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

Zhen Yuan[1,*], Ruoxiang Huang[1,*], Junchuan Yu[1], Yuran Wang[1],
Guangqi He[2], Masayoshi Tomizuka[3], Wei Zhan[3], and Ruihai Wu[3,†]
[1]Peking University [2]Carnegie Mellon University [3]University of California, Berkeley
[*]Equal contribution. [†]Corresponding author.

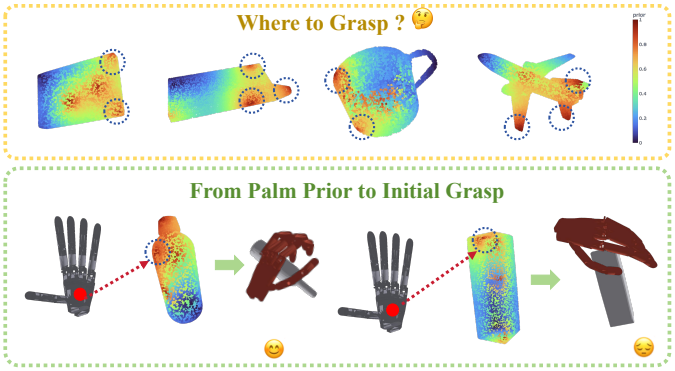

Fig. 1. **From palm priors to grasp.** Our key insight is that a grasp is not a pose: closing the pose-grasp gap requires committing to *where* on the object to place the palm. *Top (Where to Grasp?):* Where2Grasp regresses a per-point palm-anchor quality field from the object point cloud (colorbar, higher is better), with dashed circles marking high-scoring anchor candidates. *Bottom (From Palm Prior to Initial Grasp):* a selected palm anchor (red marker) conditions a palm-anchored diffusion generator that lifts the prior into an initial full-hand configuration.

***Abstract*—Learning-based dexterous grasp synthesis generalizes poorly to novel objects, with success rates collapsing even for kinematically valid, force-closure-compliant poses. We trace this to a structural blind spot: *a pose underspecifies a grasp*. Resolving this underspecification requires two palm-side signals absent from pose-only synthesis—an explicit palm anchor on the object surface, and whole-hand tactile feedback at the candidate pose. We decompose grasping for a 20-DoF hand into three composable terms (pose, palm anchor, tactile refinement) and instantiate each. Across in-distribution, unseen-category, and real-world objects, our method exceeds 75% grasp success on unseen categories, improving over the strongest baseline by 8.0 points with a seen–unseen gap of only 0.17 points—an order of magnitude smaller than that of any baseline we compare against.**

I. Introduction

Learning-based dexterous grasp synthesis has made rapid progress on training distributions [7, 14, 10], but generalization to novel objects remains brittle: success rates collapse sharply on out-of-distribution geometries, even when generated hand poses are themselves kinematically clean and force-closure compliant. We diagnose this fragility as a structural blind spot: current pipelines optimize for valid hand poses, but a pose underspecifies a grasp. Existing methods formulate the problem as pose synthesis under force-closure objectives [17], implicitly assuming pose validity entails physical success; yet a pose satisfying both constraints can still slip under gravity if its contact with the object is not properly anchored. We call this grasp underspecification: the specification "a kinematically valid, force-closure-compliant pose" is insufficient to determine a physically successful grasp.

Quantifying the underspecification. Three measurements on our DexGraspNet-based benchmark [11] support this diagnosis. First, the GraspQP force-closure energy correlates only weakly with downstream lift success (Spearman $\rho = -0.12$), so quasi-static pose quality alone is a poor predictor of physical outcome. Second, when grasps on a fixed object are projected onto a palm-anchor latent space, success rates vary sharply by region (83% in the best palm-anchor region vs. 14% in the worst), and forcing the worst palm anchor to a random one already lifts success from 57% to 72%, while picking the best anchor reaches 85% — a 13-point gain over random selection. Third, the best–worst gap across palm anchors has a median of 28 points per object, with 24% of objects showing a gap larger than 40 points. Pose quality alone is therefore insufficient; the palm commitment location on the object is itself a dominant source of grasp-success variance.

