This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DemoFunGrasp: Universal Dexterous Functional Grasping via Demonstration-Editing Reinforcement Learning

Chuan Mao[1] Haoqi Yuan[1,2] Ziye Huang[1,2] Chaoyi Xu[1,2] Kai Ma[1] Zongqing Lu[1,2,†]
[1]Peking University [2]BeingBeyond

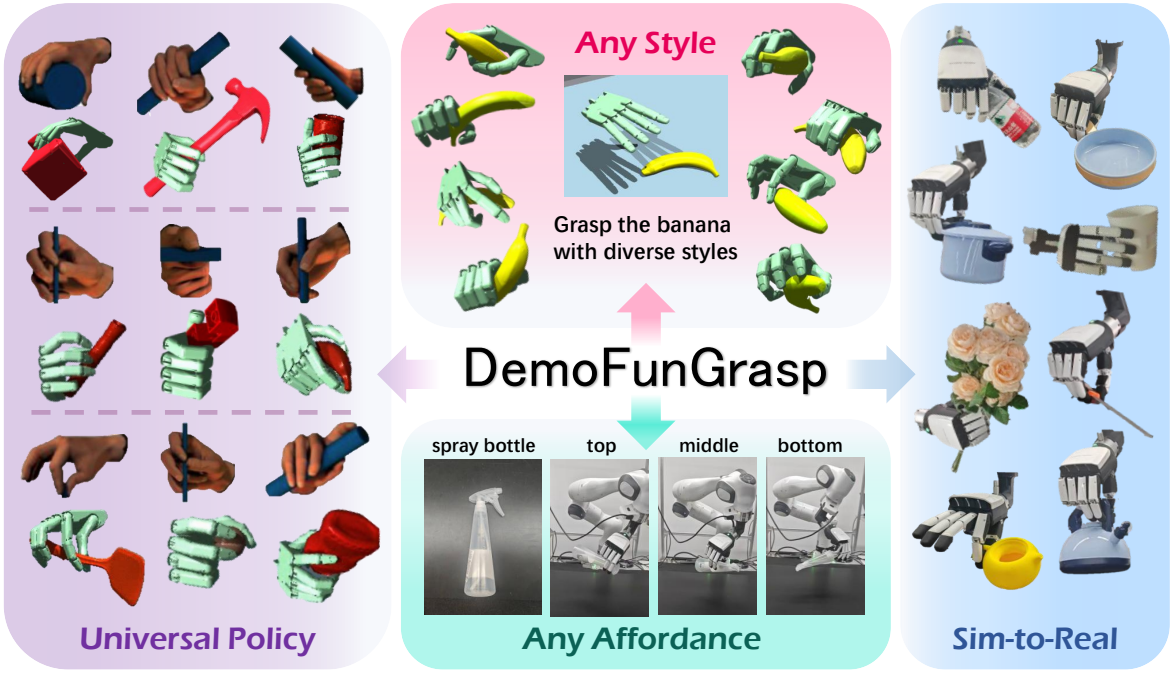


Figure 1. **DemoFunGrasp** is a reinforcement learning framework for universal dexterous functional grasping. The learned policy generalizes to unseen combinations of objects and functional grasping conditions, and achieves zero-shot sim-to-real transfer. For the same object, the policy can produce diverse grasps by adjusting the grasping style and affordance.

Abstract

Reinforcement learning (RL) has achieved great success in dexterous grasping, significantly improving grasp performance and generalization from simulation to the real world. However, fine-grained functional grasping, which is essential for downstream manipulation tasks, remains underexplored and faces several challenges: the complexity of specifying goals and reward functions for functional grasps across diverse objects, the difficulty of multi-task RL exploration, and the challenge of sim-to-real transfer. In this work, we propose ***DemoFunGrasp*** *for universal dexterous functional grasping. We factorize functional grasping conditions into two complementary components — grasping style and affordance — and integrate them into an RL framework that can learn to grasp diverse objects with functional grasping conditions. To address the multi-task optimization challenge, we leverage a single grasping demonstration and reformulate the RL problem as one-step demonstration editing, substantially enhancing sample efficiency and performance. Experimental results in both simulation and the real world show that DemoFunGrasp generalizes to unseen combinations of objects, affordances, and grasping styles, outperforming baselines in both success rate and functional grasping accuracy. In addition to strong sim-to-real capability, by incorporating a vision-language model (VLM) for planning, our system achieves autonomous instruction-following grasp execution.*

† Corresponding author: <lu@beingbeyond.com>

1. Introduction

Manipulation is a core capability of embodied intelligence, with grasping serving as its indispensable foundation. An effective grasp not only provides stability but also enables subsequent tool-use and manipulation tasks. Consequently, many studies [15, 30, 37, 41, 43] have explored learning robust grasping policies to achieve adaptive, closed-loop grasping in real-world tabletop settings. However, most existing approaches focus solely on achieving mechanical stability and do not consider the functional requirements of the grasp.

Functional grasping naturally highlights the advantages of dexterous robotic hands over parallel grippers. With a higher number of degrees of freedom (DoFs), dexterous hands can adapt to diverse object geometries and support a broad range of grasping styles. Recent studies have explored synthesizing human-like grasping poses by generating datasets [5, 7, 19, 33] or by learning from large-scale human data [3, 12, 14]. However, these methods typically rely heavily on human supervision, and the limited availability of high-quality data restricts their generalization to unseen objects. Furthermore, the generated poses are usually executed through open-loop planning, which limits their practicality in real-world tabletop scenarios. Other works [1, 36] leverage reinforcement learning (RL) in large-scale simulation to obtain closed-loop functional grasping policies. However, the high dimensionality of a dexterous hand's action space, together with the multi-task optimization challenge introduced by diverse objects and functional grasping styles, creates substantial difficulties for RL algorithm design and significantly limits the performance.

