BEYOND FINGERTIPS: FROM PALM PRIORS TO TACTILE REFINEMENT FOR GENERALIZABLE DEXTEROUS GRASPING

Anonymous authors
Paper under double-blind review

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 refined by an actor–critic policy using whole-hand tactile observations organized by the hand's kinematic graph. Our tactile infrastructure and refinement policy share this morphology-parameterized graph interface, allowing the same policy architecture to be instantiated and trained separately for hands with different link counts, degrees of freedom, and sensor layouts without redesigning the tactile representation. Experiments on two hand morphologies, using independently trained policies and hand-specific data budgets, validate this architectural portability. On in-distribution and unseen object categories, our method achieves 79.30% grasp success on unseen objects, outperforming the strongest baseline by 8.0 percentage points. Without real-world grasp training, the method further attains zero-shot aggregate success rates of 43.3% on the Wuji Hand and 31.7% on LinkerHand O6, evaluated on separate household-object sets.

1 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.

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 using the 20-DoF Wuji Hand as the primary embodiment. 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 pretrained 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-rollout-validated whole-

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

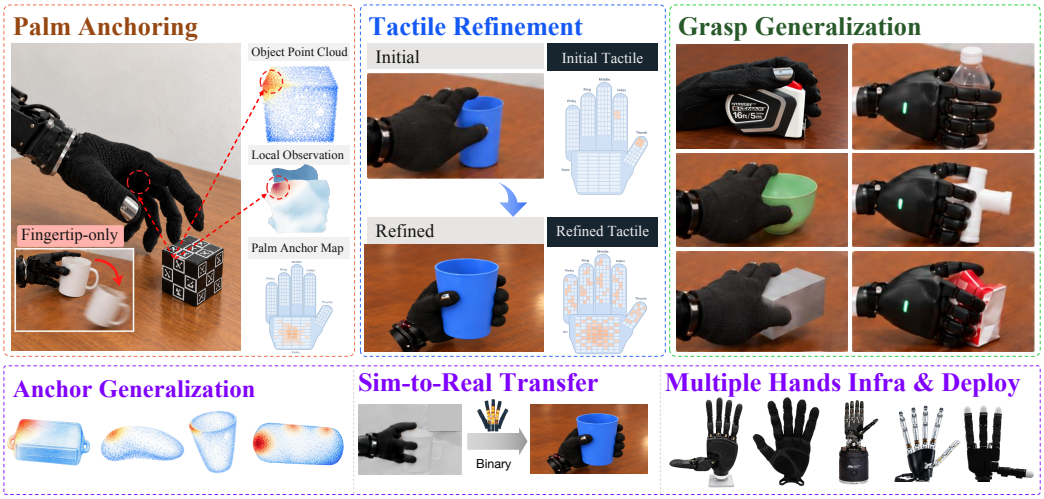

Figure 1: **Beyond fingertips: palm-anchored, tactile-refined grasping.** Fingertip-only contact can yield unstable grasps, whereas palm engagement provides stable whole-hand support. Guided by palm-anchor prior, our framework generates the grasp and refines its contact distribution using whole-hand tactile feedback.

hand tactile dataset that we release. We instantiate this tactile infrastructure on five hand morphologies for large-scale data collection, and separately train the same graph-policy architecture on two hand morphologies using simulation corpora of different scales. This validates architectural portability without assuming shared policy weights or matched data budgets across hands. Across in-distribution objects, unseen-category test sets, and real-world household items, our method substantially improves grasp success over pose-only baselines and supports zero-shot transfer to real hardware.

Why touch is complementary. 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.

2 Related Work

Dexterous grasp synthesis and contact representations. Classical grasp planning evaluates force closure and wrench-space quality under kinematic and frictional constraints (Nguyen, 1986; Ferrari & Canny, 1992). Large-scale learning systems such as DexGraspNet and UniDexGrasp (Wang et al., 2023; Xu et al., 2023) instead learn object-conditioned distributions over high-dimensional hand poses, while DGTR predicts diverse grasp sets in one forward pass and improves their quality with test-time adaptation (Xu et al., 2024). ContactDB and ContactPose provide dense human–object contact annotations (Brahmbhatt et al., 2019; 2020); contact-conditioned generators such as GenDexGrasp (Li et al., 2023) and ContactGen (Liu et al., 2023) use full contact maps or hand-part assignments to guide grasp synthesis. In contrast, Where2Grasp isolates palm placement as a per-point object-surface quality field, leaving finger contacts to the downstream pose generator.

