifting HaMeR into the same performance tier as the top-ranked approaches, *without requiring any specific architecture modification*. On HO-3D v2, which emphasizes hand-object interactions, the same trend holds. WiLoR w/ AnyHand attains the best overall results, while HaMeR also improves substantially, with PA-MPJPE of HaMeR w/ AnyHand reduced from 7.7 mm to 7.47 mm. In contrast, training with RenderIH [32] under the same protocol yields smaller improvements on both benchmarks, highlighting the quality and effectiveness of AnyHand. Qualitative examples in Appx. Fig. 6 further illustrate that training with AnyHand helps reduce errors in challenging poses and occluded joints, which are common failure modes for RGB-based methods. More visual results are in Appx. A.2.

Out-of-domain results. To evaluate the out-of-domain generalization ability, we directly evaluate performance on the HO-Cap [65] benchmark without any fine-tuning, whose images are entirely from unseen sources with a clear domain shift. As reported in Tab. 4, augmenting training with AnyHand’s synthetic data improves both WiLoR and HaMeR, and yields a notable ranking change: HaMeR w/ AnyHand attains PA-MPJPE of 4.66 mm, slightly outperforming WiLoR w/ AnyHand, which has PA-MPJPE of 4.69 mm. This is particularly interesting because HaMeR is overall weaker than WiLoR on HO-Cap under the original training setup. The fact suggests that the performance gap between the two architectures is not fixed, but can be influenced by the training data distribution.

Analysis. Two key observations emerge from these experiments. Firstly, adding AnyHand’s synthetic data in the original training data corpora provides consistent improvements across two representative ViT-based baselines (HaMeR and WiLoR), suggesting that the gains are not tied to a single pipeline

Table 4: Comparison with HaMeR [52] and WiLoR [55] on the HO-Cap benchmark [65] as an in-the-wild case. Better results are **bolded**.

Method	AUC _j ↑	PA-MPJPE ↓	F@5mm ↑	F@15mm ↑
HaMeR [52]	0.901	4.94	0.621	0.984
WiLoR [55]	0.899	5.02	0.615	0.981
HaMeR w/ AnyHand (Ours)	0.907	4.66	0.662	0.985
WiLoR w/ AnyHand (Ours)	0.906	4.69	0.652	0.987

design. The stronger gains on HaMeR are consistent with its smaller training dataset, whereas WiLoR’s additional refinements and larger training corpus leave less room for improvement. Moreover, on HO-Cap, the performance gains from adding AnyHand are substantially larger than the differences between architectures, suggesting that improvements from training data can outweigh architectural differences. Overall, these findings support our thesis that scaling training data quality, quantity, and diversity is a stronger lever than iterating on architectures alone.

4.3 Ablations on AnyHand

We ablate our AnyHand for in-domain and out-of-domain performance in Fig. 4 and Tab. 5. We keep the HaMeR [52] architecture and training protocol fixed, and only vary the data configurations.

Scaling of adding AnyHand in training. Fig. 4 studies the effect of synthetic-data scaling by varying the amount of synthetic data while keeping the real-data portion fixed. The results show that augmenting HaMeR with AnyHand yields a substantial performance boost over the no-synthetic baseline across all benchmarks. On the in-domain benchmarks (FreiHAND and HO-3D v2), the small dataset ($\frac{1}{3}$ on FreiHAND, $\frac{2}{3}$ on HO-3D) already provides the main gains, while increasing it to full size only provides a modest improvement. However, on the out-of-domain HO-Cap benchmark, we see a more consistent improvement as the synthetic budget increases, suggesting that scaling up synthetic data may be particularly beneficial for improving robustness to domain shifts. This is likely because the larger synthetic dataset covers a wider range of poses, shapes, textures, and backgrounds that better match the diversity of *unseen* real-world scenarios. This provides a strong motivation for investing in large-scale synthetic data generation in future work.

Sim-and-real mixed-training recipe. The scaling experiment above shows that adding AnyHand improves performance, but it does not answer how synthetic and real data should be mixed. To study this, we fix the total training set size to 2.7M samples, matching HaMeR’s original training budget, and vary the proportion of AnyHand. For example, when AnyHand accounts for 25% of the 2.7M data, the remaining 75% samples are drawn from HaMeR’s original training corpus and scaled down proportionally. All models are trained for the same number of steps for a controlled comparison.

As shown in Fig. 5, using AnyHand alone is not sufficient: the 100% AnyHand setting performs worse than the mixed settings and also degrades relative to HaMeR’s original recipe. This confirms that AnyHand is best used as complementary supervision rather than as a full replacement for real training data. In contrast, all mixed settings improve over the 0% baseline, showing that AnyHand consistently benefits HaMeR under the same training budget. Notably, the 50% and 75% mixtures achieve strong performance across the evaluated benchmarks, and in several cases surpass the WiLoR reference

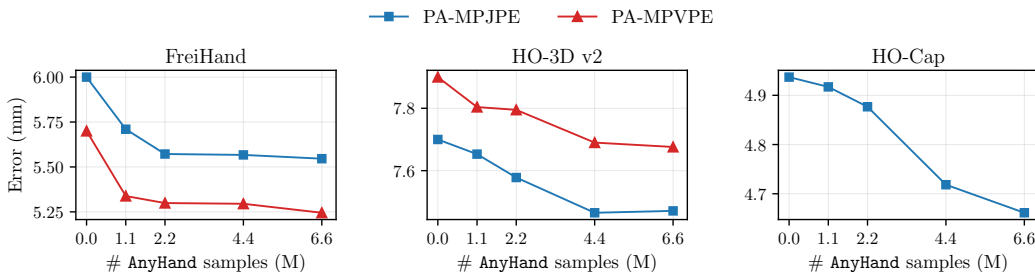


Figure 4: **Scaling HaMeR training with AnyHand in addition.** We retrain HaMeR [52] while keeping its original real-data training set fixed, and vary the number of additional AnyHand samples used. We report PA-MPJPE and PA-MPVPE on FreiHAND [81], HO-3D v2 [20], and HO-Cap [65]. Adding AnyHand in training consistently reduces error, with gains saturating beyond ~ 2 –4M samples.