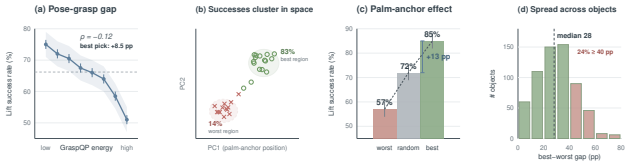

Fig. 2. **Failure analysis of the pose-grasp gap.** GraspQP energy is only weakly correlated with final lift success, while the latent palm-anchor analysis reveals strong regional variation in success rates, highlighting palm-anchor selection as a key source of dexterous grasp performance variation.

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.

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. [16], 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 [8] 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 [15]:

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

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

where π 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 [9] 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. [6], training the denoiser to recover the clean target from the noised iterate $x_t = \sqrt{\bar\alpha_t}\,x_0 + \sqrt{1-\bar\alpha_t}\,\epsilon$ under an MSE loss; we found x_0-prediction more stable than ϵ-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 [5] replaces the conditioning with a zero vector with probability p_{cfg}; at inference we combine conditional and unconditional predictions as

$$\hat q_0^{(t)} = \hat q_0^{(t)}(\emptyset) + w_{\text{cfg}}\left(\hat q_0^{(t)}(c) - \hat q_0^{(t)}(\emptyset)\right), \quad w_{\text{cfg}} \geq 1, \tag{3}$$

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 [12, 1], 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 [3], 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 [13], SimSiam contrastive [4], and force-alignment objectives, then

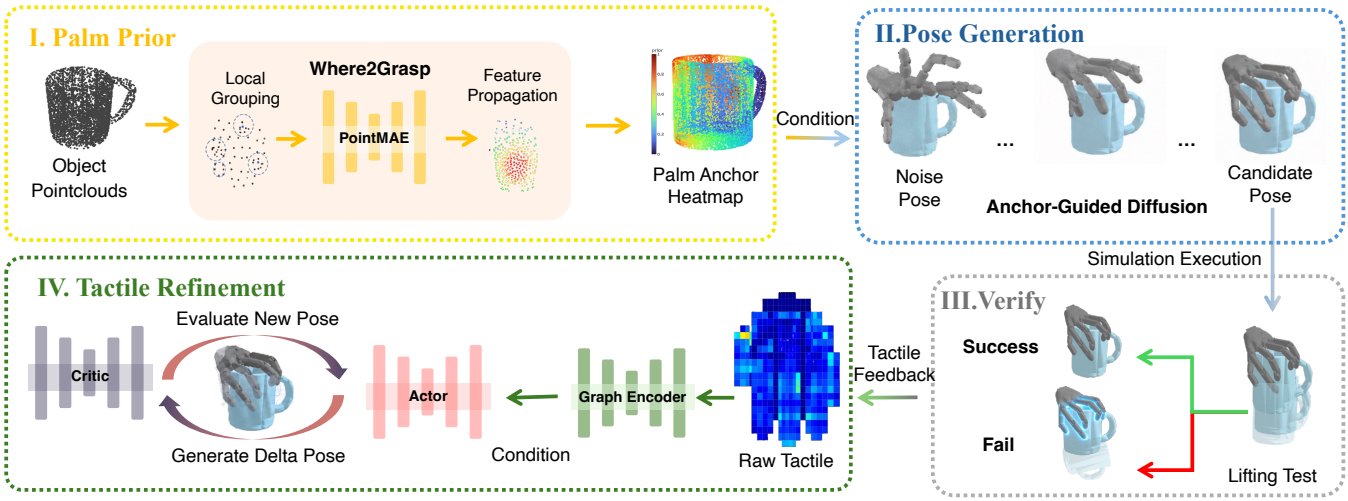

Fig. 3. **Overview of the proposed palm-anchored tactile refinement pipeline.** Given an object point cloud, Where2Grasp predicts a dense palm-anchor quality field over the observed surface. The selected high-quality anchor conditions a hierarchical diffusion generator that first samples the palm placement and then the full dexterous hand configuration. Candidate grasps are physically verified to obtain whole-hand tactile signatures, which are encoded by a tactile graph network and passed to an actor–critic refiner that applies a bounded residual correction before final execution.