To tackle these challenges, we propose **DemoFunGrasp**, an RL framework that learns functional grasping policies capable of generalizing across diverse objects, grasp styles, and affordance conditions. We first decompose each functional grasping condition into two components: *affordance* and *grasping style*. The affordance specifies the region of the object to grasp (where to grasp), and the grasping style specifies the reference hand pose (how to grasp). They provide a complete description of the intended functional grasp. To learn a universal policy across all objects and functional conditions, we introduce an RL framework that incorporates these conditions into both the policy observations and the reward function. Diverse objects, affordances, and grasping styles are sampled in parallel simulation to train a universal policy. To address the multi-step and multi-task exploration challenge of this RL problem, we build on recent advances in demonstration-editing RL [39]. We collect one grasping demonstration and train a policy that outputs a wrist transformation (determining where to grasp) and a hand-style adaptation delta action (determining how to grasp). These actions edit the object-centric robot actions in the demonstration, which is then replayed for trial-and-error learning. This formulation reduces the problem to a single-step RL task and tightly links the action space with the functional grasping conditions, leading to significantly improved sample efficiency.

Experiments on 3,200 DexGraspNet objects [31] show that DemoFunGrasp achieves state-of-the-art performance, with higher affordance accuracy and greater style diversity compared with prior work. We further demonstrate the extensibility of our approach by applying it to multiple dexterous hand embodiments without hyperparameter tuning, achieving an overall success rate above 77% on human-intended grasping styles and affordances. To enable sim-to-real transfer, we collect 30k successful rollouts in simulation using the RL policy and distill them into an RGB-based policy through imitation learning. The resulting vision-based policy transfers to a real robot in a zero-shot manner and achieves a 71% success rate. Finally, by integrating a vision-language model (VLM) with the DemoFunGrasp policy, we construct an autonomous grasping system that reaches an average real-world success rate of 64.4% for functional grasping guided by language instructions.

Our main contributions are summarized as follows:

- We introduce DemoFunGrasp, a framework for universal dexterous functional grasping that can handle a wide range of objects under diverse affordance and grasping style.
- We integrate grasping affordances and styles into the observation space, reward function, and action space of demonstration-editing RL. This provides informative learning signals and significantly improves sample efficiency, allowing the policy to address the challenging multi-task optimization problem in functional grasping.
- Compared with prior work, our approach not only achieves higher success rates but also follows diverse grasping styles and affordances more accurately. We demonstrate that the learned policy can act as a low-level executor for VLMs, enabling autonomous language-guided functional grasping.

Code and project page are available at project page.

2. Related Work

2.1. Dexterous Functional Grasping

Functional grasping targets grasps aligned with an object's intended use rather than mere stability. Early works built human grasp taxonomies and mapped grasp forms to task intent [10, 29], providing foundations for functional synthesis. Recent methods exploit human priors and semantic cues to generate diverse functional grasps via optimization [7], data-driven learning [11, 31], cross-category contact transfer [35], and demonstration retargeting [14]. These generation frameworks are typically open-loop and fragile in tabletop execution. RL-based approaches learn closed-

loop functional policies [1, 36] but face large multi-task exploration spaces that hinder optimization and generalization. Other works use large VLMs for semantic conditioning [12, 32], yet rely on heavy annotations and inherit VLM robustness issues. We propose an efficient RL framework that learns a universal functional-grasp policy for diverse objects and conditions, enabling reliable sim-to-real transfer.

2.2. Affordance-Based Grasping

Affordance-based grasping aims to localize functionally meaningful regions that reflect intended object use. Classical methods [2, 8, 9, 17, 28] predict affordance maps from visual input, revealing grasp-supporting structures such as handles or rims. Recent works extend this paradigm to dexterous hands and multimodal reasoning [22, 23, 33, 34]. However, most approaches predominantly emphasize perception and affordance inference, while providing limited consideration for closed-loop control or execution reliability, which restricts their effectiveness in real-world manipulation settings. Our work bridges the gap between affordance perception and grasping execution by learning an end-to-end RL policy.

2.3. RL for Dexterous Manipulation

RL has shows strong capability for dexterous manipulation in simulation [24–26], but early methods exhibit limited generalization and weak sim-to-real transfer. Subsequent work improves data efficiency and robustness through techniques such as residual learning [15], human-in-the-loop adaptation [20], curriculum learning [30, 37, 41], and cross-embodiment transfer [38]. Demonstration-augmented approaches further reduce learning complexity by leveraging a small number of human demonstrations and formulating residual RL [44, 45] or one-step demonstration-editing RL [39], which alleviates the exploration burden. Works on sim-to-real transfer and tactile feedback [13, 18] further enhance real-world reliability in contact-rich scenarios. Building on these advances, our approach integrates demonstration editing RL to address the multi-task optimization challenge, learning a style- and affordance-conditioned closed-loop policy that tightly couples functional semantics with dexterous control.

3. Method

3.1. Problem Formulation

We formulate dexterous functional grasping as a one-step Markov Decision Process (MDP) following Yuan et al. [39]. Under the state-based setting, the agent observes:

$$(\mathbf{s}_r, \mathbf{s}_o, \mathbf{x}_o, p_{\text{afford}}, l_{\text{style}}).$$

Here, $\mathbf{s}_r$ denotes the robot end-effector 6D pose and $\mathbf{s}_o$ denotes the target object pose. $\mathbf{x}_o$ represents the complete object point cloud. $p_{\text{afford}} \in \mathbb{R}^3$ indicates the affordance point in 3D space, such as a cup handle or a lid edge, representing the desired functional contact region. l_{style} is a one-hot vector encoding the grasping style category.