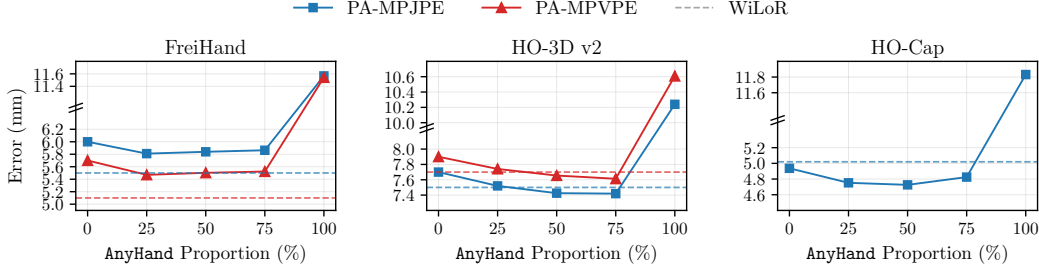


Figure 5: **Effect of the training-data mixing ratio.** HaMeR is trained with a fixed budget of 2.7M samples, matching its original data recipe, while varying the proportion of AnyHand. The remaining samples are drawn from HaMeR’s original training corpus and scaled down proportionally. The dashed line indicates WiLoR’s original result as a comparison. The results provide a rough estimate of the impact of different mixture ratios under a fixed training budget.

Table 5: **Ablations on AnyHand components on the FreiHAND [81] benchmark.** In each experiment, the HaMeR [52] model is trained with different data configurations while other settings are identical. Here “w/ interpolated pose” means using interpolated poses from real datasets instead of diffusing from the default DPoser-Hand [42].

HaMeR [52] Variant	PA-MPJPE ↓	PA-MPVPE ↓	F@5mm ↑	F@15mm ↑
HaMeR [52]	6.000	5.700	0.7850	0.9900
w/ AnyHand	5.545	5.246	0.8110	0.9930
w/ AnyHand-Interact	5.850	5.525	0.7940	0.9919
w/ AnyHand-Single	5.626	5.286	0.8109	0.9930
w/o realistic arm textures	5.686	5.312	0.8096	0.9926
w/ interpolated poses	5.707	5.335	0.8056	0.9926

despite using the simpler HaMeR architecture and less training data, suggesting that mixing a large proportion of synthetic data with real data is an effective strategy for improving hand pose estimation.

Ablations on other variants. As summarized in Tab. 5, we further ablate key design choices in AnyHand. First, dropping either the *Single* branch or the *Interact* branch degrades performance, suggesting that when the target benchmarks contain a mix of single-hand and hand-object interaction cases, jointly training with both single and interaction samples yields the best overall results. Second, removing arm texture slightly hurts performance, suggesting that realistic arm appearance (beyond geometry alone) provides useful contextual cues and improves generalization. Third, replacing diffusion-based pose synthesis with poses interpolated from real data leads to a consistent performance drop, indicating that diffusion provides more effective pose diversity for this mixed training scheme.

Summary. These ablations collectively show that the gains from AnyHand are not simply due to increased training data. Scaling synthetic data shows diminishing returns in-domain, while still improving out-of-domain performance. The mixing experiment indicates that synthetic data complements rather than replaces real data. In addition, the component ablations demonstrate that specific design choices contribute to performance. Overall, these findings suggest that both data quality and quantity contribute to the improvements in Sec. 4.2.

5 Conclusions

We have introduced AnyHand, a large-scale synthetic dataset that provides diverse hand scenes with rich annotations. By incorporating AnyHand into the original training sets of state-of-the-art hand pose estimation models, we demonstrated consistent improvements across multiple benchmarks, validating the effectiveness of our synthetic data for enhancing RGB-based approaches. We have also proposed a novel RGB-D architecture that incorporates a lightweight depth fusion module, and showed that it outperforms prior RGB-D approaches on hand pose estimation. Overall, our results open a promising avenue for advancing hand pose estimation by improving the *quality, diversity, and modality coverage* of training data, rather than solely focusing on architectural innovations.

References

- [1] Apple. Apple Vision Pro. <https://www.apple.com/apple-vision-pro/>, 2024. Accessed: 2026-03-03.
- [2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, pages 1067–1076, 2019.
- [3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024.
- [4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023.
- [5] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019.
- [6] Dan Casas and Marc Comino-Trinidad. SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In *BMVC*, 2023.
- [7] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021.
- [8] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *ICCV*, pages 12929–12938, 2021.
- [9] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, pages 20544–20554, 2022.
- [10] Wencan Cheng, Hao Tang, Luc Van Gool, and Jong Hwan Ko. Handdiff: 3d hand pose estimation with diffusion on image-point cloud. In *CVPR*, pages 2274–2284, 2024.
- [11] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *CoRL*, 2024.
- [12] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787. Springer, 2020.
- [13] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *NeurIPS*, 36:35799–35813, 2023.
- [14] 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 *CVPR*, pages 13142–13153, 2023.
- [15] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. In *IROS*, pages 12248–12255. IEEE, 2025.
- [16] 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.
- [17] 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 *CVPR*, pages 12943–12954, 2023.

