RAGrasp: A Retrieval-Augmented Framework with Diversity-Aware Modeling for Dexterous Grasp Generation

Zhuo Yang[1], Sanping Zhou[1*], Sen Wang[1], Jingyi Tian[1], Jiayi Li[1], Gang Hua[2], Le Wang[1]

[1] State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
[2] Amazon.com, Inc., Bellevue, WA 98004, USA
`https://yz-alpha.github.io/ragrasp/`

Abstract. Dexterous grasping is a fundamental ability for robots to perform complex manipulation tasks. However, simultaneously maximizing grasp success rates and maintaining grasp-type diversity remains a significant challenge. In response, we introduce RAGrasp, a retrieval-augmented framework that comprises two core components: a Geometry-Aware hand pose Initialization strategy(GAI) and a Diversity-Aware grasp-type Selection network(DAS). First, GAI constructs a distribution-aligned retrieval dataset and leverages object geometric similarity to retrieve high-quality initial samples, significantly improving grasp success rates. Second, DAS adopts a retrieval-augmented dual-head classifier that formulates grasp-type selection as a collaborative optimization problem of single-label and multi-label grasp-type classification, improving diversity and allowing flexible diversity adjustment. Extensive experiments show that RAGrasp yields an 8.2% relative improvement in success rate on BODex, alongside average relative gains of 13.0% and 11.9% on Dexonomy under type-unconditioned and type-conditioned settings.

Keywords: Dexterous Grasping · Retrieval-Augmented Generation · Generative Model

1 Introduction

Dexterous grasping, a fundamental skill for performing complex manipulation tasks, has gained increasing attention in robotics and embodied AI [18,33,36,42]. Existing methods can be broadly categorized into two groups: analytical methods and learning-based methods. Analytical methods [11, 17, 21, 25, 38] formulate grasp synthesis as a mathematical optimization problem. However, they typically assume complete knowledge of the object geometry, limiting their applicability to partial observations inherent in real-world. Learning-based methods [23, 44, 45, 49, 51] can handle partial observations and can be further divided

* Corresponding author.

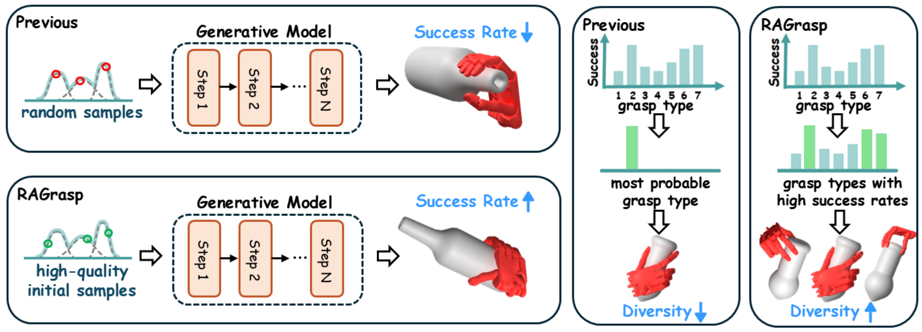

Fig. 1: Motivation of our pipeline. (Left) There are some high-quality initial samples, such as those aligned with the target object's geometry and grasp type, that are more likely to produce physically plausible grasp poses. (Right) For many objects, multiple grasp types are feasible and can achieve high success rates; consequently, selecting a grasp type beyond the single most probable grasp type can notably enhance diversity while maintaining a high success rate.

into three categories: regression-based, generation-based and hybrid methods. Regression-based methods [20, 45] directly learn a deterministic mapping from object representation to grasp poses. However, such deterministic mappings struggle to represent the inherently multimodal distribution of feasible grasps. Generative models [14, 19] capture these complex distributions more effectively. Generation-based approaches [15, 51] therefore employ them to jointly generate the hand's global pose and joint angles. However, because the global pose and joint angles are strongly coupled during generation, gradients from joint-angle prediction can perturb the optimization of the global pose, which is crucial for producing physically plausible grasps. Consequently, recent hybrid methods [3,4] decouple this process, employing generative modeling for global pose generation followed by a regression network for predicting joint angles.

Despite these advancements, current generation-based and hybrid methods largely overlook the role of hand pose initialization strategies in grasp generation. High-quality initial samples, such as those aligned with the object geometry and grasp type, exhibit a high probability of generating physically plausible grasp poses, as illustrated in Fig. 1 (Left). This motivates us to design a hand pose initialization strategy to identify these high-quality initial samples.

