# Beyond Fingertips: From Palm Priors to Tactile Refinement for Generalizable Dexterous Grasping

Zhen Yuan[2*] Ruoxiang Huang[3*] Junchun Yu[4] Yuran Wang[3] Guanqi He[5] Masayoshi Tomizuka[1]
Wei Zhan[1] Ruihai Wu[1†]

[1]UC Berkeley [2]University of Hong Kong [3]Peking University [4]Chinese University of Hong Hong [5]CMU

***Abstract*— Learning-based dexterous grasp synthesis has shown strong performance on training distributions, but often generalizes poorly to novel objects even when the generated hand poses satisfy kinematic and force-closure constraints. We argue that pose validity alone is insufficient for reliable grasp execution because it does not explicitly account for palm placement or the contact state established after execution. To address this limitation, we formulate dexterous grasping using three complementary components: pose generation, palm-anchor prediction, and tactile refinement. Our method first predicts a dense palm-anchor quality field over the object surface and uses the selected anchor to condition a hierarchical diffusion model for full-hand grasp generation. The resulting grasp is then refined using whole-hand tactile feedback from 16 hand links through a tactile graph encoder and an actor-critic refinement module. Experiments on in-distribution and unseen object categories show that our method achieves 79.30 % grasp success on unseen objects, outperforming the strongest baseline by 8.0 percentage points while maintaining a 0.17-point seen-to-unseen performance gap. The method further transfers zero-shot to real-world household objects without real-world grasp training.**

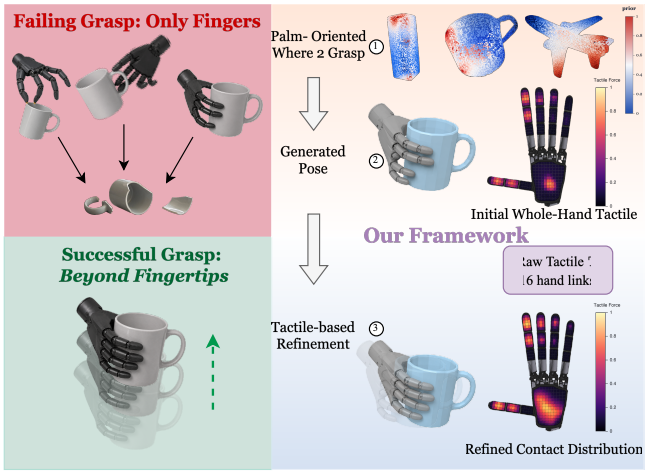



Fig. 1. **Beyond fingertips: palm-anchored, tactile-refined grasping.** Fingertip-only contact can yield unstable grasps, whereas palm engagement provides stable whole-hand support. Our framework predicts a palm-anchor prior, generates an initial grasp, and uses tactile feedback from 16 hand links to refine the contact distribution.

## I. Introduction

Learning-based dexterous grasp synthesis often generalizes poorly to novel objects, even when generated poses satisfy kinematic and force-closure constraints. We argue that reliable grasping additionally requires explicit palm placement and feedback on the contact state after execution. Accordingly, we decompose grasping into palm-anchor prediction, pose generation, and tactile refinement. Our method predicts a dense palm-anchor field, generates full-hand grasps with anchor-conditioned diffusion, and refines them using whole-hand tactile feedback from 16 hand links. On unseen object categories, our method achieves 79.30% grasp success, outperforming the strongest baseline by 8.0 points with only a 0.17-point seen-to-unseen gap. It further transfers zero-shot to real-world household objects without real-world grasp training.

