HUMAN-ROBOT ALIGNMENT FOR DEXTEROUS HAND RECOGNITION WITH GEOMETRY-ALIGNED SYNTHETIC PAIRS

Anonymous authors
Paper under double-blind review



Figure 1: **Generative approach for dexterous robot hand recognition.** Detection and pose estimation on synthetic robot hands (top) and real robot hands from RealDex (bottom). Human domain models (HaMeR, WiLoR) fail to recognize anthropomorphic robot hands due to the domain gap. Our method recognizes them without training on real robot data.

ABSTRACT

Anthropomorphic robot hands imitate human hand kinematics, but the appearance gap degrades the performance of SOTA human-hand detectors and pose estimators, while the scarcity of labeled robot-hand images limits finetuning and reduces accuracy for human hands. We present SynDex, which adapts pretrained human-hand models to robot hands without any real robot training images. SynDex renders depth and normal maps from annotated human hands and generates paired human-robot images from the same geometry. A vision-language model checks each candidate against its real reference, yielding 47K pixel-aligned human–robot pairs. A frozen human-domain teacher then supervises a student. Student features on each robot image regress teacher features on the paired human image, while a preservation term on real human images limits forgetting. SynDex improves the WiLoR's robot-hand detection AP_{50} from 9.55 to 38.45 on RealDex and from 2.93 to 60.66 on HRDexDB, and lowers the robot-hand PA-MPJPE of HaMeR and WiLoR by up to 1.85 mm, while keeping human-hand pose error within 0.4 mm of the original models on FreiHAND and HO3Dv2.

1 INTRODUCTION

Reliable visual perception of dexterous robot hands is crucial for emerging physical AI applications such as visual teleoperation (Qin et al., 2022), imitation learning from videos (Wang et al., 2024), and spatially aware human-robot interaction (Robinson et al., 2023). Proprioceptive sensing with static camera-to-robot calibration serves well in controlled environments, but it degrades under mechanical wear and sensor drift (Klingensmith et al., 2016; Hu et al., 2023), and in safety-critical interaction with humans where the internal software state may not reflect the actual physical state (Zhou et al., 2025a). Vision-based perception is therefore necessary for safe and resilient operation.

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.

Pose estimation and detection of parallel-jaw grippers have progressed substantially (Labbé et al., 2021; Goswami et al., 2025), but these techniques do not transfer to dexterous anthropomorphic hands. State-of-the-art robotic hands such as the Shadow Hand (Company, 2005) contain more than 20 actuated joints and degrees of freedom (DoF), a pose space orders of magnitude larger than the 4 to 6 DoF of conventional grippers (Chi et al., 2022). The resulting self-occlusion and visual ambiguity make direct supervised training impractical (Gao et al., 2024).

In contrast, the human hand domain offers highly accurate detectors (Li et al., 2022; Potamias et al., 2025) and pose estimators (Pavlakos et al., 2024; Potamias et al., 2025) trained on millions of annotated images (Zimmermann et al., 2019; Moon et al., 2020). However, reusing them for anthropomorphic robot hands faces two main challenges. (i) **Training data scarcity:** real dexterous robot data are difficult to scale as capture environments are confined to fixed laboratories (Lim et al., 2026), and 3D pose annotation requires costly multi-view capture (Richardson et al., 2025). (ii) **Human-robot domain gap:** human and robot hands differ in appearance, material, and surface geometry, so backbones trained on human images do not transfer zero-shot to robot pixels (Fig. 1). Render-based simulation lacks photorealism and scene diversity (Tremblay et al., 2018; Lu et al., 2023; Zhao et al., 2025). Diffusion-based generation provides photorealism but is prone to structural artifacts such as wrong finger counts or flipped palm orientation.

We address both challenges with SynDex, a vision backbone alignment framework that adapts pretrained human hand models without real robot training data. First, a multi-conditional diffusion model (Li et al., 2024) conditioned on depth and normal maps rendered from real human hand data generates human-robot image pairs conditioned on the same hand geometry. A VLM-driven filter then automatically curates each generated candidate against its real reference, yielding tens of thousands of curated pairs without per-image annotation or manual effort. Second, a feature alignment loss matches student features on the robot view to frozen teacher features on the paired human view, transferring the human domain prior onto robot pixels, while a knowledge preservation term anchored on real human images prevents the student from collapsing onto the synthetic data distribution. The framework requires zero real robot annotations and generalizes across robot topologies and morphologies, including the Shadow and Inspire hands.

Our primary contributions are as follows:

- We introduce a scalable paired data generation methodology that couples geometry-conditioned diffusion synthesis of pose-aligned human-robot image pairs with VLM-driven automatic filtering.
- We propose a cross-domain feature alignment strategy that maps robot features into the human feature space, enabling knowledge transfer across architectures and tasks without manual annotations.
- We demonstrate our framework on two downstream tasks. It improves anthropomorphic robot hand detection from 9.55 to 38.45 AP_{50} on RealDex and from 2.93 to 60.66 AP_{50} on HRDexDB, and improves pose estimation by up to 1.9 mm PA-MPJPE, while preserving human-domain accuracy.

2 Related Work

Domain adaptation (DA). Recent DA transfers a model trained on a labeled source domain to an unlabeled target domain via unsupervised (Du et al., 2024; Phan et al., 2024), source-free (Tang et al., 2024; Huang et al., 2025), test-time (Niu et al., 2024; Hoang et al., 2024), or foundation-model adaptation (Wei et al., 2024; Khanna et al., 2024). Closer to robotics, EgoBridge (Punamiya et al., 2025) and UOT (Cheng et al., 2025) align human and robot representations through optimal transport, HRAlign (Zhou et al., 2025b) adapts pretrained encoders with image-level contrastive alignment on paired demo videos, PCD (Wu et al., 2025a) employs training-free contrastive decoding, and SlotMIM (Wen et al., 2025) learns object-centric representations via masked image modeling. Our framework uses shared geometric conditioning to generate synthetic human-robot pairs with spatial correspondence, enabling dense feature alignment.

Synthetic data for hand vision. Render-based methods (Gao et al., 2022; Lin et al., 2021a; Moon et al., 2023; Zhao et al., 2025) synthesize geometrically accurate samples but lack photorealism.

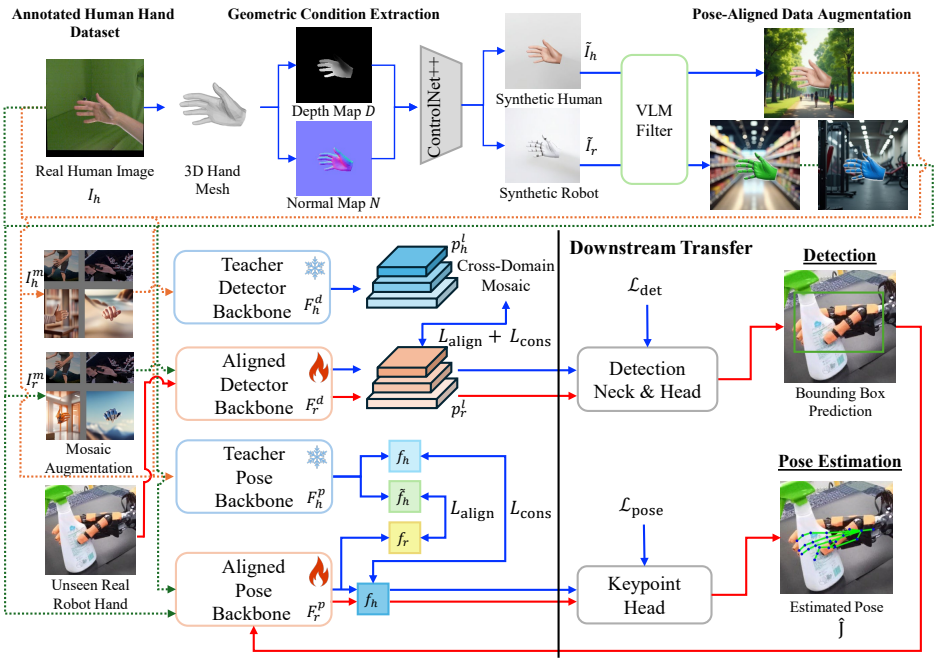

Figure 2: **Overview of our vision backbone alignment framework.** A geometry-conditioned diffusion pipeline (left) generates pose-aligned synthetic human ($\tilde{I}_h$) and robot ($\tilde{I}_r$) image pairs from real human images and 3D geometric cues. A frozen teacher backbone F_h and a trainable student backbone F_r are aligned through a feature alignment loss $\mathcal{L}_{\text{align}}$ on the synthetic pair, with a knowledge preservation term anchored on real human images I_h. The student is then transferred to hand detection and pose estimation via task-specific heads. Blue: training flow. Red: inference flow. Orange: (frozen) pretrained backbone input flow. Green: our (trainable) backbone input flow.

Generative methods based on GANs (Mueller et al., 2018; Baek et al., 2020) or diffusion models (Hu et al., 2022; Narasimhaswamy et al., 2024; Zhang et al., 2024; Chen et al., 2025) improve visual fidelity but remain confined to single-domain human hand generation. None supports geometry-aligned cross-embodiment pair generation, which we obtain by conditioning a multi-conditional diffusion model on a shared 3D geometric source. Recent work uses LLMs and VLMs for automatic data construction and refinement (Bensabath et al., 2026; Jeong et al., 2026), and we follow this principle for automatic quality control.

Gripper pose estimation and detection. Prior work focuses on parallel-jaw grippers with limited articulation. RoboPose (Labbé et al., 2021), HPE (Ban et al., 2024), and RoboPEPP (Goswami et al., 2025) achieve strong pose results on the DREAM benchmark (Lee et al., 2020), and synthetic datasets such as SimRGBD-Franka (Tian et al., 2024) enable render-based detection training (Tian et al., 2024; Wu et al., 2025b). However, dexterous hand recognition further requires recognizing complex kinematics under substantial self-occlusion and appearance variation by different viewpoints.

Human hand pose estimation. Monocular 3D hand pose estimation has advanced rapidly (Chen et al., 2022; Pavlakos et al., 2024; Potamias et al., 2025), built on the parametric hand model (Romero et al., 2017) and large-scale real datasets (Zimmermann et al., 2019; Hampali et al., 2020). GraphHR-Net (Zhang et al., 2025) explores whole-body estimation for teleoperation, yet accurate estimation for dexterous robot hands remains largely unaddressed. Our method transfers these models to unseen dexterous hands without any real robot data.