- [18] 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 *CVPR*, pages 17461–17474, 2025.
- [19] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024.
- [20] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020.
- [21] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020.
- [22] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, pages 11090–11100, 2022.
- [23] 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*, pages 11807–11816, 2019.
- [24] 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*, pages 11807–11816, 2019.
- [25] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Mesh4d: 4d mesh reconstruction and tracking from monocular video. In *CVPR*, 2026.
- [26] 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*, pages 758–767, 2023.
- [27] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.
- [28] 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.
- [29] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pages 4990–5000, 2020.
- [30] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, pages 10138–10148, 2021.
- [31] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *CVPR*, pages 6991–7003, 2025.
- [32] 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*, pages 20395–20405, 2023.
- [33] Lijun Li, Li’an Zhuo, Bang Zhang, Liefeng Bo, and Chen Chen. Diffhand: End-to-end hand mesh reconstruction via diffusion models. *arXiv preprint arXiv:2305.13705*, 2023.
- [34] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Puppet-master: Scaling interactive video generation as a motion prior for part-level dynamics. In *ICCV*, 2025.
- [35] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Dso: Aligning 3d generators with simulation feedback for physical soundness. In *ICCV*, 2025.

- [36] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021.
- [37] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, 2021.
- [38] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, pages 14687–14697, 2021.
- [39] Xingyu Liu, Pengfei Ren, Yuchen Chen, Cong Liu, Jing Wang, Haifeng Sun, Qi Qi, and Jingyu Wang. Sa-fusion: multimodal fusion approach for web-based human-computer interaction in the wild. In *Proceedings of the ACM Web Conference 2023*, pages 3883–3891, 2023.
- [40] Xingyu Liu, Pengfei Ren, Yuanyuan Gao, Jingyu Wang, Haifeng Sun, Qi Qi, Zirui Zhuang, and Jianxin Liao. Keypoint fusion for rgb-d based 3d hand pose estimation. In *AAAI*, volume 38, pages 3756–3764, 2024.
- [41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIG-GRAPH Asia)*, 34(6):248:1–248:16, October 2015.
- [42] Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, et al. Dposer-x: Diffusion model as robust 3d whole-body human pose prior. In *ICCV*, pages 9988–9997, 2025.
- [43] Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. Dexmachina: Functional retargeting for bimanual dexterous manipulation. *arXiv preprint arXiv:2505.24853*, 2025.
- [44] Meta. Meta Quest 3. <https://www.meta.com/quest/quest-3/>, 2023. Accessed: 2026-03-03.
- [45] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020.
- [46] Gyeongsik Moon, Shou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, pages 548–564. Springer, 2020.
- [47] 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.
- [48] 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 *CVPR*, pages 554–563, 2023.
- [49] Chaoyi Pan, Changhao Wang, Haozhi Qi, Zixi Liu, Homanga Bharadhwaj, Akash Sharma, Tingfan Wu, Guanya Shi, Jitendra Malik, and Francois Hogan. Spider: Scalable physics-informed dexterous retargeting. *arXiv preprint arXiv:2511.09484*, 2025.
- [50] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handocnet: Occlusion-robust 3d hand mesh estimation network. In *CVPR*, pages 1496–1505, 2022.
- [51] 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 *CVPR*, 2019.
- [52] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, pages 9826–9836, 2024.

- [53] Poly Haven. Poly Haven HDRIs. <https://polyhaven.com/hdri>, 2026. Previously known as HDRI Haven. CC0 licensed HDRI environment maps. Accessed: 2026-05-02.
- [54] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *CVPR*, pages 4670–4680, June 2023.
- [55] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *CVPR*, pages 12242–12254, 2025.
- [56] Chen Qian, Xiao Sun, Yichen Wei, Xiaou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *CVPR*, pages 1106–1113, 2014.
- [57] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420. IEEE, 2009.
- [58] Pengfei Ren, Yuchen Chen, Jiachang Hao, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Two heads are better than one: Image-point cloud network for depth-based 3d hand pose estimation. In *AAAI*, volume 37, pages 2163–2171, 2023.
- [59] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.
- [60] Liangzhi Shi, Yulin Liu, Lingqi Zeng, Bo Ai, Zhengdong Hong, and Hao Su. Learning adaptive dexterous grasping from single demonstrations. In *IROS*, pages 9456–9463. IEEE, 2025.
- [61] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *ECCV*, pages 211–228. Springer, 2020.
- [62] 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, pages 101–114. Wiley Online Library, 2015.
- [63] SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. 2025. URL <https://arxiv.org/abs/2511.16624>.
- [64] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025.
- [65] Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. *arXiv preprint arXiv:2406.06843*, 2024.
- [66] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025. URL <https://arxiv.org/abs/2507.02546>.
- [67] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, pages 17868–17879, 2024.
- [68] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. In *ICCV*, 2025.
- [69] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

- [70] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, June 2025.
- [71] Pengfei Xie, Wenqiang Xu, Tutian Tang, Zhenjun Yu, and Cewu Lu. Ms-mano: Enabling hand pose tracking with biomechanical constraints. In *CVPR*, pages 2382–2392, 2024.
- [72] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 35:38571–38584, 2022.
- [73] 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 *ICCV*, 2021.
- [74] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *CVPR*, pages 2750–2760, 2022.
- [75] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.
- [76] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating grasping motions for diverse objects at scale. In *ECCV*, 2024.
- [77] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024.
- [78] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019.
- [79] 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*, pages 12255–12265, 2025.
- [80] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017.
- [81] 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*, pages 813–822, 2019.