Resolving this underspecification requires supplying the missing variables. We identify two palm-side signals absent from current pipelines: an explicit palm anchor on the object surface, and whole-hand tactile feedback at the candidate pose. Together these decompose grasping into three composable terms—pose, palm anchor, and tactile refinement—which we realize in a single pipeline for a 20-DoF dexterous hand. Where2Grasp, a per-point palm-anchor affordance regressor built on a transformer point encoder with a listwise ranking objective, predicts a continuous palm-anchor quality field on the object surface. The selected anchor conditions a palm-anchored hierarchical diffusion generator that emits the full 29-D hand configuration. A self-supervised tactile graph encoder spanning all 16 hand links summarizes the contact signature at each candidate, and a bounded-residual actor–critic refiner produces the final pose under a learned grasp-success critic. Realizing the tactile branch at scale required a new tactile simulator that enforces strict observation–dynamics decoupling, on top of which we curate a physics-validated whole-hand tactile dataset that we release. Across in-distribution objects, unseen-category test sets, and real-world household items, our method substantially improves grasp success and disturbance robustness over pose-only baselines, exceeding 75% grasp success on unseen-category objects and transferring zero-shot to real hardware.

**Why touch is essential.** Geometry can suggest *where* a grasp should be placed, but it cannot determine the contact state that actually emerges after execution. Small errors in pose, object geometry, or contact mechanics can change which links engage the object and whether the resulting support is stable. Whole-hand tactile feedback therefore plays a complementary role to the palm prior: the prior proposes a geometrically plausible grasp, while touch reveals the realized contact distribution and enables corrective action.

In this sense, tactile sensing is not merely an additional perception modality, but the feedback signal that closes the loop between grasp generation and physical interaction.

## II. Method

### A. Pipeline Overview

We resolve grasp underspecification by supplying the two missing palm-side variables in sequence. Given an object point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ with $N = 1024$, the pipeline operates in three stages: (i) Where2Grasp predicts a continuous palm-anchor quality field on the observed surface; (ii) a hierarchical diffusion generator samples a palm position conditioned on the anchor field and then a full hand configuration conditioned on the sampled palm; (iii) candidate grasps are physically verified in simulation, and an actor–critic refiner uses the resulting whole-hand tactile signature to apply a bounded residual correction. We parameterize the hand state as $q = (x, r, \theta) \in \mathbb{R}^{29}$ with palm position $x \in \mathbb{R}^3$, palm orientation $r \in \mathbb{R}^6$ in the continuous rotation representation of Zhou et al. [1], and joint angles $\theta \in \mathbb{R}^{20}$.

### B. Palm-Anchor Prediction

Where2Grasp maps the object point cloud to a per-point palm-anchor score $s_i \in [0, 1]$ estimating the suitability of each observed surface point as a palm anchor, with higher values indicating more suitable anchors. The architecture is a transformer point encoder followed by a feature-propagation decoder and a sigmoid head, instantiated with the Point-MAE backbone [2] so that publicly released ShapeNet pretraining can be re-used. We supervise it with an analytic target field $s_i^*$ computed from a closed-form local-geometry grasp-quality functional, combining a masked Huber regression loss with a Plackett–Luce listwise ranking loss [3]:

$$\mathcal{L}_{w2g} = \lambda_{\text{reg}}\,\mathcal{L}_{\text{reg}} + \lambda_{\text{rank}}\,\mathcal{L}_{\text{rank}}, \tag{1}$$

$$\mathcal{L}_{\text{rank}} = -\frac{1}{|\mathcal{M}|}\sum_{i=1}^{|\mathcal{M}|}\log\frac{\exp(s_{\pi(i)})}{\sum_{j\geq i}\exp(s_{\pi(j)})}, \tag{2}$$

where $\pi$ sorts the valid targets $\{s_i^* : i \in \mathcal{M}\}$ in descending order so that higher-quality anchors are ranked first, and $\mathcal{M}$ is the mask of surface points for which the analytic target is defined under the current rigid augmentation. The ranking term encourages the ordering of candidate anchors to remain consistent across object categories, while $\mathcal{L}_{\text{reg}}$ calibrates the absolute score values. At inference, the score field conditions the downstream sampler in two complementary ways: we draw palm anchors stochastically with a temperature-controlled softmax $\Pr(i) \propto \exp(\beta\, s_i)$ that favors high-quality points while retaining diversity, and we forward the per-point score as a feature channel to the palm denoiser to bias attention toward high-quality regions.

