# R$^2$D-GRASP: REWARD-REFINED DIFFUSION FOR PHYSICALLY RELIABLE DEXTEROUS GRASPING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Dexterous grasp diffusion models learn plausible and diverse grasps from pose supervision, but pose labels do not reveal whether small wrist or finger variations will lead to different physical outcomes during execution. We propose R²D-Grasp, a physics-feedback post-training framework for a wrist-conditioned grasp DDPM. R²D-Grasp samples independent denoising prefixes and generates multiple suffixes from each shared intermediate state. Simulation outcomes define a prefix-level global advantage and a within-prefix local advantage. The global advantage updates shared prefix transitions, while suffix transitions receive the combined global and local advantages, aligning coarse grasp-mode selection and local contact refinement with different denoising stages. At inference, a hand-frame geometric energy ranks wrist candidates according to object alignment and envelope compatibility. Across UniDexGraspNet, RealDex, MultiDex, and DexGRAB, R²D-Grasp achieves average success rates of 87.0% in free-space stability tests and 89.3% in tabletop grasping. When trained only on UniDexGraspNet, it reaches an average zero-shot success rate of 85.9% on the other three datasets.

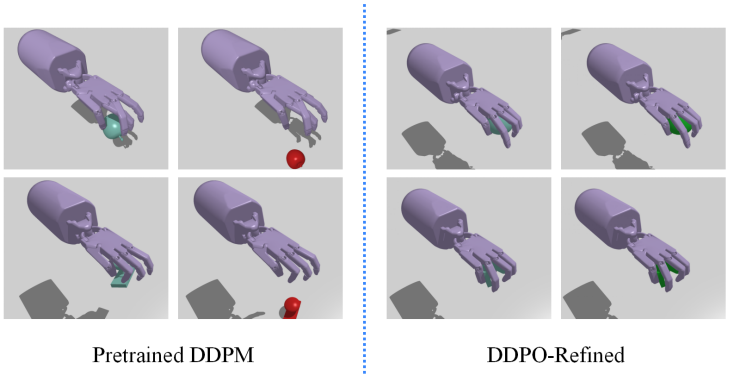



Figure 1: Physics-feedback post-training improves the execution stability of generated grasps.

## 1 INTRODUCTION

Dexterous grasping allows robots to pick up and manipulate everyday objects in unstructured environments. Anthropomorphic multi-fingered hands offer greater flexibility than parallel grippers. They also provide richer contact patterns and can adapt to objects with different shapes and sizes.

Learning-based methods generate multi-fingered grasps from large-scale pose data Wang et al. (2023); Xu et al. (2023); Liu et al. (2023). Diffusion models further capture high-dimensional and multimodal grasp distributions Huang et al. (2023); Weng et al. (2024); Lu et al. (2024). Physical constraints and contact guidance can improve geometric plausibility Jiang et al. (2021a); Zhong et al. (2025); Zhao et al. (2024). However, pose supervision does not describe the outcome of dynamic execution. Millimeter-scale changes in the wrist or fingertips can alter contacts, enclosure, and force directions. Geometrically similar poses may therefore produce different grasp outcomes.

Reinforcement learning uses execution feedback and has enabled dexterous manipulation and in-hand reorientation Rajeswaran et al. (2018); Andrychowicz et al. (2020). Dynamic grasping methods optimize wrist and finger control through curriculum learning, policy distillation, residual learn-



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.

ing, or trajectory editing Wan et al. (2023); Wang et al. (2025); Huang et al. (2025); Yuan et al. (2025). Their high-dimensional action spaces make exploration difficult. Reward-based diffusion post-training offers another route. DDPO and DPPO formulate reverse denoising as a sequential decision process Black et al. (2024); Ren et al. (2025). TreeGRPO further compares branches with shared denoising prefixes Ding & Ye (2026). However, TreeGRPO applies its combined advantage across the complete generation trajectory. Reverse diffusion contains decisions at different granularities. Early steps determine the overall grasp mode, whereas late steps adjust local contacts. Applying the same credit assignment to both stages obscures this distinction.

Motivated by this distinction, we propose $R^2$D-Grasp, a post-training framework that refines a grasp diffusion model with physical rewards. We start from a conditional DDPM pretrained with pose supervision Ho et al. (2020). We organize its reverse diffusion process as a hierarchical tree with shared prefixes and independent suffixes. The mean return under each shared prefix is normalized across prefixes. This produces a global advantage. The return deviation of each branch is normalized within the corresponding prefix group. This produces a local advantage. The two signals optimize prefix and suffix transitions, respectively. During inference, the wrist module proposes complete 6D wrist poses. A geometric alignment energy ranks these candidates using object-center alignment and the object envelope along the current hand orientation. The conditional DDPM keeps the approach orientation of each retained candidate. It then generates the final wrist translation and finger configuration.

We evaluate $R^2$D-Grasp on UniDexGraspNet Xu et al. (2023), RealDex Liu et al. (2024), MultiDex Li et al. (2023), and DexGRAB Taheri et al. (2020); Zhong et al. (2025). The evaluation covers free-space grasp stability, tabletop dynamic grasping, and zero-shot cross-dataset transfer. $R^2$D-Grasp achieves an average free-space success rate of $87.0\%$ and an average tabletop success rate of $89.3\%$ across the four datasets. When trained only on UniDexGraspNet, it achieves an average zero-shot success rate of $85.9\%$ on the other three datasets. We also study the roles of hierarchical physical rewards and geometric wrist-pose filtering. Our analysis includes shared-prefix diagnostics, post-training strategy comparisons, and wrist-strategy ablations.

Our main contributions are as follows:

- We propose $R^2$D-Grasp, which combines a pose-supervised conditional grasp diffusion model with hierarchical physical feedback from simulation. The framework improves the dynamic reliability of generated grasps while limiting the shift from the pretrained distribution.
- We introduce a local branch comparison mechanism based on shared intermediate denoising states. It uses sparse physical rewards to construct relative optimization signals from contact-reliability differences among similar grasps.
- We conduct systematic experiments under a unified dynamic grasping protocol. We evaluate grasp success, local physical sensitivity, changes in the generated distribution, and generalization across objects and datasets.