frozen. A critic head $Q(q \mid z_t, \mathcal{P}) = \sigma(g_{\text{crit}}(h))$ predicts grasp success and an actor head produces a bounded residual update $\Delta q = \tau \cdot \tanh(g_{\text{act}}(h))$, giving $q' = q + \Delta q$. Given the binary simulator outcome $y \in \{0, 1\}$, the actor is trained with

$$\mathcal{L}_{\text{actor}} = -w_Q\, Q(q') + w_c \, \|q' - q\|_\Lambda^2 + w_a \mathbb{1}\{y = 1\} \, \|q' - q\|_2^2, \tag{4}$$

where Λ 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 [11] 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.

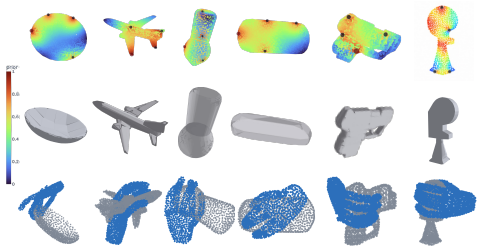

Fig. 4. **Qualitative simulation cases across diverse objects.** Six representative objects spanning seen and unseen categories from the 1,000-object DexGraspNet benchmark. *Top:* the per-point palm-anchor prior field predicted by Where2Grasp (colorbar, higher is better), with dark markers indicating 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 [7]	39.56	43.56
GraspQP [17]	34.67	21.50
DexDiffuser [14]	25.66	21.64
UniDexGrasp++ [10]	73.60	71.30
$\pi_{0.5}$ [2]	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.

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

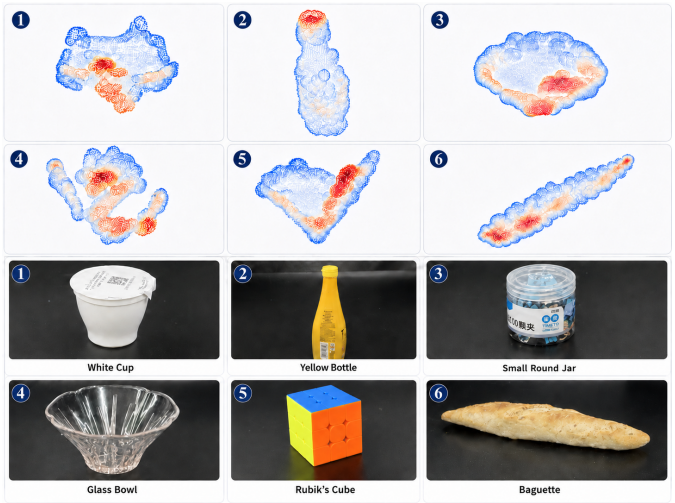

Fig. 5. **Real-world partial point clouds used for sim-to-real deployment.** The RGB-D observations contain only the visible object surface and include sensor noise, boundary discontinuities, and missing back-side geometry. These partial point clouds are the direct inputs to Where2Grasp during deployment, enabling dexterous grasp generation without explicit shape completion.

D. Real-World Transfer

We further evaluate the pipeline on a real dexterous hand in a zero-shot setting, without any real-world grasping data for training. Given partial point clouds captured by an RGB-D camera (Figure 5), the model directly predicts grasp poses and executes them on out-of-distribution household objects. Because Where2Grasp scores palm graspability per surface point, occluded geometry simply does not enter the candidate set rather than corrupting it, so no shape-completion module is required at inference and the anchor field degrades gracefully with viewpoint occlusion. The generated grasps adapt well to unseen object geometries and achieve stable grasping across diverse real-world cases (Figure 6); the contact patterns recovered from the 16 physical tactile sensors are physically consistent with those produced in simulation. This indicates that the observation–dynamics decoupling in our simulator and the kinematic-tree tactile encoder together produce a tactile representation that transfers across the sim-to-real gap without re-calibration.