Moreover, selecting an appropriate grasp type is also critical for achieving high grasp success rates and diversity. As a pioneering effort, Dexonomy [3] introduces a multi-type grasp dataset and trains a classifier to infer grasp types from the object's point cloud. However, this classifier predicts only the most probable grasp type, collapsing multiple feasible grasp modes into a single prediction and thereby limiting grasp diversity. To mitigate this, we propose a novel grasp-type selection network that enables more diverse grasp generation while maintaining a high success rate, as illustrated in Fig. 1 (Right).

In this work, we present RAGrasp, a novel retrieval-augmented framework composed of two key components: a Geometry-Aware hand pose Initialization strategy (GAI) and a Diversity-Aware grasp-type Selection network (DAS).

Geometry-Aware Hand Pose Initialization Strategy. GAI improves grasp generation by anchoring the initial latent state to geometrically relevant priors. First, a mobius normalizing flow [22] is trained to model global hand rotation on the $SO(3)$ manifold, generating a large set of candidate grasps for training objects. These candidates are rigorously filtered via physics-based simulation [37] to retain only stable grasps. Crucially, we utilize model-generated grasps rather than dataset ground truths to eliminate distribution shifts between the training pipeline and the inference model. During inference, for each object in the test set, we retrieve several similar objects from the training set by computing feature similarity. The successful grasps associated with the retrieved objects are subsequently inverted through the trained flow model to obtain their exact representations in the base distribution. Because the retrieved objects share similar geometry with the test object, their representations in the base distribution are naturally better aligned with the test object's geometry than random initial samples, thereby serving as high-quality initial samples that facilitate the generation of physically plausible grasp poses.

Diversity-Aware Grasp-Type Selection Network. To improve diversity, DAS formulates grasp-type selection as a collaborative optimization problem of single-label and multi-label grasp-type classification. Specifically, DAS takes as input the global object feature and the grasp-type success statistics. Moreover, the network adopts a dual-head architecture. The first head models grasp-type feasibility as a multi-label classification task, predicting whether the success probability of each grasp type exceeds a threshold τ. Grasp types whose probabilities exceed τ are randomly selected, naturally promoting diversity. The second head performs single-label classification to identify the most probable grasp type. When none of the grasp-type probabilities predicted by the first head exceed the threshold τ, the prediction from the second head is selected, thereby ensuring a reliable grasp choice and maintaining high success rates.

In summary, our main contributions are as follows:

- We propose GAI, a Geometry-Aware hand pose Initialization strategy based on distribution-aligned retrieval, which retrieves model-generated successful grasps instead of ground-truth grasps to provide geometry-aware initial samples for stable and physically plausible grasp generation.
- We introduce DAS, a Diversity-Aware grasp-type Selection network that selects diverse grasp types with high success rates, mitigating grasp-type collapse while enabling flexible control over the diversity level.
- Extensive evaluations demonstrate that RAGrasp achieves state-of-the-art performance across large-scale benchmark datasets. Furthermore, we validate our method through real-world robotic grasping experiments.

2 Related Work

2.1 Dexterous Grasp Generation

Dexterous grasp generation methods can be broadly categorized into analytical and learning-based approaches. Analytical methods [4, 6, 26, 39, 41] formulate grasp synthesis as mathematical optimization problems. However, these approaches typically assume access to complete object geometry. Learning-based methods [13, 43, 45, 50, 51, 53] can be divided into regression-based, generation-based and hybrid approaches. Regression-based models [20,45] predict deterministic grasp poses and therefore often exhibit limited diversity. Generation-based methods explicitly learn a grasp distribution to represent multi-modal solutions. For example, diffusion-based approaches [15, 51] generate diverse grasps via iterative denoising. Moreover, hybrid methods combine generative modeling for global pose estimation with regression networks for joint configuration prediction. For instance, Dexonomy [3] employs a mobius normalizing flow to model global rotation and translation, followed by a regression network to predict joint angles, effectively leveraging the strengths of both paradigms. Nevertheless, current generation-based and hybrid methods overlook the impact of hand pose initialization strategies on grasp generation quality.

2.2 Initialization Strategies in Generative Models

Recent studies [2, 5, 24, 27] report that initial noise in diffusion models [9, 14, 29, 30, 34] can substantially influence semantic alignment with text prompts and the perceptual quality of synthesized images. Several works have explored strategies for optimizing or selecting the initial noise. InitNO [12], for instance, optimizes the initial noise to improve text–image alignment. Noise selection and refinement strategies are further explored in [1, 27], demonstrating that certain noise patterns consistently yield higher semantic fidelity and visual quality. This effect is formalized as the golden noise phenomenon in [52], which learns noise prompts to transform random Gaussian noise into semantically enriched latent representations. However, these strategies are primarily developed for diffusion-based architectures and are not directly applicable to normalizing flow-based models. In particular, mobius normalizing flow, which is well suited for modeling global rotation and translation in dexterous grasping, lacks a mechanism for latent initialization.

