

EvolvingGrasp: Evolutionary Grasp Generation via Efficient Preference Alignment

Yufei Zhu^{1,*}, Yiming Zhong^{1,*}, Zemin Yang¹, Peishan Cong¹,
Jingyi Yu¹, Xinge Zhu^{2,†}, Yuexin Ma^{1,†}

¹ ShanghaiTech University ² The Chinese University of Hong Kong

<https://evolvinggrasp.github.io/>

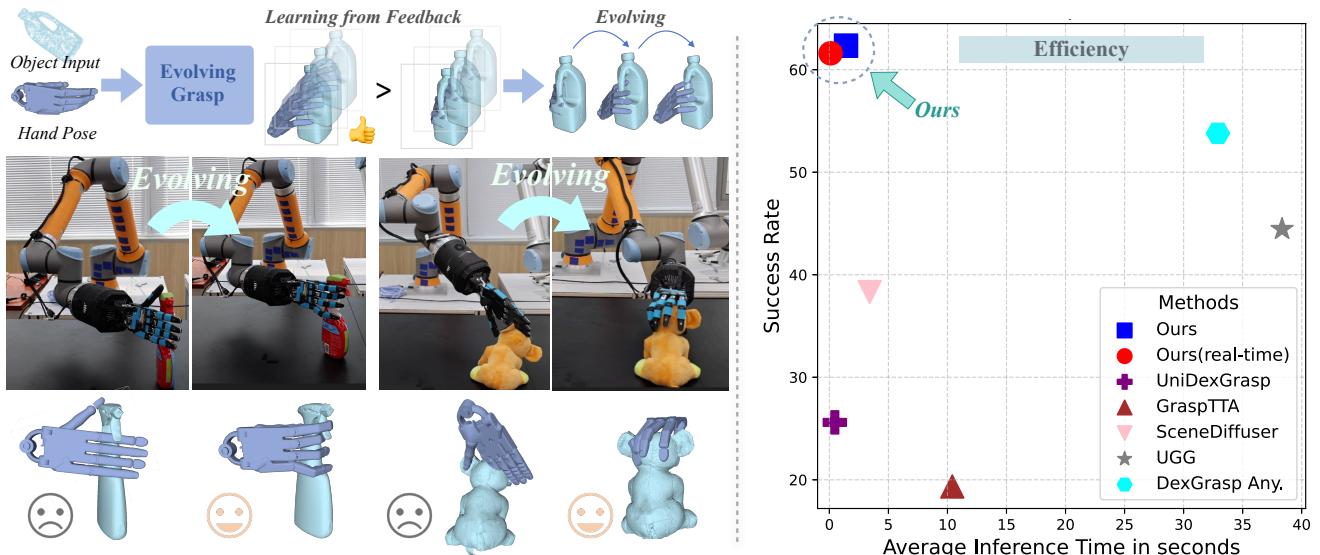


Figure 1. The left part illustrates **EvolvingGrasp**, an approach akin to evolution, where it enables the model to learn from experience and iteratively refine its grasping strategy. The right part demonstrates its efficiency and effectiveness

Abstract

Dexterous robotic hands often struggle to generalize effectively in complex environments due to the limitations of models trained on low-diversity data. However, the real world presents an inherently unbounded range of scenarios, making it impractical to account for every possible variation. A natural solution is to enable robots learning from experience in complex environments—an approach akin to evolution, where systems improve through continuous feedback, learning from both failures and successes, and iterating toward optimal performance. Motivated by this, we propose *EvolvingGrasp*, an evolutionary grasp gen-

eration method that continuously enhances grasping performance through efficient preference alignment. Specifically, we introduce Handpose-wise Preference Optimization (HPO), which allows the model to continuously align with preferences from both positive and negative feedback while progressively refining its grasping strategies. To further enhance efficiency and reliability during online adjustments, we incorporate a Physics-aware Consistency Model within HPO, which accelerates inference, reduces the number of timesteps needed for preference fine-tuning, and ensures physical plausibility throughout the process. Extensive experiments across four benchmark datasets demonstrate state-of-the-art performance of our method in grasp success rate and sampling efficiency. Our results validate that *EvolvingGrasp* enables evolutionary grasp generation, ensuring robust, physically feasible, and preference-aligned grasping in both simulation and real scenarios.

^{1*} Equal contribution.

^{2†} Corresponding author. This work was supported by NSFC (No.62206173), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo).

1. Introduction

Dexterous robotic grasping has made significant strides in embodied manipulation, enabling more adaptive and precise interactions compared to traditional grippers. Existing grasping methods [5, 9, 12, 16, 18–20, 30, 32–34, 43, 48, 49, 51] generally fall into two categories: optimization-based approaches [5, 9, 30, 33], which refine hand poses to achieve force-closure states, learning-based approaches [27, 41, 42, 44], which directly map object features to grasp poses through regression, probabilistic modeling, and generative-based methods [13, 14, 25, 26, 43], which utilize diffusion model to estimate the distribution of hand poses. Recent advances, such as DexGrasp Anything [52], have further introduced physics-based constraints to improve grasp feasibility. However, a fundamental limitation persists—limited generalization. These methods, trained on limited datasets, struggle to adapt to complex environments. This challenge is exacerbated by an inherent property of the real world: **its unbounded diversity**. The vast range of object shapes, materials, and environmental conditions makes it impractical to predefine an exhaustive set of grasping strategies. Without the ability to adapt in deployment, grasping models remain constrained, failing to handle variations beyond their training distribution. To overcome this, a natural approach is to enable evolutionary grasp generation, where the system **learns from experience** (*i.e.*, both failures and successes), refining its grasping strategy through iterative improvements based on real-world interactions. This process not only enhances generalization but also allows for preference alignment, ensuring that grasping behaviors adapt to task-specific requirements and user-defined preferences. However, achieving efficient evolutionary refinement is non-trivial, as many existing learning-based approaches rely on slow, compute-intensive updates, particularly in diffusion-based models that require numerous iterative steps and physics simulations.

To address these challenges, we propose EvolvingGrasp, an evolutionary grasp generation framework that efficiently refines grasp strategies through preference alignment while maintaining physical plausibility. At its core, we introduce **Handpose-wise Preference Optimization** (HPO), a novel method that reformulates preference alignment [2, 38, 39, 50, 53] as a posterior probability optimization problem, encouraging generated grasps to converge toward preferred distributions while diverging from non-preferred ones. Notably, the proposed HPO is an extension of Direct Preference Optimization (DPO) [31], where it is also, to the best of our knowledge, the first to incorporate the DPO into the dexterous grasp. To further improve efficiency, we integrate HPO into a **Physics-Aware Consistency Model** (including two parts, *i.e.*, **Physics-Aware Distillation** for training and **Physics-Aware Sampling** for inference), which pretrains a

diffusion model and distills it into a lightweight, few-step sampling model. This enables both **rapid inference and efficient preference fine-tuning**, significantly reducing the number of required sampling steps and optimization iterations. Additionally, we introduce three physics-aware constraints to ensure the stability, realism, and feasibility of generated grasp poses—surface pulling force to maintain stable contact, external penetration repulsion force to prevent object penetration, and self-penetration repulsion force to avoid inter-finger collisions.

Extensive experiments across four benchmark datasets demonstrate that EvolvingGrasp achieves state-of-the-art results, significantly improving grasp success rate, sampling efficiency, and physical plausibility, with **30x speedup** over existing methods, demonstrating robust generalization across simulated and real-world benchmarks. Our contributions can be summarized as follows:

- We introduce EvolvingGrasp, an efficient evolutionary grasp generation framework that enables iterative refinement, addressing the challenge of generalizing to diverse and unstructured real-world environments.
- Efficient preference alignment is achieved through Handpose-wise Preference Optimization (HPO), which formulates grasp adaptation as a posterior probability optimization problem, enabling the model to iteratively converge toward preferred grasp distributions.
- We propose a Physics-Aware Consistency Model (PCM) that accelerates preference alignment by reducing sampling steps while enforcing geometric consistency and physical feasibility through structured constraints.
- Extensive experiments across four benchmark datasets demonstrate that our method achieves state-of-the-art grasp success rates, physical plausibility, and sampling efficiency. Furthermore, it enables real-time grasp generation with minimal computational overhead, achieving comparable performance to gradient-based methods.