3 METHOD

Our framework has two components (Fig. 2). A geometry-conditioned generative pipeline with VLM-driven automatic filtering synthesizes and verifies pose-aligned human-robot image pairs without

manual annotation or curation. A feature alignment step then maps the robot view of each pair into the pretrained human feature space while preserving the human domain prior.

3.1 Geometry-Aligned Paired Data Synthesis

Paired human-robot image synthesis. Existing hand image generation methods (Zhang et al., 2024; Chen et al., 2025) condition primarily on 2D pose, which underdetermines the 3D geometry of the hand. Without depth or surface normal cues, they cannot distinguish the palm side from the dorsum and often generate incorrect orientations. We instead generate cross-domain pairs conditioned on shared 3D geometry to preserve articulation across human and robot views.

Given a real human hand image I_h with 3D pose annotation from a large-scale human dataset (Zimmermann et al., 2019), we render a depth map $D \in \mathbb{R}^{H\times W}$ and a normal map $N \in \mathbb{R}^{H\times W\times 3}$ from the underlying 3D hand geometry (Kolotouros et al., 2019) and feed both as conditioning signals to a multi-conditional diffusion model (Li et al., 2024). The model produces a synthetic human image $\tilde{I}_h$ and a synthetic robot image $\tilde{I}_r$ conditioned on the same depth and normal maps, with domain-specific differences in material, color, and surface appearance. The generative process introduces natural variation in lighting and background, improving downstream robustness. We provide prompt design and augmentation details in the appendix.

VLM-driven automatic filtering. Image generators are prone to generate artifacts such as extra or fused fingers and flipped palm orientation. For scalable synthetic data construction, we use an open-weight vision-language model (Qwen Team, 2026) to filter the images automatically. For each input sample and its geometric conditions, we synthesize ten candidates per embodiment and apply a two-stage filter. (i) *Rule-based verification*: the VLM inspects each candidate with its real reference image and rejects candidates whose palm-side orientation or visible finger count disagrees with the reference. This pairwise protocol reformulates an ambiguous filtering decision into clear palm/dorsum classification and finger counting tasks. (ii) *Best-sample selection*: surviving candidates compete in iterative mini-batch rounds based on pose agreement with the reference and photorealism. Prompts and other filtering details are provided in the appendix. The entire generation and filtering pipeline runs on a single 24GB GPU without human intervention, so it is effectively scalable. In total, the pipeline yields $\sim$47K synthetic image pairs.

The shared geometric source has two advantages for training. First, it provides spatial correspondence between $\tilde{I}_h$ and $\tilde{I}_r$, enabling our second contribution of pixel-level feature alignment learning. Second, it is annotation-free, since the ground truth poses of I_h remain valid for $\tilde{I}_h$.

3.2 Feature Alignment Learning

Let F_h denote a teacher backbone pretrained on a large-scale human dataset and F_r a trainable student initialized from the same weights. The teacher remains frozen throughout training. From a synthetic pair we extract backbone output features $\mathbf{f}_h = F_h(\tilde{I}_h) \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{f}_r = F_r(\tilde{I}_r) \in \mathbb{R}^{C\times H\times W}$. We define the alignment loss as the squared L2 distance, with stop-gradient $\mathrm{sg}()$ on the teacher:

$$\mathcal{L}_{\text{align}} = \left\|\mathbf{f}_r - \mathrm{sg}(\mathbf{f}_h)\right\|_F^2. \tag{1}$$

Shared geometric conditioning encourages corresponding spatial locations in $\mathbf{f}_h$ and $\mathbf{f}_r$ to represent the same hand regions across domains. Minimizing $\mathcal{L}_{\text{align}}$ therefore pushes the student toward a representation that factors out domain-specific appearance while preserving the human articulation prior.

Why L2 distance? A contrastive objective as in HRAlign (Zhou et al., 2025b) is a natural alternative. $\tilde{I}_h$ and $\tilde{I}_r$ establish spatial correspondence, under the same depth and normal conditions, so the pair is kinematically aligned. Under an isotropic Gaussian assumption with fixed variance for the feature residual $\mathbf{f}_r - \mathrm{sg}(\mathbf{f}_h)$, L2 loss is equivalent to the maximum likelihood criterion. Unlike contrastive loss, this objective requires neither a temperature parameter nor a choice of the number of negative samples. We verify this choice empirically in Tab. 7.

Knowledge preservation using real human features. Learning from synthetic robot pixels alone risks forgetting the human prior encoded in the initial weights. To preserve source domain performance, we anchor the student to the teacher on real human images I_h drawn from the human

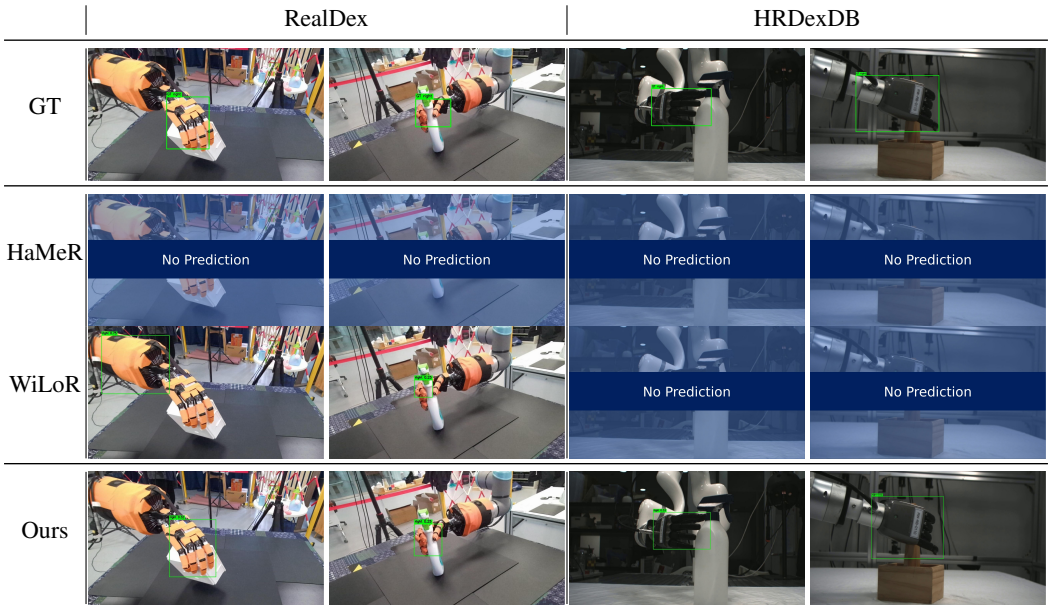

Figure 3: Qualitative robot-hand detection comparisons on RealDex (columns 1–2) and HRDexDB (columns 3–4). HaMeR and WiLoR often fail to detect the robot hands, while our method successfully localizes them across both datasets.

dataset:

$$\mathcal{L}_{\text{cons}} = \left\| F_r(I_h) - \text{sg}(F_h(I_h)) \right\|_F^2. \tag{2}$$

Analogous to source replay in continual learning, this term regularizes the student's features on the original human distribution to reduce forgetting during adaptation.

3.3 Downstream Task Transfer

For each downstream task, we train the student using feature alignment, knowledge preservation, and task supervision:

$$\mathcal{L} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{task}}\mathcal{L}_{\text{task}}. \tag{3}$$

For hand detection, we adapt the pretrained WiLoR detector in a dual-backbone teacher-student configuration. The frozen teacher serves as the human-domain reference; the student is initialized from the same pretrained weights and jointly optimized during adaptation; and multi-scale student features are aligned with the corresponding frozen teacher features at the P3 to P5 levels. $\mathcal{L}_{\text{task}}$ combines standard detection supervision, including bounding box regression, left/right hand classification, and a distribution focal loss, with auxiliary 2D keypoint supervision. For hand pose estimation, we use the 3D hand keypoint regression head of state-of-the-art pose estimators (Pavlakos et al., 2024; Potamias et al., 2025), and $\mathcal{L}_{\text{task}}$ is its standard 3D and 2D pose supervision with an adversarial term. At inference, we remove the teacher backbone and deploy only the student backbone with the task-specific head.

4 Experiments

Our experiments evaluate transfer to real dexterous robot hands without real robot training data, applicability to detection and pose estimation, and the effects of adaptation on human hand performance.

4.1 Experimental Setup

Synthetic paired data. We generate pose-aligned pairs $(\tilde{I}_h, \tilde{I}_r)$ using ControlNet++ (Li et al., 2024) conditioned on depth and normal maps rendered from real human datasets (e.g., FreiHAND (Zimmermann et al., 2019), HO3D (Hampali et al., 2020)). Each generated candidate is then screened by the VLM filter. We report generation, prompt design, and filtering details in the appendix.

Table 1: Hand detection results on RealDex and HRDexDB human/robot, reported as AP_{50}/mAP. Higher is better. HaMeR†: Detectron2 and ViTPose.

Method	RealDex	HRDexDB	
	Robot ↑	Human ↑	Robot ↑
HaMeR†	1.48/0.52	88.44/47.69	0.10/0.05
WiLoR	9.55/2.13	**99.25/60.25**	2.93/1.13
WiLoR + **Ours**	**38.45/9.87**	94.65/54.22	**60.66/23.07**

Table 2: Hand pose estimation results on RealDex and HRDexDB human/robot. Metric is PA-MPJPE (mm). Lower is better.