Appendix

A More evaluations of AnyHand on RGB-only settings

A.1 Training corpora and synthetic data proportions

We provide additional details on the original training corpora used by the two evaluated frameworks. HaMeR [52] is trained on approximately 2.72M samples, which are predominantly real images, with only about 61K synthetic samples from RHD [80], corresponding to roughly 2.3% of its training data. WiLoR [55] expands the training corpus to approximately 4.19M samples by including additional datasets such as ARCTIC [17], BEDLAM [4], ReInterHand [47], and HOT3D [3]. However, synthetic data still accounts for a relatively small portion of its training corpus, around 18.7%.

After enhancing the original training corpora with AnyHand, the proportion of synthetic data increases substantially. For HaMeR [52], adding AnyHand raises the synthetic-data proportion from approximately 2.3% to 71.5%. For WiLoR [55], the proportion increases from approximately 18.7% to 68.4%. This represents, to the best of our knowledge, the first attempt to train transformer-based foundation-style hand mesh reconstruction models with such a large proportion of synthetic data. These results therefore provide an important empirical study of whether large-scale synthetic RGB-D supervision can effectively improve modern hand pose estimators.

A.2 Qualitative comparisons



Figure 6: **Qualitative visualizations.** WiLoR w/ AnyHand vs. WiLoR on FreiHAND [81] and AnyHand Test Set. Left to right: input, GT, WiLoR w/ AnyHand, WiLoR. Adding synthetic data improves fine-grained pose estimation, particularly fingertip bending and finger joint angles (as boxed), yielding meshes that better match the image evidence.

Fig. 6 provides qualitative comparisons between the original WiLoR and WiLoR trained with AnyHand on FreiHAND and AnyHand test examples. The improvements are most visible in challenging cases involving bent fingers, unusual hand poses, and partial occlusions. Compared with the original WiLoR, WiLoR w/ AnyHand better matches the image evidence around fingertips and finger joints, producing meshes with more accurate finger bending and hand shape. These examples support



Figure 7: **Additional in-the-wild qualitative comparisons.** We compare WiLoR trained with AnyHand against the original WiLoR and HaMeR on more real-world images. WiLoR w/AnyHand shows better mesh-to-image alignment, more accurate hand scale and shape recovery, and more faithful articulation for challenging hand-object interactions and viewpoint changes.

our quantitative results and show that AnyHand helps reduce common RGB-based failure cases, such as over-smoothed fingers, inaccurate joint angles, and poor alignment under occlusion.

To further illustrate the qualitative prediction quality of models trained with AnyHand in addition, we provide additional visual comparisons in Fig. 7. The figure includes seven real-image examples

Table 6: **Comparison with the state-of-the-art on the FreiHAND benchmark [81].** Top results are emphasized in **top1**, **top2**, and **top3**. Notably, augmented training with AnyHand yields a **7.6%** PA-MPJPE improvement for HaMeR and a **1.9%** improvement for WiLoR.

Method	PA-MPJPE ↓	PA-MPVPE ↓	F@5mm ↑	F@15mm ↑
I2L-MeshNet [45]	7.4	7.6	0.681	0.973
Pose2Mesh [12]	7.7	7.8	0.674	0.969
I2UV-HandNet [8]	6.7	6.9	0.707	0.977
METRO [36]	6.5	6.3	0.731	0.984
Mesh Graphormer [37]	5.9	6.0	0.764	0.986
MobRecon [9]	5.7	5.8	0.784	0.986
AMVUR [26] _{CVPR23}	6.2	6.1	0.767	0.987
HaMeR [52] _{CVPR24}	6.0	5.7	0.785	0.990
Hamba [16] _{NeurIPS24}	5.7	5.3	0.806	0.992
WiLoR [55] _{CVPR25}	5.5	5.1	0.825	0.993
HaMeR w/ AnyHand	5.5 ₄₅	5.2 ₄₆	0.811	0.993
WiLoR w/ AnyHand	5.3 ₉₄	5.0 ₄₆	0.827	0.994

from the HO-Cap [65] dataset, two from the HO-3D evaluation set [20], and one in-the-wild web image. These examples allow us to examine model behavior on both standard benchmark data and less controlled real-world imagery.

As shown in Fig. 7, WiLoR w/ AnyHand more consistently recovers plausible palm width, finger thickness, and overall hand extent compared with the original WiLoR and HaMeR. The improvement is especially clear in hand-object interaction cases, where WiLoR w/ AnyHand better preserves fine articulation details such as finger bending angles, fingertip placement, and the projected contour of the hand. Overall, these visualizations suggest that training with AnyHand not only improves standard pose-estimation accuracy, but also helps produce more realistic meshes with better image-space alignment in challenging real-world cases.

A.3 Full performance on FreiHAND

We report the complete version of Tab. 2 from the main paper in Tab. 6.

A.4 Full performance on HO-3D

We report the complete version of Tab. 3 from the main paper in Tab. 7.

B Assessing AnyHand on RGB-D settings

RGB-D architecture. Unlike the RGB-only setting, where we focus on evaluating the impact of AnyHand, the RGB-D setting requires an architectural change to fuse depth, as illustrated in Fig. 8. We use dual embedding branches to tokenize RGB and depth, followed by a lightweight bidirectional cross-attention module that exchanges information between the two modalities at corresponding image patches. The fused tokens are then concatenated with task tokens and passed through the remaining transformer blocks for 3D hand pose estimation. Note that, the fusion module is designed to be lightweight and modular, allowing it to be easily integrated into existing ViT-based architectures like WiLoR [55] with minimal changes.

Training. We conduct two training variants: *Real-only* trains on the real RGB-D datasets HO-3D [20] and DexYCB [7], while *Real + AnyHand* further trains together with our proposed AnyHand, following the same recipe and protocol as in the RGB experiments.

Evaluation. We evaluate the models on HO-3D v2 [20] and report on (1) scale-translation aligned MPJPE (STA-MPJPE), and (2) Procrustes-aligned MPJPE (PA-MPJPE), where for both metrics, lower is better.