Tactile sensing, simulation, and feedback. Vision-based sensors such as GelSight (Yuan et al., 2017) and DIGIT (Lambeta et al., 2020) enable fine-grained contact sensing, while TACTO (Wang et al., 2022) and TacSL (Akinola et al., 2025) provide scalable simulation of high-resolution visuotactile units. Touch has also been used to close the grasping loop: MAT adjusts coarse multi-finger

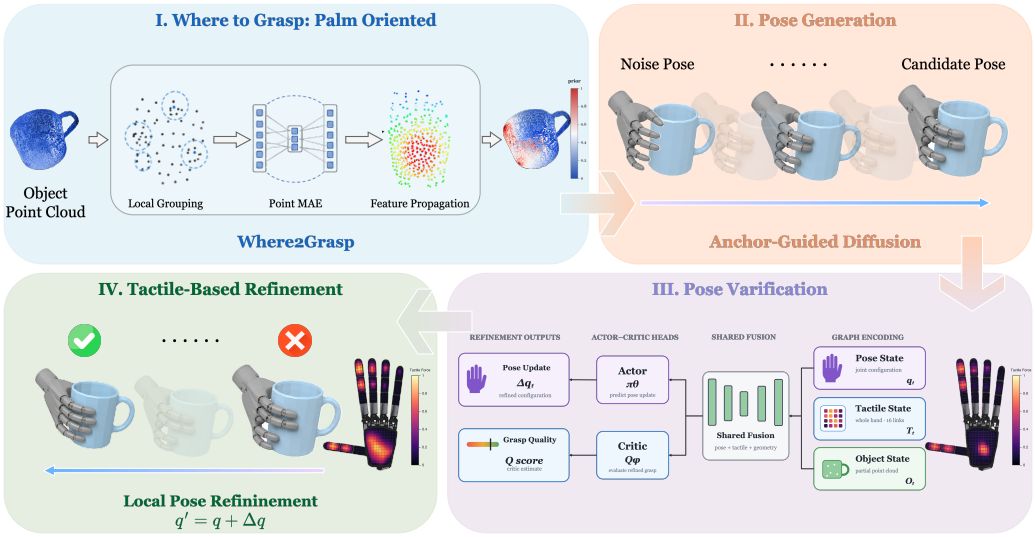

Figure 2: **Overview of the proposed palm-anchored grasp generation and tactile refinement pipeline.** *I. Where to Grasp:* Given an object point cloud, Where2Grasp predicts a dense palm-anchor quality field and selects a high-quality grasping region. *II. Pose Generation:* The selected anchor conditions 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. The lower inset enlarges stage III for readability.

grasps using tactile and proprioceptive feedback (Wu et al., 2020), and recent Contact-Grounded Policy predicts coupled robot-state and tactile trajectories for contact-rich dexterous manipulation (Xu et al., 2026). Our setting focuses on refining generated static grasp poses from morphology-indexed whole-hand tactile observations, supported by a passive SDF-based simulator instantiated across heterogeneous hand morphologies.

Diffusion and recent generative grasp models. Diffusion models have been applied to trajectory planning (Janner et al., 2022) and visuomotor action generation (Chi et al., 2023). In grasp synthesis, SE(3)-DiffusionFields (Urain et al., 2023) learns an SE(3) cost field for joint grasp and motion optimization, while DexDiffuser and DexGrasp Anything generate dexterous hand configurations (Weng et al., 2024; Zhong et al., 2025). Recent approaches introduce complementary priors: GeoDexGrasp decouples root rotation and joint generation using object-centric geometry (Han et al., 2026), KPGrasp uses hand keypoints and flow matching (Huang et al., 2026), and DexEvolve refines seed grasps in a high-fidelity simulator before distilling them into a diffusion model (Zurbrügg et al., 2026). These methods strengthen open-loop grasp generation; our method instead predicts an object-surface palm-anchor field and uses realized whole-hand tactile observations to correct execution-time contact uncertainty.