Method	RealDex	HRDexDB	
	Robot ↓	Human ↓	Robot ↓
HaMeR	15.08	6.24	15.32
WiLoR	14.42	5.64	15.96
HaMeR + **Ours**	**14.24**	**5.67**	**13.67**
WiLoR + **Ours**	**14.13**	**5.43**	**14.11**

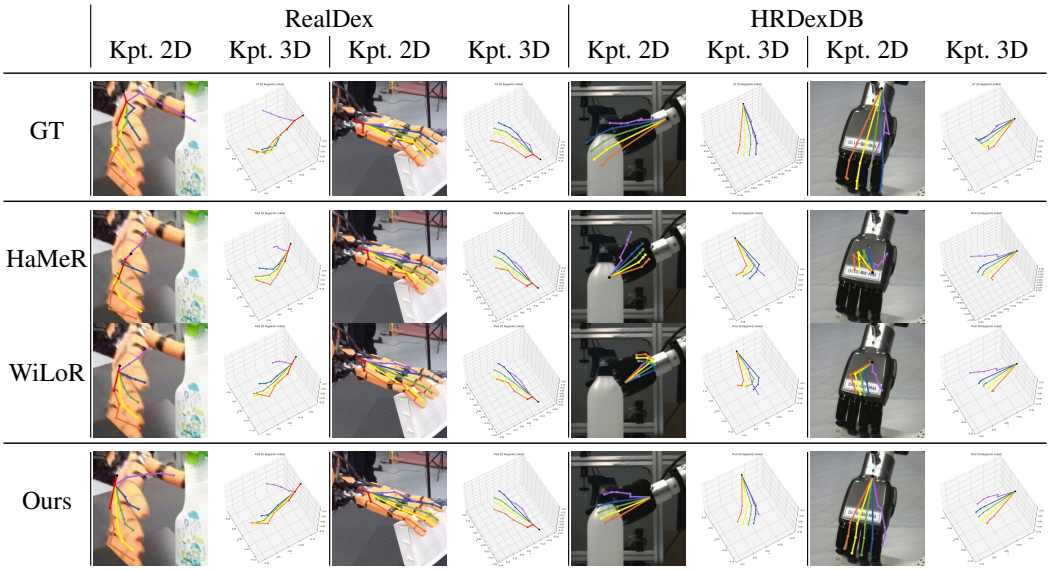

Figure 4: Qualitative robot hand pose estimation on RealDex (columns 1–4) and HRDexDB (columns 5–8). Each sample shows a 2D keypoint overlay and a 3D keypoint visualization.

Source domain (human) benchmarks. We evaluate hand detection on COCO-WholeBody (Jin et al., 2020) and Oxford-Hands (Mittal et al., 2011), following (Potamias et al., 2025), and hand pose estimation on FreiHAND (Zimmermann et al., 2019) and HO3Dv2 (Hampali et al., 2020) following (Pavlakos et al., 2024; Potamias et al., 2025).

Target domain (robot) benchmarks. We assess zero-shot transfer on RealDex (Liu et al., 2024) and HRDexDB (Lim et al., 2026). RealDex contains multi-view RGB-D sequences of a teleoperated dexterous hand grasping 52 objects, from which we sample 8,612 test images. HRDexDB captures 100 objects grasped by both human and dexterous robot hands with shared capture setups, so its human/robot splits measure domain-specific performance under matched conditions. We use 13,466 robot and 4,389 human test images. The appendix provides benchmark construction details.

Baselines. For human hand detection, we compare against MediaPipe (Zhang et al., 2020), OpenPose (Cao et al., 2019), ViTDet (Li et al., 2022), and the human domain detectors of HaMeR (Pavlakos et al., 2024) and WiLoR (Potamias et al., 2025). For human hand pose estimation, we compare against METRO (Lin et al., 2021b), Mesh Graphormer (Lin et al., 2021c), MobRecon (Chen et al., 2022), AMVUR (Jiang et al., 2023), HaMeR (Pavlakos et al., 2024), and WiLoR (Potamias et al., 2025). On the robot benchmarks, the WiLoR detector and the HaMeR and WiLoR pose estimators serve as zero-shot baselines. At evaluation, we retarget predicted hand poses to Shadow and Inspire hand kinematics. We adapt GMR (Araujo et al., 2025) to infer keypoints in a training-free manner, and then compute the pose metrics. We apply the same procedure to all methods.

Table 3: Comparison with the state of the art on FreiHAND (PA-MPJPE in mm, F@5, F@15). Color coding follows Tab. 5.

Method	PA-MPJPE ↓	F@5 ↑	F@15 ↑
METRO	6.5	0.731	0.984
Mesh Graphormer	5.9	0.764	0.986
MobRecon	5.7	0.784	0.986
AMVUR	6.2	0.767	0.987
HaMeR	6.0	0.785	0.990
WiLoR	**5.5**	**0.825**	**0.993**
HaMeR + **Ours**	6.1	0.778	0.990
WiLoR + **Ours**	5.9	0.797	0.991

Table 4: Comparison with the state of the art on HO3Dv2 (PA-MPJPE in mm, F@5, F@15). Color coding follows Tab. 5.

Method	PA-MPJPE ↓	F@5 ↑	F@15 ↑
METRO	10.4	0.484	0.946
MobRecon	9.2	0.538	0.957
AMVUR	8.3	0.608	0.965
HaMeR	7.7	0.635	0.980
WiLoR	**7.5**	**0.646**	**0.983**
HaMeR + **Ours**	7.8	0.628	0.980
WiLoR + **Ours**	**7.5**	0.643	0.982

Table 5: Human hand detection results on COCO-WholeBody and Oxford-Hands. **Green** and blue indicate the best and second best results. HaMeR†: Detectron2 and ViTPose.

Method	Size (Mb)	FPS	COCO-Whole		Oxford-Hands	
			AP_{50} ↑	mAP ↑	AP_{50} ↑	mAP ↑
MediaPipe (Zhang et al., 2020)	25	25	15.43	3.72	8.72	1.80
OpenPose (Cao et al., 2019)	141	29	37.05	9.06	20.74	4.41
ViTDet (Li et al., 2022)	1400	1	41.64	13.21	67.56	29.77
WiLoR (Potamias et al., 2025)	25	138	**62.48**	25.97	82.64	**48.98**
WiLoR + **Ours**	25	138	60.15	**32.56**	**82.94**	37.28

Metrics. For detection, we follow the COCO protocol and report mean Average Precision (mAP) alongside AP_{50}. For pose estimation, we follow HaMeR and WiLoR and report Procrustes-Aligned MPJPE (PA-MPJPE) in mm and F-scores at 5 mm and 15 mm thresholds (F@5, F@15).

4.2 Zero-Shot Comparison with the State of the Art

Robot hand detection. Table 1 summarizes detection on the RealDex and HRDexDB test sets. Despite training without any real robot image or annotation, our adapted detector lifts WiLoR from 9.55 to 38.45 AP_{50} on RealDex and from 2.93 to 60.66 AP_{50} on HRDexDB, while HaMeR† shows almost no transfer on either benchmark. On the matched HRDexDB human split, the adapted detector retains 94.65 AP_{50} compared with 99.25 for the original WiLoR, showing a smaller human-domain cost for larger robot-domain gains.

Figure 3 shows qualitative comparisons. On RealDex and HRDexDB, HaMeR and WiLoR either fail to detect the robot hands or misplace the boxes. Our adapted detector localizes the robot hand in all four examples, demonstrating improved cross-domain and cross-embodiment generalization.

Robot hand pose estimation. Table 2 compares zero-shot pose estimation with HaMeR and WiLoR on the RealDex test set and the paired human/robot splits of HRDexDB. The paired HRDexDB benchmark exposes the embodiment gap. On the same objects and capture setups, the zero-shot baselines are about 9 to 10 mm worse on robot hands (15.32 and 15.96 mm) than on human hands (6.24 and 5.64 mm) on HRDexDB-robot. Our alignment reduces the robot hand error of HaMeR from 15.08 to 14.24 mm on RealDex and from 15.32 to 13.67 mm on HRDexDB, and that of WiLoR from 14.42 to 14.13 mm and from 15.96 to 14.11 mm, while also improving human hand accuracy to 5.67 and 5.43 mm. Aligning robot features to the frozen human feature space thus narrows the embodiment gap while preserving the rich kinematic structure learned from human data, without any real robot annotation.

Figure 4 shows qualitative results. The baselines misestimate hand scale, confuse digits under self-occlusion, and degrade under motion blur, whereas our adapted models localize individual joints accurately across diverse configurations.

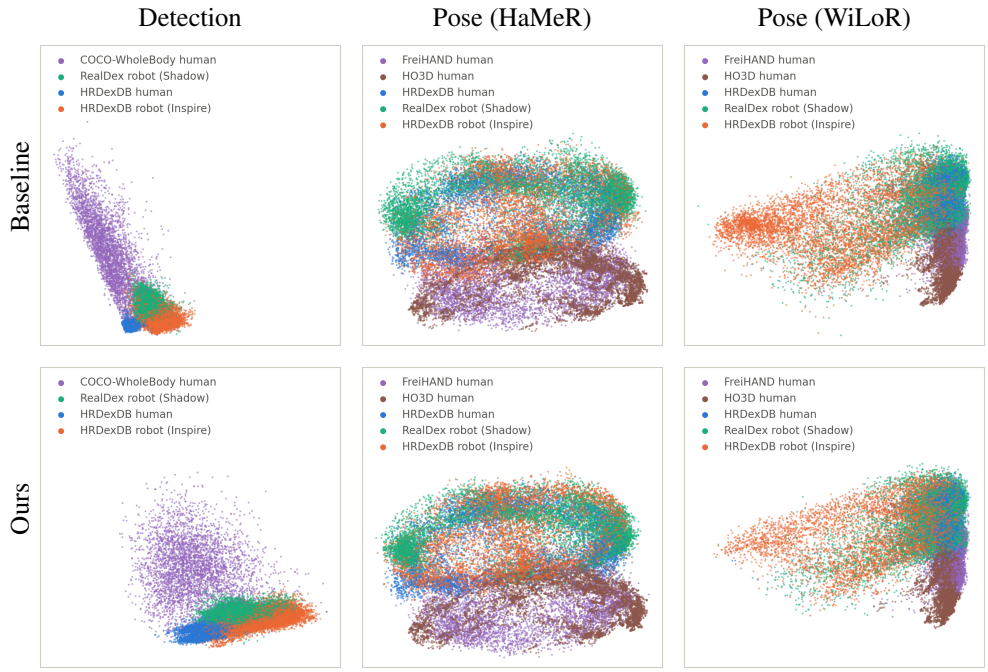

Figure 5: Feature distribution comparison with PCA. Our method pulls the distant robot features toward the human distribution.

