# BiDexAffordance: Learning Collaborative Affordances for Efficient Bimanual Dexterous Grasping

Peiqi Liu[1,2], Jingwen Li[1], Zeyuan Chen[1], Yue Chen[1], Shuqi Zhao[2], Yuanpei Chen[1], Chenfeng Xu[2], Masayoshi Tomizuka[2], Wei Zhan[2], Ruihai Wu[2]
[1]Peking University [2]University of California, Berkeley

***Abstract*—Bimanual dexterous grasping is essential for manipulating large, heavy, or awkward objects, yet it remains difficult to scale because two high-DoF hands create a large coupled search space. We present *BiDexAffordance*, an affordance-driven framework that learns collaborative contact priors from simulated bimanual interactions and uses them to guide physics-based grasp refinement. Instead of directly regressing full dual-hand configurations, our method predicts two object-surface affordance maps: a first-hand map that identifies promising initiation regions and a conditional second-hand map that captures cooperative contact choices given the first hand. The selected keypoint pair initializes a lightweight optimizer, reducing search while retaining physical validation. Across 170 objects from 11 categories, BiDexAffordance improves simulated bimanual grasp success over optimization-only and diffusion-plus-optimization baselines, generalizes to unseen instances and categories, and transfers to a dual-arm real-robot setup.**

## I. Introduction

Bimanual dexterous manipulation is increasingly relevant as robotic hands become more accessible and reliable. Many everyday objects cannot be grasped robustly by a single hand: pots, trays, appliances, and wide containers often require coordinated contacts, force balance, and collision-free placement of two hands. Prior bimanual work has shown the promise of dual-hand coordination [4, 2, 8], but scalable grasp synthesis remains challenging because the configuration space grows with both hand poses and all finger joints.

Existing pipelines expose a core tradeoff. Optimization-based approaches such as BimanGrasp [8] can validate contacts physically and apply to arbitrary geometry, but they are slow and sensitive to initialization. Learning-based approaches can produce poses in one forward pass, but directly modeling the joint distribution over two dexterous hands often requires large training sets and may generalize poorly to novel objects. This paper studies a hybrid alternative: use learning to predict compact object-centric contact priors, then use optimization to recover physically plausible full-hand configurations.

We introduce **BiDexAffordance**, a framework for efficient bimanual dexterous grasp generation from object point clouds. The key representation is a pair of collaborative bimanual affordance maps defined on the object surface. A first-hand map captures regions that are likely to support a successful bimanual grasp, while a second-hand map is conditioned on the selected first-hand contact and captures compatibility between the two hands. This sequential factorization avoids directly searching over all contact pairs and creates an interface between scalable affordance learning and physics-based control or planning.

Our contributions are: (i) an affordance-guided bimanual grasp synthesis framework that combines learned cooperative contact priors with lightweight physics-based refinement; (ii) a sequential first-hand/conditional-second-hand affordance formulation for high-DoF bimanual grasping; and (iii) simulation and real-robot evaluations showing improved success, generalization, and optimization efficiency.

## II. Method

### *A. Problem Setup*

Given an object point cloud $O = \{x_i\}_{i=1}^N$, the goal is to compute a bimanual grasp

$$G = \{T_1, \theta_1, T_2, \theta_2\},$$

where $T_h = [R_h, t_h] \in SE(3)$ and $\theta_h$ denote the wrist pose and joint state of hand $h \in \{1, 2\}$. A valid grasp should be stable under gravity, avoid severe object–hand and hand–hand penetration, and coordinate the two hands on compatible object regions. In simulation we use ShadowHand models; on the real robot we retarget the generated grasp to a dual-arm setup with dexterous hands.

### *B. Affordance Supervision from Simulated Grasps*

We generate candidate bimanual grasps with a physics-based synthesis pipeline and validate them in Isaac Gym [5]. For each successful grasp, we sample an object point set $O$ and extract one surface contact keypoint per hand. Let $\mathcal{F}$ be sampled hand-surface points transformed into the grasp pose. We find nearby object points according to the normalized hand–object distance and compute a contact centroid by inverse-distance weighting. The centroid is projected back to the sampled object surface to obtain a discrete keypoint $x_h^*$.