3 Method

3.1 Pipeline Overview

Figure 2 summarizes the pipeline. 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 three learned components are arranged into four operational stages: (i) Where2Grasp predicts a continuous palm-anchor quality field and selects an object-surface anchor $a \in \mathcal{P}$; (ii) a hierarchical diffusion generator samples a hand palm position conditioned on a and then the remaining orientation and joint state conditioned on the sampled palm, yielding a full hand configuration; (iii) in simulation experiments, each candidate is physically executed to obtain a whole-hand tactile obser-

vation; and (iv) an actor–critic refiner applies a bounded residual correction. For an embodiment e with d_e actuated joints, we parameterize the hand state as $q^{(e)} = (x, r, \theta) \in \mathbb{R}^{9+d_e}$ with palm position $x \in \mathbb{R}^3$, palm orientation $r \in \mathbb{R}^6$ in the continuous rotation representation of Zhou et al. (2019), and joint angles $\theta \in \mathbb{R}^{d_e}$. For the primary 20-DoF Wuji Hand, $q \in \mathbb{R}^{29}$.

3.2 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 (Pang et al., 2022) so that publicly released ShapeNet pretraining can be re-used. We supervise it with a physics-rollout-derived target field. For each sampled surface anchor p_i, we instantiate the same number R of grasp trials whose palm is placed at that anchor and execute them in simulation. Let $y_{i,r} \in \{0, 1\}$ denote whether trial r satisfies the simulator success criterion. The target anchor quality is the empirical success frequency

$$s_i^* = \frac{1}{R} \sum_{r=1}^{R} y_{i,r}, \qquad i \in \mathcal{M}, \tag{1}$$

Thus, anchors that repeatedly yield successful grasps receive brighter target scores, whereas rarely successful anchors receive lower scores. The network combines a masked Huber regression loss with a Plackett–Luce listwise ranking loss (Xia et al., 2008):

$$\mathcal{L}_{w2g} = \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}, \tag{2}$$

with the masked regression term

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathcal{H}_\delta(s_i - s_i^*), \tag{3}$$

where $\mathcal{H}_\delta$ is the Huber penalty with threshold δ. The listwise term is

$$\mathcal{L}_{\text{rank}} = -\frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \log \frac{\exp(s_{\pi(i)})}{\sum_{j=i}^{|\mathcal{M}|} \exp(s_{\pi(j)})}, \tag{4}$$

where $\pi : \{1, \ldots, |\mathcal{M}|\} \to \mathcal{M}$ 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 set of sampled surface anchors with recorded rollouts. The ranking term preserves the target-induced relative ordering of candidate anchors within each object, while $\mathcal{L}_{\text{reg}}$ calibrates their absolute score values. At inference, the score field conditions the downstream sampler in two complementary ways: we draw palm anchors stochastically using the inverse-temperature softmax $\Pr(i) \propto \exp(\beta\, s_i)$ that favors high-quality points while retaining diversity, set the selected surface anchor to $a = p_i$, and forward the per-point score as a feature channel to the palm denoiser.

3.3 Palm-Anchored Pose Generation

The pose generator is a two-stage conditional diffusion model that first samples the hand palm position x conditioned on the selected surface anchor a, and then samples (r, θ) conditioned on x; the two outputs are assembled as $q = (x, r, \theta)$. Both stages share a denoising architecture: a PointNet++ encoder (Qi et al., 2017) 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 clean-sample prediction (commonly called x_0-prediction) (Ho et al., 2020). Writing $u = x$ for the

first stage and $u = (r, \theta)$ for the second, each denoiser recovers the clean target from the noised iterate

$$u_t = \sqrt{\bar{\alpha}_t}\, u_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{5}$$

For diffusion stage $k \in \{1, 2\}$ with stage-specific conditioning c_k, the training objective is

$$\mathcal{L}_{\text{diff}}^{(k)} = \mathbb{E}_{u_0, t, \epsilon} \left[\left\| \hat{u}_{0,\phi_k}(u_t, t, c_k) - u_0 \right\|_2^2 \right]. \tag{6}$$

We found clean-sample prediction more stable than ϵ-prediction at the state dimensionalities considered here.