Table 7: **Comparison with the state-of-the-art on the HO-3D v2 benchmark [20].** Top results are emphasized in **top1**, **top2**, and **top3**. Using AnyHand reduces PA-MPJPE by **3.0%** for HaMeR and by **1.9%** for WiLoR.

Method	AUC _j ↑	PA-MPJPE ↓	AUC _v ↑	PA-MPVPE ↓	F@5 ↑	F@15 ↑
Liu et al. [38]	0.803	9.9	0.810	9.5	0.528	0.956
HandOccNet [50]	0.819	9.1	0.819	8.8	0.564	0.963
I2UV-HandNet [8]	0.804	9.9	0.799	10.1	0.500	0.943
HONotate [21]	0.788	10.7	0.790	10.6	0.506	0.942
Hasson et al. [23]	0.780	11.0	0.777	11.2	0.464	0.939
ArtiBoost [74]	0.773	11.4	0.782	10.9	0.488	0.944
Pose2Mesh [12]	0.754	12.5	0.749	12.7	0.441	0.909
I2L-MeshNet [45]	0.775	11.2	0.722	13.9	0.409	0.932
METRO [36]	0.792	10.4	0.779	11.1	0.484	0.946
MobRecon [9]	-	9.2	-	9.4	0.538	0.957
KeypointTrans [22]	0.786	10.8	-	-	-	-
AMVUR [26]	0.835	8.3	0.836	8.2	0.608	0.965
HaMeR [52]	0.846	7.7	0.841	7.9	0.635	0.980
Hamba [16]	0.850	7.5	0.843	7.7	0.648	0.982
WiLoR [55]	0.851	7.5	0.846	7.7	0.646	0.983
HaMeR w/ AnyHand	0.851	7.4 ₇₁	0.847	7.6 ₇₆	0.645	0.984
WiLoR w/ AnyHand	0.853	7.3 ₅₅	0.848	7.6 ₂₄	0.649	0.984

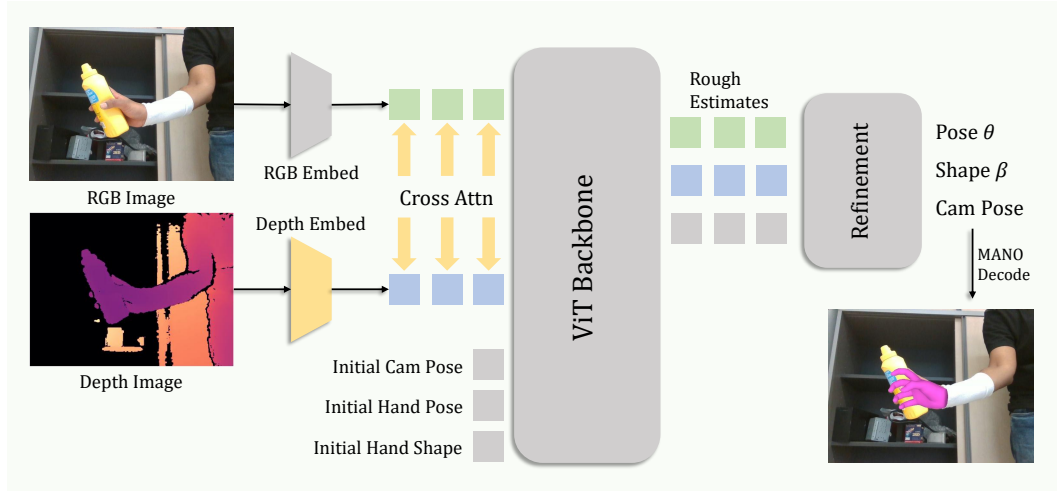


Figure 8: **Workflow of AnyHandNet-D.** Built upon WiLoR [55]’s RGB-only pipeline, we add a lightweight depth fusion module (highlighted in **yellow**). RGB and depth are embedded into parallel token sequences, followed by a bidirectional RGB-Depth cross-attention. The fused tokens are concatenated with other task tokens and fed into the ViT backbone and refinement head, whose outputs are finally MANO decoded.

Results. The quantitative results are reported in Tab. 8. Our method consistently outperforms prior RGB-D approaches on both the STA-MPJPE (no rotation alignment) and PA-MPJPE (with rotation alignment) metrics, indicating that the improvement is not merely from more accurate global hand orientation, but also a more accurate articulated hand structure after removing global similarity transforms. Compared to Keypoint-Fusion [40], our model reduces the STA-MPJPE from 1.87 cm to 1.09 cm, a relative error reduction of approximately 41.7%. Moreover, even the real-only variant remains competitive and already surpasses prior RGB-D baselines on STA and PA, indicating that the fusion module is robust on its own, while mixed training with large-scale synthetic depth provides an additional boost. As an ablation, removing RGB-D cross-attention leads to worse convergence and higher error, suggesting its effectiveness and importance in hinting the backbone to jointly consider RGB and depth cues on a given hand-like region.