## 2 Related Work

### 2.1 Learning-Based Dexterous Grasping

Analytical simulators such as GraspIt! established contact-based grasp evaluation Miller & Allen (2004). Large-scale synthetic and real-world datasets later supported geometry-conditioned dexterous grasp generation Wang et al. (2023); Liu et al. (2024); Taheri et al. (2020); Li et al. (2023). Factorized and representation-based models capture wrist and finger structure Xu et al. (2023); Liu et al. (2023), while diffusion models represent multimodal grasp distributions Huang et al. (2023); Weng et al. (2024); Lu et al. (2024); Zhang et al. (2024). Physical constraints and contact guidance further improve geometric feasibility Jiang et al. (2021a); Zhong et al. (2025); Zhao et al. (2024). Reinforcement learning supports dexterous manipulation Rajeswaran et al. (2018); Andrychowicz et al. (2020), and dynamic methods optimize wrist and finger control through learned policies, curricula, distillation, residual experts, or trajectory editing Qin et al. (2022); Zhang et al. (2025); Wan et al. (2023); Wang et al. (2025); Huang et al. (2025); Yuan et al. (2025). Candidate 6-DoF poses are commonly obtained through point-cloud evaluation, generative sampling, scene annotations, or

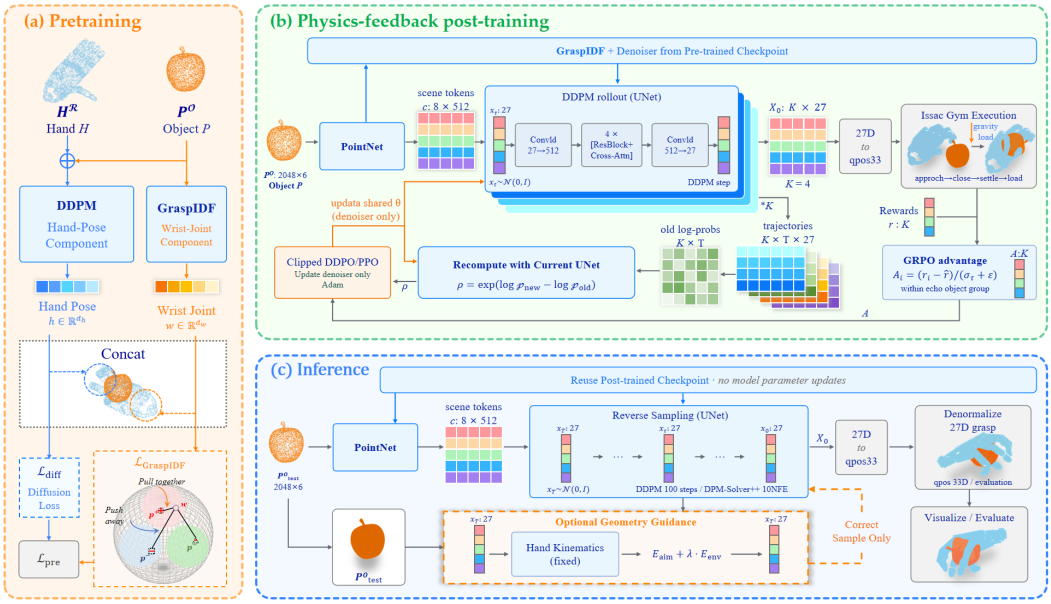



Figure 2: Overview of the R$^2$D-Grasp pipeline. (a) Pose-supervised pretraining learns the conditional grasp distribution. (b) Physics-feedback post-training evaluates sampled grasps in Isaac Gym and updates the denoiser with relative reward signals. (c) Inference reuses the post-trained checkpoint without parameter updates and applies geometry-aware selection before grasp execution.

graspability prediction Liang et al. (2019); Mousavian et al. (2019); Fang et al. (2020); Sundermeyer et al. (2021); Breyer et al. (2021); Jiang et al. (2021b). R$^2$D-Grasp retains diffusion-based pose generation but uses simulated execution outcomes for post-training. Its hand-frame envelope energy ranks wrist candidates by closing width and approach depth.

### 2.2 Reinforcement Learning for Diffusion Models

DDPM, DDIM, and score-based models provide discrete and continuous formulations of reverse denoising Ho et al. (2020); Song et al. (2021a;b). Diffusion models have since been applied to trajectory planning and visuomotor control Janner et al. (2022); Chi et al. (2023). DDPO formulates denoising as a finite-horizon decision process, and DPPO extends this formulation to continuous-control policies Black et al. (2024); Ren et al. (2025). DSRL instead keeps the base diffusion policy fixed and learns to select latent noise Wagenmaker et al. (2025). TreeGRPO constructs shared-prefix trees and propagates relative advantages across their trajectories Ding & Ye (2026). R$^2$D-Grasp places the tree at a fixed boundary in reverse diffusion. Mean returns across prefix groups define global advantages for prefix transitions, while relative returns within each group define local advantages for suffix transitions. The same physical outcomes therefore supervise overall grasp modes and local contact variations at their corresponding denoising stages.

## 3 Method

Figure 2 summarizes the three stages of R$^2$D-Grasp. We first pretrain a wrist-conditioned grasp DDPM from pose data, refine its reverse process with hierarchical physical rewards, and apply geometry-aware wrist selection during inference.

### 3.1 Wrist-Conditioned Grasp Prior

Given an object point cloud $O$, we represent a grasp as

$$h = (R, t, q), \qquad x_0 = (t, q) \sim p_\theta(x_0 \mid O, R). \tag{1}$$

Here, $R \in SO(3)$ is the wrist approach rotation, $t \in \mathbb{R}^3$ is the final wrist translation, and $q \in \mathbb{R}^{24}$ contains the Shadow Hand joint angles. A wrist-pose module proposes complete 6-DoF candidates Xu et al. (2023):

$$\mathcal{T}(O) = \left\{ \widetilde{T}_i = (R_i, \widetilde{t}_i) \right\}_{i=1}^{C}. \tag{2}$$

R$^2$D-Grasp retains only $R_i$ as the condition of the grasp model. The candidate translation $\widetilde{t}_i$ is used for geometric selection and approach initialization, but it is not the final wrist position. The conditional diffusion model instead generates $x_0 = (t, q)$. This decomposition lets one approach rotation support multiple wrist translations and finger configurations.

We pretrain the conditional DDPM with the standard noise-prediction objective Ho et al. (2020):

$$\mathcal{L}_{\mathrm{DDPM}}(\theta) = \mathbb{E}_{x_0,s,\epsilon}\Big[\big\|\epsilon - \epsilon_\theta\big( \sqrt{\bar{\alpha}_s} x_0 + \sqrt{1-\bar{\alpha}_s}\epsilon, s, O, R\big)\big\|_2^2\Big]. \tag{3}$$

Here, $s$ is the diffusion step, $\bar{\alpha}_s$ is the cumulative noise schedule, and $\epsilon$ is Gaussian noise. The denoiser learns to predict the noise added to the wrist translation and finger configuration $x_0 = (t, q)$, conditioned on the object geometry $O$ and approach rotation $R$. Its reverse transitions $p_\theta(x_{s-1} \mid x_s, O, R)$ define the pretrained grasp prior. We then execute generated grasps in simulation. Each rollout includes approach, closure, and lifting, and returns a binary outcome $r = \mathcal{E}(O, R, x_0) \in \{0, 1\}$. This physical reward is used to refine the pretrained diffusion policy. Appendix A.1 gives the complete diffusion and inference details.