2. Related Work

2.1. Dexterous Grasping

Dexterous grasp generation aims to produce diverse and high-quality grasping poses for robotic hands to interact with objects effectively. Recent works can be categorized into regression-based [18, 19, 32] and generation-based methods [13, 14, 25–27, 41, 42]. Regression-based methods directly predict grasping parameters from the input object, but they often suffer from mode collapse issues that limits output diversity. Generation-based methods, though capable of producing varied solutions, face efficiency challenges. SceneDiffuser [13], UGG [25], and DexGrasp Anything [52] require multiple sampling steps to generate diverse grasping poses, and DexGrasp Anything [52] additionally incurs computational overhead by calculating phys-

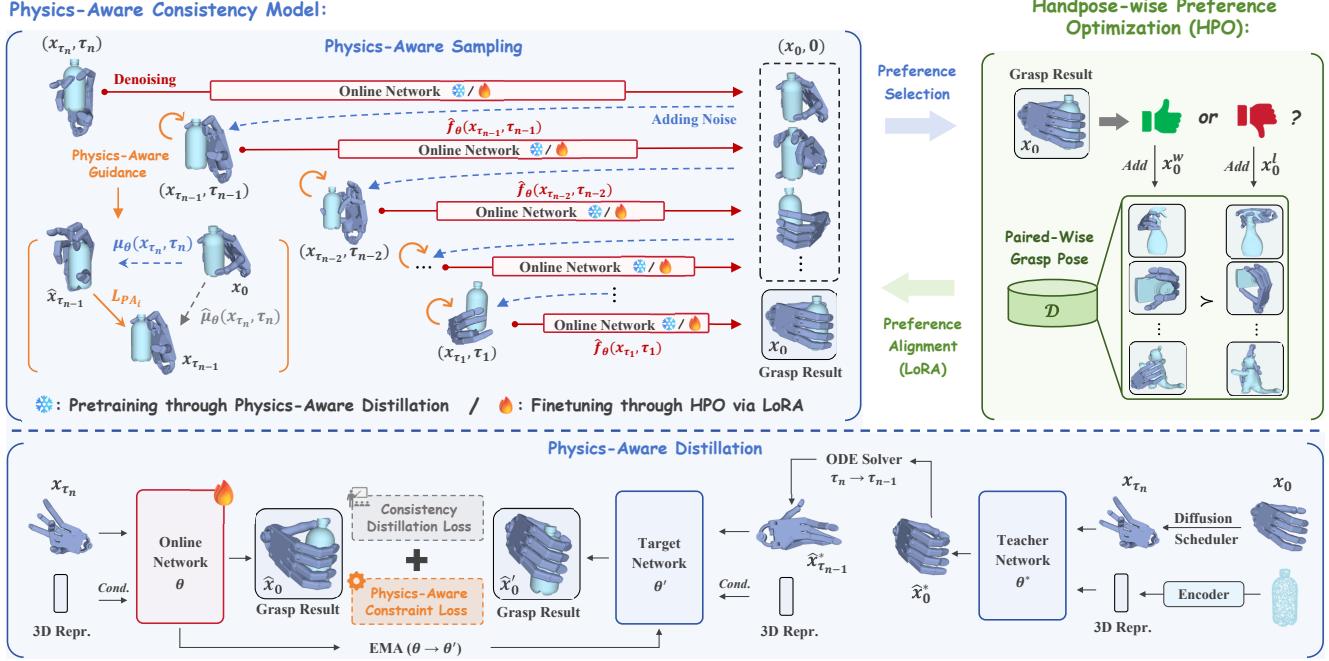


Figure 2. Overview of EvolvingGrasp. The evolutionary process begins with the human preference guidance, where Handpose-wise Preference Optimization (HPO, highlighted in the green rectangle) is employed to facilitate preference alignment. These grasp poses are generated by the Physics-Aware Consistency Model (shown in blue rectangles), including Sampling and Distillation mechanism, to ensure the sampling efficiency and the physical plausibility. In this way, EvolvingGrasp, an efficient evolutionary grasp generation framework is proposed to enable the grasp model iteratively converge toward preferred distributions.

ical constraint losses and performing gradient updates at each sampling iteration. However, existing methods cannot generate grasping poses aligned with human habits due to their lack of evolutionary adaptation and preference alignment. They also struggle to balance physical plausibility with computational efficiency.

2.2. Accelerating Diffusion Models

Accelerating dexterous grasping is vital for enhancing real-time adaptability. A series of diffusion-based acceleration methods have emerged, making real-time and efficient generation possible. DDIM [35] accelerates inference by reformulating the stochastic denoising process into a deterministic ODE solver. The DPM-Solver [23, 24] series further designs efficient high-order ODE solvers, achieving comparable quality with only $10 \sim 20$ steps. However, their performance degrades significantly when sampling steps are reduced to $2 \sim 4$. Consistency works [15, 22] support few-step sampling while preserving quality. CTM [15] extends CM to reduce cumulative errors and discretization inaccuracies. sCMs [22] unifies flow matching and diffusion frameworks to avoid discretization errors and hyperparameter tuning. Therefore, we employ the consistency model to enable efficient evolutionary refinement, reducing inference timesteps while maintaining result quality.

2.3. Preference Finetuning for Diffusion Models

Preference alignment play a crucial role in enabling generative models to learn from user feedback, optimize generation strategies, and progressively improve output quality. Some researchers utilize it into diffusion models to better align with human preferences in image generation area. These methods can be categorized into two types, finetuning with reward model [1, 4, 6, 8] and reward-free finetuning [10, 39, 46]. The former includes DDPO [1] and DPOK [8] which treat the denoising process of diffusion models as an MDP and finetune using multiple reward models. Finetuning without reward model includes D3PO [46] and Diffusion-DPO [39], which extend the theoretical framework of DPO [31] to multi-step MDPs. These approaches learn an optimal reward model and then use it to refine the sampling strategy, making them a relatively more cost-effective alternative. However, RL-based fine-tuning for diffusion models still requires backpropagation at every sampled timestep, making it highly time-consuming. Applying this approach directly to grasping tasks is impractical for real-world applications, where efficient refinement is crucial. Therefore, we propose a faster preference alignment fine-tuning approach that reduces backpropagation steps, improving grasping performance while aligning it with human preferences.

3. Methodology

3.1. Overview

To achieve efficient evolutionary grasp generation that aligns with human preferences while maintaining both efficiency and physical plausibility, we propose Evolving Grasp, as illustrated in Fig. 2. The evolutionary process begins with preference alignment, where we introduce Handpose-Wise Preference Optimization (HPO) (Section 3.3). HPO formulates grasp adaptation as a posterior probability optimization problem, allowing the model to iteratively converge toward preferred grasp distributions. However, directly applying HPO suffers from slow sampling and inefficiencies in preference alignment, limiting its practicality for real-time grasp refinement. To address these challenges, we incorporate a Physics-Aware Consistency Model (PCM) (Section 3.4), which leverages a consistency model framework to enhance efficiency by reducing the number of required sampling steps. While this improves inference speed, naive consistency-based sampling may still generate physically implausible grasp poses. To ensure geometric consistency and physical feasibility, PCM integrates a physics-aware distillation and sampling mechanism, which enforces structured physical constraints on the generated poses.

3.2. Problem Formulation

Given the object point cloud representation $O \in \mathbb{R}^{N \times 3}$, our goal is to generate dexterous grasp poses with high success rate and low penetration from the posterior distribution $P(x | O)$, where $x = \{x_i\}_{i=1}^n$. Specifically, the pose parameters contain three parts, joint angles of the hand $\theta_h \in \mathbb{R}^{24}$, global translation $T_{global} \in \mathbb{R}^3$, and global rotation $R_{global} \in SO(3)$.