In the second stage, the object point cloud is recentered at the sampled palm position x to exploit translation equivariance, and the conditioning is augmented with an MLP embedding of x. Classifier-free guidance dropout at training (Ho & Salimans, 2022) replaces the conditioning with a zero vector with probability p_{cfg}; at inference we combine conditional and unconditional predictions as

$$\hat{u}_0^{(t)} = \hat{u}_0^{(t)}(\emptyset) + w_{\text{cfg}} \left(\hat{u}_0^{(t)}(c) - \hat{u}_0^{(t)}(\emptyset) \right), \quad w_{\text{cfg}} \geq 1, \tag{7}$$

where c denotes the stage-specific conditioning. With K sampled surface anchors and M diffusion samples per anchor, the generator yields $K \cdot M$ candidate grasps per object.

3.4 Whole-Hand Tactile Simulation Infrastructure

Learning tactile grasping across embodiments requires tactile observations that are (i) available over the whole grasping surface of the hand, not only at the fingertips, (ii) physically consistent with the simulated contact geometry, and (iii) comparable across hands whose morphology, actuation and sensor density differ. We therefore built a tactile simulation infrastructure on the PhysX-based Isaac Lab stack (Mittal et al., 2025) that equips five dexterous hands—Allegro, Shadow, Ability, Wuji, and LinkerHand O6—with distributed tactile skins covering the palm and finger links. The infrastructure validates each configured sensor surface in simulation and records tactile signals passively and synchronously during large-scale grasp rollouts.

Existing simulators such as TACTO and TacSL provide efficient models of high-resolution visuotactile units (Wang et al., 2022; Akinola et al., 2025). Our infrastructure instead targets morphology-indexed, link-level tactile fields distributed across an entire dexterous hand, while enforcing strict observation–dynamics decoupling: SDF-based sensor queries are read-only and do not write forces or state into the contact solver. Tactile observation therefore does not alter the simulated dynamics by construction. For the primary embodiment, the 20-DoF Wuji Hand, 16 sensor-bearing links covering the palm and fingers share a single tactile-field abstraction; the other hands use morphology-specific link and sensor counts. SDF queries across all parallel environments are batched into a single GPU call.

For each embodiment e, we represent the hand by a morphology graph $\mathcal{G}^{(e)} = (\mathcal{V}^{(e)}, \mathcal{E}^{(e)})$, where nodes index sensor-bearing links and edges are induced by the hand's kinematic tree. The tactile simulator emits link-local observations in this node order, and the refinement policy consumes the same topology and node-aligned observations. This shared graph interface provides the semantic alignment between sensing and policy: changing the hand replaces the morphology specification rather than the tactile representation.

For the primary Wuji Hand, we roll out more than 500 diffusion-sampled candidate grasps per object on the 1,000-object benchmark. Each candidate is paired with its simulator stability label and a whole-hand tactile record at the configured 16-link resolution. All task-specific components are trained only on the 800 seen-object split; no grasp, tactile, or physics-rollout samples from the 200 category-disjoint unseen objects are used for task-specific training. The external Point-MAE initialization described above uses generic ShapeNet geometry pretraining rather than grasp supervision from this benchmark. The LinkerHand O6 study analogously uses a separate 400-seen/100-unseen

object split, as specified in Section 4.4. Unlike prior contact-affordance corpora (Brahmbhatt et al., 2019), our dataset jointly aligns generator-distribution grasps, physics-rollout stability labels, and dense whole-hand tactile records on the *same* samples.

3.5 Tactile Graph Encoding and Refinement

Each candidate grasp is executed in simulation until physical settling, producing a whole-hand tactile observation $\mathcal{T}^{(e)}$ consisting of link-local tactile grids attached to the nodes of $\mathcal{G}^{(e)}$. For the primary Wuji Hand, these correspond to the palm and 15 finger links. Each link-local grid is first mapped to a fixed-width node token, avoiding a global concatenation whose dimension would depend on the sensor layout. A tactile graph encoder then propagates information along the kinematic tree and applies graph-level pooling to obtain a fixed-dimensional tactile embedding z_t, independent of the embodiment-specific number of nodes. The encoder is pretrained with masked autoencoding (He et al., 2022), a non-contrastive SimSiam objective (Chen & He, 2021), and a simulator-derived force-alignment objective, then frozen.