Table 8: **Comparison with RGB-D methods on the HO-3D v2 benchmark.** We report STA-MPJPE, and PA-MPJPE (in cm; ↓). Top results are emphasized in top1, top2, and top3. AnyHandNet-D achieves the best overall results, and the real-only variant also surpasses prior RGB-D methods. Ablation on Xttn is also included. **Note that baselines are trained in a dataset-specific setting, while AnyHandNet-D targets cross-dataset generalization without benchmark-specific retraining.**

Input	Method	STA-MPJPE ↓	PA-MPJPE ↓
RGB-D	DiffHand [33]	-	0.93
D&PCL	IPNet [58]	2.01	0.87 ₁
RGB-D	Keypoint-Fusion [40]	1.87	0.94 ₃
RGB-D	AnyHandNet-D (Real-only)	1.20 ₁	0.89 ₁
	Eval w/ MoGe-2 [66] Estimated Depth	1.18 ₄	0.80 ₂
RGB-D	AnyHandNet-D (Real + AnyHand)	1.09 ₇	0.81 ₄
	Eval w/ MoGe-2 [66] Estimated Depth	1.06 ₈	0.79 ₃
	w/o RGB-D Cross Attention	1.21 ₁	0.85 ₅

Estimation with missing depth. In real-world applications, depth maps are not always available. We thus additionally evaluate our RGB-D model on HO-3D v2 by replacing the ground-truth depth maps with depth estimated from RGB using MoGe-v2 [66]. Surprisingly, this yields even better performance than using the ground-truth depth: STA-MPJPE improves from 1.09 cm to 1.06 cm, and PA-MPJPE improves from 0.81 cm to 0.79 cm. This is likely because ground-truth HO-3D depth maps are heavily quantized and contain missing values, whereas MoGe-v2 produces smoother, denser estimated depths that may better match the synthetic training distribution, where background depth is also MoGe-based.

C AnyHand dataset details

C.1 Synthetic data generation pipeline

We provide additional implementation details of AnyHand generation below.

FOV range. For each rendered view, the camera field of view is sampled uniformly from 30° to 40°.

Camera distance distribution. We sample the hand-camera distance from Gaussian distributions with means of 0.6m, 0.7m, and 1.0m, each with a standard deviation of 0.1m. This produces both close-up and distant views while keeping the hand at a realistic scale in the image.

Viewpoint sampling strategy. During rendering, the MANO hand is placed at the origin and a forearm mesh is attached at the wrist. We then sample a camera center at the chosen distance along a random 3D viewing direction, and orient the camera to look at the origin. This allows the hand-arm pair to be observed from diverse viewpoints in 3D space.

Lighting configuration. We use at most five lights per scene. For each scene, we randomly choose the number of lights from one to five. The ambient illumination is first set to roughly match the dominant color tone of the sampled background, as described in the main text, to maintain visual consistency between the rendered foreground and background. The remaining lights are randomly chosen from three types: point, directional, and spot. For each light, we randomize its placement and associated parameters, and assign either a vivid color with some probability or a near-white to slightly warm tone otherwise. We also randomly enable or disable shadows and vary the corresponding shadow ranges. This strategy increases illumination diversity while preserving overall scene coherence.

Processing GraspXL sequences. GraspXL [76] provides hand-object interaction sequences, where the hand approaches an object, establishes contact, and then releases it. For each sample in AnyHand-Interact, we first randomly select an object and then randomly choose one interaction sequence associated with that object. Since our goal is to capture occluded hands during active manipulation, we select the exact frame where the hand is grabbing on the object. This frame is then used for rendering the hand-object sample.

Table 9: **Summary of the AnyHand generation pipeline and dataset statistics.** The table summarizes the main design choices, rendering settings, and annotations used in AnyHand.

Component	Summary
Hand shapes	47,438, sampled from real datasets
Hand poses	On-the-fly samples from DPoser-Hand [42]
Hand textures	10,240 unique textures
Arm textures	254 textures
Backgrounds	1,270 indoor images and HDRI maps
Objects	>500K objects
Camera FOV	Randomly sampled from 30° to 40°
Camera distance	Mean 0.6m / 0.7m / 1.0m, std 0.1m
Lights	1 to 5 random lights per scene
Views per scene	2
AnyHand-Single	1.25M scenes, 2.5M images
AnyHand-Interact	2.05M scenes, 4.1M images
Stored annotations	RGB, depth, mask, bbox, camera, 3D pose/shape

The original GraspXL sequences are built on simplified, texture-free object meshes for efficient physics simulation. To improve visual realism, we retrieve the corresponding textured object meshes from Objaverse [14] and align them to the simplified meshes used in GraspXL. Specifically, we scale the textured mesh to match the size of the simulation mesh and use it during rendering, while preserving the hand pose and object placement from the selected GraspXL frame. The remaining rendering process, including camera sampling, lighting, background composition, and depth generation, follows the same pipeline as the single-hand branch.

In summary, the details of AnyHand are listed in Tab. 9.

C.2 Additional visualization of AnyHand

We provide additional qualitative examples from AnyHand in Fig. 9. For each sample, we show the rendered RGB image together with its corresponding 3D hand mesh, 2D joint annotations, and depth map. These examples illustrate the diversity of AnyHand in hand pose, camera viewpoint, appearance, and interaction setting, while highlighting its rich multi-modal annotations.

D Additional discussion on synthetic hand datasets

Overview. Tab. 1 summarizes representative synthetic datasets for 3D hand pose estimation. Here, we provide a more detailed comparison along several data-design axes: pose source, hand-shape and appearance diversity, object coverage, rendering variation, arm context, and depth supervision. This axis-based comparison is important because prior datasets often improve one specific aspect of synthetic hand data, but do not jointly provide million-scale RGB-D supervision, hand-object occlusion, explicit arm context, broad appearance variation, and diverse 3D annotations.

Early synthetic RGB-D supervision. RHD [80] is one of the earliest synthetic resources for RGB-based 3D hand pose estimation. It is built from 20 animated characters performing 39 actions, with a split over characters and actions to reduce train-test overlap. It provides RGB images, segmentation masks, 21-keypoint 2D/3D annotations, camera intrinsics, and depth maps, making it an important early reference for aligned synthetic supervision. Its main limitation is not annotation completeness, but scale and coverage: the dataset contains only 44k images and focuses on isolated hands without object-induced occlusion or explicit hand–forearm continuity. Therefore, RHD establishes the feasibility of synthetic RGB-D supervision, but it is not designed for the current data-scaling regime of foundation-style hand reconstruction.