The policy outputs an action $a = (\Delta T, \Delta q, k)$, where ΔT updates the end-effector pose, and Δq together with k modulates hand joint scaling coefficient relative to the target grasping style, conditioned on the object geometry.

For vision-based imitation learning, we use the following observation:

$$(\mathbf{s}_r, \mathbf{r}, c_{\text{afford}}, l_{\text{style}}),$$

where $\mathbf{r} = (R, G, B)$ and c_{afford} denotes the 2D projection of the affordance point in the RGB image. The visual policy predicts the end-effector pose and hand joint angles (t, r, q) in a closed-loop manner, providing smooth control based on raw image input.

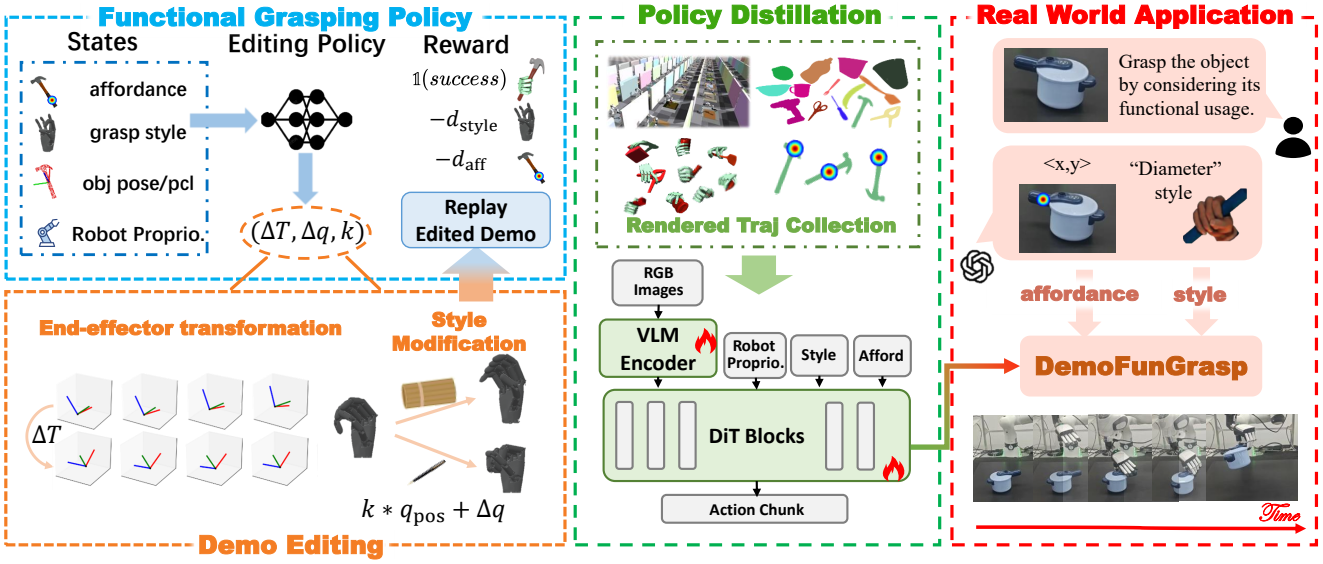
The state-based policy, which has access to privileged geometric information, accelerates RL and provides high-quality expert data. The vision-based policy emphasizes perceptual robustness for sim-to-real transfer.

3.2. One-Step Demonstration-Editing RL

Unlike conventional multi-step RL, which suffers from unstable optimization and high sample complexity in high-DoF manipulation, our framework simplifies functional grasping into a one-step RL problem. We observe that grasp success and functional accuracy can be effectively optimized by editing a single high-quality demonstration with residual actions learned through policy updates. This substantially reduces exploration difficulty and training cost while preserving the naturalistic grasp behavior.

Demonstration Representation. Each demonstration $D = \{(\mathbf{p}_t^{\text{ee-obj}}, \mathbf{q}_t^{\text{ref-hand}})\}_{t=0}^{T_D}$, which are collected by teleoperation, records the full end-effector–to–object trajectory and the corresponding hand joint sequence. At timestep t, $\mathbf{p}_t$ denotes the end-effector pose and $\mathbf{q}_t$ denotes the hand joint position. Unlike conventional three-stage static grasp data (pre-, grasp-, and post-contact), we preserve the continuous temporal evolution, including the timing of finger closure and the subtle compliance behaviors during contact formation. This temporal richness allows our one-step editing algorithm to interpolate motion in a physically plausible manner. The interpolation details are provided in Sec. 3.3.

Given a demonstration, the policy predicts $\{\Delta T, \Delta q, k\}$ as residual corrections to the base trajectory. The edited grasp is then executed once, and the reward metric outcome provides the RL signal for policy optimization algorithm. In this way, the policy effectively learns a smooth mapping from conditioned inputs to small corrective actions to improve grasp stability and contact precision. This "one-step editing" mechanism reformulates RL into a tractable refinement process rather than full motion synthesis, significantly reducing the difficulty of dexterous manipulation learning.



- For Chuan Mao: 需要你修改的点

Figure 2. Overview of **DemoFunGrasp**. (1) **Demonstration editing:** A source demonstration is adapted through end-effector transformation and object-geometry–aware hand style adjustment. (2) **Functional grasping policy learning:** An affordance- and style-conditioned one-step RL policy is trained. (3) **Vision-based imitation:** The learned policy is transferred to RGB observations for closed-loop, vision-based execution. (4) **Real-world deployment:** The vision-based policy is guided by a Vision-Language Model (VLM) for autonomous planning and execution.