The graph encoder, pooling operator, and actor–critic design are reused across embodiments. Only the morphology graph and the hand-state input/output dimensions are instantiated from the corresponding hand specification. We train a separate policy instance for each evaluated hand; thus, "shared" refers to the graph interface and network architecture, not to shared learned weights across embodiments.

Let $c_{\mathcal{P}}$ denote the object feature and $\phi_q(q)$ a hand-state embedding. We fuse these with the tactile embedding as $h(q, z_t, \mathcal{P}) = f_{\text{fuse}}(\phi_q(q), z_t, c_{\mathcal{P}})$ before the actor and critic heads; when scoring a proposal q', the hand-state branch is recomputed at q' while z_t and $c_{\mathcal{P}}$ remain fixed.

Each critic Q_m in an ensemble of size C predicts grasp success as

$$Q_m(q \mid z_t, \mathcal{P}) = \sigma(g_{\text{crit},m}(h(q, z_t, \mathcal{P}))), \qquad \bar{Q}(q \mid z_t, \mathcal{P}) = \frac{1}{C}\sum_{m=1}^{C} Q_m(q \mid z_t, \mathcal{P}). \tag{8}$$

An actor head produces a bounded residual update

$$\Delta q = \boldsymbol{\tau} \odot \tanh(g_{\text{act}}(h(q, z_t, \mathcal{P}))), \qquad q' = q + \Delta q. \tag{9}$$

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

$$\mathcal{L}_{\text{actor}} = -w_Q\, \bar{Q}(q' \mid z_t, \mathcal{P}) + w_c \, \|q' - q\|_\Lambda^2 + w_a \mathbb{1}\{y = 1\} \, \|q' - q\|_2^2, \tag{10}$$

where $\boldsymbol{\tau}$ gives component-wise residual bounds and $\|v\|_\Lambda^2 = v^\top \Lambda v$ for a positive diagonal matrix Λ that balances translation, rotation, and joint components. The third term is a *successful-sample preservation regularizer* that discourages unnecessary movement on already-successful inputs, preventing collapse onto a single high-score mode of the critic. The fusion module and critic ensemble are optimized with a class-weighted BCE on the simulator label y. During actor optimization, the tactile encoder, fusion module, and critics are held fixed, while gradients through $\bar{Q}(q' \mid z_t, \mathcal{P})$ with respect to the proposed state are retained; therefore, only the actor parameters are updated. 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 score, and an unrefined no-op candidate is included to provide a principled refusal option.

4 Experiment

4.1 Setup

For the primary Wuji Hand benchmark, we evaluate on 1,000 object instances from DexGraspNet (Wang et al., 2023), partitioned by category into 800 instances from training categories and 200 instances from held-out categories, 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

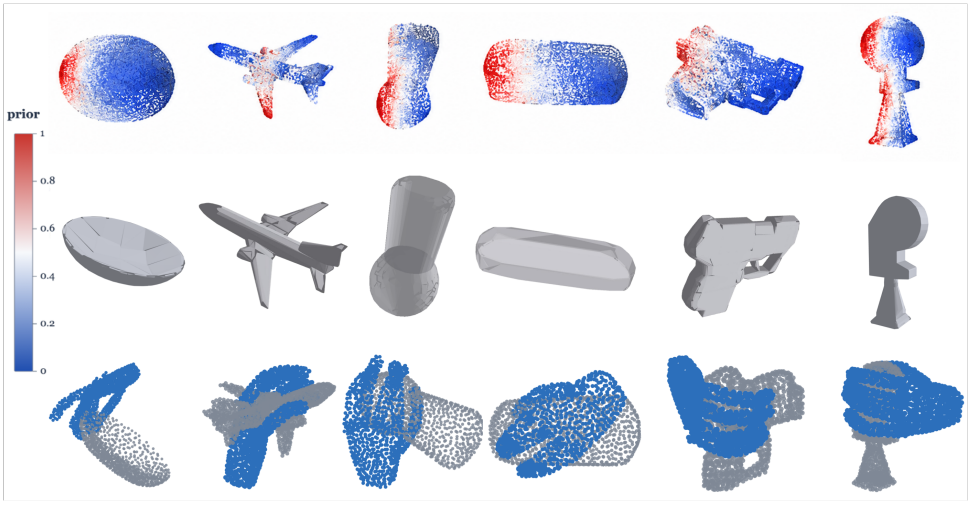

Figure 3: **Qualitative simulation cases across diverse objects.** Six representative objects spanning seen and unseen categories from the 1,000-object DexGraspNet benchmark. *Top:* the per-point palm-anchor prior field predicted by Where2Grasp (colorbar, higher is better), with dark markers indicating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