### C. Palm-Anchored Pose Generation

The pose generator is a two-stage conditional diffusion model that first samples a palm position and then a full hand configuration conditioned on the sampled palm. Both stages share a denoising architecture: a PointNet++ encoder [4] produces a global object feature $c_{\mathcal{P}}$, which is concatenated with stage-specific conditioning and fed to a FiLM-conditioned residual MLP denoiser. We adopt the $x_0$-prediction parameterization of Ho et al. [5], training the denoiser to recover the clean target from the noised iterate

$$x_t = \sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon \tag{3}$$

under an MSE loss; we found $x_0$-prediction more stable than $\epsilon$-prediction at the low action dimensionalities considered here.

In the second stage, the object point cloud is recentered at the sampled palm anchor $x$ to exploit translation equivariance, and the conditioning is augmented with an MLP embedding of $x$. Classifier-free guidance dropout at training [6] replaces the conditioning with a zero vector with probability $p_{\text{cfg}}$; at inference we combine conditional and unconditional predictions as

$$\hat{q}_0^{(t)} = \hat{q}_0^{(t)}(\emptyset) + w_{\text{cfg}}\left(\hat{q}_0^{(t)}(c) - \hat{q}_0^{(t)}(\emptyset)\right), \quad w_{\text{cfg}} \geq 1, \tag{4}$$

yielding $K \cdot M$ candidate grasps per object.

### D. Tactile Simulation and Refinement

Realizing the tactile branch at scale required infrastructure beyond existing tactile simulators [7, 8], which are either single-finger or conflate observation and dynamics channels. We address this with a tactile simulator that enforces strict observation–dynamics decoupling: sensors are read-only and never inject forces into the contact solver, so the same grasp produces identical physical outcomes whether sensors are mounted or not. Sixteen sensors covering the palm and finger links share a single force-field abstraction, and SDF queries across all parallel environments are batched into a single GPU call. Built on this simulator, we curate a physics-validated grasp dataset of over 1,000 objects with more than 500 diffusion-sampled grasps each, where successful and failed grasps carry whole-hand tactile signatures at matched density. Unlike prior contact-affordance corpora [9], our dataset jointly aligns generator-distribution grasps, physics-rollout stability labels, and dense whole-hand tactile records on the *same* samples.

Each candidate grasp is executed in simulation until physical settling, producing a whole-hand tactile observation $\mathcal{T}$ consisting of dense pressure grids on the palm and on each of the 15 finger links. A tactile graph encoder propagates information along the kinematic tree of the hand and pools the 16 link tokens into a tactile embedding $z_t$; the encoder is pretrained self-supervised with masked autoencoding [10], SimSiam contrastive [11], and force-alignment objectives, then frozen.