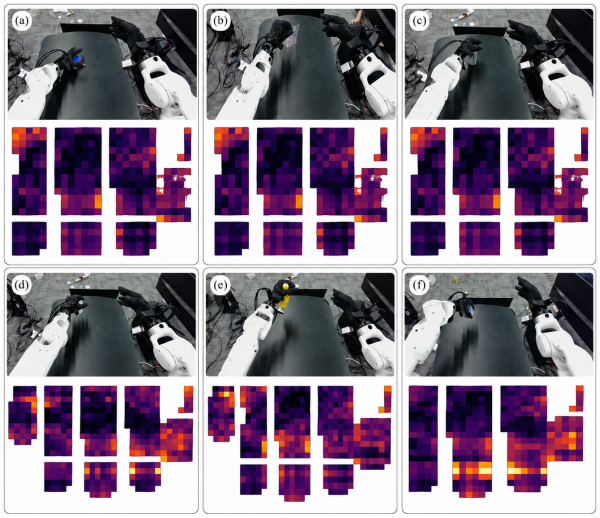

Fig. 6. **Qualitative real-world grasps with whole-hand tactile readout.** Zero-shot execution on six out-of-distribution household objects, shown in panels (a) to (f), using a real dexterous hand with no real-world grasping data used for training. For each case, *top:* a third-person view of the executed grasp; *bottom:* the corresponding whole-hand tactile signature recorded from the 16 physical sensors (palm and 15 finger links) and flattened to the sensor layout, where warmer colors denote higher normal contact force. The sim-trained pipeline yields stable grasps and physically consistent contact patterns across diverse geometries, indicating effective sim-to-real transfer.

IV. Conclusion and Limitation

A. 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.

B. Limitation

Our evaluation is confined to a single 20-DoF hand under static tabletop conditions, leaving cross-embodiment transfer, cluttered scenes, and language- or task-conditioned grasping to future work. The tactile refinement loop is closed only in simulation: the signed-distance tactile readout does not model the deformable contact mechanics of physical sensors, and closing it on real tactile hardware without re-calibration remains open. Refinement also requires one round of physical verification at inference, an overhead that a learned stability proxy would reduce.

References

[1] Iretiayo Akinola, Jie Xu, Jan Carius, Dieter Fox, and Yashraj Narang. Tacsl: A library for visuotactile sensor simulation and learning. *IEEE Transactions on Robotics*, 41:2645–2661, 2025. doi: 10.1109/TRO.2025.3547267.

[2] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 17–40. PMLR, 2025.

[3] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8701–8711, 2019. doi: 10.1109/CVPR.2019.00891.

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. doi: 10.1109/CVPR46437.2021.01549.

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[7] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8068–8074, 2023.

[8] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision – ECCV 2022*, volume 13662 of *Lecture Notes in Computer Science*, pages 604–621. Springer, 2022. doi: 10.1007/978-3-031-20086-1_35.

[9] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[10] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He 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)*, pages 3891–3902, 2023.

[11] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *IEEE International Conference on Robotics and Automation*, pages 11359–11366, 2023.

[12] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, 7(2): 3930–3937, 2022. doi: 10.1109/LRA.2022.3146945.

[13] Zihao Wei, Chen Wei, Jieru Mei, Yutong Bai, Zeyu Wang, Xianhang Li, Hongru Zhu, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders are secretly efficient learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 7986–7995, 2024. doi: 10.1109/CVPRW63382.2024.00797.

[14] Zehang Weng, Haofei Lu, Danica Kragic, and Jens Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models. *IEEE Robotics and Automation Letters*, 9(12):11834–11840, 2024. doi: 10.1109/LRA.2024.3498776.

[15] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1192–1199, 2008.

[16] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019.

[17] René Zurbrügg, Andrei Cramariuc, and Marco Hutter. Graspqp: Differentiable optimization of force closure for diverse and robust dexterous grasping. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 2583–2602. PMLR, 2025.