### 3.2 Hierarchical Physical-Reward Refinement

R$^2$D-Grasp places a fixed branch boundary at diffusion step $\tau$ to separate mode-level and contact-level grasp decisions. For each condition $(O, R)$, we sample $M$ independent prefixes and copy the intermediate state $x_\tau^m$ of prefix $m$ into $K$ suffixes, following the shared-prefix sampling principle Ding & Ye (2026):

$$\mathcal{B}_m = \left\{ x_0^{m,k} \right\}_{k=1}^{K}, \qquad x_0^{m,k} \sim p_\theta(x_0 \mid x_\tau^m, O, R). \tag{4}$$

Different prefixes represent different overall grasp modes. Branches in the same set $\mathcal{B}_m$ share the early denoising history and differ only after $x_\tau^m$. Their differences therefore concentrate on local wrist and contact variations. After physical execution, each leaf receives a reward $r_{m,k}$. We compose the prefix-level and within-prefix comparisons directly from these rewards:

$$\begin{aligned}
\bar{r}_m &= \frac{1}{K}\sum_{k'=1}^{K} r_{m,k'}, \\
A_m^{\mathrm{g}} &= \frac{\bar{r}_m - \dfrac{1}{M}\displaystyle\sum_{m'=1}^{M} \bar{r}_{m'}}{\sqrt{\dfrac{1}{M}\displaystyle\sum_{m'=1}^{M}\left(\bar{r}_{m'} - \dfrac{1}{M}\displaystyle\sum_{m''=1}^{M}\bar{r}_{m''}\right)^2 + \delta}}, \\
A_{m,k}^{\mathrm{l}} &= \frac{r_{m,k} - \bar{r}_m}{\sqrt{\dfrac{1}{K}\displaystyle\sum_{k'=1}^{K}\left(r_{m,k'} - \bar{r}_m\right)^2 + \delta}}.
\end{aligned} \tag{5}$$

Here, $m, m', m'' \in \{1, \ldots, M\}$ index independent prefixes, while $k, k' \in \{1, \ldots, K\}$ index suffixes under a prefix. Each $r_{m,k} \in \{0, 1\}$ is the physical outcome of one leaf, $\bar{r}_m$ is the mean outcome of prefix $m$, and $\delta > 0$ is a small numerical stabilizer. The global advantage compares prefix means across the $M$ trees. The local advantage compares the $K$ outcomes that share prefix $m$.

For a shared transition at $s > \tau$, all suffixes under prefix $m$ reuse the same state pair $(x_s^m, x_{s-1}^m)$ and therefore have the same score gradient. Since the local advantages are centered within that prefix, their gradient contribution cancels:

$$\begin{aligned} &\sum_{k=1}^{K} A^{\mathrm{l}}_{m,k} \nabla_\theta \log p_\theta\big(x_{s-1}^m \mid x_s^m, O, R\big) \\ &\quad = \left(\sum_{k=1}^{K} A^{\mathrm{l}}_{m,k}\right) \\ &\qquad \cdot \nabla_\theta \log p_\theta\big(x_{s-1}^m \mid x_s^m, O, R\big) = 0, \qquad s > \tau. \end{aligned} \tag{6}$$

The local residual therefore cannot rank a shared prefix or increase its sampling probability. We use $A^{\mathrm{g}}_m$ to update prefix transitions and introduce $A^{\mathrm{l}}_{m,k}$ only after branching. The time-dependent advantage is

$$A_{m,k,s} = \begin{cases} A^{\mathrm{g}}_m, & s > \tau, \\ A^{\mathrm{g}}_m + \lambda A^{\mathrm{l}}_{m,k}, & s \le \tau. \end{cases} \tag{7}$$

This assignment associates mode-level comparison with early denoising and contact-level comparison with late denoising.

For unified notation, we write $x_s^{m,k} = x_s^m$ on the shared prefix $s > \tau$. All reverse transitions then use a single probability ratio

$$\rho_{m,k,s}(\theta) = \frac{p_\theta(x_{s-1}^{m,k} \mid x_s^{m,k}, O, R)}{p_{\theta_{\mathrm{old}}}(x_{s-1}^{m,k} \mid x_s^{m,k}, O, R)}. \tag{8}$$

Let the clipped surrogate for a transition be

$$\ell_\epsilon(\rho, A) = \min(\rho A, \mathrm{clip}(\rho, 1 - \epsilon, 1 + \epsilon)A)\,.$$

The complete objective exposes the two temporal credit regimes in a single expectation Schulman et al. (2017); Black et al. (2024); Ren et al. (2025):

$$\mathcal{L}_{\mathrm{R^2D}} = -\mathbb{E}_{O,R}\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{s=\tau+1}^{S} \ell(\rho_{m,s}, A^{\mathrm{g}}_m) + \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{s=1}^{\tau} \ell\big(\rho_{m,k,s}, A^{\mathrm{g}}_m + \lambda A^{\mathrm{l}}_{m,k}\big)\right], \tag{9}$$

The first sum updates shared prefix transitions with global credit. The second updates suffix transitions with both global and local credit. Appendix A.2 derives the prefix cancellation and unified objective.

### 3.3 Geometry-Aware Wrist Selection

The initial approach pose strongly affects grasp execution. At inference time, we therefore rank each wrist candidate $\widetilde{T}_i = (R_i, \widetilde{t}_i)$ with a training-free geometric energy before sampling the final grasp. The aiming energy measures the wrist distance and the lateral offset of the object center in the palm plane:

$$E_{\mathrm{aim},i} = \mathrm{rank}\big(\|R_i(-\widetilde{t}_i)\|_2\big) + \lambda_l \,\mathrm{rank}\left(\sqrt{(a_c^\top R_i(-\widetilde{t}_i))^2 + (a_f^\top R_i(-\widetilde{t}_i))^2}\right). \tag{10}$$

The envelope energy measures whether the object fits the hand under the current orientation:

$$E_{\mathrm{env},i} = \left[\frac{w_{i,c} - W_{\max}}{W_{\max}}\right]_+^2 + \lambda_d \left[\frac{w_{i,p} - D_{\max}}{D_{\max}}\right]_+^2. \tag{11}$$

Here, $a_c$ and $a_f$ are the finger closing and extension axes, and $[z]_+ = \max(0, z)$. The quantities $w_{i,c}$ and $w_{i,p}$ are robust object extents along the closing axis $a_c$ and approach axis $a_p$. The limits $W_{\max}$ and $D_{\max}$ describe the closing width and approach depth that the hand can accommodate. The operator $\mathrm{rank}(\cdot)$ is the percentile rank within the candidate pool. The two terms are combined as

$$E_{\mathrm{geo},i} = E_{\mathrm{aim},i} + \lambda_e \,\mathrm{rank}(E_{\mathrm{env},i}). \tag{12}$$

The aiming term favors candidates aligned with the reachable palm region, while the envelope term rejects orientations that exceed the hand's closing or approach capacity. Lower-energy candidates provide the approach rotations used by the conditional DDPM, which then predicts the final wrist translation and finger configuration. In tabletop grasping, the same criterion ranks the yaw directions allowed by the palm-down task constraint.

# 4 Experiments

The experiments answer four questions. (1) Does physical post-training improve both the stability of generated grasp configurations and complete tabletop execution? (2) Can suffix variations after a shared denoising prefix produce distinct physical outcomes? (3) How do global prefix credit and local suffix credit contribute to the improvement? (4) Does the learned policy transfer to object distributions that are absent from training?

## 4.1 Experimental Setup

We conduct all experiments in Isaac Gym Makoviychuk et al. (2021) with a Shadow Hand that has 24 finger degrees of freedom and a floating 6-DoF wrist. The model takes 2,048 object points with surface normals as input. We evaluate on UniDexGraspNet Xu et al. (2023), RealDex Liu et al. (2024), MultiDex Li et al. (2023), and DexGRAB Taheri et al. (2020); Zhong et al. (2025). All reported results use the designated test objects. Simulation outcomes are used during physical post-training but are unavailable to every method at test time.