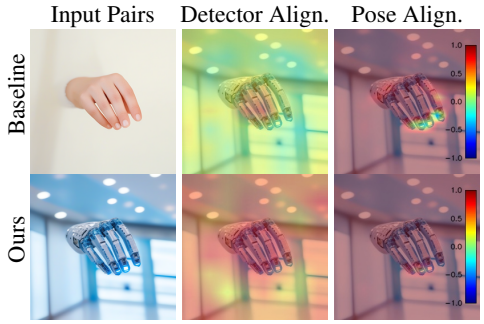

Figure 6: Feature alignment visualized by cosine similarity. Green: low alignment. Red: high alignment. Our method aligns better on the fingertips.

Table 6: Ablation study on RealDex val. Detection metric is AP_{50} and pose metric is PA-MPJPE (mm). Baseline: human model without adaptation. $\mathcal{L}_{\text{task}}$: naive task-only finetuning.

Variant	Detection WiLoR ↑	Pose HaMeR ↓	Pose WiLoR ↓
Baseline	7.64	18.4	17.0
+ $\mathcal{L}_{\text{task}}$	30.05	18.2	17.0
+ $\mathcal{L}_{\text{align}}$	36.40	17.7	16.8
+ $\mathcal{L}_{\text{cons}}$ (Full)	**40.92**	**17.1**	**16.4**

4.3 Source Domain Preservation

A standard concern with adapting a pretrained backbone is catastrophic forgetting of the source distribution, which we measure on four human benchmarks. On COCO-WholeBody (Tab. 5), our method decreases AP_{50} modestly by 2.33 points while improving mAP by 6.59 points, so localization under stricter IoU thresholds actually improves. On Oxford-Hands, although our method lowers mAP, our method slightly improves AP_{50}, maintaining human domain performance. On FreiHAND (Tab. 3), WiLoR+Ours reaches 5.9 mm PA-MPJPE versus 5.5 mm for the baseline WiLoR, and on HO3Dv2 (Tab. 4) the adapted model ties the baseline at 7.5 mm and matches its F@15. The preservation term $\mathcal{L}_{\text{cons}}$ protects the source representation, while the alignment loss shifts it toward the robot domain.

Table 7: Distance metric ablation on RealDex val. Detection metric: AP_{50}, pose metric: PA-MPJPE (mm).

Distance metric	AP_{50} ↑	PA-MPJPE ↓
L1	30.59	16.7
KL divergence	30.94	16.7
Contrastive	30.67	16.9
L2 (ours)	**40.92**	**16.4**

Table 8: Quantitative spatial correspondence between synthetic human-robot pairs. Mask IoU: SAM3 predictions with the GT box. Depth F-score: DA3-large predictions within GT hand mask.

Pair	Mask IoU ↑	Depth F@10 mm ↑
GT–real	73.8	49.8
GT–syn. human	75.0	50.4
GT–syn. robot	71.2	49.7
Syn. human–syn. robot	**77.0**	**90.1**

4.4 Ablation Study

Table 6 cumulatively adds each loss term on the RealDex val. Task-only finetuning ($\mathcal{L}_{\text{task}}$) raises detection AP_{50} from 7.64 to 30.05 but changes little on pose (0.0/0.2 mm decrease). Adding $\mathcal{L}_{\text{align}}$ increases AP_{50} by 6.35 points to 36.40, and $\mathcal{L}_{\text{cons}}$ by a further 4.52 points to 40.92, a total gain of 10.87 points over $\mathcal{L}_{\text{task}}$. For pose, the two terms reduce PA-MPJPE by 1.1 mm on HaMeR and by 0.6 mm on WiLoR, and each term lowers the error on both estimators. Therefore, alignment losses improve pose accuracy, whereas task-only finetuning suffers from the domain gap.

In Tab. 7, we compare four candidate distance metrics: L1, L2, KL divergence, and contrastive InfoNCE as in HRAlign (Zhou et al., 2025b). L2 performs best on both tasks, supporting its adoption as our dense alignment criterion.

The appendix provides generalization results on other robots such as Allegro Hand, loss-weight ablations, failure mode analysis, and feature activation analyses.

4.5 Feature Alignment Analysis

To validate the correspondence fidelity between the synthetic human-robot pair images, we provide quantitative measures in Tab. 8. We infer hand masks with SAM3 (Carion et al., 2026) prompted by the GT box and report mask IoU, and we infer depth maps with DA3-large (Lin et al., 2025) and report depth F@10mm within the GT hand mask. The GT-real pair provides the model accuracy on real images, and both GT-synthetic pairs reach similar scores. Therefore, our generation pipeline does not introduce meaningful geometry degradation. The synthetic human-robot pair shows the highest match. This within-pair correspondence agrees with our feature alignment framework.

In Fig. 5, we visualize the human-robot feature alignment across benchmarks using PCA. Many robot features appear far from the human distribution in the baseline PCA projections with severe skewness, whereas our method shows better alignment between the domains.

Figure 6 compares cosine similarity maps of the baseline and our method on both tasks. The baseline exhibits weak and poorly localized similarity, whereas our method concentrates coherent responses on the foreground robot hand while suppressing background clutter, indicating that the alignment loss directs attention toward the hand region and encourages domain-invariant representations.

5 Conclusion

We presented SynDex, a vision backbone alignment framework that transfers a human hand prior to dexterous robot hands with zero real robot data. A geometry-conditioned diffusion pipeline with VLM-driven automatic filtering yields pose-aligned human-robot pairs at scale, and these pairs supervise a teacher-student feature alignment regularized by human prior preservation. Across multiple backbones and two tasks, the framework improves detection and pose estimation performance across benchmarks with different robot hands, including Shadow and Inspire, while preserving human source accuracy.

Limitations and future work. Our method assumes anthropomorphic hands such as Shadow, Allegro, and Inspire. Generalization to parallel-jaw grippers and integration of teleoperation and imitation learning are natural directions for future work.

AI USE STATEMENT

Generative models are part of our research pipeline. We used a depth- and normal-conditioned diffusion model (SDXL with a multi-condition ControlNet) to generate the synthetic human and robot hand images, FLUX.1 Kontext to edit robot-hand color and background, and the open-weight VLM Qwen3.5-27B to filter and select generated samples (Sec. 3.1, App. A.1–A.5). We used LLM-based tools to edit the manuscript for grammar and readability. We did not use any generative AI to propose the method, design experiments, or interpret results. We have reviewed all AI-assisted work and take responsibility for the final content of this paper.

ETHICS STATEMENT

We have read the ICLR Code of Ethics and address it as follows.

Contribute to Society and to Human Well-being. The proposed method adapts pretrained human models to dexterous hands without real robot data. Reliable perception of robot hands improves visual teleoperation, imitation learning from video, and safe human-robot interaction.

Uphold High Standards of Scientific Excellence. We evaluate all methods on the same benchmark with the same evaluation protocol. We also report accuracy on four human benchmarks, including results where the proposed method improves performance.

Avoid Harm. The proposed framework outputs hand bounding boxes and hand joint positions and does not recognize identity. Our method is intended to assist safety monitoring in robotics applications and avoid harm.

Be Honest, Trustworthy and Transparent. Section A.4 details the VLM filtering process, including the prompts and rejection rates. The AI Use Statement states all generative models in our framework. Section 5 states the scope of SynDex, which assumes anthropomorphic hands.

Be Fair and Take Action not to Discriminate. The proposed framework recognizes non-human robot hands and therefore does not include demographic attributes. Our method consists of open-source models that are inclusive and accessible.

Respect the Work Required to Produce New Ideas and Artifacts. We cite all evaluation benchmarks and datasets used to generate synthetic pairs. We use datasets and pretrained models that comply with the license agreements.

Respect Privacy. We do not collect new real human images. The diffusion model does not receive a real image as an input, so the synthetic hand images do not contain the identity of the source subjects. We run the VLM filtering locally, so we do not transmit any information externally.

Honour Confidentiality. We use no confidential data and developed it with publicly available data and models.

REPRODUCIBILITY STATEMENT

Section 3 defines the feature alignment loss, the knowledge preservation term, and the full training objective. Section A specifies the geometric conditions, prompts, dataset size, and evaluation benchmark setup. Section A.4 reports the specific VLM model and the full prompts. In Sec. A.7, we provide the details of the optimizer, learning rate, weight decay, batch size, iteration count, and hardware for each task. We report the loss-weight experiment in Tab. 10. The framework uses only open-source or open-weight models; no closed-source models are used. We include the anonymized source code in the supplementary material.

REFERENCES

Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3D hand poses interacting objects. In *CVPR*, pp. 6121–6131, 2020.

Shikun Ban, Juling Fan, Xiaoxuan Ma, Wentao Zhu, Yu Qiao, and Yizhou Wang. Real-time holistic robot pose estimation with unknown states. In *ECCV*, pp. 1–17. Springer, 2024.

Léore Bensabath, Mathis Petrovich, and Gül Varol. Text-driven 3D hand motion generation from sign language data. In *CVPR*, 2026.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE PAMI*, 43(1):172–186, 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 *ICLR*, volume 2026, pp. 138846–138923, 2026.

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 *CVPR*, 2025.

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*, 2022.

Shuo Cheng, Liqian Ma, Zhenyang Chen, Ajay Mandlekar, Caelan Garrett, and Danfei Xu. Generalizable domain adaptation for sim-and-real policy co-training. *arXiv preprint arXiv:2509.18631*, 2025.

I-Ting Chi, Teeranoot Chanthasopeephan, and Dung-An Wang. Design of a parallel gripper based on topology synthesis and evolutionary optimization. *Journal of Mechanisms and Robotics*, 14(2): 021008, 2022.

Shadow Robot Company. Shadowrobot. `https://www.shadowrobot.com/dexterous-hand-series/`, 2005.

Zhekai Du, Xinyao Li, Fengling Li, Ke Lu, Lei Zhu, and Jingjing Li. Domain-agnostic mutual prompting for unsupervised domain adaptation. In *CVPR*, pp. 23375–23384, 2024.

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, 2022.

Kun Gao, Haoyang Zhang, Xiaolong Liu, Xinyi Wang, Liang Xie, Bowen Ji, Ye Yan, and Erwei Yin. Challenges and solutions for vision-based hand gesture interpretation: A review. *Computer Vision and Image Understanding*, 248:104095, 2024.

Raktim Gautam Goswami, Prashanth Krishnamurthy, Yann LeCun, and Farshad Khorrami. Robopepp: Vision-based robot pose and joint angle estimation through embedding predictive pre-training. In *CVPR*, pp. 6930–6939, 2025.