Reward Function. The total reward combines functional alignment, style consistency, and contact smoothness:

$$r = \lambda_{\text{afford}} r_{\text{afford}} + \lambda_{\text{close}} r_{\text{close}} + \lambda_{\text{qpos}} r_{\text{qpos}} + r_{\text{success}}.$$

r_{success} rewards successful grasps, r_{afford} and r_{close} encourage spatial alignment with the desired affordance point, while r_{qpos} preserves stylistic consistency. Together, these terms ensure that edited grasps remain human-like and functional while adapting to novel object geometries. Their definition is in the following section.

3.3. Style-Aware Demonstration Editing

Styles Selection. We adopt the grasp taxonomy of Feix et al. [10] as an initial reference, and empirically refine it through large-scale simulation experiments. Our experiments reveal two key observations: (1) when normalized for object scale, multiple grasp styles converge to nearly identical hand configurations and contact patterns in successful cases, indicating redundancy; (2) for table-top manipulation, some grasping styles are physically infeasible, yielding consistently low success rates due to workspace constraints or unstable wrist orientation. Based on these findings, we prune the grasp styles to retain a representative subset, removing redundancies and styles with consistently poor performance during training.

Consequently, we adopt nine representative styles for the Shadow Hand (e.g., palmar pinch, lateral, small diameter), and four for the Inspire Hand, whose lower degrees of freedom naturally constrain its expressive range. Each selected style is parameterized by a canonical joint configuration $\mathbf{q}_{\text{pos}}$ and a contact mask that specifies the intended contact fingers. Examination of various human functional grasping patterns suggests that a functional grasp is determined once the hand's style, contact points, and the object's affordance point are specified.

Geometry-Conditioned Hand Pose Editing. Because grasp feasibility is highly sensitive to local geometry, we allow residual scaling and joint adjustments $(k, \Delta q)$ to adapt the hand posture according to the sampled object geometry $\mathbf{x}_o$. Our target hand joint configuration is $\mathbf{q}^*_{\text{pos}} = k \cdot \mathbf{q}_{\text{pos}} + \Delta\mathbf{q}$. Although the editing relies only on a single demonstration, this parameterization supports generalization to unseen object shapes via the learned geometric embedding. To balance the success rate and style intention, we design style reward:

$$r_{\text{qpos}} = \exp\left(-\|\mathbf{q}_{\text{pos}} - \mathbf{q}^*_{\text{pos}}\|_2\right).$$

Motion Interpolation. Given an edited target configuration $\mathbf{q}^*_{\text{pos}}$, we interpolate the motion along the reference trajectory using a fractional coefficient along the reference trajectory:

$$\mathbf{q}_i = \mathbf{q}_0^{\text{ref}} + f \cdot (\mathbf{q}_i^{\text{ref}} - \mathbf{q}_0^{\text{ref}}), \quad f = \frac{\mathbf{q}^*_{\text{pos}} - \mathbf{q}_0^{\text{ref}}}{\mathbf{q}_{T_l}^{\text{ref}} - \mathbf{q}_0^{\text{ref}}}$$

This interpolation produces smooth joint trajectories without requiring collision checks and preserves temporal coherence between fingers and object motion.

3.4. Affordance Conditions

Sampling Strategy. To provide rich functional supervision, it is necessary to sample diverse affordance points on the objects' surface that are physically feasible for grasping. We estimate an affordance likelihood distribution over the object point cloud using surface-normal alignment and the object's canonical pose after each environment randomly reset. In each training episode, a single affordance candidate is sampled from this distribution. This stochastic conditioning enables the policy to discover multiple valid functional grasp configurations for the same object.

Reward Shaping. We adopt a hierarchical reward consisting of a sparse proximity term r_{afford} and a dense alignment term r_{close}:

$$r_{\text{afford}} = \mathbb{I}\left(\text{success}\right)\mathbb{I}\left(d^{T-1} < \frac{\text{obj}_{\text{bb}}}{\gamma}\right)\exp\left(-d^{T-1}\right),$$

$$r_{\text{close}} = \mathbb{I}\left(d_{\text{min}}^{0:T-1} < \text{threshold}\right)$$

Here, $d_{\text{min}}^{0:T-1}$ denotes the minimum distance between the affordance point and the hand contact point along the trajectory, independent of grasp success. obj_{bb} denotes the length of the longest edge of the object's mesh bounding box, and γ is a hyperparameter. This object-scale normalization improves affordance accuracy across objects of different sizes.

The hierarchical reward design encourages early-stage exploration toward the affordance region via r_{close}, and fine-grained affordance alignment in later stages via r_{afford}, providing complementary guidance essential to stable and precise functional grasping.

3.5. Vision-Based Sim-to-Real Transfer

Data Collection. After convergence, we deploy the trained policy to collect large-scale demonstrations by executing successful grasps across a diverse set of object instances. Each sample consists of synchronized robot state observations, RGB images, functional conditions and control actions, forming a dataset for visual imitation learning.

Network Architecture. We evaluate three architectural variants to predict continuous action sequences (t, r, q): (i) ACT [42] with a pretrained ViT encoder, (ii) a ViT-based diffusion policy, and (iii) a diffusion transformer equipped with a Vision-Language Model encoder [4]. Our experiments show that using a VLM encoder with a DiT backbone achieves the best performance, as it effectively better captures the multi-modal structure of grasp strategies and better models the inherent uncertainty across different object instances. This benefit is most pronounced when different grasp styles correspond to distinct affordance semantics.