We use two complementary execution protocols. In free space, the hand approaches from a pre-grasp pose, closes, and stabilizes before the grasp is tested under gravity and perturbations from four horizontal directions. This protocol measures the physical stability of the generated configuration without table contact or a task-specific controller. In the tabletop protocol, an object is released with a random orientation and grasping starts after it settles. The hand then approaches, closes, lifts the object by $0.15\,\mathrm{m}$, and holds it. A trial succeeds if the object is raised by more than $0.05\,\mathrm{m}$, remains in the hand, and moves by no more than $0.10\,\mathrm{m}$ relative to the hand during holding. We report the fraction of successful trials as the success rate. We use matched object sets, initial object states, success criteria, and trial budgets whenever the methods admit a common evaluation interface.

All diffusion-based variants share the same DDPM backbone Ho et al. (2020), pretrained initialization, and Isaac Gym rollout budget. The default hierarchical sampler uses two shared prefixes and four suffixes per prefix. R$^2$D-Grasp (3D$\rightarrow$Table) projects the free-space post-trained model into the tabletop action space. R$^2$D-Grasp (Table) instead performs physical post-training in the tabletop environment. Table 1(b) trains a separate model on each dataset, while Table 3 fixes the training source to UniDexGraspNet. Appendix B gives the complete execution schedules, optimization settings, and hyperparameters.

Table 1: Grasp success rates (%) in free space and on the tabletop. R$^2$D-Grasp achieves the best result on all four datasets under both evaluation settings.

*(a) Free-space grasp stability*

| Method | UniDex | RealDex | MultiDex | DexGRAB |
|---|---|---|---|---|
| GraspTTA Jiang et al. (2021a) | 63.8 | 42.0 | 62.5 | 55.0 |
| DexGraspAnything Zhong et al. (2025) | 78.6 | 82.0 | 85.5 | 72.0 |
| **R$^2$D-Grasp** | **85.3** | **85.1** | **91.6** | **85.8** |

*(b) Tabletop dynamic grasping*

| Method | UniDex | RealDex | MultiDex | DexGRAB |
|---|---|---|---|---|
| UniDexGrasp Xu et al. (2023) | 84.5 | 65.0 | 72.0 | 67.0 |
| DemoGrasp Yuan et al. (2025) | 89.2 | 82.0 | 90.5 | 82.0 |
| R$^2$D-Grasp (3D→Table) | 82.0 | 79.7 | 84.0 | 77.0 |
| **R$^2$D-Grasp (Table)** | **90.8** | **88.4** | **93.0** | **84.9** |

### 4.2 Free-Space and Tabletop Grasping

Table 1 reports success rates under the two protocols. Panel (a) isolates the stability of the generated grasp configuration. Panel (b) evaluates the complete tabletop sequence and retains the native control pipeline of each dynamic baseline.

R$^2$D-Grasp achieves the highest free-space success rate on all four datasets. It improves over DexGraspAnything by 6.7, 3.1, 6.1, and 13.8 points on UniDexGraspNet, RealDex, MultiDex, and DexGRAB, respectively. The gains appear across all four object distributions and are largest on DexGRAB. Because this protocol evaluates the final configuration under a common perturbation test, the comparison isolates grasp stability from tabletop motion and controller design.

The 3D→Table model reaches 77.0–84.0% without tabletop post-training. This result measures how far the free-space grasp prior can be extended through task projection alone. R$^2$D-Grasp (Table) raises success by 8.8, 8.7, 9.0, and 7.9 points over this transferred model on the four datasets. It also exceeds the strongest dynamic baseline by 1.6, 6.4, 2.5, and 2.9 points, respectively. The two R$^2$D-Grasp variants therefore distinguish direct deployment of a free-space model from physical adaptation to tabletop constraints.

### 4.3 Reward-Refinement Analysis

Figure 3 examines whether suffix branches retain outcome-relevant variation after sharing a reverse-diffusion prefix. The branching ratio $\rho$ denotes the fraction of denoising steps shared within each group. Across the tested ratios, grasp success remains between 89.34% and 90.59%. Even at $\rho = 0.85$, 50.0% of the groups contain both successful and failed grasps. Thus, late denoising steps can still alter the physical outcome under a largely fixed grasp mode.

Figure 4 provides qualitative examples. Longer shared prefixes preserve the global approach pattern, while suffix branches mainly vary the wrist translation and finger configuration.

Figure 4 provides qualitative examples. Longer shared prefixes preserve the global approach pattern, while suffix branches mainly vary the wrist translation and finger configuration.

Physical post-training provides the largest change in success. Relative to supervised DDPM, R$^2$D-Grasp improves by 17.3 points on UniDexGraspNet and 30.1 points on RealDex. Full-Trajectory DDPO and Local-only refinement also improve over the supervised model, but each omits one level of credit. Combining global prefix credit with local suffix credit gives a further 1.7 and 2.4 points over Full-Trajectory DDPO, and 4.5 and 6.7 points over Local-only refinement. The comparison shows that the prefix-level and suffix-level objectives contribute complementary information.

The remaining columns examine whether the gain is accompanied by a large change in the sampled grasp distribution. Full-Trajectory DDPO and R$^2$D-Grasp have approximate KL values of 0.15 and 0.12 relative to the supervised DDPM. The sampled KL is estimated after training and is distinct from the 0.03 threshold used to limit each policy update. R$^2$D-Grasp therefore attains the higher

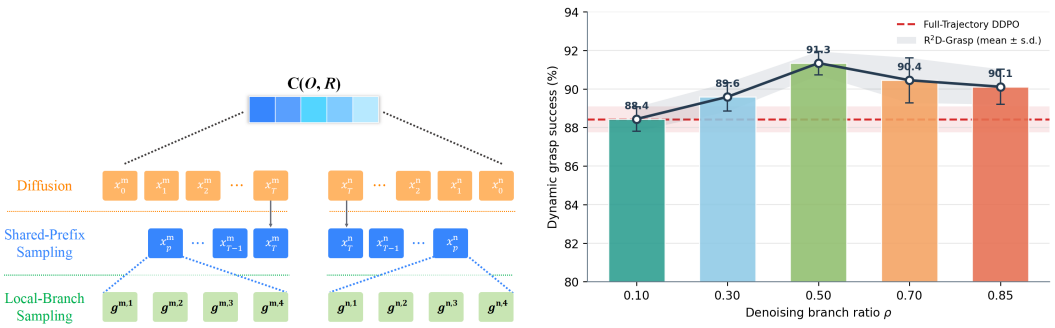



Figure 3: Shared-prefix branching and branch-ratio analysis. (a) Local branches share a denoising prefix and differ after the branch point. (b) Dynamic grasp success of models trained with different branch ratios on RealDex. All settings use matched objects, wrist conditions, and initial noise, and are evaluated over five runs.