Each keypoint supervises a dense heatmap

$$\mathcal{A}_h(x_i) = \exp\left(-\frac{\|x_i - x_h^*\|^2}{2\sigma^2}\right),$$

where $\sigma$ scales with the object bounding-box diagonal. For the first hand, multiple successful grasps of the same object

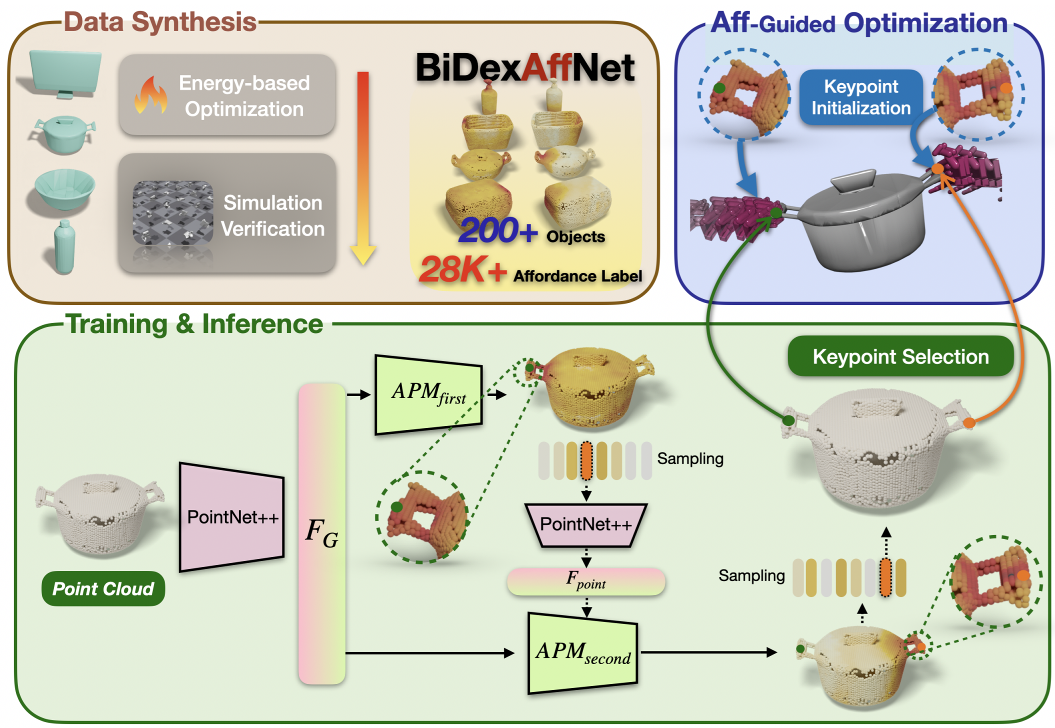



Fig. 1. **BiDexAffordance pipeline.** We synthesize bimanual affordance labels from simulated grasps, predict first-hand and conditional second-hand affordance maps from an object point cloud, and use the selected cooperative keypoints to initialize physics-based refinement.

are aggregated into a multi-center target. For the second hand, each training sample stores a first-hand contact and the compatible second-hand heatmap.

### C. Collaborative Affordance Prediction

BiDexAffordance predicts cooperative contacts sequentially. A PointNet++ backbone [7] extracts per-point features $f_i$ from $N = 8192$ surface points. The first affordance prediction model outputs

$$\mathcal{A}_{\text{first}}(x_i) = \sigma(g_{\text{first}}(f_i)).$$

We select a first-hand keypoint $x^*_{\text{first}}$ by argmax for maximum success or by softmax sampling for diversity. The second model conditions on this selected contact:

$$\mathcal{A}_{\text{second}}(x_i \mid x^*_{\text{first}}) = \sigma(g_{\text{second}}([f_i; \phi(x_i, x^*_{\text{first}})])),$$

where $\phi$ contains relative position and distance features. This factorization approximates the joint collaborative affordance as

$$\mathcal{A}_{\text{bi}}(x_{\text{first}}, x_{\text{second}}) = \mathcal{A}_{\text{first}}(x_{\text{first}}) \cdot \mathcal{A}_{\text{second}}(x_{\text{second}} \mid x_{\text{first}}),$$

reducing contact-pair search from quadratic over surface points to two linear surface predictions. Both heads are trained with per-point $L_1$ heatmap loss.