Domain Randomization. To narrow the sim-to-real gap, we perform aggressive domain randomization in IsaacGym, varying object and table textures, lighting directions and intensities, camera extrinsics, and initial object and robot poses.

Affordance Projection. Each 3D affordance point is projected onto the image plane using camera intrinsics and extrinsics, providing an explicit spatial cue for the visual encoder.

4. Experiments

We conduct experiments to answer the following questions: (1) How does our method perform in terms of grasp success rate, affordance alignment, and grasp-style diversity? (2) How do the reward design and editing mechanism influence functional grasp performance? (3) How effectively does our method transfer from simulation to real-world settings?

4.1. Experimental Settings

Train/Eval Settings. We train our policy in IsaacGym[21] using reinforcement learning. For state-based training, we construct a mixed-object dataset by combining YCB [6] and DexGraspNet [31], yielding a broad distribution of object geometries. During training, object poses are randomized within a $50 \times 50\,\text{cm}$ square region; affordance points and desired hand styles are selected randomly. We use PPO [27] as the underlying RL algorithm.

Once the state-based policy is obtained, we deploy it in simulation to collect 30k trajectories with randomization. These trajectories serve as supervision for training the vision-based policy. To evaluate the vision-based policy, we select three object categories for simulation and real-world experiments, respectively (Appendix B).

Evaluation Metrics. In the absence of fully comparable baselines for functional grasping, we benchmark our policy against existing grasp-generation and grasp-policy methods using four metrics:

1. **Grasp Success Rate (GSR)** is the fraction of successful grasps over the evaluation set. A grasp is considered successful if the object is lifted by at least 10 cm and remains stable for 20 time steps.
2. **Success Affordance Distance (SAD)** (simulation only) measures how closely the executed grasp aligns with the intended functional region. It is computed as the Euclidean distance between the target affordance point and style-aware contact point at the time of lift.
3. **Style Diversity (SD)** (simulation only) quantifies the diversity of successful grasping strategies by computing the average Euclidean distance between all pairs of successful hand joint configurations.
4. **Style Accuracy (SA)** (simulation only) measures the fraction of successful grasps that match the conditioned style:

$$SA = \frac{\text{style match \& success}}{\text{success}},$$

where a *style match* occurs if the executed grasp adopts the same style as the target condition. Higher SA indicates stronger adherence to the intended grasp style.

Real-World Settings. We conduct real-world experiments using a 6-DoF Inspire Hand (with 6 active and 6 passive joints) mounted on a 7-DoF Franka 3 arm. For evaluating vision-based policies, we employ two RealSense D435i cameras positioned diagonally around the table. The complete workspace configuration is shown in Appendix A.

4.2. Affordance-Conditioned Grasping

Existing works in affordance-based grasping lack a unified metric for execution accuracy. To evaluate DemoFunGrasp's performance across functional regions, we compare the diversity of successful grasp regions and mean success affordance distance with DemoGrasp [39], a state-of-the-art method.

We collect contact data from 1,000 simulated environments and visualize the contact point distributions in Fig. 3. As shown, DemoGrasp tends to focus on the tightest or most geometrically stable region of the object (e.g., purple points in the right hammer image). In contrast, DemoFunGrasp, conditioned on randomly sampled affordance points, generates diverse contact locations across different functional parts such as handles, rims, or edges, thereby enabling genuinely functional grasps.

We further quantify this effect by computing the mean distance between the achieved contact point and the target affordance point in successful trials, as reported in Table 1. Across both object categories seen and unseen in the training set, DemoFunGrasp reduces this affordance distance by over 3 cm compared to DemoGrasp, demonstrating its stronger alignment with the intended functional regions.

Table 1. **Mean Success Affordance Distance(SAD) (cm) between Achieved Contact Region and Target Affordance Point.**

Model	Train	Seen Cat.	Unseen Cat.
DemoGrasp	6.29	6.27	6.20
ours	**3.03**	**3.02**	**3.21**

4.3. Grasping Style Diversity

DemoFunGrasp avoids ambiguous or physically implausible hand poses that RL-only baselines sometimes produce (e.g., extending the middle finger while contracting others), which rarely occur in natural manipulation or functional grasps.

To quantify diversity and functional plausibility, we compare our method against UniDexGrasp [37], a state-of-the-art grasp-generation and RL pipeline.

Table 2. **Grasp Success Rate (GSR) and Style Diversity (SD) compared to an Unconditioned Baseline.**

Method	Seen Cat.		Unseen Cat.	
	GSR↑	SD↑	GSR↑	SD↑
UniDexGrasp	74.3	1.00	70.8	1.00
Ours	76.26	**1.48**	71.65	**1.44**
Ours (best style)	**81.74**	**1.48**	**78.68**	1.36

Table 2 shows that conditioning on style and affordance preserves GSR while substantially increasing style diversity. During training, average GSRs are 77.2% (DemoFunGrasp) and 79.4% (UniDexGrasp). Although slightly lower on the training set, our method generalizes better to unseen categories, indicating stronger robustness. By modeling local geometric features, DemoFunGrasp keeps GSR nearly unchanged across test domains and attains almost 1.5× the style diversity of UniDexGrasp, producing a broader range of physically meaningful, functionally distinct grasps.

Across the 9 tested styles, GSR ranges from 70.4% to 77.9% (mean 75.0%), demonstrating the robustness of style-conditioned grasping. We also report per-object performance for the *best-performing style*: for each instance, we replay all style candidates and record the best-performing one. Selecting only the best style, however, reduces diversity—especially on unseen categories—suggesting a compact subset of high-quality styles tends to dominate robustness.