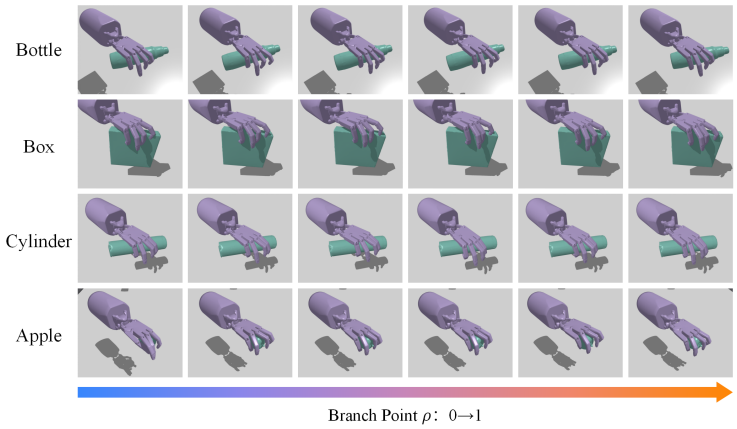



Figure 4: Qualitative shared-prefix samples at different denoising branch ratios $\rho$. Each row shares the same object and wrist condition.

success rate with the smaller of the two measured distribution shifts. Its joint distance is 0.26, compared with 0.42 for supervised DDPM, so the samples are more concentrated under a fixed object and wrist condition. The invalid-candidate rate separately decreases from 10.8% for supervised DDPM to 2.5% for R$^2$D-Grasp. The improved success rate thus coincides with fewer invalid pre-grasps rather than a larger overall shift from the supervised sampling distribution.

### 4.4 Generalization and Geometry-Aware Wrist Selection

We train all methods on UniDexGraspNet and evaluate them directly on MultiDex, RealDex, and DexGRAB without target-domain updates. The execution protocol and success criteria match Table 1(b).

R$^2$D-Grasp achieves the best result on the source set and all three target datasets. It exceeds the strongest baseline by 3.4, 12.7, and 12.4 points on MultiDex, RealDex, and DexGRAB, respectively. These results show that the UniDexGraspNet-trained policy transfers across object collections without target-domain refinement.

Table 4 further isolates wrist proposal and geometry-aware selection. In free space, replacing random candidates with the wrist predictor improves success by 17.4 points on MultiDex and 21.6 points on RealDex. Geometry-aware selection adds another 4.0 and 6.0 points, bringing the learned candidates within 0.7 and 1.3 points of the GT-initialized oracle. On the tabletop task, geometry-aware selection also outperforms random and long-axis strategies under both post-training settings. The larger wrist-related differences on RealDex are consistent with its higher proportion of elongated and flat objects, whose feasible grasps are more sensitive to the approach direction.

Table 2: Comparison of physical post-training strategies under the free-space stability protocol. SR is the grasp success rate (%). Approximate KL is measured from inference samples relative to the supervised DDPM. R$^2$D-Grasp achieves the highest SR on both datasets.

| | SR (%) | | Sample statistics | | |
|---|---|---|---|---|---|
| Method | UniDex | RealDex | Joint dist. | KL ↓ | Invalid (%) ↓ |
| Supervised DDPM Ho et al. (2020) | 68.0 | 55.0 | 0.42 | 0.00 | 10.8 |
| Full-Trajectory DDPO Black et al. (2024) | 83.6 | 82.7 | 0.31 | 0.15 | 2.9 |
| Local-only refinement | 80.8 | 78.4 | – | – | – |
| **R$^2$D-Grasp** | **85.3** | **85.1** | 0.26 | **0.12** | **2.5** |

Table 3: Zero-shot cross-dataset tabletop grasp success (%). All methods are trained on UniDexGraspNet and evaluated without target-domain updates.

| | Source domain | Unseen target domains | | |
|---|---|---|---|---|
| Method | UniDex | MultiDex | RealDex | DexGRAB |
| UniDexGrasp Xu et al. (2023) | 84.0 | 72.0 | 67.0 | 58.0 |
| DemoGrasp Yuan et al. (2025) | 89.2 | 85.6 | 70.9 | 72.7 |
| **R$^2$D-Grasp (Table)** | **90.8** | **89.0** | **83.6** | **85.1** |

Table 4: Ablation of wrist proposal and geometry-aware selection. (a) Free-space candidates. (b) Tabletop strategies under 3D or tabletop post-training. Values are success rates (%); bold marks the best non-oracle result in each setting.

*(a) Free-space wrist candidates*

| Wrist source | | | Geometry | SR (%) | |
|---|---|---|---|---|---|
| Random | GT (oracle) | Predictor | Energy | MultiDex | RealDex |
| ✓ | | | ✓ | 74.2 | 63.5 |
| | ✓ | | | 92.3 | 86.4 |
| | | ✓ | | 87.6 | 79.1 |
| | | ✓ | ✓ | **91.6** | **85.1** |

*(b) Tabletop wrist strategies*

| | Wrist strategy | | | SR (%) | |
|---|---|---|---|---|---|
| Post-training | Random | Long-axis | Energy | MultiDex | RealDex |
| Free-space | ✓ | | | 60.2 | 54.3 |
| Free-space | | ✓ | | 65.7 | 71.8 |
| Free-space | | | ✓ | **84.0** | **79.7** |
| Tabletop | ✓ | | | 82.3 | 62.1 |
| Tabletop | | ✓ | | 84.1 | 78.2 |
| Tabletop | | | ✓ | **93.0** | **88.4** |

## 5 Conclusion

Pose-supervised grasp diffusion models do not directly optimize execution reliability, while terminal rewards provide coarse credit across denoising steps. R$^2$D-Grasp organizes reverse diffusion into shared prefixes and independent suffixes. Global advantages optimize overall grasp modes, and local advantages refine contact decisions. During inference, geometric energy reranks wrist candidates before the conditional DDPM generates final grasps.

Experiments on four datasets show gains in free-space stability, tabletop dynamic grasping, and zero-shot transfer. Compared with full-trajectory post-training, R$^2$D-Grasp achieves higher success with a smaller sampled distribution shift and fewer invalid candidates. Ablations show complementary gains from wrist prediction and geometric filtering, especially on the anisotropic objects in RealDex. These results support stage-specific credit assignment for refining grasp diffusion models with physical feedback.

## REFERENCES


Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petroni, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020. doi: 10.1177/0278364919887447. URL https://doi.org/10.1177/0278364919887447.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

Michel Breyer, Jen Jen Chung, Lionel Ott, Roland Siegwart, and Juan Nieto. Volumetric grasping network: Real-time 6 DOF grasp detection in clutter. In *IEEE International Conference on Robotics and Automation*, 2021. URL https://arxiv.org/abs/2101.01132.

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*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://arxiv.org/abs/2303.04137.

Zheng Ding and Weirui Ye. TreeGRPO: Tree-advantage GRPO for online RL post-training of diffusion models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.08153.

Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. GraspNet-1Billion: A large-scale benchmark for general object grasping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11444–11453, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Fang_GraspNet-1Billion_A_Large-Scale_Benchmark_for_General_Object_Grasping_CVPR_2020_paper.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2301.06015.