### D. Affordance-Guided Refinement

The predicted keypoints initialize two hands near compatible object regions. For each selected contact $x^*_h$, we define an approach direction from the contact toward the object centroid $c_O$,

$$a_h = \frac{c_O - x^*_h}{\|c_O - x^*_h\|}.$$

The palm normal is aligned with $a_h$, a random twist can be sampled around this axis, and the palm center is placed at a fixed offset outside the object. The resulting wrist pose initializes a bimanual optimizer using distance, force-closure, penetration, inter-hand, and joint-limit terms from prior differentiable grasp synthesis [8, 9]. Thus, learning proposes where the hands should cooperate, while optimization handles geometry, collisions, and physical feasibility.

## III. Experiments

### A. Setup

We evaluate on 170 objects across 11 categories with 28K validated bimanual grasps. Objects come from SAPIEN, PartNet, ShapeNet, and BimanGrasp-style assets [10, 6, 1, 8]. We split object instances within each category into seen and unseen subsets, and also test two fully unseen categories. For each object we generate 1,024 candidates and validate grasps in Isaac Gym by simulating gravity along six axis-aligned directions for 100 steps. A grasp is successful if the object remains stably held and maximum penetration depth is below 0.1 cm.

We compare against three baselines: Uni2Bim, which independently optimizes two single-hand grasps; BimanGrasp, a bimanual optimizer with random initialization; and

| Method | Seen | Unseen | Steps |
|---|---|---|---|
| Uni2Bim | 4.7 | 4.3 | 10K |
| BimanGrasp | 14.5 | 13.9 | 10K |
| DDPM+Opt | 15.8 | 12.5 | 1K |
| Ours (Joint) | 16.1 | 17.6 | 1K |
| Ours (Uni) | 35.7 | 34.7 | 1K |
| **Ours (Full)** | **36.9** | **35.9** | 1K |

TABLE I
AVERAGE SIMULATED GRASP SUCCESS RATE (%) ACROSS 11 OBJECT CATEGORIES.

| Setting | AP | IoU@1% | IoU@5% |
|---|---|---|---|
| First hand | 0.71 | 0.59 | 0.62 |
| Second hand (GT cond.) | 0.63 | 0.64 | 0.69 |
| Second hand (pred. cond.) | 0.48 | 0.52 | 0.61 |

TABLE II
AFFORDANCE HEATMAP QUALITY ON HELD-OUT OBJECTS.

DDPM+Opt, which predicts poses with a diffusion prior [3] before optimization. We also test two ablations: a joint contact predictor without our sequential factorization, and a unimanual optimizer initialized by our affordance prior.

### B. Results

Table I shows that BiDexAffordance improves both seen and unseen success over optimization-only and diffusion-plus-optimization baselines. The gap is especially large on unseen instances: the full method reaches 35.9% success, compared with 13.9% for BimanGrasp and 12.5% for DDPM+Opt. This supports the main design choice: predicting object-centric cooperative contact regions is easier to transfer than directly modeling a high-dimensional dual-hand pose distribution. Equally important, the seen-to-unseen drop is small (36.9% to 35.9%), while DDPM+Opt drops from 15.8% to 12.5%. This suggests that the learned representation captures reusable object-surface support structure rather than memorizing category-specific hand poses.

The ablations clarify where the gain comes from. Replacing the sequential conditional structure with a joint predictor gives only 17.6% unseen success, suggesting that directly modeling paired contacts is data-inefficient. Conversely, using our affordance initialization with a unimanual optimizer already raises Uni2Bim from 4.3% to 34.7%, showing that learned cooperative regions substantially simplify the downstream search even before full bimanual refinement.

Figure 3 evaluates optimization efficiency. Baselines without learned affordance priors need many iterations and converge to lower success. Our method reaches useful grasps with as few as 10 refinement steps and continues to improve with larger budgets. This behavior is important for future large-scale pipelines where many object hypotheses, task candidates, or policy rollouts must be evaluated quickly.

Beyond the 11-category instance split, BiDexAffordance also transfers to two fully unseen categories: it achieves 35.8% success on lamps and 39.5% on baskets, outperforming BimanGrasp (14.8%, 18.6%) and DDPM+Opt (11.3%, 15.9%). These category-level results are important because they test whether the model has learned relational contact patterns, such as opposing support and complementary reach, rather than only instance-level geometry.