A critic head

$$Q(q \,|\, z_t, \mathcal{P}) = \sigma(g_{\text{crit}}(h)) \tag{5}$$

predicts grasp success and an actor head produces a bounded residual update

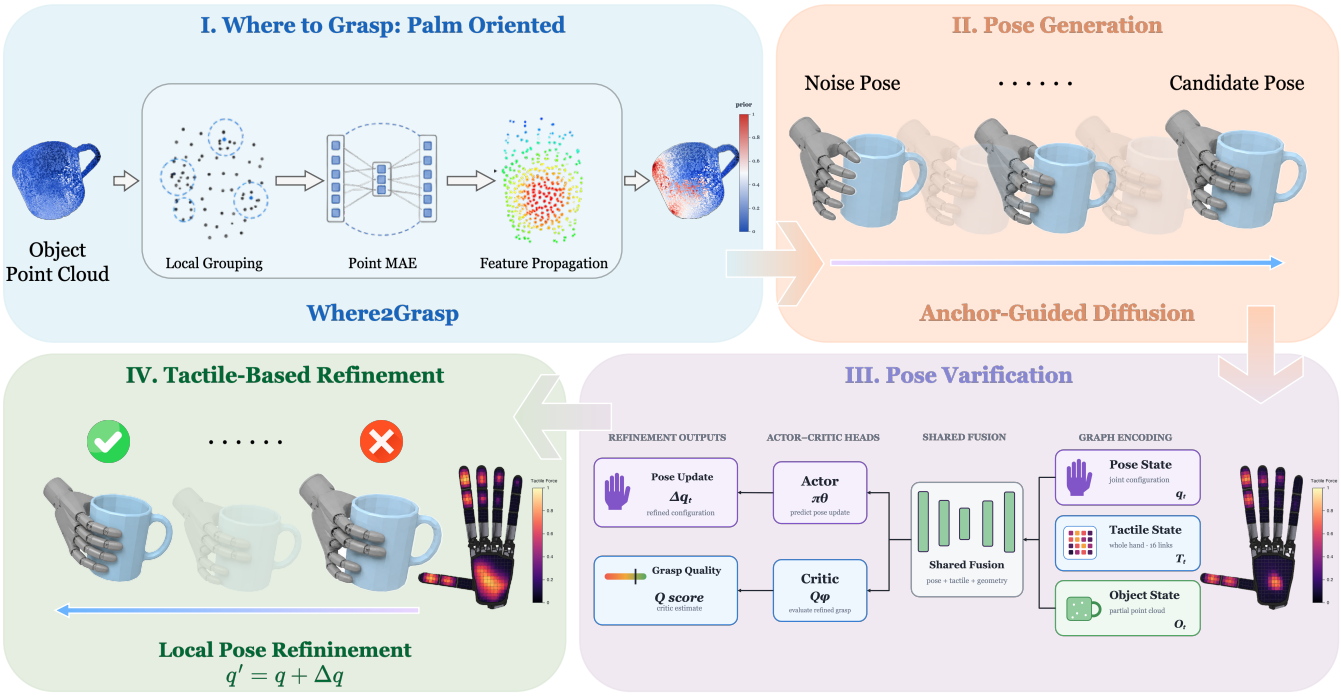



Fig. 2. **Overview of the proposed palm-anchored grasp generation and tactile refinement pipeline.** *I. Where to Grasp:* Given an object point cloud, Where2Grasp predicts a dense palm-anchor quality field and selects a high-quality grasping region. *II. Pose Generation:* The selected anchor conditions an anchor-guided diffusion model to generate an initial dexterous grasp pose. *III. Pose Verification:* The candidate grasp is physically verified to obtain whole-hand tactile feedback, which is jointly encoded with the pose and object states by the actor–critic refiner. *IV. Tactile-Based Refinement:* Bounded local updates, $q' = q + \Delta q$, progressively refine the candidate toward a stable full-hand grasp.

$$\Delta q = \tau \cdot \tanh(g_{\text{act}}(h)), \qquad q' = q + \Delta q. \tag{6}$$

Given the binary simulator outcome $y \in \{0, 1\}$, the actor is trained with

$$\mathcal{L}_{\text{actor}} = -w_Q\, Q(q') + w_c \,\|q' - q\|_\Lambda^2 + w_a \mathbb{1}\{y = 1\}\, \|q' - q\|_2^2, \tag{7}$$

where $\Lambda$ scales components by physical units. The third term is a *success-anchor regularizer* that discourages unnecessary movement on already-successful inputs, preventing collapse onto a single high-score mode of the critic; the actor gradient is detached from the shared backbone and critic head, so actor updates affect only the actor. The critic is trained jointly under a class-weighted BCE on the simulator label $y$. At test time the actor is applied iteratively in two phases (joint-only then full-pose); refinements are accepted only when they strictly improve an ensemble mean critic, and an unrefined no-op candidate is included to provide a principled refusal option.