Given ground truth samples from the data distribution $\pi(x_0)$, the noise schedule weight α , the goal of diffusion models is to fit the GT data distribution. The objective of training diffusion model is as follows:

$$\mathbb{E}_{t \sim [0, T], x_0, \epsilon \sim N(0, I)} \left[\|\epsilon - \epsilon_\theta(x_t, t, O)\|^2 \right] \quad (1)$$

where $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. Given the noise related parameter σ_t and the mean from x_t to x_{t-1} , the reverse process of diffusion model is as follows:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \quad (2)$$

3.3. Handpose-wise Preference Optimization

To achieve preference alignment in the evolutionary adaptation process, we introduce the HPO, an extension of direct preference optimization (DPO). In DPO, the fundamental assumption is that we have access to data generated by the model, with human annotators providing the corresponding

preferences. Specifically, for a given object O , we observe pairs of grasp poses x_0^w and x_0^l , where the set x_0^w is preferred over x_0^l . These preferences can be modeled using the Bradley-Terry model, which expresses the probability of one sample being preferred over another as:

$$p_{BT}(x_0^w \succ x_0^l) = \sigma(r(c, x_0^w) - r(c, x_0^l)) \quad (3)$$

The training objective of DPO is to maximize the likelihood of the observed preferences, which can be formulated as a binary classification problem. Specifically, the whole model is finetuned to minimize the following loss function:

$$L_{BT}(\theta) = -\mathbb{E}_{(x_0^w, x_0^l) \sim \mathcal{D}} \log \sigma \left(\begin{array}{c} \beta \mathbb{E}_{x_{1:T}^w \sim \pi_\theta(x_{1:T}^w | x_0^w)} \\ x_{1:T}^l \sim \pi_\theta(x_{1:T}^l | x_0^l) \end{array} \right) \left[\log \frac{\pi_\theta(x_{0:T}^w)}{\pi_{ref}(x_{0:T}^w)} - \log \frac{\pi_\theta(x_{0:T}^l)}{\pi_{ref}(x_{0:T}^l)} \right] \quad (4)$$

where \mathcal{D} is the paired-wise grasp pose dataset. Following [39], we can get the upper bound of Eq. (4) as:

$$L_{BT}(\theta) \leq - \begin{array}{c} \beta \mathbb{E}_{(x_0^w, x_0^l) \sim \mathcal{D}, n \sim \mathcal{U}(1, N), x_{n-1,n}^w \sim \pi_\theta(x_{n-1,n}^w | x_0^w)} \\ \log \sigma \left(\log \frac{\pi_\theta(x_{n-1}^w | x_n^w)}{\pi_{ref}(x_{n-1}^w | x_n^w)} - \log \frac{\pi_\theta(x_{n-1}^l | x_n^l)}{\pi_{ref}(x_{n-1}^l | x_n^l)} \right) \end{array} \quad (5)$$

where $\pi_\theta(x_{n-1} | x_n)$ is the probability of sampling x_{n-1} from x_n , which can be computed by:

$$\pi_\theta(x_{n-1} | x_n) = \frac{1}{\sqrt{2\pi\sigma_n}} \exp\left(-\frac{(x_n - \mu_\theta(x_n, n))^2}{2\sigma_n^2}\right) \quad (6)$$

As far as we know, HPO is the first to integrate DPO into grasp pose generation, from which we extend DPO to a more flexible form. In HPO, there is no requirement to maintain an equal number of preferred and non-preferred grasp poses, which enables more adaptive preference learning. The objective of HPO is introduced as follows:

$$L_{HPO} = \mathbb{E}_{x_0^i \sim \mathcal{D}, n \sim \mathcal{U}(1, N), x_{n-1,n}^i \sim \pi_\theta(x_{n-1,n}^i | x_0^i)} \log \sigma \left(\sum_{i=1}^{N_{suc}} \log \frac{\pi_\theta(x_{n-1}^i | x_n^i)}{\pi_{ref}(x_{n-1}^i | x_n^i)} - \sum_{j=1}^{N_{fail}} \log \frac{\pi_\theta(x_{n-1}^j | x_n^j)}{\pi_{ref}(x_{n-1}^j | x_n^j)} \right) \quad (7)$$

where N_{suc} and N_{fail} represent the number of preferred and non-preferred poses. HPO optimizes grasp pose generation by quantifying the probabilistic divergence between successful grasps (preferred samples) and failed grasps (non-preferred samples), thereby driving the model towards human-preferred behaviors. This process inherently involves the dynamic adjustment and maximization of reward signals: preferred grasp poses are reinforced due to

their higher probability, while non-preferred poses are suppressed, guiding the model to progressively discard ineffective strategies.

Preferred grasp selection can be conducted through either simulation-based evaluation or human-in-the-loop selection. In the simulation-based approach, poses that achieve success across all six directional evaluations are classified as preferred samples, while the remaining poses are treated as negative samples. Alternatively, in the human selection process, grasp poses that align with human intuition and habitual preferences are designated as preferred samples, while the others are considered non-preferred. Finally, we finetune the whole model via LoRA [11] using Eq. (7) to align the model with preferences.

3.4. Physics-Aware Consistency Model

3.4.1. Consistency Model

Directly applying HPO faces two major efficiency challenges: generating poses requires over hundreds of timesteps per inference, and preference fine-tuning demands a large number of backpropagation steps. To address these inefficiencies, we adopt the consistency model framework to accelerate both sampling process and preference alignment. The core idea of the consistency model is to learn a mapping from any point along the ODE trajectory back to its starting point, which corresponds to the data distribution. Given a Probability Flow ODE trajectory [36] $\{x_{\tau_n}\}_{\tau_n \in [0, T]}$, a consistency model f_θ is defined as:

$$f_\theta : (x_{\tau_n}, \tau_n, O) \mapsto x_0 \quad (8)$$

where x_0 is the initial point of the trajectory, O is the observation as condition. Due to space limitations, we omitted condition O in other parts of the paper, except for the appendix. The self-consistency property ensures that:

$$f_\theta(x_{\tau_n}, \tau_n) = f_{\theta'}(x_{\tau'_n}, \tau'_n) \quad \forall \tau_n, \tau'_n \in [0, T] \quad (9)$$

The consistency model needs to meet a key boundary condition: when $t = 0$, the model output should be the input itself, that is, $f_\theta(x_0, 0, O) = x_0$. This can be achieved as the following way:

$$f_\theta(x_{\tau_n}, \tau_n) = c_{\text{skip}}(\tau_n)x_{\tau_n} + c_{\text{out}}(\tau_n)F_\theta(x_{\tau_n}, \tau_n) \quad (10)$$

where $c_{\text{skip}}(t)$ and $c_{\text{out}}(t)$ are differentiable functions that satisfy $c_{\text{skip}}(\epsilon) = 1$ and $c_{\text{out}}(\epsilon) = 0$. $F_\theta(x, t)$ denotes a deep neural network that predicts \hat{x}_0 .

$$F_\theta(x_{\tau_n}, \tau_n) = \frac{1}{\sqrt{\bar{\alpha}_{\tau_n}}} \left(x_{\tau_n} - \sqrt{1 - \bar{\alpha}_{\tau_n}} \epsilon_\theta(x_{\tau_n}, \tau_n) \right) \quad (11)$$

Consistency distillation leverages pre-trained diffusion model to condense multi-step sampling into a more efficient

few-step inference process. The process involves generating pairs of adjacent points on the PFODE trajectory using numerical ODE solvers and minimizing the difference between the model's outputs for these pairs. The loss function for consistency distillation is defined as:

$$\mathcal{L}_{CD} = \mathbb{E} \left[d \left(f_\theta(x_{\tau_n}, \tau_n), f_{\theta'}(\hat{x}_{\tau_{n-1}}^*, \tau_{n-1}) \right) \right] \quad (12)$$

where O is omitted for simplicity, $\hat{x}_{\tau_{n-1}}^*$ is computed using a numerical ODE solver, $d(\cdot, \cdot)$ is a L_2 distance metric, θ and θ' represent the online network and target network parameters. In addition, the parameters of θ' are updated by the exponential moving average (EMA) of the parameters of θ . $\hat{x}_{\tau_{n-1}}^*$ can be computed as:

$$\hat{x}_{\tau_{n-1}}^* \leftarrow \sqrt{\bar{\alpha}_{\tau_{n-1}}} F_\theta(x_{\tau_n}, \tau_n) + \sqrt{1 - \bar{\alpha}_{\tau_{n-1}}} \epsilon \quad (13)$$

During sampling, we utilize the consistency function to directly generate the final sample and the quality of the generated sample can be enhanced through an iterative process that alternates between denoising and injecting noise. Given the sequence of timesteps $S \in \{\tau_i \mid \tau_0 = 0, \tau_{N-1} = T, \tau_i < \tau_{i+1} \text{ for } i = 0, 1, \dots, N-1\}$, the adding-noise process can be formulated as:

$$\hat{x}_{\tau_{n-1}} = \mu_\theta(x_{\tau_n}, \tau_n) + \sigma_{\tau_n} \epsilon \quad (14)$$

where $\mu_\theta = \sqrt{\bar{\alpha}_{\tau_{n-1}}} f_\theta(x_{\tau_n}, \tau_n)$, $\sigma_{\tau_n} = \sqrt{1 - \bar{\alpha}_{\tau_{n-1}}}$. Subsequently, we conduct the prediction of the final sample utilizing the trained consistency function once more.