Many existing methods ignore object–table interactions or constrain object orientation while allowing arbitrary hand approaches, which complicates fair tabletop comparisons. To highlight differences, we visualize hand-joint embeddings with t-SNE (Fig. 4) and evaluate three settings: UniDexGrasp, Dexonomy, and ours (using 9 initial styles from Dexonomy). After RL, styles remain well separated (blue convex hulls), confirming our conditioning preserves style distinctiveness; slight cluster shifts reflect optimization toward more stable, contact-consistent postures.

4.4. Ablation Study

We conduct an ablation study on the state-based policy to evaluate the contributions of reward design and style adaptation mechanisms. Table 3 summarizes the results, showing that each component plays a critical role in the overall performance.

Affordance Reward. Removing the affordance reward slightly increases the grasp success rate but substantially increases the success affordance distance, indicating that the policy tends to grasp at positions that are less functionally meaningful. This validates that the affordance reward effectively guides the policy toward target functional regions rather than merely maximizing stability.



Figure 3. The contact points distribution. Left: DemoFunGrasp, right DemoGrasp.

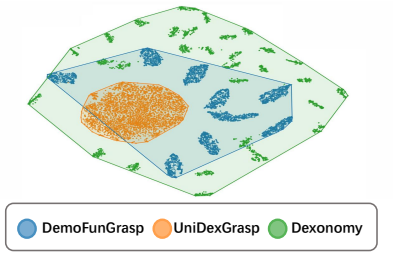


Figure 4. Grasp style diversity of UniDexGrasp (an RL method), Dexonomy (a grasp synthesis method), and ours.

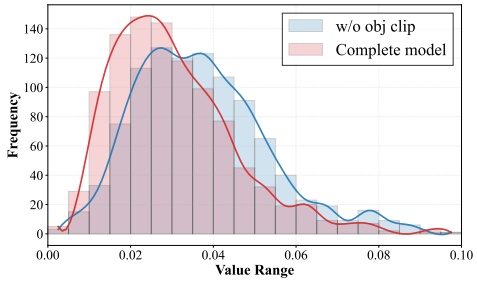


Figure 5. The affordance distance of the complete model and the model without object size aware clipping.

Table 3. **Comprehensive Ablation Study on Model Components.**

Method	GSR↑	SAD↓	SA↑
DemoFunGrasp	77.04	3.02	94.74
Component Ablation			
- w/o afford reward	78.28	4.60	95.02
- w/o obj size clipping	**81.67**	3.63	93.28
- w/o close reward	76.44	3.79	90.58
- w/o qpos reward	74.38	**2.98**	95.36
- w/o style disturbance	58.67	3.61	**100**

Object Size Clipping. When the clipping threshold does not account for object size, large objects tend to yield lower affordance rewards, resulting in suboptimal contact accuracy. Introducing object size clipping mitigates the training instability caused by the wide variation in object sizes. As shown in Fig. 5, the affordance distance distribution indicates that the size normalization parameter γ effectively functions as a clipping factor, alleviating performance bottlenecks in large-object affordance alignment.

Style Disturbance. Eliminating style disturbance severely harms performance reducing GSR to 58.67%. This highlights the importance of style adaptation: perturbation-based exploration ensures both robust and diverse grasping behaviors.

With the additional filters—*affordance distance*$<$*4 cm* and *style match*—the complete model achieves a balanced performance across grasp success, functional accuracy, and style fidelity, attaining the highest success rate of 60.55%.

4.5. RGB-based Policy Evaluation

We first evaluate our RGB-based policy in simulation on a set of objects (Appendix B.3) by randomly sampling affordances and styles to measure generalization capability. The grasp success rate and success affordance distance on the test set are 81.2% and 3.79cm, respectively. Demonstrations are provided in Appendix F.1.

To validate zero-shot sim-to-real transfer, we deploy the RGB policy directly in the real world. Objects are categorized into three groups (Appendix B.4) based on their functionality and geometry.

To generate functional affordance points automatically, we employ the state-of-the-art VLM *Embodied-R1* [40], with DemoFunGrasp used as the low-level control policy. The **Object Functional Grounding** prompt is used to extract affordance points reflecting object usage; details and comparisons to other VLMs are provided in Appendix E.

We report the following metrics: **Intended Affordance Score (IAS)** = 1 if the grasped point matches the target affordance, 0 otherwise; **Intended Style Score (ISS)** = 1 if the final hand configuration matches the conditioned style, 0 otherwise. The results are shown in Table 4.

In simulation, kitchen objects yield lower performance than food items, primarily due to small or thin objects (e.g., forks) occluding the camera view when the hand approaches.

In real-world experiments, we first evaluate on a set of 9 representative daily objects. DemoFunGrasp achieves a 71% grasp success rate (GSR) for human-chosen affordances and 64% for VLM-predicted affordances, demonstrating robustness in functional grasp generation.

To further assess generalization, we expand the real-world evaluation to 33 additional objects covering a broader range of geometries, including highly irregular and partially deformable items. Across 33×3 trials, the policy achieves a 70.7% success rate (70 successful grasps), indicating stable performance beyond the original benchmark set.

Main failure modes include: (i) affordance regions too small for precise grasping; (ii) objects outside the training distribution; and (iii) VLM predictions misaligned due to challenging object poses or inaccessible regions.

Overall, our method demonstrates strong performance on diverse daily objects, achieving high success rates with accurate affordance grounding and style adherence. Fig. 6 illustrates representative trials, while additional demonstrations are provided in Appendix F.2.