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

Trung-Hieu Hoang, Duc Minh Vo, and Minh N. Do. Persistent test-time adaptation in recurring testing scenarios. In *NeurIPS*, 2024.

Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. *NeurIPS*, 35:23805–23817, 2022.

Junlei Hu, Dominic Jones, and Pietro Valdastri. Coordinate calibration of a dual-arm robot system by visual tool tracking. In *ICRA*, pp. 11468–11473. IEEE, 2023.

Yuyang Huang, Yabo Chen, Junyu Zhou, Wenrui Dai, Xiaopeng Zhang, Junni Zou, Hongkai Xiong, and Qi Tian. Diffusion-driven progressive target manipulation for source-free domain adaptation. *arXiv preprint arXiv:2510.25279*, 2025.

Uyoung Jeong, Yihalem Yimolal Tiruneh, Hyung Jin Chang, Seungryul Baek, and Kwang In Kim. Thom: Generating physically plausible hand-object meshes from text. In *CVPR Findings*, pp. 3653–3664, June 2026.

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.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, pp. 196–214. Springer, 2020.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Samar Khanna, Medhanie Irgau, David B Lobell, and Stefano Ermon. Explora: Parameter-efficient extended pre-training to adapt vision transformers under domain shifts. *arXiv preprint arXiv:2406.10973*, 2024.

Matthew Klingensmith, Siddartha S Sirinivasa, and Michael Kaess. Articulated robot motion for simultaneous localization and mapping (arm-slam). *RA-L*, 1(2):1156–1163, 2016.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, pp. 2252–2261, 2019.

Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic. Single-view robot pose and joint angle estimation via render & compare. In *CVPR*, 2021.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Timothy E Lee, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Oliver Kroemer, Dieter Fox, and Stan Birchfield. Camera-to-robot pose estimation from a single image. In *ICRA*, pp. 9426–9432. IEEE, 2020.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *ECCV*, pp. 129–147. Springer, 2024.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, pp. 280–296. Springer, 2022.

Jongbin Lim, Taeyun Ha, Mingi Choi, Jisoo Kim, Byungjun Kim, Subin Jeon, and Hanbyul Joo. HRDexDB: A large-scale dataset of dexterous human and robotic hand grasps. *arXiv preprint arXiv:2604.14944*, 2026.

Fanqing Lin, Connor Wilhelm, and Tony Martinez. Two-hand global 3D pose estimation using monocular RGB. In *WACV*, 2021a.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021b.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021c.

Yumeng Liu, Yaxun Yang, Youzhuo Wang, Xiaofei Wu, Jiamin Wang, Yichen Yao, Sören Schwertfeger, Sibei Yang, Wenping Wang, Jingyi Yu, et al. Realdex: Towards human-like grasping for robotic dexterous hand. *arXiv preprint arXiv:2402.13853*, 2024.

Jingpei Lu, Florian Richter, and Michael C Yip. Markerless camera-to-robot pose estimation via self-supervised sim-to-real transfer. In *CVPR*, pp. 21296–21306, 2023.

Arpit Mittal, Andrew Zisserman, and Philip HS Torr. Hand detection using multiple proposals. *BMVC*, 2011.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, pp. 548–564. Springer, 2020.

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

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 *CVPR*, 2018.

Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. In *CVPR*, 2024.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes, 2024. URL https://arxiv.org/abs/2404.01650.

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.

Hoang Phan, Lam Tran, Quyen Tran, and Trung Le. Enhancing domain adaptation through prompt gradient alignment. *NeurIPS*, 37:45518–45551, 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, June 2025.

Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppili, Lawrence Y Zhu, Simar Kareer, Judy Hoffman, and Danfei Xu. Egobridge: Domain adaptation for generalizable imitation from egocentric human data. In *Human to Robot: Workshop on Sensorizing, Modeling, and Learning from Humans*, 2025.

Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *RA-L*, 7(4):10873–10881, 2022.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Benjamin A Richardson, Felix Grüninger, Lukas Mack, Joerg Stueckler, and Katherine J Kuchenbecker. Isyhand: A dexterous multi-finger robot hand with an articulated palm. *arXiv preprint arXiv:2509.26236*, 2025.

Nicole Robinson, Brendan Tidd, Dylan Campbell, Dana Kulić, and Peter Corke. Robotic vision for human-robot interaction and collaboration: A survey and systematic review. *ACM Transactions on Human-Robot Interaction*, 12(1):1–66, 2023.

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.

Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *CVPR*, pp. 23711–23720, 2024.

Yang Tian, Jiyao Zhang, Guowei Huang, Bin Wang, Ping Wang, Jiangmiao Pang, and Hao Dong. Robokeygen: robot pose and joint angles estimation via diffusion-based 3D keypoint generation. In *ICRA*, pp. 5375–5381. IEEE, 2024.

Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018.

Jun Wang, Yuzhe Qin, Kaiming Kuang, Yigit Korkmaz, Akhilan Gurumoorthy, Hao Su, and Xiaolong Wang. Cyberdemo: Augmenting simulated human demonstration for real-world dexterous manipulation. In *CVPR*, pp. 17952–17963, 2024.

Youzhuo Wang, Jiayi Ye, Chuyang Xiao, Yiming Zhong, Heng Tao, Hang Yu, Yumeng Liu, Jingyi Yu, and Yuexin Ma. Dexh2r: A benchmark for dynamic dexterous grasping in human-to-robot handover. *arXiv preprint arXiv:2506.23152*, 2025.

Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *CVPR*, pp. 28619–28630, 2024.

Xin Wen, Bingchen Zhao, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. A data-centric revisit of pre-trained vision models for robot learning. In *CVPR*, pp. 12143–12154, 2025.

Shihan Wu, Xu Luo, Ji Zhang, Junlin Xie, Jingkuan Song, Heng Tao Shen, and Lianli Gao. Policy contrastive decoding for robotic foundation models. *arXiv preprint arXiv:2505.13255*, 2025a.

Tianshu Wu, Jiyao Zhang, Shiqian Liang, Zhengxiao Han, and Hao Dong. Foundation feature-driven online end-effector pose estimation: A marker-free and learning-free approach. In *ICRA*, 2025b. doi: 10.1109/ICRA55743.2025.11128327.

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

Mingyu Zhang, Qing Gao, Yuanchuan Lai, Ye Zhang, Tao Chang, and Yulan Guo. 3D whole-body pose estimation using graph high-resolution network for humanoid robot teleoperation. In *ICRA*, pp. 1090–1096, 2025. doi: 10.1109/ICRA55743.2025.11128446.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

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.

Enshen Zhou, Qi Su, Cheng Chi, Zhizheng Zhang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and He Wang. Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection. In *CVPR*, pp. 6919–6929, 2025a.

Jiaming Zhou, Teli Ma, Kun-Yu Lin, Zifan Wang, Ronghe Qiu, and Junwei Liang. Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. In *CVPR*, 2025b.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russel, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019. URL "https://lmb.informatik.uni-freiburg.de/projects/freihand/".

Supplementary Material

Contents

A Implementation Details **16**

A.1 Geometry-Conditioned Diffusion 16

A.2 Diffusion Generation Process 17

A.3 Prompt Design 17

A.4 VLM-Driven Image Filtering 18

A.5 Robot Hand Appearance Diversity 18

A.6 Evaluation Benchmark Details 19

A.7 Model and Training Details 20

B Analysis **20**

B.1 Feature Activation Visualization 20

B.2 Pose Stability Analysis 20

B.3 Failure Mode Analysis 21

C Additional Experiments **25**

D Additional Qualitative Results **25**

D.1 Qualitative Detection Results 25

D.2 Qualitative Pose Estimation Results 25

D.3 Results on Other Robot Hands 25

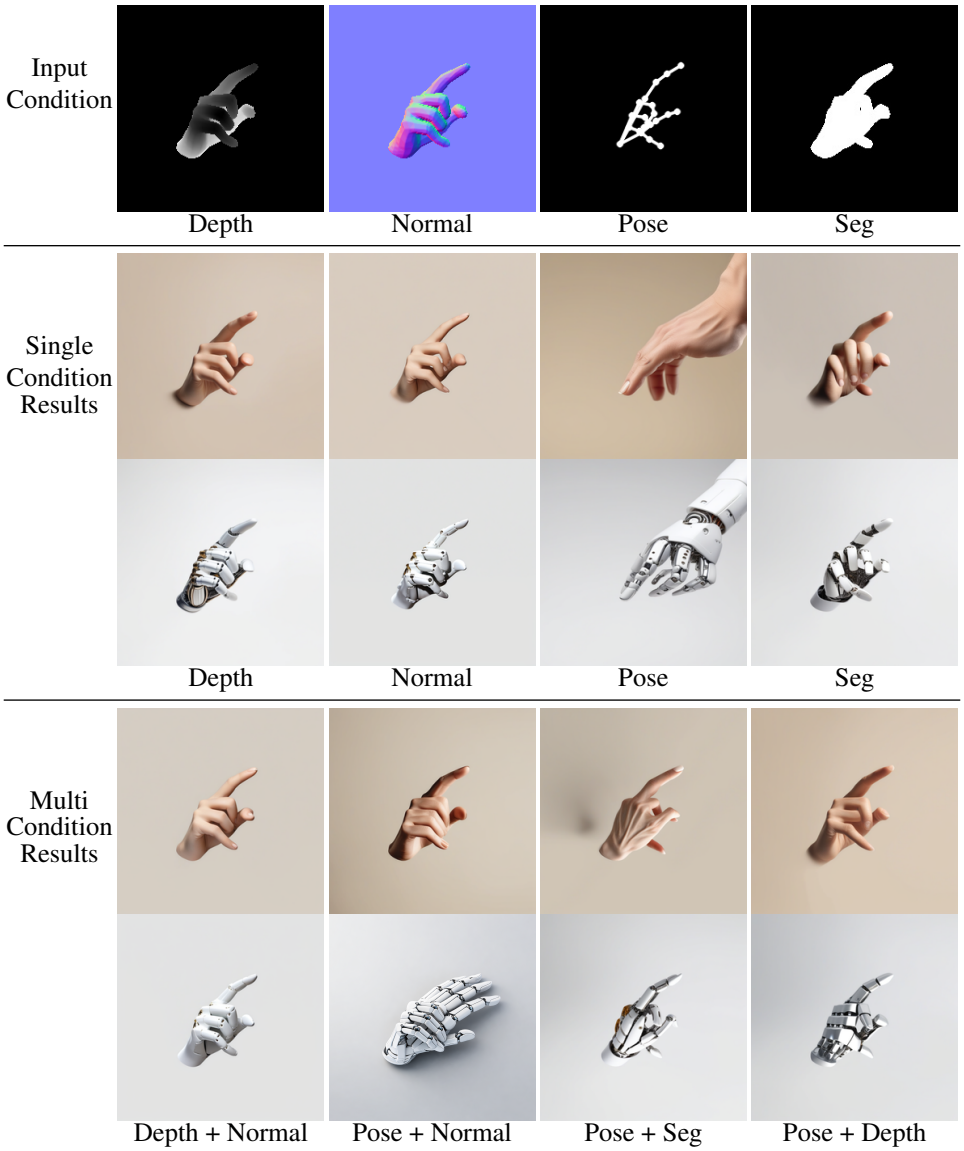

Figure 7: Comparison of image generation conditions.