Ziye Huang, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Efficient residual learning with mixture-of-experts for universal dexterous grasping. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02475.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, volume 162, pp. 9902–9915, 2022. URL https://proceedings.mlr.press/v162/janner22a.html.

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *IEEE/CVF International Conference on Computer Vision*, pp. 11107–11116, 2021a. URL https://openaccess.thecvf.com/content/ICCV2021/html/Jiang_Hand-Object_Contact_Consistency_Reasoning_for_Human_Grasps_Generation_ICCV_2021_paper.html.

Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-DoF grasp detection via implicit representations. In *Robotics: Science and Systems*, 2021b. doi: 10.15607/RSS.2021.XVII.024. URL https://arxiv.org/abs/2104.01542.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. GenDexGrasp: Generalizable dexterous grasping. In *IEEE International Conference on Robotics and Automation*, pp. 8068–8074, 2023. URL https://arxiv.org/abs/2210.00722.

Hongzhuo Liang, Xiaojian Ma, Shuang Li, Michael Görner, Song Tang, Bin Fang, Fuchun Sun, and Jianwei Zhang. PointNetGPD: Detecting grasp configurations from point sets. In *IEEE International Conference on Robotics and Automation*, pp. 3629–3635, 2019. doi: 10.1109/ICRA.2019.8794435. URL https://arxiv.org/abs/1809.06267.

Qingtao Liu, Yu Cui, Qi Ye, Zhengnan Sun, Haoming Li, Gaofeng Li, Lin Shao, and Jiming Chen. Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. *arXiv preprint arXiv:2303.09806*, 2023. URL https://arxiv.org/abs/2303.09806.

Yumeng Liu, Yaxun Yang, Youzhuo Wang, Xiaofei Wu, Jiamin Wang, Yichen Yao, Sören Schwertfeger, Sibei Yang, Wenping Wang, Jingyi Yu, Xuming He, and Yuexin Ma. Realdex: Towards human-like grasping for robotic dexterous hand. In *International Joint Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2402.13853.

Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. UGG: Unified generative grasping. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2311.16917.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High performance GPU-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. URL https://arxiv.org/abs/2108.10470.

Andrew T. Miller and Peter K. Allen. GraspIt!: A versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. doi: 10.1109/MRA.2004.1371616. URL https://doi.org/10.1109/MRA.2004.1371616.

Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-DOF GraspNet: Variational grasp generation for object manipulation. In *IEEE/CVF International Conference on Computer Vision*, pp. 2901–2910, 2019. doi: 10.1109/ICCV.2019.00299. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Mousavian_6-DOF_GraspNet_Variational_Grasp_Generation_for_Object_Manipulation_ICCV_2019_paper.html.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Qi_PointNet_Deep_Learning_CVPR_2017_paper.html.

Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Hao Su, and Xiaolong Wang. Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation. In *Conference on Robot Learning*, pp. 594–605, 2022. URL https://openreview.net/forum?id=tJE1Yyi8fUX.

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. In *Robotics: Science and Systems*, 2018. doi: 10.15607/RSS.2018.XIV.049. URL https://www.roboticsproceedings.org/rss14/p49.html.

Allen Z. Ren, Justin Lidard, Lars L. Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.00588.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *IEEE International Conference on Robotics and Automation*, 2021. URL https://arxiv.org/abs/2103.14127.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, pp. 581–600, 2020. URL https://grab.is.tue.mpg.de/.

Andrew Wagenmaker, Mitsuhiko Nakamoto, Yunchu Zhang, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. *arXiv preprint arXiv:2506.15799*, 2025. URL https://arxiv.org/abs/2506.15799.

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 *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2304.00464.

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*, 2023. URL https://arxiv.org/abs/2210.02697.

Wenbo Wang, Fangyun Wei, Lei Zhou, Xi Chen, Lin Luo, Xiaohan Yi, Yizhong Zhang, Yaobo Liang, Chang Xu, Yan Lu, Jiaolong Yang, and Baining Guo. Unigrasptransformer: Simplified policy distillation for scalable dexterous robotic grasping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.02699.

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.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2303.00938.

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. URL https://arxiv.org/abs/2509.22149.

Hui Zhang, Zijian Wu, Linyi Huang, Sammy Christen, and Jie Song. RobustDexGrasp: Robust dexterous grasping of general objects from single-view perception. *arXiv preprint arXiv:2504.05287*, 2025. URL https://arxiv.org/abs/2504.05287.

Zhengshen Zhang, Lei Zhou, Chenchen Liu, Zhiyang Liu, Chengran Yuan, Sheng Guo, Ruiteng Zhao, Marcelo H. Ang Jr., and Francis E. H. Tay. Dexgrasp-diffusion: Diffusion-based unified functional grasp synthesis pipeline for multi-dexterous robotic hands. *arXiv preprint arXiv:2407.09899*, 2024. URL https://arxiv.org/abs/2407.09899.

Fuqiang Zhao, Dzmitry Tsetserukou, and Qian Liu. Graingrasp: Dexterous grasp generation with fine-grained contact guidance. *arXiv preprint arXiv:2405.09310*, 2024. URL https://arxiv.org/abs/2405.09310.

Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.08257.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html.

# A Additional Method and Implementation Details

## A.1 Conditional Grasp Diffusion

R$^2$D-Grasp represents a grasp as $h = (R, t, q)$ and models the remaining grasp variables as $x_0 = (t, q)$ after conditioning on the object point cloud $O$ and wrist rotation $R$. For a clean sample $x_0$, the forward diffusion process at step $s$ is

$$q(x_s \mid x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_s}x_0, (1-\bar{\alpha}_s)I\big)\,, \tag{13}$$

or equivalently

$$x_s = \sqrt{\bar{\alpha}_s}x_0 + \sqrt{1-\bar{\alpha}_s}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{14}$$

The denoising network predicts the injected noise from $(x_s, s, O, R)$. This prediction parameterizes the reverse transition

$$p_\theta(x_{s-1} \mid x_s, O, R) = \mathcal{N}(\mu_\theta(x_s, s, O, R), \Sigma_s)\,. \tag{15}$$

The point-cloud encoder maps 2,048 object points with surface normals to a geometry feature $f_O$. The wrist rotation is represented by its continuous 6D encoding (Zhou et al., 2019). We combine the object feature, wrist feature, and diffusion-step embedding with the noisy grasp state in the denoising network. The resulting conditional model learns the distribution of final wrist translations and finger configurations under a fixed approach rotation.

The wrist module first proposes complete candidates $\widetilde{T}_i = (R_i, \widetilde{t}_i)$. The candidate translation $\widetilde{t}_i$ initializes the approach and supports geometric evaluation. It is not copied into the final grasp. After candidate selection, the diffusion model retains $R_i$ as its condition and generates a new pair $(t, q)$. This separation allows one approach rotation to support multiple final wrist positions and contact configurations.