Synthetic hand-object interaction. ObMan [24] introduced synthetic hand-object manipulation with 3D hand and object meshes, segmentation masks, and depth maps, and improves realism by rendering embodied hands rather than floating-hand synthesis. However, its diversity and scale remain limited by modern standards. ObMan contains 2,772 ShapeNet objects from 8 everyday categories and 21K grasp configurations, resulting in a relatively constrained space of object shapes, hand poses, and interaction patterns. Its hand appearance is based on 176 high-resolution hand



Figure 9: **Qualitative visualizations of AnyHand.** Additional examples from AnyHand, showing the rendered RGB images together with their corresponding 3D hand meshes, 2D joint annotations, and depth maps. These examples illustrate the diversity of poses, appearances, viewpoints, and interaction scenarios covered by the dataset.

textures from 20 subjects, and its rendering relies on LSUN/ImageNet background images and Blender compositing rather than diverse HDR environment illumination. In contrast, AnyHand-Interact provides million-scale hand-object supervision with broader pose, texture, object, occlusion, and lighting diversity.

Two-hand interaction data. Re:InterHand [47] focuses on reducing the appearance gap in two-hand interaction data. It captures interacting-hand motions from 10 subjects in a multi-camera studio with 170 synchronized cameras and 469 lights, and relights the tracked hands using 2,144 high-resolution environment maps. Its data therefore provides realistic two-hand poses and relighted appearances, but its scope is limited to hand–hand interaction. It does not include manipulated objects, object-induced occlusions, aligned depth maps, or arm context around the wrist. Its subject-level shape diversity is also bounded by the 10 captured subjects. Therefore, Re:InterHand addresses the realism of two-hand images, but does not target scalable hand-object reconstruction or RGB-D hand pose learning.

RenderIH [32] also targets two-hand interaction data. It starts from InterHand2.6M poses and augments them with random joint offsets to increase pose variation and reduce unrealistic hand–hand interpenetration. It renders 1.2M images with varied viewpoints, 300 HDR panoramas, and 30 base hand textures with additional shader-level skin-tone perturbations. However, because its pose generation is based on InterHand2.6M, its pose distribution remains closely tied to the original InterHand pose space. In addition, RenderIH does not model arm or wrist context: the hand meshes terminate at an open wrist boundary. This omission may affect training quality, as Zhao et al. [79] show that arm context is a considerable factor in reducing the synthetic-to-real gap for hand pose estimation. Therefore, while RenderIH improves the visual diversity of two-hand renderings, it does not address wrist-boundary artifacts or provide the broader arm context needed for robust real-image generalization. Thus, RenderIH is better viewed as an augmentation and rendering pipeline for two-hand interaction poses, rather than a newly generated large-scale hand dataset with a substantially different pose source.

In contrast, AnyHand targets a different setting. It provides scalable RGB-D supervision for both single-hand and hand-object scenes, with jointly rendered hand, forearm, object, and depth annotations. Compared with Re:InterHand and RenderIH, AnyHand explicitly covers hand-object manipulation, object-induced occlusion, aligned depth supervision, and arm context, while also using a larger appearance source with 10,240 diffused hand appearances and 254 forearm textures.

Relation to synthetic-to-real gap analysis. A closely related recent work is Zhao et al. [79], which studies the synthetic-to-real gap in 3D hand pose estimation using benchmark-matched synthetic counterparts. Unlike the datasets discussed above, their work is primarily a controlled analysis framework rather than a large-scale training resource. They explicitly decompose factors such as hand texture, background, arm context, object occlusion, pose distribution, and skeleton topology. This objective is different from ours: Zhao et al. aim to minimize confounding factors and stay close to target benchmark distributions, whereas AnyHand aims to build a scalable synthetic data-generation pipeline that extends beyond the coverage of existing real datasets.

The difference is reflected in pose, appearance, and scale. Zhao et al. align their synthetic pose distribution with target real datasets such as FreiHAND and Dex-YCB by fitting NIMBLE to MANO annotations. Therefore, their synthetic data is largely derived from existing benchmark poses and can be viewed as benchmark-conditioned rendering and augmentation, rather than a new dataset with an independent pose source. They use NIMBLE-based hand textures generated from 38 photo-realistic hand texture assets and 669 HDRI scenes. By contrast, AnyHand samples poses on the fly from the DPoser-Hand diffusion prior trained on multiple real datasets, and uses Handy-based texture generation to obtain 10,240 unique hand appearances. In addition, AnyHand attaches textured forearm geometry using 254 SMPLitex body textures and renders hand–forearm geometry jointly in simulation, which avoids the boundary artifacts that can arise from compositional arm pasting.

Takeaway. Overall, prior synthetic hand datasets have made important progress along different axes: RHD establishes early aligned RGB-D supervision; ObMan introduces synthetic hand-object interaction with depth and segmentation; Re:InterHand improves realism for relighted two-hand interaction; RenderIH improves synthetic two-hand pose optimization and photorealistic rendering; and Zhao et al. provide a controlled analysis of the synthetic-to-real gap. AnyHand is designed to address a complementary but increasingly important gap: a simulation-native, million-scale RGB-D training resource that jointly covers diverse hand poses, large hand-shape and texture variation, explicit forearm context, large-scale object-induced occlusion, realistic rendering variation, and aligned geometric annotations.