Table 4. **Simulation and Real-world Results with Human-Annotated/VLM-Predicted Grasping Conditions.**

Metrics	Simulator (human)			Real World (human)			Real World (VLM)		
	Food	Kitchen	Tool	Daily item	Small tool	Large tool	Daily item	Small tool	Large tool
GSR↑	22/25	18/25	20/25	11/15	11/15	10/15	12/15	10/15	7/15
IAS↑	0.96	0.86	0.80	0.80	0.93	0.80	0.87	0.67	0.40
ISS↑	1.00	0.82	0.92	0.80	1.00	0.93	0.93	0.87	0.87

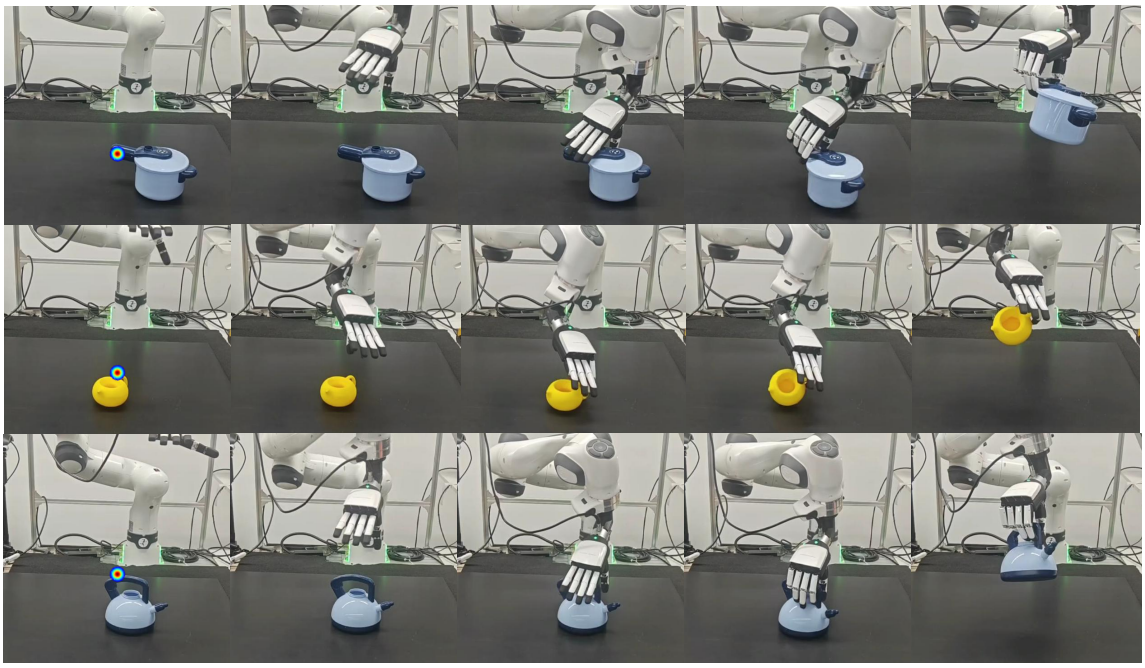


Figure 6. Real-world experimental demonstrations. The policy grasps a toy pot (first row), a teapot (second row), and a kettle (third row).

4.6. Additional Analysis

Cross-Embodiment Evaluation. We further assess the generality of our RL-based training framework across different robotic embodiments. Detailed experimental results and analysis are presented in Appendix C. Our findings indicate that DemoFunGrasp consistently achieves high success rates when trained independently on robotic hands with varying degrees of freedom (DoF) and style capacities, demonstrating that the framework can be readily applied to diverse embodiments without any hyperparameter tuning.

Base Algorithm Comparison. In addition to reinforcement learning, we also evaluate several baseline strategies. The comparative results and corresponding analyses are provided in Appendix G.

5. Conclusion and Limitations

We present a framework for universal dexterous functional grasping by integrating style- and affordance-conditioned policies with an efficient RL method. Our approach achieves robust zero-shot sim-to-real performance and flexible adaptation to varied object geometries and functionalities. The integration of Vision-Language Models further enables autonomous and human-like planning and execution in real-world settings. We hope that DemoFunGrasp serves as a step toward more versatile and semantically aware dexterous manipulation systems.

However, the current performance of DemoFunGrasp still falls short of human capability, largely due to the absence of dynamic pre-grasp manipulation and in-hand adjustment. Although the policy can grasp diverse objects across different affordance regions and styles, its accuracy remains at the centimeter scale, indicating room for further improvement. Future work may explore learning dynamic adjustment policies and incorporating tactile sensing to improve accuracy.

Acknowledgments

This work was supported by NSFC in part under Grant 62450001 and 62476008. The authors would like to thank the anonymous reviewers for their valuable comments and advice.

References

[1] Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping. *arXiv preprint arXiv:2312.02975*, 2023. 2, 3

[2] Paola Ardón, Eric Pairet, Ronald PA Petrick, Subramanian Ramamoorthy, and Katrin S Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578, 2019. 3

[3] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023. 2

[4] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 5

[5] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 2

[6] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics & Automation Magazine*, 2015. 5, 1

[7] Jiayi Chen, Yubin Ke, Lin Peng, and He Wang. Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. *Robotics: Science and Systems*, 2025. 2

[8] Renaud Detry, Dirk Kraft, Oliver Kroemer, Leon Bodenhagen, Jan Peters, Norbert Krüger, and Justus Piater. Learning grasp affordance densities. *Paladyn*, 2(1):1–17, 2011. 3

[9] Haonan Duan, Peng Wang, Yayu Huang, Guangyun Xu, Wei Wei, and Xiaofei Shen. Robotics dexterous grasping: The methods based on point cloud and deep learning. *Frontiers in Neurorobotics*, 15:658280, 2021. 3

[10] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems*, 46(1):66–77, 2015. 2, 4