A Implementation Details

A.1 Geometry-Conditioned Diffusion

We generate pose-aligned human-robot image pairs using a diffusion model conditioned on geometric cues derived from the ground-truth hand meshes (Romero et al., 2017). For each hand instance, we project the hand mesh using the ground-truth camera intrinsics to render geometric maps from the original viewpoint. In particular, we extract depth and normal maps, which provide dense geometric information about the hand pose and shape. These maps serve as conditioning inputs for a ControlNet-based Stable Diffusion XL (SDXL) pipeline (Li et al., 2024), enabling the diffusion model to preserve the underlying hand geometry while synthesizing photorealistic appearances.

Various geometric conditions can be used to guide diffusion models, including depth, normal, segmentation, and pose. Among these options, we employ depth and normal maps since they provide complementary geometric cues. As illustrated in Fig. 7, this combination provides stable geometric guidance and results in more consistent hand poses than alternative conditioning signals.



Figure 8: Generation results under different prompt designs. The first column shows results without background prompts ("no bg prompt"), while the remaining columns include background descriptions ("bg prompt").

To ensure high-quality samples, we filter the generated images with the VLM-driven procedure. For each GT input, we generate 10 candidate images per domain and retain at most one verified result, yielding 13,002 image pairs. We subsequently apply color and background augmentation to robot images using Flux.1-Kontext (Labs et al., 2025), prompting with four color prompts (red, orange, green, blue) and 33 diverse background descriptions to generate diverse colored robot hands. The final dataset comprises 13,002 synthetic human hand images and 47,299 synthetic robot hand images.

A.2 Diffusion Generation Process

We generate human-robot image pairs using a diffusion pipeline conditioned on geometric maps. The pipeline consists of the following generation steps:

- Render depth and normal maps from the hand mesh.
- Provide these maps as ControlNet conditioning inputs.
- Generate images using scene-specific prompts with randomized seeds.

The generation process is repeated for both human-hand and robot-hand prompts, producing geometry-aligned human and robot images with diverse appearances while preserving the same underlying hand pose.

A.3 Prompt Design

To encourage realistic rendering while preserving the underlying hand geometry, we design prompts that explicitly describe the hand appearance and the surrounding environment, as shown in Fig. 8.

For human hand synthesis, we use prompts such as:

> *"one realistic human right hand, photorealistic texture, natural pose, bright lighting, sharp focus on hands, background is a gym, slightly blurred equipment"*

For robot hand synthesis, we replace the semantic descriptor with a robot-hand description while preserving the same scene structure. Example prompts include:

> *"one realistic robot right hand, photorealistic texture, natural pose, bright lighting, sharp focus on hand, background is a gym, slightly blurred equipment"*

By keeping the environmental descriptions consistent across human and robot prompts, we encourage the diffusion model to generate image pairs that share similar background contexts but differ in hand appearance.

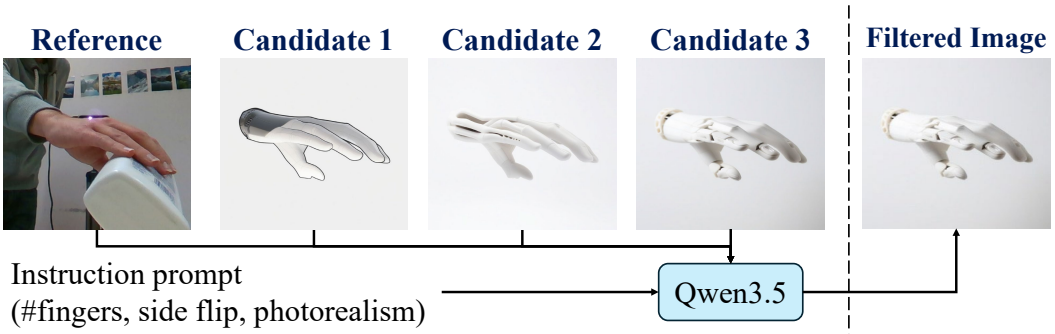

Figure 9: Stage 2 of the VLM automatic filtering process.

```
The FIRST image is a REAL reference photo of a right hand.
The SECOND image is a SYNTHETIC rendering of a right {hand_type} in the
same pose.
Report two attributes as JSON:
- side_align:  true if the same palm side is visible; otherwise false.
- n_fingers:  true if the number of visible fingers is the same.  Count
partial, fused, blurred, and phantom/extra fingers.
Respond with ONLY: {"side_align":  true/false, "n_fingers":
true/false}
```

Figure 10: Prompt for the stage 1 VLM filtering. {hand_type}: "human hand" or "robotic hand".

A.4 VLM-Driven Image Filtering

We automate quality control with Qwen3.5-27B (Qwen Team, 2026) with 4-bit quantization. For each geometric condition, the ten candidates per domain pass a two-stage filter. We accept a single finalist only if it passes a final photorealism gate. Frames with no acceptable candidate are discarded entirely, ensuring overall training data quality.

Stage 1: rule-based verification. The VLM receives each candidate side by side with its real reference image I_h and answers two verification queries: whether the visible palm-side orientation of the candidate agrees with the reference, and whether the visible finger count agrees (Fig. 10). Candidates failing either check are rejected. Judging relative to the reference avoids ambiguous absolute palm/dorsum decisions. 1.4% of the images are rejected at this stage.

Stage 2: best-sample selection. Illustration of stage 2 is provided in Fig. 9. Surviving candidates compete in small-group tournament rounds. In each round, the VLM views a group of candidates together with the reference and selects the sample with the best pose agreement and photorealism (Fig. 11). Winners advance until a single finalist remains, and the finalist is accepted only if it passes a final photorealism gate. If no candidate survives, the frame is discarded entirely. At this stage, we reject 95% of images, selecting only the best sample from every 10 candidates.

A.5 Robot Hand Appearance Diversity

To improve detector generalization across various robotic hands, we intentionally generate robot images with diverse visual appearances. While the geometric condition (depth and normal maps) fixes the underlying hand pose, the diffusion model produces robot hands with different visual characteristics such as color, material, texture, and surface details.

We achieve this diversity by varying the text prompts and random seeds during generation. The prompts describe different robot hand styles and environments, enabling the synthesis of robot hands with a wide range of visual properties. The resulting dataset contains robot hands with diverse appearances (colors and materials) while preserving the same underlying hand geometry.

```
The FIRST image is a REAL reference photo of a right hand.
The next {n} images are SYNTHETIC candidate renderings of a right
{hand_type} in the same pose, numbered 1 to {n} in the order shown.
Choose the SINGLE BEST candidate (except the reference photo), judged
by ALL of these criteria:
1.  Same side as the reference - palm side orientation (e.g., facing
to the viewer, facing to the side, invisible/self-occluded) must
match.
2.  Photorealistic - penalize (cartoon, comic, sketch, painting
styles) and (unrealistic texture, vague, blur).
Respond with ONLY a single integer from 1 to {n}, and nothing else.
```

Figure 11: Prompt for the stage 2 VLM filtering. {hand_type}: "human hand" or "robotic hand". {n}: 1–3.



Figure 12: **Robot hand appearance diversity.** Generated robot hands under identical geometric conditioning exhibit diverse colors, materials, and lighting conditions while preserving the same pose. This diversity encourages geometry-aware representations rather than reliance on specific visual appearances.

In addition to the appearance diversity introduced during generation, we apply color augmentation during training by randomly perturbing brightness, contrast, and saturation. This encourages the model to focus on geometry and structural cues rather than appearance-specific features.

Figure 12 shows examples of generated robot hands with different visual appearances under the identical geometric condition. Although the visual appearance varies substantially, the underlying pose remains identical due to the shared geometric conditioning. This diversity helps the detector learn representations that generalize across different robot hand designs and visual conditions.

A.6 Evaluation Benchmark Details

For source domain (human) benchmarks, we follow COCO-style evaluation.

RealDex (Liu et al., 2024) contains 2,630 multi-view RGB-D sequences of a teleoperated dexterous hand grasping 52 objects. We sample every 5th frame, yielding 8,592 validation and 8,612 test images.

HRDexDB (Lim et al., 2026) comprises 2.1K sequences covering 100 objects grasped by a human and by four dexterous hands (two Inspire and two Allegro variants). The human and robot splits share the same objects and capture setups, so they measure domain-specific performance under matched conditions. To construct the benchmark, we reserve 21 objects ($\sim$20%) as an object-level test split,

Table 9: Joint angle velocity evaluation of the pose estimation results. Metric: average velocity (rad/s).

Method	RealDex
HaMeR	0.23
WiLoR	0.26
HaMeR + Ours	**0.20**
WiLoR + Ours	**0.21**

keep 7 camera views selected by farthest-point sampling, and sample every 30th frame, yielding $13{,}466$ robot-hand (Inspire) and $4{,}389$ human-hand test images.

For detection, we derive a tight bounding box around the projected ground-truth hand geometry with a small padding. For pose estimation, we use the ground truth crop and 3D joints.

A.7 Model and Training Details

For the downstream detection experiments, we employ a single-stage detection pipeline (Khanam & Hussain, 2024). Similar to (Potamias et al., 2025), our detector regresses bounding boxes for the left and right hands separately. Following the baselines, images are resized to 480×480 resolution. The CSPDarknet backbone produces multi-scale feature maps, which are then processed through the feature pyramid neck module and the detection head.