object remains above its initial height. We compare against six methods spanning contact-map, optimization, diffusion, RL, and foundation-VLA paradigms: GenDexGrasp, GraspQP, DexDiffuser, DexGrasp Anything, UniDexGrasp++, and $\pi_{0.5}$, with learning-based baselines re-trained or fine-tuned on the seen split when applicable. All methods are evaluated using the same physical success criterion. The primary metric is grasp success rate (SR), computed as the mean per-object pass rate. The cross-embodiment LinkerHand O6 study uses a separate hand-specific split of 400 seen and 100 unseen objects, as detailed in Section 4.4.

4.2 Main Results

Table 1 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 is consistent with improved cross-category generalization rather than only an in-distribution gain. 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 (Figure 3).

Table 1: Grasp success rate (%) on 1,000 DexGraspNet object instances: 800 from training categories and 200 from held-out categories.

Method	Training-Cat. Objects	Held-Out-Cat. Objects
GenDexGrasp (Li et al., 2023)	39.56	43.56
GraspQP (Zurbrügg et al., 2025)	34.67	21.50
DexDiffuser (Weng et al., 2024)	25.66	21.64
DexGrasp Anything (Zhong et al., 2025)	53.38	41.75
UniDexGrasp++ (Wan et al., 2023)	73.60	71.30
$\pi_{0.5}$ (Black et al., 2025)	52.24	40.28
Ours	**79.47**	**79.30**

4.3 ABLATIONS

Table 2 ablates the core modules and tactile coverage. Removing the palm-anchor prior drops unseen-category SR by 3.20 points, with a larger effect on unseen than seen, supporting its value for cross-category transfer. Removing tactile refinement causes the largest drop on both splits (-5.80 / -6.22 points), showing that tactile-conditioned refinement contributes the largest ablation gain. 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. The fingertip-only baseline retains Where2Grasp and palm-anchor conditioning, but removes tactile observations from the palm and non-fingertip links. It therefore isolates the benefit of whole-hand tactile coverage while keeping the geometric grasp prior unchanged.

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 2: Ablation results on the simulation benchmark. SR is reported on seen and unseen splits. Parenthesized values denote percentage-point changes relative to the full pipeline.

Variant	Training-Cat. Objects	Held-Out-Cat. Objects
Full pipeline	**79.47**	**79.30**
w/o Where2Grasp	77.13 (-2.34)	76.10 (-3.20)
Fingertip-only Tactile	74.05 (-5.42)	73.88 (-5.42)
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)

4.4 CROSS-EMBODIMENT PORTABILITY

We evaluate whether the tactile refinement architecture can be instantiated on hands with different kinematics and sensing layouts. The study includes the 20-DoF Wuji Hand used in the main benchmark and LinkerHand O6. For each embodiment, sensor-bearing links form graph nodes and the kinematic tree induces the edges. The same graph encoder and actor–critic design is instantiated with morphology-specific topology and state dimensions, while the two policy instances are trained independently without shared weights. Wuji uses the 800-training-category/200-held-out-category split in Section 4.1, whereas LinkerHand uses a smaller hand-specific 400/100 split. We therefore assess portability through within-hand ablations, not absolute cross-hand rates. On LinkerHand, tactile refinement raises SR from 60.58% to 69.74% on training-category objects and from 57.15% to 67.43% on held-out-category objects. The palm prior and graph encoder also improve both splits; complete results are reported in Appendix A.