E Compute resources

Our main training experiments were conducted on a GPU cluster. Unless otherwise specified, each main run used a single node with 8 NVIDIA H100 GPUs, 128 CPU threads from dual Intel Xeon Platinum 8350C CPUs, and approximately 1 TiB of system memory. Each full training run took approximately 5–7 days under this setup. We used the same evaluation pipeline for all reported checkpoints to ensure consistent comparison across baselines and variants.

F Limitations

While AnyHand provides large-scale and diverse RGB-D supervision for hand pose estimation, it also has several limitations that suggest useful directions for future work.

First, hand poses are sampled from DPoser-Hand [42], a learned diffusion prior trained on a finite collection of real datasets. Although we apply biomechanical filtering (BMC loss [73]) to remove implausible poses, some anatomically unlikely configurations may still pass through the filter. Future work may further improve pose sampling by incorporating stronger anatomical constraints or by adapting the pose prior to specific downstream domains.

Second, our composition pipeline estimates background depth and composites the rendered hand onto 2D background images, rather than placing the hand inside a full 3D scene. This prioritizes scalability but means the background depth near hand boundaries may not be perfectly consistent. In practice, this is acceptable because most hand pose estimation pipelines operate on tightly cropped hand regions, where the learning signal is dominated by the hand itself. Rendering complete 3D scenes at this scale would require substantially higher asset and computation costs, which would reduce the efficiency of large-scale dataset construction. A useful future direction is to develop an efficient scene-level rendering pipeline for large synthetic hand datasets.

Finally, AnyHand is built by integrating existing tools for pose sampling, texture generation, and rendering. Its primary contribution is as a large-scale data resource for the community. Beyond the dataset itself, we provide systematic studies on data scaling, real-synthetic mixing, and component design that can inform future synthetic data efforts for hand pose estimation. Leveraging AnyHand to train a unified foundation model for both RGB and RGB-D hand pose estimation is a promising direction that we leave for future exploration.

G Benchmark on AnyHand test set

We evaluate state-of-the-art RGB methods, including HaMeR [52] and WiLoR [55], as well as their variants trained with AnyHand (as mentioned in the RGB experiments), on AnyHand as a benchmark.

As shown in Tab. 10, the original HaMeR and WiLoR models exhibit non-trivial generalization to AnyHand, despite not being trained on this data distribution. However, adding AnyHand in training consistently improves performance across all four splits and across both pose and mesh metrics, suggesting that exposure to AnyHand during training substantially reduces the domain gap to the AnyHand test set.

The gains are particularly pronounced on the hand-object interaction splits (AnyHand-Interact). On AnyHand-Interact-EnvMap and AnyHand-Interact-Indoor, adding AnyHand in training reduces MPJPE from roughly 19–26 mm to about 5–6 mm, while raising F@5 from around 0.09–0.11 to about 0.56–0.61. These improvements correspond to large gains in both pose accuracy and mesh reconstruction quality under challenging interaction scenarios. On the single-hand splits (AnyHand-Single), the improvements are smaller but still consistent across metrics, demonstrating that the benefits of AnyHand extend beyond interaction-heavy cases. Overall, these results support the use of AnyHand both as a scalable training source and as a reference benchmark for future methods.

Table 10: **Benchmarking methods on AnyHand**. Best results in each sub-table are shown in **bold**. For F-scores, we report both pre-alignment and post-alignment values. These results serve as reference values for future methods utilizing AnyHand.

Method	Pose Metrics						Mesh Metrics			
	AUC _j ↑	MPJPE ↓	AUC _{j_pa} ↑	PA-MPJPE ↓	AUC _{j_sta} ↑	STA-MPJPE ↓	F@5 ↑	F@15 ↑	F-al@5 ↑	F-al@15 ↑
AnyHand-Single-EnvMap										
HaMeR	0.600	21.290	0.800	10.090	0.624	19.840	0.098	0.552	0.293	0.854
WiLoR	0.567	23.486	0.778	11.199	0.588	22.088	0.090	0.518	0.245	0.817
HaMeR w/ AnyHand	0.701	14.980	0.890	5.500	0.805	9.780	0.139	0.625	0.593	0.967
WiLoR w/ AnyHand	0.715	14.290	0.904	4.820	0.811	9.530	0.140	0.658	0.681	0.973
AnyHand-Single-Indoor										
HaMeR	0.573	23.020	0.788	10.690	0.596	21.460	0.093	0.523	0.265	0.836
WiLoR	0.542	25.355	0.763	11.974	0.557	24.171	0.086	0.496	0.217	0.791
HaMeR w/ AnyHand	0.755	12.400	0.857	7.160	0.792	10.560	0.191	0.744	0.455	0.929
WiLoR w/ AnyHand	0.731	13.780	0.821	9.020	0.752	12.740	0.182	0.712	0.346	0.881
AnyHand-Interact-EnvMap										
HaMeR	0.638	19.250	0.801	10.030	0.664	17.860	0.107	0.582	0.260	0.840
WiLoR	0.579	23.304	0.772	11.523	0.603	21.946	0.091	0.507	0.200	0.789
HaMeR w/ AnyHand	0.891	5.480	0.933	3.360	0.892	5.460	0.601	0.956	0.827	0.986
WiLoR w/ AnyHand	0.890	5.510	0.934	3.300	0.892	5.440	0.607	0.951	0.829	0.986
AnyHand-Interact-Indoor										
HaMeR	0.595	21.990	0.795	10.360	0.642	19.380	0.096	0.526	0.247	0.828
WiLoR	0.545	25.594	0.768	11.757	0.583	23.332	0.085	0.464	0.195	0.778
HaMeR w/ AnyHand	0.887	5.690	0.932	3.400	0.889	5.590	0.579	0.952	0.883	0.986
WiLoR w/ AnyHand	0.882	5.940	0.931	3.440	0.885	5.770	0.560	0.944	0.817	0.985