For the downstream dexterous robot pose estimation, we follow the standard hand pose estimation pipelines (Pavlakos et al., 2024; Potamias et al., 2025). We resize the images to 256×256 resolution and feed them into the ViT-H backbone. The processed backbone features are used to regress 3D hand keypoints via the transformer decoder head. We additionally adapt the pose retargeting pipeline (Araujo et al., 2025) to obtain physically constrained 3D robot hand keypoints. We apply the same procedure to all the baselines to ensure fairness.

For vision backbone alignment training, we use a batch size of 8 for both real and synthetic data. We use the AdamW optimizer throughout the experiments. For the detection task, we use a single RTX 3090 GPU for the experiments. We apply an initial learning rate of 1.0×10^{-2}, a weight decay of 5.0×10^{-4}, and 2 workers. Total training takes 350,000 iterations and takes approximately 2 days. For the pose estimation task, we use a single A6000 GPU for the experiments. We use an initial learning rate of 1.0×10^{-5}, a weight decay of 1.0×10^{-4}, and 4 workers. Total training takes 250,000 iterations and takes approximately 2 days.

B Analysis

B.1 Feature Activation Visualization

We first examine the feature activations produced by the backbone on detection and pose estimation tasks. Given a human image $\tilde{I}_h$and a robot image $\tilde{I}_r$generated from the identical geometric conditioning, we compare the feature responses of the baseline and our aligned models.

Figure 13 visualizes the feature activation maps for both networks. Compared to the baseline, our method produces more concentrated responses for both the human and robot hands, indicating that the proposed alignment encourages the network to learn more consistent hand representations across domains.

B.2 Pose Stability Analysis

We validate the stability of the pose estimation pipeline in Tab. 9. To quantify temporal jitter, we compute the average joint-angle velocity over consecutive frames. Results show that our method, when applied to both HaMeR (Pavlakos et al., 2024) and WiLoR (Potamias et al., 2025) baselines, produces more stable pose trajectories on RealDex (Liu et al., 2024). Although our method is not

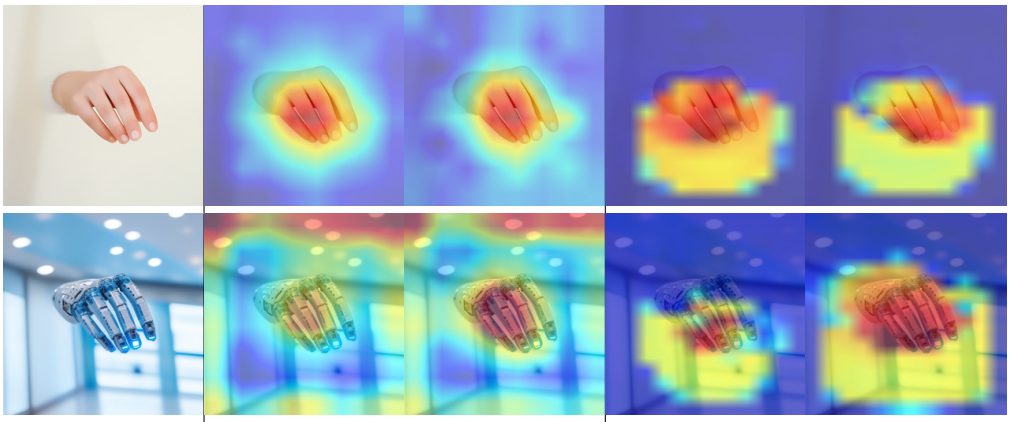

Figure 13: **Feature activation visualization.** Our method produces more focused responses on the robot hand region compared to the baseline, indicating improved cross-domain feature alignment.

Table 10: Loss weight ablations on RealDex-val pose. All metrics are in mm.

Alignment Loss Weight		**Task Loss Weight**	
Weight	PA-MPJPE ↓	Weight	PA-MPJPE ↓
2.0	16.9	1.0	17.5
1.0	16.9	0.1	18.0
0.5	17.4	0.01	16.9
0.1	17.2	0.001	17.1

explicitly trained to enforce temporal smoothness, it exhibits less pose jitter as compared to the baselines.

B.3 Failure Mode Analysis

We provide failure mode analysis in multiple aspects.

Occlusion. In Tab. 11, we place synthetic occluders inside the ground-truth box and vary the fraction of the area occluded. Under the heaviest occlusion we test, our detector still scores 8.47 on RealDex and 15.07 on HRDexDB, which is higher than what the baseline achieves without occlusion (4.00 and 2.93). At light occlusion, the relative loss is also smaller for ours.

Camera viewpoint. In Tabs. 12 and 13, we report every view separately. Hand scale is the diagonal of the ground-truth box divided by the diagonal of the input image. The baseline spans a factor of 8.6 across the RealDex views (2.79 to 23.87) and a factor of 36 across the HRDexDB views (0.24 to 8.68), while our method spans 2.2 and 3.0. Our detector improves every view and stays within a narrow band, so the improvement is not attributed from a specific viewpoint. On both benchmarks, the baseline is weakest on the view with the largest mean hand scale, RealDex cam1 at 0.193 and HRDexDB 23173282 at 0.381.

Hand scale. In Sec. B.3, we report the performance by the quartiles of hand scale. On RealDex, the baseline loses significant amount of accuracy from the smallest to the largest quartile (22.75 to 3.00), while ours maintains between 32.69 and 45.48. On HRDexDB, both models fall in the largest quartile, which indicates that the scale sensitivity is inherited from the baseline detector. Our weakest quartile (37.36) exceeds the baseline's strongest (3.18) by an order of magnitude.

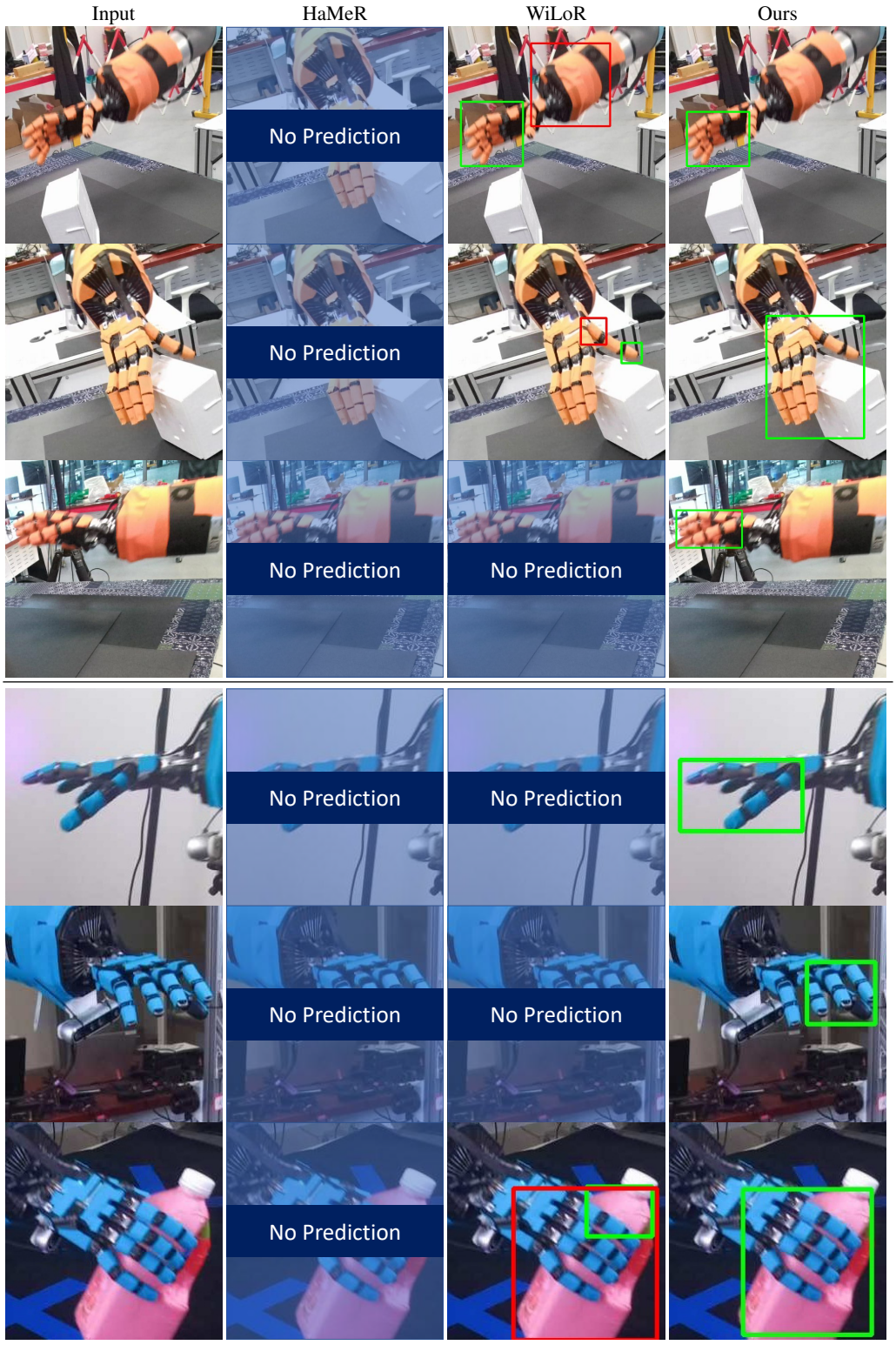

Figure 14: Additional qualitative results for the dexterous robot hand detection on RealDex (Rows 1–3) and DexH2R (Rows 4–6). The red box indicates the left hand, and the green box indicates the right hand.