## III. Experiment

### A. Setup

We evaluate on 1,000 objects from DexGraspNet [12] split at the category level into 800 seen and 200 unseen, executed on the 20-DoF hand under a settle-close-hold protocol: joints close until contact stabilizes, gravity is applied along $-\hat{z}$, and the grasp is held for two seconds, with success $S = 1$ if the object remains above its initial height. We compare against five methods spanning contact-map, diffusion, RL, and foundation-VLA paradigms: GenDexGrasp, GraspQP, DexDiffuser, UniDexGrasp++, and $\pi_{0.5}$, all re-trained or fine-tuned on the seen split under a matched evaluation budget. The primary metric is grasp success rate (SR), computed as the mean per-object pass rate.

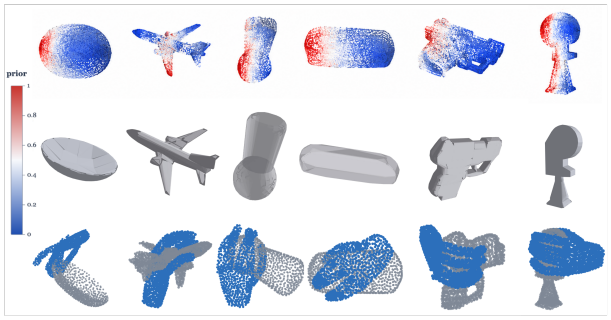



Fig. 3. **Qualitative simulation cases across diverse objects.** Six representative objects spanning seen and unseen categories from the 1,000-object DexGraspNet benchmark. *Top:* the per-point palm-anchor prior field predicted by Where2Grasp (colorbar, higher is better), with dark markers indicating sampled palm anchors. *Middle:* object geometry. *Bottom:* final tactile-refined grasp.

### B. Main Results

Table I reports SR on both splits. Our pipeline attains the highest success rate on both $\mathcal{D}_{\text{seen}}$ (79.47%) and $\mathcal{D}_{\text{unseen}}$ (79.30%), exceeding the strongest baseline (UniDexGrasp++) by 5.87 points on seen and 8.00 points on unseen. The seen–unseen gap is only 0.17 points—an order of magnitude smaller than that of every baseline, which we read as direct evidence of genuine cross-category generalization

rather than headline-number improvement. Qualitatively, the Where2Grasp prior concentrates on graspable regions across object categories, and the sampled palm anchors yield stable whole-hand grasps on both seen and unseen shapes.

**Table I.** Grasp success rate (%) on the simulation benchmark of 1,000 DexGraspNet objects (800 seen + 200 unseen categories).

| Method | Seen Objects | Unseen Objects |
|---|---|---|
| GenDexGrasp [13] | 39.56 | 43.56 |
| GraspQP [14] | 34.67 | 21.50 |
| DexDiffuser [15] | 25.66 | 21.64 |
| UniDexGrasp++ [16] | 73.60 | 71.30 |
| $\pi_{0.5}$ [17] | 52.24 | 40.28 |
| **Ours** | **79.47** | **79.30** |

### C. Ablations

Table II ablates the three core modules. Removing the palm-anchor prior drops unseen-category SR by 3.20 points, with a larger effect on unseen than seen, confirming its value for cross-category transfer. Removing tactile refinement causes the largest drop on both splits ($-5.80$ / $-6.22$ points), showing that closed-loop tactile feedback is the main source of robustness. Replacing the kinematic-tree graph encoder with an MLP over per-link tokens drops unseen SR by 4.00 points, validating the kinematic inductive bias.

Removing tactile refinement causes the largest degradation among all ablations. This suggests that touch is most valuable after a plausible grasp has already been generated: it reveals execution-time contact uncertainty that geometry alone cannot observe, allowing the policy to correct the realized physical interaction rather than merely refine a geometric pose.