3 Method

In this section, we present RAGrasp, a novel framework that leverages retrieval for high-quality dexterous grasping. We first commence with a brief background on mobius normalizing flow. Then, we delve into the key designs of RAGrasp, including a Geometry-Aware hand pose Initialization strategy(GAI) and a Diversity-Aware grasp-type Selection network(DAS).

3.1 Preliminaries

Mobius Normalizing Flow. Mobius normalizing flow [22] is proposed to model rotations directly on the SO(3) manifold. This avoids the many-to-one mappings that often occur when representing rotations as 6D vectors in Euclidean space. It involves two complementary transformations tailored for the manifold structure of SO(3), enabling expressive yet invertible mappings on the rotation manifold.

The mobius transformation $f_\omega : \mathbb{S}^D \to \mathbb{S}^D$ is parameterized by a vector $\omega \in \mathbb{R}^{D+1}$ satisfying $\|\omega\| < 1$ and defined as

$$f_\omega(c) = \frac{1 - \|\omega\|^2}{\|c - \omega\|^2}(c - \omega) - \omega, \tag{1}$$

where $c \in \mathbb{S}^D$ represents a point on the unit hypersphere. To further alleviate the difficulty of learning a global rotation of distributions on SO(3), Mobius normalizing flow introduces an affine transformation defined as

$$g(q) = \frac{Wq}{\|Wq\|}, \tag{2}$$

where q is a unit quaternion and $W \in \mathbb{R}^{4\times 4}$ is an invertible matrix. This mapping linearly scales and rotates the quaternion through W, followed by normalization to ensure the output remains on the unit sphere and satisfies the antipodal constraint $g(-q) = -g(q)$.

By alternately applying the mobius transformation on rotation matrices and the affine transformation on quaternions, mobius normalizing flow can effectively model arbitrary rotation distributions on SO(3). This enables accurate global hand rotation modeling, which is crucial for achieving high grasp success rates in dexterous grasping.

3.2 Overview

Learning from large-scale grasp datasets provides a promising way for data-driven methods to learn diverse and robust grasp patterns across a wide range of objects. This way employs a dataset $\mathcal{D} = \{\mathcal{O}_i^{\text{partial}}, \mathcal{H}_i\}_{i=1}^{N_\mathcal{D}}$, containing $N_\mathcal{D}$ partial object point clouds $\mathcal{O}^{\text{partial}}$. Each $\mathcal{O}^{\text{partial}}$ is coupled with a grasp set $\mathcal{H} = \{H_i\}_{i=1}^{N_\mathcal{H}}$, whose size is $N_\mathcal{H}$. Specifically, a grasp H_i consists of three poses, including the pre-grasp pose $\boldsymbol{h}_p$, grasp pose $\boldsymbol{h}_g$, and squeeze pose $\boldsymbol{h}_s$:

$$H_i = \{\boldsymbol{h}_p,\, \boldsymbol{h}_g,\, \boldsymbol{h}_s\}, \tag{3}$$

The pre-grasp pose $\boldsymbol{h}_p$ is adopted for collision-free motion planning, followed by the grasp pose $\boldsymbol{h}_g$ for establishing stable contact, and the squeeze pose $\boldsymbol{h}_s$ for applying forces through hand–object interactions. Moreover, each pose is composed of three parts:

$$\boldsymbol{h} = (\boldsymbol{\theta},\ \boldsymbol{R},\ \boldsymbol{t}), \tag{4}$$

where $\boldsymbol{\theta} \in \mathbb{R}^k$ denotes the joint angles of the hand, and k represents the degrees of freedom (DoFs). For example, in the adopted ShadowHand model, the hand has $k = 22$ DoFs in total, including five DoFs for both the thumb and the little finger, and four DoFs for each of the index, middle, and ring fingers. In addition, $\boldsymbol{R} \in \text{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$ represent the global rotation and translation of the hand.