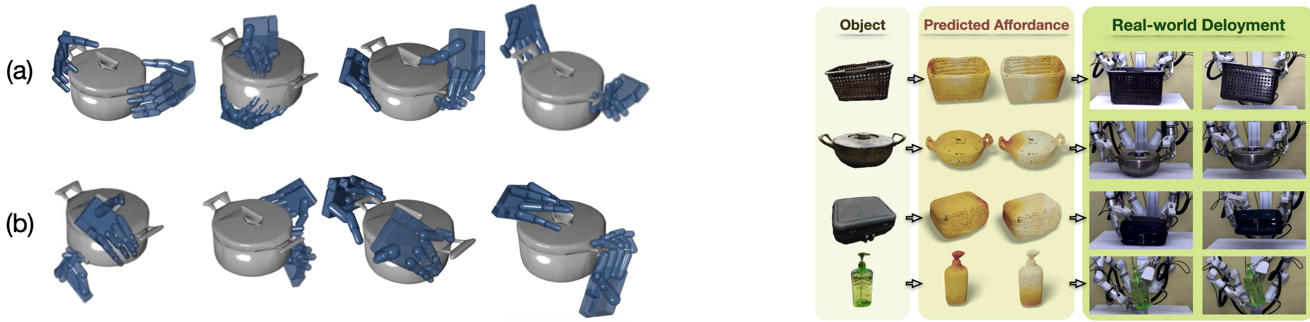
*a) Real-world deployment:* We deploy the pipeline on a dual-arm system with two 6-DoF arms, dexterous hands, and an RGB-D camera. After camera-to-robot calibration, object points are sampled from the observation, affordance maps are predicted, reachable grasps are filtered by inverse kinematics, and the robot executes a pre-grasp, grasp, and lift sequence. Across 9 objects and 90 trials, BiDexAffordance succeeds in 42 trials, compared with 24 for BimanGrasp and 22 for DDPM+Opt. These results suggest that the learned contact prior survives perception noise and calibration error, while also revealing current limitations in reachability and real-world execution. This real-world protocol is deliberately stricter than offline visualization: a candidate must pass partial observation, arm kinematics, hand-object collision, actuation, and lifting dynamics before it is counted as successful.

### C. Additional Analysis

*a) Affordance prediction quality:* To separate representation learning from downstream optimization, we evaluate the predicted heatmaps against the dense labels used for training. Each sampled object-surface point is treated as an instance, with the heatmap value as confidence. We report average precision after thresholding the ground-truth heatmap at $\tau = 0.5$, and IoU after binarizing both prediction and target by their top-$k$% points. The first-hand affordance reaches 0.71 AP, indicating that the model reliably finds promising initiation regions. The conditional second-hand predictor reaches 0.63 AP under ground-truth conditioning and 0.48 AP under predicted conditioning, showing that most degradation comes from first-stage sampling error rather than an inability to model cooperative contacts.

*b) Headroom and supervision source:* We also quantify the headroom of the pipeline by replacing predicted affordance keypoints with ground-truth affordance labels while keeping the same optimizer and step budget. The oracle variant reaches 37.4% total success, only modestly above our 35.9%, suggesting that optimization and grasp stability are now comparable bottlenecks to affordance prediction. In a separate morphology-transfer study on 10 ARCTIC objects, supervision from human interactions reaches 39.7% success, while robot-specific synthesized affordance supervision reaches 58.2%. This gap supports a central motivation for our approach: internet-scale or human video priors are valuable, but bimanual dexterity still benefits from embodiment-aware simulation and contact validation.

*c) Failure modes:* Manual inspection of failed grasps reveals three recurring causes: incorrect affordance regions, collision or penetration during refinement, and collision-free but unstable grasps that slip or drop under gravity. The




**(a)** Diverse generated bimanual grasps

**(b)** Real-world affordance and execution

Fig. 2. **Qualitative results.** BiDexAffordance preserves grasp diversity while transferring predicted affordance-guided grasps to real bimanual robot execution.