## A.2 Hierarchical Physical-Reward Refinement

Let $\mathbf{x}_{S:0} = (x_S, \ldots, x_0)$ denote a reverse-diffusion trajectory. A branch boundary $\tau$ divides its transition likelihood into a prefix and a suffix:

$$\begin{aligned} p_\theta(\mathbf{x}_{S:0} \mid O, R) = &\prod_{s=\tau+1}^{S} p_\theta(x_{s-1} \mid x_s, O, R) \\ &\cdot \prod_{s=1}^{\tau} p_\theta(x_{s-1} \mid x_s, O, R). \end{aligned} \tag{16}$$

The first product generates a shared state $x_\tau^m$ for prefix $m$. We copy this state and independently sample $K$ suffixes. Repeating the process for $M$ prefixes produces

$$\mathcal{B}_m = \{x_0^{m,k}\}_{k=1}^{K}, \qquad x_0^{m,k} \sim p_\theta(x_0 \mid x_\tau^m, O, R). \tag{17}$$

Physical execution assigns a binary return $r_{m,k}$ to every leaf. Define the prefix mean and the mean over all prefix groups as

$$\bar{r}_m = \frac{1}{K}\sum_{k=1}^{K} r_{m,k}, \qquad \bar{r} = \frac{1}{M}\sum_{m=1}^{M} \bar{r}_m. \tag{18}$$

The centered leaf return has the exact decomposition

$$r_{m,k} - \bar{r} = \underbrace{(\bar{r}_m - \bar{r})}_{\text{between-prefix}} + \underbrace{(r_{m,k} - \bar{r}_m)}_{\text{within-prefix}}. \tag{19}$$

The first term compares complete grasp modes represented by different prefixes. The second compares contact-level alternatives that share the same early denoising history. Standardizing each component within its comparison group gives

$$\begin{aligned} A^{\mathrm{g}}_m &= \frac{\bar{r}_m - \bar{r}}{\sqrt{\frac{1}{M}\sum_{u=1}^{M}(\bar{r}_u - \bar{r})^2 + \delta}}, \\ A^{\mathrm{l}}_{m,k} &= \frac{r_{m,k} - \bar{r}_m}{\sqrt{\frac{1}{K}\sum_{j=1}^{K}(r_{m,j} - \bar{r}_m)^2 + \delta}}, \end{aligned} \tag{20}$$

where $\delta > 0$ is a numerical stabilizer. Because the local residuals share a common denominator and sum to zero within every prefix group, their aggregate score-gradient contribution to a shared transition also vanishes. They can distinguish suffixes after branching, but cannot assign preference among the shared transitions of prefix $m$. The corresponding transition-level advantage is therefore

$$A_{m,k,s} = \begin{cases} A^{\mathrm{g}}_m, & s > \tau, \\ A^{\mathrm{g}}_m + \lambda A^{\mathrm{l}}_{m,k}, & s \le \tau. \end{cases} \tag{21}$$

Let $\theta_{\mathrm{old}}$ denote the rollout policy and define

$$\rho_{m,k,s}(\theta) = \frac{p_\theta(x^{m,k}_{s-1} \mid x^{m,k}_s, O, R)}{p_{\theta_{\mathrm{old}}}(x^{m,k}_{s-1} \mid x^{m,k}_s, O, R)}. \tag{22}$$

For a transition with advantage $A$, the clipped surrogate is

$$\ell_\epsilon(\rho, A) = \min(\rho A, \mathrm{clip}(\rho, 1-\epsilon, 1+\epsilon)A)\,. \tag{23}$$

Substituting Eq. 21 into the diffusion-policy objective yields

$$\begin{aligned} \mathcal{L}_{R^2D}(\theta) = -\mathbb{E}_{(O,R)}\Bigg[\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\Bigg(&\sum_{s=\tau+1}^{S}\ell_\epsilon(\rho_{m,k,s}, A^{\mathrm{g}}_m) \\ &+ \sum_{s=1}^{\tau}\ell_\epsilon\big(\rho_{m,k,s}, A^{\mathrm{g}}_m + \lambda A^{\mathrm{l}}_{m,k}\big)\Bigg)\Bigg]. \end{aligned} \tag{24}$$

Equation 24 optimizes one diffusion policy with two temporal credit regimes. Prefix transitions are compared through the mean physical reliability of their descendant grasps. Suffix transitions additionally receive the within-prefix residual that distinguishes nearby contact outcomes.

The post-training procedure alternates between collecting branched trajectories and updating the denoiser:

1. Fix the rollout parameters $\theta_{\text{old}} \leftarrow \theta$. For each condition $(O, R)$, sample $M$ independent initial noise states and denoise each state to the branch boundary $\tau$.
2. Copy each shared state $x_\tau^m$ into $K$ branches and sample their suffixes independently under $p_{\theta_{\text{old}}}$. Retain the sampled transitions and their rollout-policy log probabilities.
3. Decode each of the $MK$ leaves into a grasp and execute it in simulation to obtain $r_{m,k} \in \{0, 1\}$. Compute the prefix means, global advantages, and local advantages using Eqs. 18 and 20.
4. Compute the transition likelihood ratios in Eq. 22 and optimize Eq. 24, assigning global credit to prefix transitions and combined global–local credit to suffix transitions.

The physical simulator supplies training rewards. Inference uses the resulting diffusion policy and geometric candidate selection without querying simulation returns.

### A.3 Geometry-Aware Wrist Selection

For a wrist candidate $\widetilde{T}_i = (R_i, \widetilde{t}_i)$, we transform the complete object point cloud $P_O$ into the candidate hand frame:

$$P_i^h = \left\{ R_i(x - \widetilde{t}_i) \mid x \in P_O \right\}. \tag{25}$$

For hand-frame axis $k$, the robust object extent is

$$w_{i,k} = Q_{95}(P_{i,k}^h) - Q_5(P_{i,k}^h). \tag{26}$$

Using percentile extents reduces the influence of isolated points. Let $a_c$, $a_f$, and $a_p$ denote the closing, finger-extension, and approach axes. The object center in the candidate hand frame is $c_i^h = R_i(-\widetilde{t}_i)$. The aiming term is

$$\begin{aligned} E_{\text{aim},i} = \; & \operatorname{rank}\big(\|c_i^h\|_2\big) \\ & + \lambda_l \operatorname{rank}\Big(\sqrt{(a_c^\top c_i^h)^2 + (a_f^\top c_i^h)^2}\Big). \end{aligned} \tag{27}$$

The first component measures wrist distance to the object center. The second measures lateral displacement within the palm plane. The envelope term evaluates whether the object fits the hand under the current orientation:

$$E_{\text{env},i} = \left[ \frac{w_{i,c} - W_{\max}}{W_{\max}} \right]_+^2 + \lambda_d \left[ \frac{w_{i,p} - D_{\max}}{D_{\max}} \right]_+^2, \tag{28}$$

where $[z]_+ = \max(0, z)$. The energy penalizes closing widths and approach depths that exceed the hand capacity. It does not directly penalize the extent along $a_f$, which allows elongated geometry to align with the finger direction. The final candidate score is

$$E_{\text{geo},i} = E_{\text{aim},i} + \lambda_e \operatorname{rank}(E_{\text{env},i}). \tag{29}$$

All rank operators compute percentile ranks within the candidate pool. The wrist module first generates 32 candidates, and Eq. 29 retains the four lowest-energy poses. The wrist-module probability is used only to resolve equal energies. For every retained rotation, the DDPM generates eight final wrist translations and finger configurations. Execution starts from a pre-grasp pose placed $0.10\,\text{m}$ behind the candidate along the approach direction and linearly interpolates to the generated final grasp.