3.4.2. Physics-Aware Distillation and Sampling

Although adopting the consistency model improves sampling and preference fine-tuning efficiency, it still generates physically implausible poses. To address this, we introduce a Physics-Aware Distillation and Sampling paradigm that enforces physical constraints during the distillation process of predicting \hat{x}_0 while ensuring that the sampling trajectories adhere to specific constraints. Following [52], we incorporate three physics-aware objectives, with the distillation objective of the consistency model formulated as follows:

$$\mathcal{L}_{PAD} = \mathcal{L}_{CD} + \sum_{i=1}^m \alpha_i L_{PA_i}(F_\theta(x_{\tau_n}, \tau_n), \epsilon_\theta) \quad (15)$$

where $L_{PA_i}(F_\theta(x_t, t), \epsilon_\theta)$ is the i^{th} physical constraint loss and $m = 3$ denotes three constraints, *i.e.*, Surface Pulling Force, External-penetration Repulsion Force, Self-Penetration Repulsion Force respectively [52]. These constraints guarantee grasping feasibility and maintain geometric accuracy in finger-to-object and inter-finger interactions. α_i is the corresponding weight parameter. We first train a diffusion model for grasp pose generation as a teacher

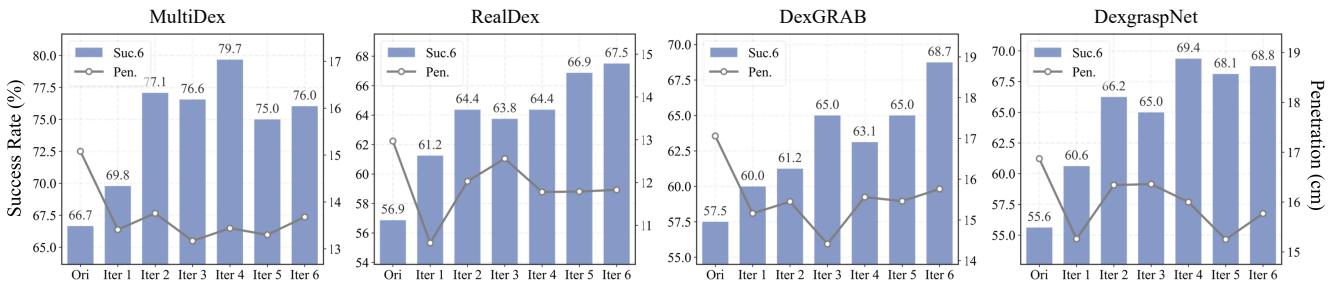


Figure 3. Mean grasping performance in terms of success rate and penetration of randomly selected 6 objects with the finetuning epoch increasing during inference optimization.

model. Then, we utilize \mathcal{L}_{PAD} to distill the teacher model into a student model.

During sampling, the consistency model estimates a clearer hand pose x_0 based on the current noisy hand pose x_{τ_n} and object information O by consistency function $f_\theta(x_{\tau_n}, \tau_n, O)$. Subsequently, physical constraints are applied to steer the sampling process, making it closer to a physically feasible grasping pose. Following [3, 7, 47], the gradient of the constraint loss is used to modify the mean from x_t to x_{t-1} :

$$\begin{aligned} \hat{\mu}_\theta(x_{\tau_n}, \tau_n) &= \mu_\theta(x_{\tau_n}, \tau_n) \\ &+ \sum_{i=1}^m \gamma_i \nabla_{x_{\tau_n}} L_{PA_i}(F_\theta(x_{\tau_n}, \tau_n), \epsilon_\theta) \end{aligned} \quad (16)$$

where γ_i is the weight parameter corresponding to each physical constraint. Moreover, we can derive a new mapping from noise to data as follows:

$$\hat{f}_\theta(x_{\tau_n}, \tau_n) = f_\theta(x_{\tau_n}, \tau_n) + \frac{1}{\sqrt{\alpha_{\tau_{n-1}}}} \sum_{i=1}^m \gamma_i \nabla_{x_{\tau_n}} L_{PA_i} \quad (17)$$

By employing the physics-aware consistency model, we derive a novel preference alignment objective based on the new sampling path. The detailed derivation is provided in Appendix B. As a result, our method can efficiently generate higher-quality poses during the evolutionary process.

4. Experiments

We first simply present experimental setup which includes datasets, evaluation metrics and baselines in Sec. 4.3.1. More details of setup including datasets and implementation details are in Appendix C. Then we provide the quantitative and qualitative results of EvolvingGrasp in Sec. 4.2, followed by ablation study about different modules and hyperparameter analyses in Sec. 4.3.2. Next, we demonstrate the evolutionary improvement of EvolvingGrasp through an experiment starting from a suboptimal model trained on a degraded dataset in Sec. 4.4. Finally, We will demonstrate EvolvingGrasp in the real-world deployment in Sec. 4.5.

4.1. Experimental Setup

Datasets. We conduct experiment on four datasets, including DexGraspNet [40], Multidex [17], Realdex [21], DexGRAB [37] respectively.

Evaluation Metrics. We use four metrics to evaluate the grasping performance. Success rate **Suc.6** measures the proportion of grasping poses where the object’s displacement does not exceed 2 cm in all six axial directions ($\pm X$, $\pm Y$, $\pm Z$), evaluating multi-directional stability. **Suc.1** measures the proportion where displacement does not exceed 2cm in at least one direction, assessing single-direction stability. **Pen.** indicates the maximum penetration depth (mm) between the hand and the object, with lower values suggesting more physically plausible grasps. Above metrics are calculated in the IssacGym simulator [28] with settings consistent with SceneDiffuser [13]. For efficiency, **Time** refers to the computational time required to generate grasping poses for a batch of objects.

Baselines. Some generative-based methods such as UniDexGrasp [45], GraspTTA [14], SceneDiffuser [13], UGG [25], and DexGrasp Anything [52] are compared on four benchmark datasets.

4.2. Main Results

To validate the effectiveness of our method in continuously enhancing grasping performance, we conduct quantitative evaluations on six randomly selected objects from each dataset, measuring both the Suc.6 metric and the degree of penetration. The results are illustrated in Fig. 3, indicating that as the number of inferences increases, the Suc.6 metric steadily improves. Although penetration exhibits some fluctuations due to increased contact area between the hand and the object, it still follows a downward trend. Additionally, we present qualitative results in Fig. 4, demonstrating how our method integrates human preferences during inference. This enables the selection of more favorable poses and progressively refines grasping poses to better align with human preferences through iteration. For instance, early generated poses may obstruct the camera lens, but after several rounds of preference-based finetuning, the model learns to produce poses that avoid blocking the lens.

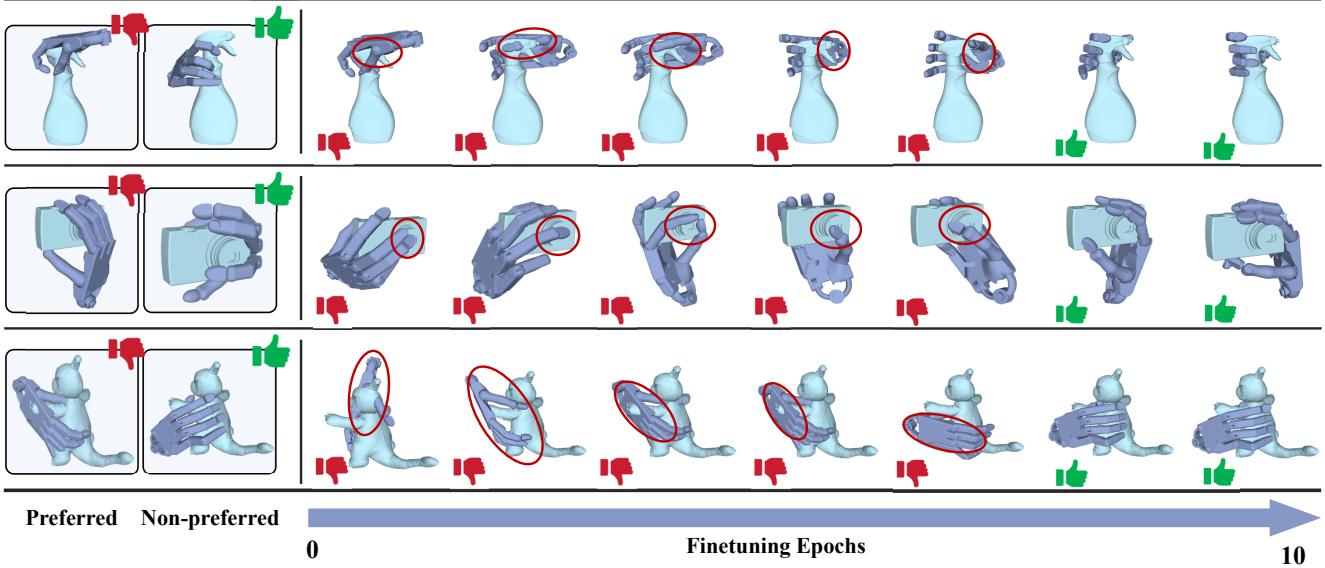


Figure 4. Evolution of robotic grasp preferences during efficient feedback-driven finetuning across 10 epochs. Top row illustrates the adjustment from hand occlusion to clear nozzle visibility. Middle row demonstrates the transition from lens obstruction to an unobstructed camera view. The bottom row shows the evolution from a top-down grasping approach to a bottom-up one, while simultaneously mitigating physical impacts.