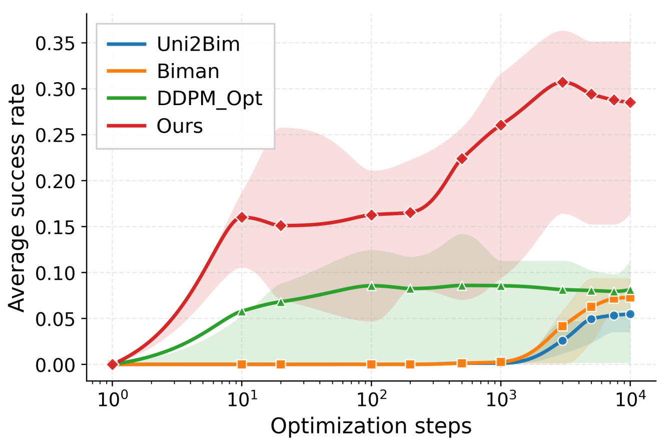



Fig. 3. Affordance-guided initialization improves success under small optimization budgets.

| Variant | Source | SR |
|---|---|---|
| BimanGrasp | random init. | 13.9 |
| Ours | predicted afford. | 35.9 |
| Ours-oracle | GT afford. | 37.4 |
| Human afford. | ARCTIC | 39.7 |
| Robot afford. | synthesized | 58.2 |

TABLE III
UPPER-BOUND AND SUPERVISION-SOURCE ANALYSES.

approximate ratio for our full method is 2:3:5. This distribution suggests two useful directions for the community: better uncertainty-aware affordance sampling to avoid poor contact regions, and tighter integration between learned priors and force/impedance-aware control so that physically plausible contacts also become dynamically robust.

*d) Relevance to scaling dexterity:* The workshop asks how dexterous manipulation can scale with larger data, better hardware, and sim-to-real pipelines. BiDexAffordance is one concrete point in that design space. It does not replace planning or control with a monolithic foundation policy; instead, it exposes a compact interface between perception and physics. Learned affordances reduce the combinatorial search over high-DoF bimanual contacts, while the optimizer and simulator keep the final grasp grounded in collision and stability checks. This decomposition is especially attractive when demonstrations are scarce, biased toward human morphology, or collected on hardware whose hands differ from the deployment platform.

## IV. Limitations and Conclusion

BiDexAffordance focuses on rigid-object lifting grasps and learns from simulation-generated, optimizer-validated supervision. This can bias the learned priors toward stable lifting rather than task- or function-specific contacts. The current pipeline also treats arm reachability as a post-filter rather than a first-class part of generation, and real-world execution remains sensitive to partial observations and calibration error.

Within these limits, the results show that collaborative affordance maps are a useful interface between scalable perception learning and physically grounded bimanual grasp synthesis. For the workshop themes of scaling dexterity with data, simulation, and sim-to-real pipelines, the main takeaway is that object-centric contact priors can substantially shrink the dual-hand search space without discarding the physical validation needed for real robots.

*a) Future directions:* The current representation is intentionally task-agnostic: it predicts where two hands can jointly hold an object, but it does not yet condition on downstream intent such as pouring, opening, handoff, or forceful manipulation. A natural next step is to combine collaborative affordance maps with language or task-conditioned constraints so that the same object can expose different bimanual contact hypotheses for different goals. Another important direction is reachability-aware generation. In our real-world system, infeasible arm poses are filtered after grasp synthesis; integrating arm kinematics, collision maps, and hardware-specific compliance during generation would reduce wasted samples and improve reliability. Finally, tactile feedback should be incorporated after initial contact. Vision and geometry are sufficient to propose candidate regions, but tactile signals can close the loop for small calibration errors, object motion, and uncertain contacts during lifting.

## REFERENCES


[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[2] Yuanpei Chen, Yiran Geng, Fangwei Zhong, et al. Bidexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. arXiv preprint arXiv:2006.11239.

[4] Franziska Krebs and Tamim Asfour. A bimanual manipulation taxonomy. *IEEE Robotics and Automation Letters*, 7(4):11031–11038, 2022.

[5] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[6] Kaichun Mo, Shilin Zhu, Angel X. Chang, et al. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[7] Charles Ruizhongtai 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*, 2017.

[8] Yanming Shao and Chenxi Xiao. Bimanual grasp synthesis for dexterous robot hands. In *IEEE International Conference on Robotics and Automation*, 2025.

[9] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, et al. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *IEEE International Conference on Robotics and Automation*, 2023.

[10] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, et al. SAPIEN: A simulated part-based interactive environment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.