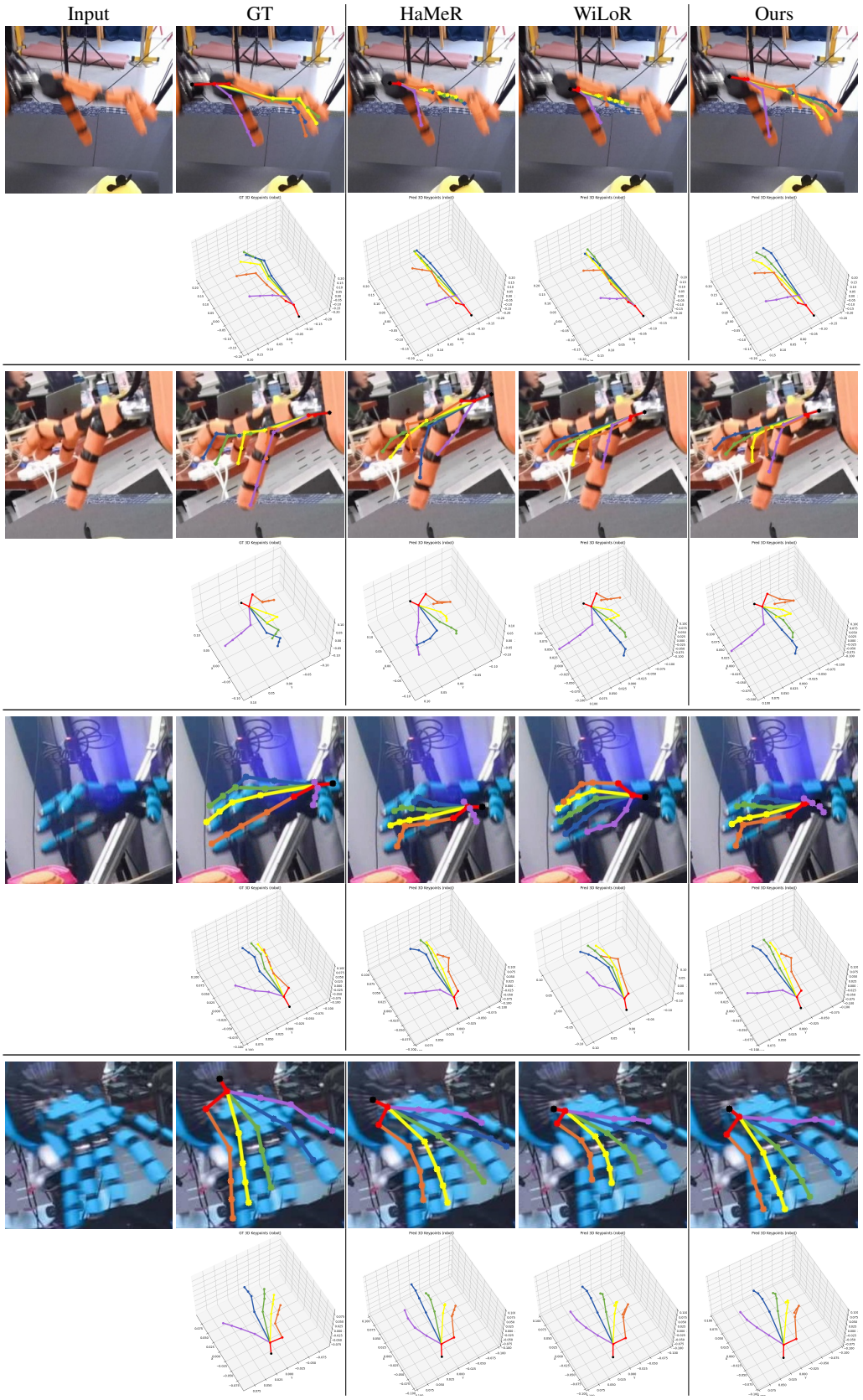

Figure 15: Additional pose estimation results on RealDex test (row 1–4) and DexH2R test (row 5–8).

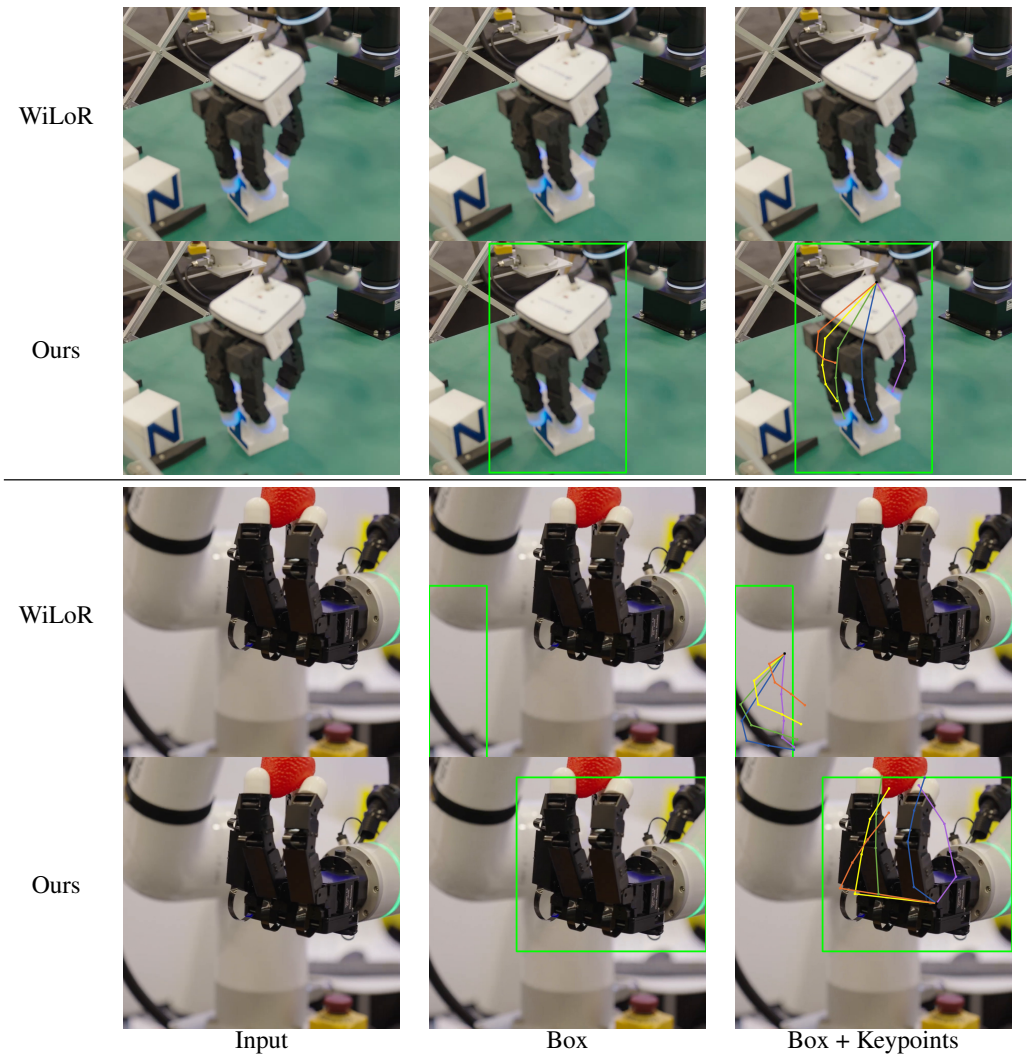

Figure 16: Qualitative results on Allegro Hand. WiLoR (top of each pair) fails detection, whereas our method (bottom of each pair) localizes the hand and predicts plausible keypoints in zero-shot.

Table 11: Occlusion analysis on RealDex and HRDexDB. Metric: AP_{50}.

Occlusion fraction	RealDex		HRDexDB	
	Baseline ↑	Ours ↑	Baseline ↑	Ours ↑
None	4.00	35.80	2.93	47.93
5–10%	2.68	27.83	1.01	38.64
10–30%	2.02	18.19	1.06	28.67
30–50%	0.82	8.47	0.74	15.07

Table 12: Camera viewpoint analysis on RealDex. Metric: AP_{50}.

Camera	Mean hand scale	Baseline ↑	Ours ↑
cam0	0.103	18.15	37.61
cam1	0.193	2.79	32.95
cam2	0.117	3.67	24.81
cam3	0.115	23.87	55.52

Table 13: Camera viewpoint analysis on HRDexDB. Metric: AP_{50}.

Camera	Mean hand scale	Baseline ↑	Ours ↑
22641005	0.363	0.38	48.91
22645026	0.204	1.37	53.10
22684737	0.296	0.31	51.75
23029839	0.250	2.00	65.36
23173282	0.381	0.24	21.70
23280285	0.285	8.68	58.12
23280286	0.307	8.29	55.54

C Additional Experiments

In Tab. 10, we present loss weight ablation results for the HaMeR baseline. We evaluate four different loss weights for $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{task}}$. For the alignment loss weight λ_{align}, we set the default value to 1.0. For the task loss weight λ_{task}, we set the default value to 0.01.

D Additional Qualitative Results

D.1 Qualitative Detection Results

We present additional qualitative results for dexterous robot hand detection in Fig. 14. The first three rows show results on the RealDex test set, and the last three rows show results on the DexH2R (Wang et al., 2025) test set. HaMeR frequently fails to detect robot hands, whereas WiLoR produces some predictions but often localizes incorrect regions. In contrast, our method produces more reliable detections across different viewpoints and hand configurations. The results indicate that the proposed human-to-robot feature alignment improves robustness in dexterous robot hand detection.

D.2 Qualitative Pose Estimation Results

In Fig. 15, we provide more qualitative results on the pose estimation task. RealDex test set results are shown in the first four rows, and DexH2R test set results are presented in the last four rows. HaMeR and WiLoR struggle to estimate the correct robot scale (rows 1–2) or precisely localize individual digits (rows 3–8). In several cases, the baselines confuse the thumb with the pinky finger (HaMeR in rows 3–4, WiLoR in rows 5–6). In contrast, our method demonstrates robustness in the challenging dexterous robot domain, maintaining accurate predictions under severe motion blur and self-occlusion.

D.3 Results on Other Robot Hands

We demonstrate qualitative results on other robot hands, specifically the Allegro Hand. In Fig. 16, we show comparative results on a YouTube video of the Allegro Hand. Since WiLoR is trained only on human hands, it fails to detect robot hands in the video. Although the Allegro Hand is a four-fingered robot hand completely unseen during training, our method successfully detects and estimates its pose. This demonstrates that our approach generalizes beyond five-fingered robot hands to other dexterous hand morphologies.

Table 14: Hand scale analysis on RealDex and HRDexDB. Quartile boundaries are 0.104, 0.121, and 0.146 on RealDex and 0.244, 0.296, and 0.343 on HRDexDB.

Hand scale quartile	RealDex		HRDexDB	
Q1, smallest	9.53	33.32	2.82	56.34
Q2	7.33	45.48	3.18	57.49
Q3	1.84	36.94	2.04	52.70
Q4, largest	1.26	32.69	0.76	37.36