[11] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[12] Jiawei He, Danshi Li, Xinqiang Yu, Zekun Qi, Wenyao Zhang, Jiayi Chen, Zhaoxiang Zhang, Zhizheng Zhang, Li Yi, and He Wang. Dexvlg: Dexterous vision-language-grasp model at scale. *arXiv preprint arXiv:2507.02747*, 2025. 2, 3

[13] Wenbin Hu, Bidan Huang, Wang Wei Lee, Sicheng Yang, Yu Zheng, and Zhibin Li. Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing. *Robotics and Autonomous Systems*, 186:104904, 2025. 3

[14] Linyi Huang, Hui Zhang, Zijian Wu, Sammy Christen, and Jie Song. Fungrasp: functional grasping for diverse dexterous hands. *IEEE Robotics and Automation Letters*, 2025. 2

[15] Ziye Huang, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Efficient residual learning with mixture-of-experts for universal dexterous grasping. *arXiv preprint arXiv:2410.02475*, 2024. 2, 3

[16] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14713–14724, 2023. 1

[17] Mia Kokic, Johannes A Stork, Joshua A Haustein, and Danica Kragic. Affordance detection for task-specific grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 91–98. IEEE, 2017. 3

[18] Toru Lin, Kartik Sachdev, Linxi Fan, Jitendra Malik, and Yuke Zhu. Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids. *arXiv preprint arXiv:2502.20396*, 2025. 3

[19] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2

[20] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 10(105): eads5033, 2025. 3

[21] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 5

[22] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 6169–6176. IEEE, 2021. 3

[23] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022. 3

[24] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafał Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *CoRR*, 2018. 3

[25] Ivaylo Popov, Nicolas Heess, Timothy Lillicrap, Roland Hafner, Gabriel Barth-Maron, Matej Vecerik, Thomas Lampe, Yuval Tassa, Tom Erez, and Martin Riedmiller. Data-efficient deep reinforcement learning for dexterous manipulation. *arXiv preprint arXiv:1704.03073*, 2017.

[26] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey

Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017. 3

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5

[28] Hyun Oh Song, Mario Fritz, Daniel Goehring, and Trevor Darrell. Learning to detect visual grasp affordance. *IEEE Transactions on Automation Science and Engineering*, 13 (2):798–809, 2015. 3

[29] Yu Sun, Eliza Amatova, and Tianze Chen. Multi-object grasping - types and taxonomy. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 777–783, 2022. 2

[30] 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*, pages 3891–3902, 2023. 2, 3

[31] 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. *arXiv preprint arXiv:2210.02697*, 2022. 2, 5, 1

[32] Yi-Lin Wei, Jian-Jian Jiang, Chengyi Xing, Xian-Tuo Tan, Xiao-Ming Wu, Hao Li, Mark Cutkosky, and Wei-Shi Zheng. Grasp as you say: Language-guided dexterous grasp generation. *Advances in Neural Information Processing Systems*, 37:46881–46907, 2024. 3

[33] Yi-Lin Wei, Mu Lin, Yuhao Lin, Jian-Jian Jiang, Xiao-Ming Wu, Ling-An Zeng, and Wei-Shi Zheng. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. *arXiv preprint arXiv:2503.07360*, 2025. 2, 3

[34] Yi-Lin Wei, Zhexi Luo, Yuhao Lin, Mu Lin, Zhizhao Liang, Shuoyu Chen, and Wei-Shi Zheng. Omnidexgrasp: Generalizable dexterous grasping via foundation model and force feedback. *arXiv preprint arXiv:2510.23119*, 2025. 3

[35] Rina Wu, Tianqiang Zhu, Xiangbo Lin, and Yi Sun. Cross-category functional grasp transfer. *IEEE Robotics and Automation Letters*, 2024. 2

[36] Tianhao Wu, Yunchong Gan, Mingdong Wu, Jingbo Cheng, Yaodong Yang, Yixin Zhu, and Hao Dong. Dexterous functional pre-grasp manipulation with diffusion policy. *arXiv preprint arXiv:2403.12421*, 2024. 2, 3

[37] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023. 2, 3, 6

[38] Haoqi Yuan, Bohan Zhou, Yuhui Fu, and Zongqing Lu. Cross-embodiment dexterous grasping with reinforcement learning. *arXiv preprint arXiv:2410.02479*, 2024. 3

[39] Haoqi Yuan, Ziye Huang, Ye Wang, Chuan Mao, Chaoyi Xu, and Zongqing Lu. Demograsp: Universal dexterous grasping from a single demonstration. *arXiv preprint arXiv:2509.22149*, 2025. 2, 3, 6

[40] Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. *arXiv preprint arXiv:2508.13998*, 2025. 7

[41] Hui Zhang, Zijian Wu, Linyi Huang, Sammy Christen, and Jie Song. Robustdexgrasp: Robust dexterous grasping of general objects. *arXiv preprint arXiv:2504.05287*, 2025. 2, 3

[42] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 5

[43] Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Zhang Chen, Tianrui Guan, Fanlian Zeng, Ka Num Lui, Yuyao Ye, Yitao Liang, et al. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. *arXiv preprint arXiv:2502.20900*, 2025. 2

[44] Bohan Zhou, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Learning diverse bimanual dexterous manipulation skills from human demonstrations. *arXiv preprint arXiv:2410.02477*, 2024. 3

[45] Kefei Zhu, Fengshuo Bai, YuanHao Xiang, Yishuai Cai, Xinglin Chen, Ruochong Li, Xingtao Wang, Hao Dong, Yaodong Yang, Xiaopeng Fan, et al. Dexflywheel: A scalable and self-improving data generation framework for dexterous manipulation. *arXiv preprint arXiv:2509.23829*, 2025. 3