4.5 REAL-WORLD TRANSFER

We evaluate zero-shot sim-to-real transfer on two physical embodiments, the Wuji Hand and LinkerHand O6, using partial RGB-D point clouds without real-world grasp training, fine-tuning, or explicit shape completion. The two hands are evaluated on different hand-specific sets of unseen household objects; consequently, their absolute success rates are reported separately and are not interpreted as a controlled comparison between embodiments. Each object is attempted 10 times. Upon object contact, the physical tactile readings are acquired and passed to the tactile-conditioned refinement policy before the final grasp is evaluated. A trial is successful only if the hand lifts the object by at least 5 cm and holds it for 2 s. Planning failures, collisions, unsuccessful lifts, and object drops are all counted as failures. Where2Grasp scores only observed surface points, and no real-world samples are used for training any component. For hand h with O_h test objects, Table 3 reports the aggregate success rate $\mathrm{SR}_h = 100 \sum_{o=1}^{O_h} s_{h,o}/(10O_h)$, where $s_{h,o}$ is the number of suc-

cessful trials for object o. Because every object has the same number of trials, this percentage is also the mean of the per-object success rates.

Table 3: Aggregate zero-shot real-world grasping results for two physical hands. Each object is attempted 10 times with tactile refinement active. Because the hands use different object sets, rates are interpreted within each hand rather than compared directly across hands.

Hand	Objects	Total Successes / Attempts	Aggregate SR (%)
Wuji Hand	6	26 / 60	43.3
LinkerHand O6	6	19 / 60	31.7

Under this protocol, the Wuji Hand succeeds in 26 of 60 trials (43.3%), while LinkerHand O6 succeeds in 19 of 60 trials (31.7%). These values demonstrate zero-shot operation on both hands but are not directly compared because the two embodiments use different object sets.

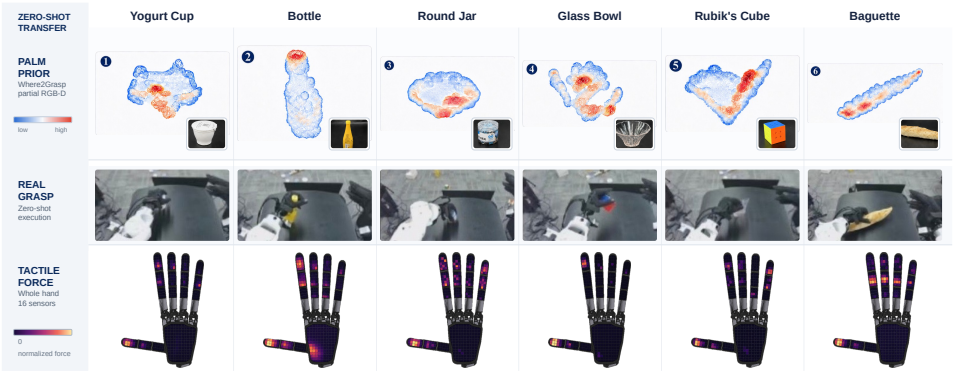

Figure 4: **Zero-shot Wuji Hand examples.** Rows show palm priors, real-world execution, and 16-sensor tactile-force measurements, without real-world training or shape completion.

5 Conclusion

We presented a palm-anchored, tactile-refined framework that addresses the gap between kinematically valid hand poses and physically successful grasps. Where2Grasp identifies promising palm-contact regions, hierarchical diffusion generates full-hand configurations conditioned on the selected anchor, and a graph-based actor–critic policy corrects the executed grasp using realized whole-hand tactile feedback. Experiments show that the palm prior and tactile refinement provide complementary improvements, reaching 79.30% success on held-out object categories in simulation. Instantiating the same graph-policy design separately on the Wuji Hand and LinkerHand O6 further demonstrates its architectural portability, while zero-shot experiments on both physical hands show that the simulation-trained pipeline can operate on unseen household objects without real-world grasp training. Together, these results support explicit palm placement and post-execution contact feedback as complementary components of generalizable dexterous grasping.

AI use statement

Generative AI tools were used for literature retrieval, language editing, consistency checking, and drafting or revising portions of the manuscript. They were not used to conduct experiments or to generate or alter datasets and reported measurements. All AI-assisted text and calculations were reviewed and verified by the authors, who take full responsibility for the paper.

REFERENCES

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.

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*, pp. 17–40. PMLR, 2025.

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*, pp. 8701–8711, 2019. doi: 10.1109/CVPR.2019.00891.

Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII*, pp. 361–378, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58600-3. doi: 10.1007/978-3-030-58601-0_22. URL https://doi.org/10.1007/978-3-030-58601-0_22.