For tabletop grasping, the task protocol fixes a palm-down orientation and the admissible wrist position. We enumerate 16 yaw directions and rank them with the envelope term. The four lowest-energy directions condition the same grasp generator. This operation projects the three-dimensional geometric criterion into the task's admissible rotation space.

Table 5: Geometry-aware selection parameters.

| Parameter | Value |
|---|---|
| Wrist candidates | 32 |
| Retained candidates | 4 |
| DDPM samples per rotation | 8 |
| $W_{\max}$ | $0.07\,\mathrm{m}$ |
| $D_{\max}$ | $0.10\,\mathrm{m}$ |
| $\lambda_l$ | 0.67 |
| $\lambda_d$ | 1 |
| $\lambda_e$ | 1 |
| Tabletop yaw candidates | 16 |
| Retained yaw directions | 4 |

# B Experimental Details

## B.1 Datasets and Evaluation Settings

We evaluate on UniDexGraspNet (Xu et al., 2023), RealDex (Liu et al., 2024), MultiDex (Li et al., 2023), and DexGRAB (Taheri et al., 2020; Zhong et al., 2025). The main per-dataset comparison trains a separate model on the corresponding training split and evaluates on its designated test objects. The zero-shot comparison instead fixes UniDexGraspNet as the only training source and directly evaluates the resulting model on the other three object collections. Target-domain objects do not participate in supervised pretraining or physical post-training in this setting.

Each method receives the same object point cloud and uses the same object set and initial physical state whenever a common evaluation interface is available. Dynamic baselines retain their native grasp generation and control pipelines. This preserves their intended execution behavior while matching the environment and success criterion.

## B.2 Execution Protocols

All experiments use Isaac Gym (Makoviychuk et al., 2021), a Shadow Hand with 24 finger degrees of freedom, and a floating 6-DoF wrist. The simulation timestep is $1/60\,\mathrm{s}$.

**Free-space stability.** The hand begins $0.10\,\mathrm{m}$ behind the target along the palm normal. It approaches for 180 frames, closes for 90 frames, and stabilizes for 180 frames. The resulting grasp is then evaluated under gravity for 300 frames and perturbed for 200 frames in each of four horizontal directions. This protocol evaluates the generated terminal configuration without table contact or a task-specific motion controller.

**Tabletop dynamic grasping.** The object is released $0.15\,\mathrm{m}$ above the table with a random orientation. Grasp execution begins after the object remains below the stability threshold for 30 consecutive frames. The hand follows the same 180-frame approach, 90-frame closure, and 180-frame stabilization schedule. It then lifts the object by $0.15\,\mathrm{m}$ over 120 frames and holds it for another 120 frames. A trial succeeds when the object rises more than $0.05\,\mathrm{m}$ above its initial table height, remains in the hand, and moves by no more than $0.10\,\mathrm{m}$ relative to the hand during holding.

## B.3 Metrics

The grasp success rate is the fraction of executed trials that satisfy the corresponding free-space or tabletop criterion. The reward-refinement analysis additionally reports three statistics computed under a fixed object and wrist condition.

For $L$ generated joint configurations, the average pairwise joint distance is

$$D_q = \frac{2}{L(L-1)} \sum_{1 \le a < b \le L} \|q^{(a)} - q^{(b)}\|_2. \tag{30}$$

Table 6: Training and implementation settings used by the reported models.

| Configuration | Value |
|---|---|
| Object points | 2,048 with normals |
| Optimizer | Adam (Kingma & Ba, 2015) |
| Learning rate | $1 \times 10^{-5}$ |
| PPO clipping range | 0.1 |
| Maximum KL per update | 0.03 |
| Shared prefixes $M$ | 2 |
| Suffixes per prefix $K$ | 4 |
| GPU | RTX 4090, 24 GB |

Approximate KL measures the shift of post-training inference samples relative to the supervised DDPM. It is computed after training and is separate from the $0.03$ threshold that limits each policy update. The invalid pre-grasp rate is the fraction of candidates that fail the palm-orientation or safety-clearance checks before execution.

For the shared-prefix diagnostic, a prefix group has mixed outcomes when its suffix set contains at least one successful and one failed grasp. For $G$ evaluated groups, the mixed-outcome ratio is

$$R_{\text{mix}} = \frac{1}{G} \sum_{g=1}^{G} \mathbf{1}\left[ \min_k r_{g,k} = 0 \ \wedge \ \max_k r_{g,k} = 1 \right]. \tag{31}$$

### B.4 Training Configuration and Controlled Comparisons

All diffusion variants use the same conditional grasp backbone (Ho et al., 2020), PointNet object encoder (Qi et al., 2017), and supervised initialization. Full-Trajectory DDPO (Black et al., 2024) and R$^2$D-Grasp use matched Isaac Gym rollout budgets. For comparisons among trained diffusion policies, we match the evaluation objects, wrist conditions, initial diffusion noise, physical states, and execution protocol. The trained transition distribution $p_\theta$ is the changing factor.

Supervised DDPM receives no physical update. Full-Trajectory DDPO assigns terminal credit along the complete reverse trajectory. Local-only refinement updates only suffix transitions with within-prefix credit. R$^2$D-Grasp applies global credit to the shared prefix and combined global–local credit to the suffix. These variants isolate the effect of temporal credit assignment while preserving the generative backbone and initialization.

The tabletop experiments contain two R$^2$D-Grasp variants. R$^2$D-Grasp (3D$\rightarrow$Table) reuses the free-space post-trained model and projects its wrist candidates into the tabletop action space. R$^2$D-Grasp (Table) performs physical post-training in the tabletop environment. Both variants use the same grasp representation and geometry-aware candidate ranking.

### B.5 Wrist Proposal and Geometric Selection Protocol

The wrist comparison separates candidate generation from candidate ranking. In free space, the candidate sources are random wrist proposals, dataset ground-truth wrist initialization, and the learned wrist predictor. Geometric ranking is applied to each source. A further configuration retains the learned predictor and replaces geometric ranking with the predictor's probability score. Holding the candidate and output budgets fixed makes the latter comparison specific to the ranking rule. The ground-truth initialization is a reference configuration that uses dataset wrist poses rather than a deployment-time proposal mechanism.

Under tabletop constraints, the palm-down orientation and admissible wrist position are fixed by the execution protocol. The comparison therefore concerns the remaining yaw direction. The three alternatives are random yaw, object long-axis guidance, and the envelope-based energy mapping described in Appendix A.3. Each alternative uses the same candidate and generated-grasp budgets. The comparison is applied separately to the free-space post-trained model and the tabletop post-

trained model, keeping the diffusion policy fixed within each comparison. This separates changes in orientation selection from changes in the post-training environment.

## C AI USE STATEMENT

In this paper, large language models (LLMs) were used to polish the text, adjust the formatting, prepare tables, provide writing suggestions for selected sections, and assist with code implementation and debugging. All AI-assisted content and code were reviewed and verified by the authors.