We conduct the experiment with comparing with other dexterous grasp generation methods in Tab. 1. Compared to other generative-based approaches, our method excels in generating high-quality, physically plausible grasping poses while significantly reducing computational time. It is worth highlighting that the proposed method achieves **30x speedup** (32s~38s v.s. 0.73s~2.7s) compared to these SOTAs. After fine-tuning, our approach achieves superior performance by leveraging its evolutionary capability through iterative refinement. Notably, our preference fine-tuning guides the model to generate poses that better align with human preferences, which may, to some extent, reduce the diversity of the generated results. We also evaluate the trade-off between efficiency and grasp quality during preference alignment with different numbers of inference steps. While increasing the steps improves performance, it inevitably incurs additional computational costs. Moreover, even without guidance during sampling, we can achieve comparable results in real time.

4.3. Ablation Study

4.3.1. Ablation on Different Modules

We investigate the impact of different modules in the physics-aware consistency model, **Physical constraints Guidance in Distillation and Sampling** from which we dubbed as **PGD** and **PGS**. We also consider the performance of HPO with and without physical guidance during sampling on test split of the multidex dataset. The results presented in Tab. 2 demonstrate the key role of physical constraints and preference finetuning to successful grasping

pose generation during preference alignment.

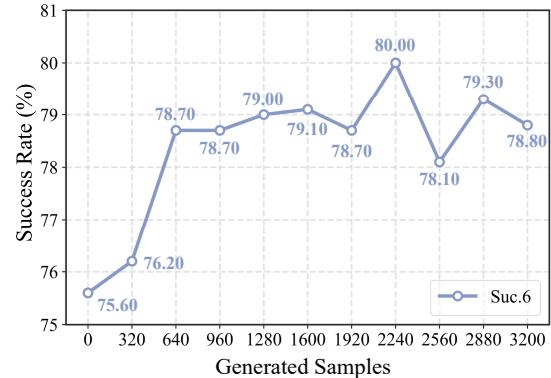


Figure 5. The relationship between the success rate of Evolving-Grasp and the number of generated samples.

4.3.2. Ablation on Evolutionary Finetuning

HPO not only effectively aligns grasping poses with human preferences at the individual object level but also enhances performance across the entire dataset. During inference across the entire dataset, we collect both successful and failed grasping samples for each object, utilizing Eq. (5) to perform lightweight finetuning of the entire model. Fig. 5 illustrates the relationship between EvolvingGrasp’s Suc.6 metric and the number of generated samples, showing that our method continuously enhances grasping performance as more samples are generated. Further ablation studies on the impact of different hyperparameters are provided in Appendix D.2.

Table 1. Grasping performance in terms of Suc.6, Suc.1, and Pen. comparison across different methods and datasets. “Step” refers to inference timestep. Bold values highlight the best results and underlined values indicate the runner-up.

Dataset \ Method	DexGraspNet				MultiDex				RealDex				DexGRAB	
	Suc.6 ↑	Suc.1 ↑	Pen. ↓	Suc.6 ↑	Suc.1 ↑	Pen. ↓	Suc.6 ↑	Suc.1 ↑	Pen. ↓	Suc.6 ↑	Suc.1 ↑	Pen. ↓	Time ↓	
UniDexGrasp [45]	33.9	70.1	31.9	21.6	47.5	13.5	27.1	59.4	39.0	20.8	55.8	37.4	0.46 ± 0.11	
GraspTTA [14]	18.6	67.8	24.5	30.3	62.8	19.0	13.3	46.4	40.1	14.4	51.0	51.4	10.41 ± 0.32	
SceneDiffuser [13]	26.6	66.9	31.0	69.8	85.6	14.6	21.7	56.1	42.0	39.1	85.0	41.1	3.41 ± 0.13	
UGG [25]	46.9	79.0	25.2	55.3	93.4	<u>10.3</u>	32.7	63.4	34.4	42.7	90.6	33.2	38.34 ± 2.31	
DexGrasp Any. [52]	53.6	90.4	21.5	72.2	96.3	9.6	34.6	71.2	23.1	56.5	91.8	28.6	32.91 ± 1.34	
Ours w/o HPO (2-step)	60.8	91.0	19.2	65.6	97.5	15.2	41.9	75.4	19.5	52.2	93.9	25.1	0.73 ± 0.03	
Ours (2-step)	62.4	90.5	19.3	65.9	97.2	15.3	44.0	77.8	<u>19.7</u>	53.3	92.5	24.3	0.73 ± 0.03	
Ours w/o HPO (4-step)	63.8	<u>93.0</u>	17.4	75.3	97.1	13.1	51.6	82.9	20.5	55.6	96.0	23.8	1.41 ± 0.07	
Ours (4-step)	<u>65.2</u>	92.7	17.2	<u>76.8</u>	<u>98.4</u>	13.0	50.6	82.5	20.3	<u>57.7</u>	95.2	23.7	1.41 ± 0.07	
Ours w/o HPO (8-step)	<u>65.2</u>	93.5	<u>16.2</u>	75.6	98.7	12.2	<u>63.6</u>	86.6	21.9	56.8	96.8	<u>23.1</u>	2.71 ± 0.08	
Ours (8-step)	65.4	92.3	15.9	80.3	98.7	12.3	64.4	89.1	21.8	60.8	96.4	22.3	2.71 ± 0.08	
Real-time (2-step)	55.2	90.5	20.0	63.7	95.0	13.8	46.5	78.9	21.4	48.9	93.3	24.8	0.06 ± 0.01	
Real-time (4-step)	59.9	90.6	19.8	64.3	96.5	11.6	58.2	<u>87.0</u>	21.1	55.4	95.4	24.1	0.10 ± 0.02	

Table 2. Ablation study on incorporating physical constraints during both training and sampling stages and the LLM module. The evaluation is conducted on Multidex when timestep is 4.

	CM	PGD	PGS	HPO	Suc.6 ↑	Suc.1 ↑	Pen. ↓	Time ↓
a	✓				60.0	94.6	14.0	0.10 ± 0.02
b	✓	✓			64.3	96.5	12.5	0.10 ± 0.02
c	✓		✓		66.2	95.6	14.9	1.41 ± 0.07
d	✓	✓		✓	67.5	96.5	11.9	0.10 ± 0.02
e	✓	✓	✓		75.3	97.1	13.1	1.41 ± 0.07
f	✓	✓	✓	✓	76.8	98.4	13.0	1.41 ± 0.07

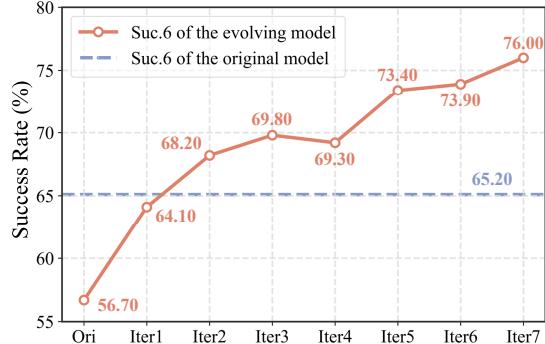


Figure 6. The variation of the mean grasping success rate of the evolving model for randomly selected objects in the multidex dataset with increasing finetuning epochs. The blue dashed line represents the grasping success rate under the original model.

4.4. Training from a Degraded Dataset

In this subsection, we explore the feasibility of enhancing model performance through feedback-driven finetuning for single-object grasping tasks, particularly when trained on suboptimal data. To begin with, we randomly add the noise on the hand pose parameter of the original multidex dataset. This perturbation creates a degraded dataset, which we then



Figure 7. Real-world deployment on Shadow Hand.

use to train and distill a suboptimal model. During the inference phase of this suboptimal model, we collect successfully grasping poses as positive samples and all other poses as negative samples. We then employ Eq. 5 to finetune the whole model with LoRA [11]. Fig. 6 illustrates the change in the average grasping success rate of the suboptimal model on randomly sampled objects from the multidex dataset as the number of fine-tuning epochs increases. The results indicate that we continuously improve the success rate during the evolutionary adaptation process and ultimately outperform the accuracy of the original model.