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

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems XIX*, 2023. doi: 10.15607/RSS.2023.XIX.026.

C. Ferrari and J. Canny. Planning optimal grasps. In *Proceedings 1992 IEEE International Conference on Robotics and Automation*, pp. 2290–2295 vol.3, 1992. doi: 10.1109/ROBOT.1992.219918.

Bing Han, Weiyuan Liu, Changlong Zhang, Chenxi Wang, Zhibin Zhao, and Zhi Zhai. GeoDexGrasp: Geometry-aware generation for data-efficient and physics-plausible dexterous grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6729–6739, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

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

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

Yuansen Huang, Jiayi Chen, Haoran Liu, Yubin Ke, Bing Han, Jiangran Lyu, Mi Yan, Li Yi, and He Wang. KPGrasp: Scalable keypoint flow matching for dexterous grasp generation, 2026. URL https://arxiv.org/abs/2606.09314.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/janner22a.html.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. doi: 10.1109/LRA.2020.2977257.

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)*, pp. 8068–8074, 2023.

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. ContactGen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20609–20620, 2023.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, et al. Isaac Lab: A gpu-accelerated simulation framework for multi-modal robot learning, 2025. URL `https://arxiv.org/abs/2511.04831`.

V.-D. Nguyen. Constructing force-closure grasps. In *Proceedings of the 1986 IEEE International Conference on Robotics and Automation*, volume 3, pp. 1368–1373, 1986. doi: 10.1109/ROBOT.1986.1087483.

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*, pp. 604–621. Springer, 2022. doi: 10.1007/978-3-031-20086-1_35.

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.

Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5923–5930, 2023. doi: 10.1109/ICRA48891.2023.10161569.

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)*, pp. 3891–3902, 2023.

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*, pp. 11359–11366, 2023.

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.

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.

Bohan Wu, Iretiayo Akinola, Jacob Varley, and Peter K. Allen. MAT: Multi-fingered adaptive tactile grasping via deep reinforcement learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 142–161. PMLR, 2020.

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*, pp. 1192–1199, 2008.

Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17933–17942, 2024.

Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, Tengyu Liu, Li Yi, and He Wang. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4737–4746, 2023. doi: 10.1109/CVPR52729.2023.00459.

Zhengtong Xu, Yeping Wang, Ben Abbatematteo, Jom Preechayasomboon, Sonny Chan, Nicholas Colonnese, and Amirhossein H. Memar. Contact-grounded policy: Dexterous visuotactile policy with generative contact grounding. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.005.

Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. GelSight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. doi: 10.3390/s17122762.

Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22584–22594, 2025. doi: 10.1109/CVPR52734.2025.02103.

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)*, pp. 5745–5753, 2019.

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*, pp. 2583–2602. PMLR, 2025.

René Zurbrügg, Andrei Cramariuc, and Marco Hutter. DexEvolve: Evolutionary optimization for robust and diverse dexterous grasp synthesis. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.059.

A Cross-Embodiment Evaluation Details

A.1 LinkerHand O6 Ablations

Table 4: LinkerHand O6 ablations in simulation on its hand-specific 400-training-category/100-held-out-category split. SR (%) and parenthesized values denote percentage-point changes relative to the full pipeline.

Variant	Training-Cat. SR	Held-Out-Cat. SR
Full pipeline	**69.74**	**67.43**
w/o Where2Grasp	66.28 (-3.46)	63.21 (-4.22)
w/o Tactile Refine	60.58 (-9.16)	57.15 (-10.28)
w/o Graph Encoder (MLP)	66.19 (-3.55)	63.31 (-4.12)

A.2 Real-World Experimental Setups

Figure 5 shows the two physical platforms, RGB-D sensing arrangements, and representative household objects used for the zero-shot experiments. Each embodiment is evaluated on its own six-object set under the protocol in Section 4.5; because the object sets differ, the resulting success rates are not treated as a direct cross-hand comparison.



(a) LinkerHand O6 setup (b) Wuji Hand setup

Figure 5: **Real-world experimental setups.** The two physical hand platforms are shown with their RGB-D sensors and representative household test objects. The hands use separate object sets for zero-shot evaluation.