**Table II.** Ablation results on the simulation benchmark. SR is reported on seen and unseen splits. Parenthesized arrows denote the change relative to the full pipeline.

| Variant | Seen Objects | Unseen Objects |
|---|---|---|
| Full pipeline | 79.47 | 79.30 |
| w/o Where2Grasp | 77.13 ↓2.34 | 76.10 ↓3.20 |
| w/o Tactile Refine | 73.67 ↓5.80 | 73.08 ↓6.22 |
| w/o Graph Encoder (MLP) | 76.66 ↓2.81 | 75.30 ↓4.00 |

### D. Real-World Transfer

We further evaluate the pipeline on a real dexterous hand in a zero-shot setting, without using any real-world grasping data for training. Given partial point clouds captured by an RGB-D camera, the model directly predicts grasp poses and executes them on out-of-distribution household objects. Because Where2Grasp evaluates palm graspability point-wise on the observed surface, occluded geometry does not enter the candidate set, eliminating the need for explicit shape completion at inference.

The generated grasps adapt to diverse unseen geometries and remain stable during real-world execution. As shown in Fig. 4, the corresponding contact patterns recorded by the 16 physical tactile sensors are physically consistent with those produced in simulation. These results indicate effective zero-shot transfer across the sim-to-real gap without re-calibration.




Fig. 4. **Zero-shot real-world transfer on unseen household objects.** Columns 1–6 correspond to six unseen objects. The *top* row shows Where2Grasp palm priors, the *middle* row shows zero-shot real-world execution, and the *bottom* row shows whole-hand tactile-force measurements from 16 physical sensors. The sim-trained pipeline transfers directly from partial RGB-D observations without real-world grasp training or explicit shape completion.

## IV. Conclusion

We formalize grasp underspecification, the gap between a kinematically valid pose and a physically successful grasp, and resolve it by decomposing a grasp into pose, palm anchor, and tactile refinement. Realized through a palm-anchor regressor, a palm-anchored hierarchical diffusion generator, and a tactile-conditioned actor–critic refiner trained against a whole-hand tactile simulator and physics-validated dataset that we release, the full pipeline attains over 75% success on unseen-category objects in simulation and transfers zero-shot to real hardware.

## REFERENCES


[1] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5745–5753.

[2] Y. Pang, W. Wang, F. E. H. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *Computer Vision – ECCV 2022*, ser. Lecture Notes in Computer Science, vol. 13662. Springer, 2022, pp. 604–621.

[3] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise approach to learning to rank: Theory and algorithm," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 1192–1199.

[4] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[6] J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022.

[7] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3930–3937, 2022.

[8] I. Akinola, J. Xu, J. Carius, D. Fox, and Y. Narang, "Tacsl: A library for visuotactile sensor simulation and learning," *IEEE Transactions on Robotics*, vol. 41, pp. 2645–2661, 2025.

[9] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8701–8711.

[10] Z. Wei, C. Wei, J. Mei, Y. Bai, Z. Wang, X. Li, H. Zhu, H. Wang, A. Yuille, Y. Zhou, and C. Xie, "Masked autoencoders are secretly efficient learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024, pp. 7986–7995.

[11] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 750–15 758.

[12] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," in *IEEE International Conference on Robotics and Automation*, 2023, pp. 11 359–11 366.

[13] P. Li, T. Liu, Y. Li, Y. Geng, Y. Zhu, Y. Yang, and S. Huang, "Gendexgrasp: Generalizable dexterous grasping," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8068–8074.

[14] R. Zurbrügg, A. Cramariuc, and M. Hutter, "Graspqp: Differentiable optimization of force closure for diverse and robust dexterous grasping," in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 305. PMLR, 2025, pp. 2583–2602.

[15] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "Dexdiffuser: Generating dexterous grasps with diffusion models," *IEEE Robotics and Automation Letters*, vol. 9, no. 12, pp. 11 834–11 840, 2024.

[16] W. Wan, H. Geng, Y. Liu, Z. Shan, Y. Yang, L. Yi, and H. Wang, "Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3891–3902.

[17] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 305. PMLR, 2025, pp. 17–40.