4.5. Real World Deployment

To verify that EvolvingGrasp can continuously enhance grasping performance and align with human preferences in real world, we deploy our model on a real ShadowHand robot, as shown in Fig. 7. The pre-grasping motion trajectory is generated based on RealDex [21]. Real-world experiments demonstrate that our method achieves successful grasping through several efficient preference finetuning when the initial grasp attempt fails. Additional video demonstrations can be found in supplementary materials.

5. Conclusion

We propose EvolvingGrasp, an evolutionary grasp generation method through efficient preference alignment. HPO is

introduced to allow the model to continuously align the performance with preference signals. We design the Physics-Aware Consistency Model to achieve both rapid inference and efficient preference finetuning while maintaining physical plausibility. Extensive experiments across four benchmarks demonstrate that our method achieves state-of-the-art results. Furthermore, we deploy our model on a real ShadowHand robot to validate its evolutionary capability in real-world scenarios.

References

- [1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 3
- [2] Yuxin Chen, Devesh K Jha, Masayoshi Tomizuka, and Diego Romero. Fdpp: Fine-tune diffusion policy with human preference. *arXiv preprint arXiv:2501.08259*, 2025. 2
- [3] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 6
- [4] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 3
- [5] Hongkai Dai, Anirudha Majumdar, and Russ Tedrake. Synthesis and optimization of force closure grasps via sequential semidefinite programming. *Robotics Research: Volume 1*, pages 285–305, 2018. 2
- [6] Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, and Matthias Grundmann. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In *CVPR*, pages 7423–7433, 2024. 3
- [7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6
- [8] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023. 3
- [9] Carlo Ferrari, John Canny, et al. Planning optimal grasps. In *IEEE International Conference on Robotics and Automation*, pages 2290–2295. IEEE, 1992. 2
- [10] Ayano Hiranaka, Shang-Fu Chen, Chieh-Hsin Lai, Dongjun Kim, Naoki Murata, Takashi Shibuya, Wei-Hsiang Liao, Shao-Hua Sun, and Yuki Mitsufuji. Human-feedback efficient reinforcement learning for online diffusion model finetuning. *arXiv preprint arXiv:2410.05116*, 2024. 3
- [11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5, 8, 13
- [12] Linyi Huang, Hui Zhang, Zijian Wu, Sammy Christen, and Jie Song. Fungrasp: Functional grasping for diverse dexterous hands. *arXiv preprint arXiv:2411.16755*, 2024. 2
- [13] Siyuan Huang, Zan Wang, Puhalo Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, pages 16750–16761, 2023. 2, 6, 8
- [14] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, pages 11107–11116, 2021. 2, 6, 8
- [15] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023. 3
- [16] Haosheng Li, Weixin Mao, Weipeng Deng, Chenyu Meng, Haoqiang Fan, Tiancai Wang, Ping Tan, Hongan Wang, and Xiaoming Deng. Multi-graspllm: A multimodal llm for multi-hand semantic guided grasp generation. *arXiv preprint arXiv:2412.08468*, 2024. 2
- [17] Puhalo 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*, pages 8068–8074. IEEE, 2023. 6
- [18] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Generating grasp poses for a high-dof gripper using neural networks. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1518–1525. IEEE, 2019. 2
- [19] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Deep differentiable grasp planner for high-dof grippers. *arXiv preprint arXiv:2002.01530*, 2020. 2
- [20] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters*, 7(1):470–477, 2021. 2
- [21] Yumeng Liu, Yaxun Yang, Youzhuo Wang, Xiaofei Wu, Jiamin Wang, Yichen Yao, Sören Schwertfeger, Sibei Yang, Wenping Wang, Jingyi Yu, et al. Realdex: Towards human-like grasping for robotic dexterous hand. *arXiv preprint arXiv:2402.13853*, 2024. 6, 8
- [22] Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024. 3
- [23] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 3
- [24] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 3
- [25] Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. Ugg: Unified generative grasping. *arXiv preprint arXiv:2311.16917*, 2023. 2, 6, 8
- [26] Jens Lundell, Enric Corona, Tran Nguyen Le, Francesco Verdoja, Philippe Weinzaepfel, Grégory Rogez, Francesc

- Moreno-Noguer, and Ville Kyrki. Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps. In *IEEE International Conference on Robotics and Automation*, pages 4495–4501. IEEE, 2021. 2
- [27] Jens Lundell, Francesco Verdoja, and Ville Kyrki. Ddgc: Generative deep dexterous grasping in clutter. *IEEE Robotics and Automation Letters*, 6(4):6899–6906, 2021. 2
- [28] 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. 6
- [29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 14
- [30] Jean Ponce, Steve Sullivan, J-D Boissonnat, and J-P Merlet. On characterizing and computing three-and four-finger force-closure grasps of polyhedral objects. In *IEEE International Conference on Robotics and Automation*, pages 821–827. IEEE, 1993. 2
- [31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 2, 3
- [32] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2
- [33] Carlos Rosales, Raúl Suárez, Marco Gabiccini, and Antonio Bicchi. On the synthesis of feasible and prehensile robotic grasps. In *IEEE International Conference on Robotics and Automation*, pages 550–556. IEEE, 2012. 2
- [34] Yanming Shao and Chenxi Xiao. Bimanual grasp synthesis for dexterous robot hands. *IEEE Robotics and Automation Letters*, 2024. 2
- [35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3
- [36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 5
- [37] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 6
- [38] Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. Sopo: Text-to-motion generation using semi-online preference optimization. *arXiv preprint arXiv:2412.05095*, 2024. 2
- [39] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, pages 8228–8238, 2024. 2, 3, 4
- [40] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzen Xu, Puha Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *IEEE International Conference on Robotics and Automation*, pages 11359–11366. IEEE, 2023. 6
- [41] Wei Wei, Daheng Li, Peng Wang, Yiming Li, Wanyi Li, Yongkang Luo, and Jun Zhong. Dvgg: Deep variational grasp generation for dextrous manipulation. *IEEE Robotics and Automation Letters*, 7(2):1659–1666, 2022. 2
- [42] 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. *arXiv preprint arXiv:2405.19291*, 2024. 2
- [43] Zehang Weng, Haofei Lu, Danica Kragic, and Jens Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models. *IEEE Robotics and Automation Letters*, 2024. 2
- [44] Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *CVPR*, pages 17933–17942, 2024. 2
- [45] Yinzen 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 *CVPR*, pages 4737–4746, 2023. 6, 8
- [46] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *CVPR*, pages 8941–8951, 2024. 3
- [47] Lingxiao Yang, Shutong Ding, Yifan Cai, Jingyi Yu, Jingya Wang, and Ye Shi. Guidance with spherical gaussian constraint for conditional diffusion. *arXiv preprint arXiv:2402.03201*, 2024. 6
- [48] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024. 2
- [49] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, pages 235–246. IEEE, 2024. 2
- [50] Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*, 2024. 2
- [51] Zhengshen Zhang, Lei Zhou, Chenchen Liu, Zhiyang Liu, Chengran Yuan, Sheng Guo, Ruiteng Zhao, Marcelo H Ang Jr, and Francis EH Tay. Dexgrasp-diffusion: Diffusion-based unified functional grasp synthesis method for multi-dexterous robotic hands. *arXiv preprint arXiv:2407.09899*, 2024. 2

- [52] Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgraspanything: Towards universal robotic dexterous grasping with physics awareness. In *CVPR*, 2025. [2](#), [5](#), [6](#), [8](#)
- [53] Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. *arXiv preprint arXiv:2502.04370*, 2025. [2](#)

Appendix

A. Pseudo Code of EvolvingGrasp

Pseudo Code of EvolvingGrasp is shown in Algorithm 1 and 2.

Algorithm 1 Physics-Aware Sampling and Handpose-Wise Preference Optimization

Require: Number of inference timesteps T , number of finetuning epochs E_{ft} , number of objects K , physical-aware consistency model ϵ_θ , test batchsize B , test time sequences $S \in \{\tau_i \mid \tau_0 = 0, \tau_{N-1} = T, \tau_i < \tau_{i+1} \text{ for } i = 0, 1, \dots, N-1\}$, differentiable functions c_{skip} and c_{out} , gradient guidance weight $\{\gamma_i\}_{i=1}^m$.

1: Copy the parameters of consistency model $\epsilon_{ref} = \epsilon_\theta$ and set ϵ_{ref} to have `requires_grad = False`.
2: **for** $e = 1 : E_{ft}$ **do**
3: # Sample grasping poses
4: **for** $k = 1 : K$ **do**
5: Choose an object O_k and sample $x_T \sim \mathcal{N}(0, \mathbf{I})$
6: **for** $i = 1 : B$ **do**
7: **for** $n = N - 1 : 0$ **do**
8: $F_\theta(x_{k,\tau_n}^i, \tau_n, O_k) = \frac{1}{\sqrt{\alpha_{\tau_n}}} \left(x_{k,\tau_n}^i - \sqrt{1 - \bar{\alpha}_{\tau_n}} \epsilon_\theta \left(x_{k,\tau_n}^i, \tau_n, O_k \right) \right)$
9: $f_\theta(x_{k,\tau_n}^i, \tau_n, O_k) = c_{skip}(\tau_n)x_{k,\tau_n}^i + c_{out}(\tau_n)F_\theta(x_{k,\tau_n}^i, \tau_n, O_k)$
10: # Sampling with Gradient Guidance:
11: $\hat{\mu}_\theta(x_{k,\tau_n}^i, \tau_n, O_k) = \sqrt{\alpha_{\tau_{n-1}}} f_\theta(x_{k,\tau_n}^i, \tau_n, O_k) + \sum_{i=1}^m \gamma_i \nabla_{x_{\tau_n}} L_{PA_i} \left(F_\theta(x_{k,\tau_n}^i, \tau_n), \epsilon_\theta \right)$
12: $\sigma_{\tau_n} = \sqrt{1 - \bar{\alpha}_{\tau_{n-1}}}$
13: $x_{k,\tau_{n-1}}^i = \hat{\mu}_\theta(x_{k,\tau_n}^i, \tau_n, O_k) + \sigma_{\tau_n} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$
14: **end for**
15: **end for**
16: # Select the Preferred Grasp Poses
17: **for** $i = 0 : B$ **do**
18: **if** x_0^i grasp object O_k matches human preference **then**
19: $h_i = 1$
20: **else**
21: $h_i = -1$
22: **end if**
23: **end for**
24: # Efficiently Feedback-driven Finetuning
25: **for** $n = N - 1 : 0$ **do**
26: **# Utilizing Fewer Timesteps for Preference Alignment.**
27: **for** $i = 1 : B$ **do**
28: with grad:
29: $\mu_\theta(x_{k,\tau_n}^i, \tau_n, O_k) = \sqrt{\alpha_{\tau_{n-1}}} f_\theta(x_{k,\tau_n}^i, \tau_n, O_k), \mu_{ref}(x_{k,\tau_n}^i, \tau_n, O_k) = \sqrt{\alpha_{\tau_{n-1}}} f_{ref}(x_{k,\tau_n}^i, \tau_n, O_k)$
30: $\pi_\theta(x_{k,\tau_{n-1}}^i \mid x_{k,\tau_n}^i, O_k) = \frac{1}{\sqrt{2\pi}\sigma_{\tau_n}} \exp\left(-\frac{(x_{k,\tau_n}^i - \mu_\theta(x_{k,\tau_n}^i, \tau_n, O_k))^2}{2\sigma_{\tau_n}^2}\right)$
31: $\pi_{ref}(x_{k,\tau_{n-1}}^i \mid x_{k,\tau_n}^i, O_k) = \frac{1}{\sqrt{2\pi}\sigma_{\tau_n}} \exp\left(-\frac{(x_{k,\tau_n}^i - \mu_{ref}(x_{k,\tau_n}^i, \tau_n, O_k))^2}{2\sigma_{\tau_n}^2}\right)$
32: **end for**
33: Update θ using gradient descent with LoRA:

$$\nabla_\theta \log \sigma \left(\sum_{i=1}^B h_i \beta \log \frac{\pi_\theta(x_{k,\tau_{n-1}}^i \mid x_{k,\tau_n}^i, \tau_n, O_k)}{\pi_{ref}(x_{k,\tau_{n-1}}^i \mid x_{k,\tau_n}^i, \tau_n, O_k)} \right)$$

34: **end for**
35: **end for**
36: **end for**

Algorithm 2 Physical-Aware Distillation

Require: Training dataset D_t , number of training epochs E_t , learning rate η , pre-trained diffusion model ϵ_θ , number of timesteps T_{dm} , distance metric $d(\cdot, \cdot)$, EMA rate μ , noise schedule $\{\alpha_t\}_{t=1}^{T_{dm}}$, physics-aware constraints weights $\{\alpha_i\}_{i=1}^m$.

- 1: Copy the parameters of the pre-trained diffusion model as the target network $\epsilon_{\theta'} = \epsilon_\theta$
- 2: **for** $e = 1 : E$ **do**
- 3: **for** $k = 1 : K$ **do**
- 4: Choose an object O_k and sample $x_0 \sim D_t$, $n \sim \mathcal{U}[1, N]$
- 5: Sample $x_{\tau_n} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{\tau_n}}x_0, (1 - \bar{\alpha}_{\tau_n})\mathbf{I})$
- 6: $F_\theta(x_{\tau_n}, \tau_n, O_k) = \frac{1}{\sqrt{\bar{\alpha}_{\tau_n}}} (x_{\tau_n} - \sqrt{1 - \bar{\alpha}_{\tau_n}}\epsilon_\theta(x_{\tau_n}, \tau_n, O_k))$
- 7: $\hat{x}_{\tau_{n-1}}^* = \sqrt{\bar{\alpha}_{\tau_{n-1}}}F_\theta(x_{\tau_n}, \tau_n, O_k) + \sqrt{1 - \bar{\alpha}_{\tau_{n-1}}}\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
- 8: $\mathcal{L}_{PAD} = \mathbb{E} \left[d \left(f_\theta(x_{\tau_n}, \tau_n), f_{\theta'} \left(\hat{x}_{\tau_{n-1}}^*, \tau_{n-1} \right) \right) \right] + \sum_{i=1}^m \alpha_i L_{PA_i} (F_\theta(x_{\tau_n}, \tau_n), \epsilon_\theta)$
- 9: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{PAD}$
- 10: $\theta' \leftarrow \text{stopgrad}(\mu\theta' + (1 - \mu)\theta)$
- 11: **end for**
- 12: **end for**

B. Proof

Defined on the new path, the proof of the upper bound is as follows:

$$\begin{aligned}
L_{BT}(\theta) &= -\mathbb{E}_{x_0^{w,l} \sim \mathcal{D}} \log \sigma \left(\beta \mathbb{E}_{x_{1:T}^{w,l} \sim \pi_\theta(x_{1:T}^{w,l} | x_0^{w,l})} \left[\log \frac{\pi_\theta(x_{0:T}^w)}{\pi_{\text{ref}}(x_{0:T}^w)} - \log \frac{\pi_\theta(x_{0:T}^l)}{\pi_{\text{ref}}(x_{0:T}^l)} \right] \right) \\
&= -\mathbb{E}_{x_0^{w,l} \sim \mathcal{D}} \log \sigma \left(\beta \mathbb{E}_{x_{1:T}^{w,l} \sim \pi_\theta(x_{1:T}^{w,l} | x_0^{w,l})} \left[\sum_{n=1}^N \log \frac{\pi_\theta(x_{\tau_{n-1}}^w | x_{\tau_n}^w)}{\pi_{\text{ref}}(x_{\tau_{n-1}}^w | x_{\tau_n}^w)} - \sum_{n=1}^N \log \frac{\pi_\theta(x_{\tau_{n-1}}^l | x_{\tau_n}^l)}{\pi_{\text{ref}}(x_{\tau_{n-1}}^l | x_{\tau_n}^l)} \right] \right) \\
&= -\mathbb{E}_{x_0^{w,l} \sim \mathcal{D}} \log \sigma \left(\beta \mathbb{E}_{x_{1:T}^{w,l} \sim \pi_\theta(x_{1:T}^{w,l} | x_0^{w,l})} N \mathbb{E}_n \left[\log \frac{\pi_\theta(x_{\tau_{n-1}}^w | x_{\tau_n}^w)}{\pi_{\text{ref}}(x_{\tau_{n-1}}^w | x_{\tau_n}^w)} - \log \frac{\pi_\theta(x_{\tau_{n-1}}^l | x_{\tau_n}^l)}{\pi_{\text{ref}}(x_{\tau_{n-1}}^l | x_{\tau_n}^l)} \right] \right) \quad (18) \\
&= -\mathbb{E}_{x_0^{w,l} \sim \mathcal{D}} \log \sigma \left(N \beta \mathbb{E}_{n, x_{\tau_n}^{w,l} \sim \pi_\theta(x_{\tau_n}^{w,l} | x_0^{w,l})} \left[\log \frac{\pi_\theta(x_{\tau_{n-1}}^w | x_{\tau_n}^w)}{\pi_{\text{ref}}(x_{\tau_{n-1}}^w | x_{\tau_n}^w)} - \log \frac{\pi_\theta(x_{\tau_{n-1}}^l | x_{\tau_n}^l)}{\pi_{\text{ref}}(x_{\tau_{n-1}}^l | x_{\tau_n}^l)} \right] \right) \\
&\leq -\mathbb{E}_{\substack{x_0^{w,l} \sim \mathcal{D}, t \sim \mathcal{U}(0,T), \\ n, x_{\tau_n}^w \sim \pi_\theta(x_{\tau_n}^w | x_0^w)}} \log \sigma \left(\beta \log \frac{\pi_\theta(x_{\tau_{n-1}}^w | x_{\tau_n}^w)}{\pi_{\text{ref}}(x_{\tau_{n-1}}^w | x_{\tau_n}^w)} - \beta \log \frac{\pi_\theta(x_{\tau_{n-1}}^l | x_{\tau_n}^l)}{\pi_{\text{ref}}(x_{\tau_{n-1}}^l | x_{\tau_n}^l)} \right)
\end{aligned}$$

where the last inequality is based on Jensen's inequality and $-\log \sigma(\cdot)$ is a strict convex function. Therefore, we use the new objective 18 to optimize the whole model with LoRA [11].

Table 3. Cross-dataset evaluation results. The highest performances are highlighted in **bold**, while the second-highest performances are indicated with underline.

Testing Dataset		DexGraspNet			MultiDex			RealDex			DexGRAB	
Training Dataset	Suc.6 ↑ Suc.1 ↑ Pen. ↓											
DexGraspNet	<u>65.2</u>	<u>92.7</u>	17.2	73.4	97.1	<u>9.7</u>	54.1	90.1	19.4	<u>58.1</u>	94.3	20.6
MultiDex	67.6	94.0	<u>19.5</u>	76.8	<u>98.4</u>	13.0	51.9	<u>88.6</u>	19.4	65.6	96.8	<u>19.5</u>
RealDex	52.2	81.1	20.7	51.5	88.1	14.0	50.6	<u>82.5</u>	20.3	46.0	80.0	18.3
DexGRAB	64.9	92.6	<u>17.1</u>	<u>75.3</u>	99.3	<u>9.9</u>	<u>53.1</u>	88.5	<u>19.7</u>	57.7	<u>95.2</u>	23.7

C. Details of Experimental Setup

C.1. Dataset Setups

DexGraspNet is a large-scale dataset for dexterous grasping, comprising 1.32 million grasp samples across 5,355 objects from 133 diverse categories. While its optimization-based generation ensures high quality and diversity, its applicability in real-world scenarios is limited.

In contrast, MultiDex focuses on a smaller set of 58 everyday objects but offers a rich variety of grasping poses for each object. This makes it an ideal dataset for studying the diversity of grasping configurations and developing methods that can generate a wide range of effective grasps for common objects.

Realdex shifts the focus to real-world applications by capturing natural human grasping behaviors. It contains 59,000 samples across 52 objects, making it highly suitable for training robots to learn human-like grasping poses. Although it covers fewer object categories, its real-world grounding allows it to effectively validate the generalization and practicality of dexterous grasping methods in real environments.

DexGRAB, derived from human hand interaction data, provides over 1.64 million grasp samples across 51 distinct objects. It offers rich grasping patterns and natural interaction behaviors, making it a valuable resource for understanding human grasping strategies. Similar to DexGraspNet, DexGRAB’s data quality is high after filtering, but its real-world applicability may also face some limitations due to its primarily simulation-based nature.

Together, these datasets offer a range of strengths and limitations, from the large-scale optimization-based approaches of DexGraspNet and DexGRAB to the real-world grounding of Realdex and the diversity-focused MultiDex. Each dataset contributes unique insights and challenges to the field of dexterous grasping research.

C.2. Implementation Details

Our EvolvingGrasp contains distillation and sampling, which are implemented using Pytorch [29] platform in one NVIDIA Tesla A40 GPU. In the distillation process, we train EvolvingGrasp for 1,000 epochs with a batch size of 1,200. During both the distillation and preference finetuning processes, the initial learning rate is set to 0.00001. For the distillation process, the learning rate remains unchanged. During inference, the success rate of the generated grasping poses is firstly evaluated. If the success rate improves, the learning rate is adaptively reduced, otherwise, it is increased accordingly. Additionally, the adjustment of the learning rate is constrained within a predefined threshold range to ensure it remains within reasonable bounds. The sampling and preference optimization processes are implemented in test split of each corresponding dataset.

Table 4. Evaluating Cross-Dataset Generalization. Model performance is compared on RealDex, with training on DexGraspNet.

Method	Suc.6 ↑	Suc.1 ↑	Pen. ↓
SceneDiffuser	16.1	52.1	29.2
GraspTTA	25.5	64.8	31.6
UGG	33.6	74.5	33.0
DexGrasp Any.	38.4	77.5	19.2
Ours w/o HPO	52.6	88.8	19.5
Ours	54.1	90.1	19.4

D. Additional Experiments

D.1. Performance of Cross Dataset

We conducted cross-validation experiments on four datasets with our method and one dataset with four methods. The results with four datasets are shown in Table 3, which demonstrate that the Physics-Aware Consistency Model trained on the Multidex dataset achieved the best performance when tested on the other datasets. The model trained and tested on the DexGRAB and DexGraspNet datasets showed moderate performance. Since Realdex is a real-world dataset with relatively lower quality, the performance of the model trained and tested on Realdex was relatively worse. The results with four methods are shown in Table 4, which illustrate that our methods can significantly improve the grasping performance on the realdex dataset compared with other methods.

D.2. More Ablation Studies

Table 5. Ablation study on different hyperparameters (i.e., the regularization weight β , the number of iterations per finetuning epoch N_{ft}). We report the results under 2, 4, and 8 steps during sampling.

T	β	Suc.6 ↑	Suc.1 ↑	Pen. ↓	N_{ft}	Suc.6 ↑	Suc.1 ↑	Pen. ↓
2	0.1	65.6	97.5	15.2	1	65.9	97.2	15.3
	0.5	63.4	96.8	15.2	3	63.4	97.1	15.1
	1.0	65.9	97.2	15.3	5	66.2	96.9	15.2
	2.0	65.9	97.2	15.3	10	65.3	97.5	15.3
4	0.1	75.9	97.1	13.1	1	76.8	98.4	13.0
	0.5	77.1	97.2	13.1	3	75.9	97.8	13.0
	1.0	76.8	98.4	13.0	5	77.5	97.5	13.2
	2.0	77.2	97.2	13.2	10	75.9	97.5	13.1
8	0.1	79.4	97.8	12.2	1	80.3	98.7	12.3
	0.5	76.8	97.8	12.1	3	76.5	98.4	12.2
	1.0	80.3	98.7	12.3	5	78.8	98.1	12.2
	2.0	78.7	97.5	12.2	10	80.0	98.1	12.2

Impact of different hyperparameters. A comprehensive analysis of different hyperparameters (i.e., regularization weight β , number of finetuning N_{ft} every epoch, number of timesteps T) to the performance during preference align-

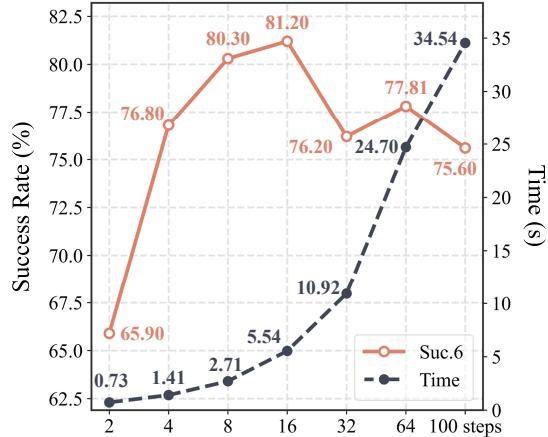


Figure 8. The effect of different sampling steps on grasping performance. The red solid line represents the grasping success rate, while the black dashed line denotes the time consumption.

ment is reported in Table 5 and Fig. 8. Table 5 demonstrates that when the sampling time steps is relatively small, such as 2 or 4 steps, increasing the number of finetuning iterations N_{ft} and raising the value of the regularization coefficient β can enhance the model’s performance. Conversely, when the sampling steps is larger, employing fewer N_{ft} and a smaller β value helps maintain the model at a high performance level. Fig. 8 shows that as the number of sampling steps increases, the grasping performance first improves and then declines. The highest grasping success rate is achieved when the sampling step is set to 16. The potential reason is that during the multi-step sampling process, each step introduces minor errors in noise handling. These errors may be masked in early steps but accumulate over time, eventually degrading the sample quality.