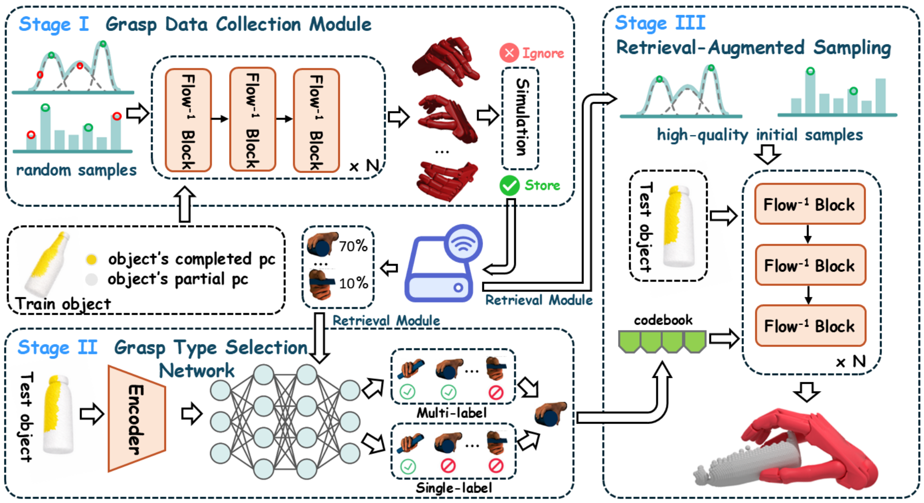

Fig. 2: Overview of RAGrasp. The framework comprises two key components, GAI and DAS. GAI incorporates a grasp data collection module that constructs a distribution-aligned retrieval dataset in Stage I and performs retrieval-augmented sampling in Stage III. For grasp-type selection in Stage II, DAS adopts a dual-head architecture, where a multi-label head estimates grasp-type feasibility for diverse grasp-type selection, while a single-label head serves as a fallback to select the most probable grasp type.

3.3 Geometry-Aware Hand Pose Initialization Strategy

Point Cloud Completion. To improve retrieval quality in the subsequent object retrieval module, RAGrasp incorporates a point cloud completion network [47] to reconstruct complete object point clouds from partial observations, as shown in Fig. 2. Given a partially observed object $\mathcal{O}_i^{\text{partial}}$, the network reconstructs a completed point cloud $\mathcal{P}_i^{\text{pred}}$ that yields a more complete and view-invariant geometric representation. This consistent shape description mitigates viewpoint-induced discrepancies and improves retrieval robustness.

Pose Generation Model Training. After obtaining the predicted entire point clouds $\mathcal{P}_i^{\text{pred}}$ from the point cloud completion network, we follow [3] to train a mobius normalizing flow model for grasp pose generation. Each $\mathcal{P}_i^{\text{pred}}$ is first encoded into a global object feature $\boldsymbol{f}_i$ using a sparse 3D convolutional backbone implemented with MinkowskiEngine [8]. A learnable codebook [40] is adopted to encode grasp types into a latent embedding. Conditioned on $\boldsymbol{f}_i$ and the grasp type embedding, the mobius normalizing flow generates the global rotation $\boldsymbol{R}_s^{\text{pred}} \in \mathrm{SO}(3)$ and translation $\boldsymbol{t}_s^{\text{pred}} \in \mathbb{R}^3$ of the squeeze pose, while additional MLPs are adopted to predict the corresponding joint angles $\boldsymbol{\theta}_s^{\text{pred}}$, pre-grasp pose $\boldsymbol{h}_p^{\text{pred}}$, and grasp pose $\boldsymbol{h}_g^{\text{pred}}$. We define the predicted squeeze pose $\boldsymbol{h}_s^{\text{pred}}$ and the generated grasp H_i^{pred} as:

$$\boldsymbol{h}_s^{\text{pred}} = (\boldsymbol{\theta}_s^{\text{pred}}, \boldsymbol{R}_s^{\text{pred}}, \boldsymbol{t}_s^{\text{pred}}), \quad H_i^{\text{pred}} = \{\boldsymbol{h}_p^{\text{pred}}, \boldsymbol{h}_g^{\text{pred}}, \boldsymbol{h}_s^{\text{pred}}\}, \tag{5}$$

where $\boldsymbol{h}_p^{\text{pred}}$, $\boldsymbol{h}_g^{\text{pred}}$, and $\boldsymbol{h}_s^{\text{pred}}$ represent the predicted pre-grasp pose, grasp pose, and squeeze pose, respectively.

Grasp Data Collection. To enable retrieval-augmented initialization, a high-quality grasp dataset is required to support reliable retrieval performance. After training the pose generation model, we employ it to generate a grasp set $\mathcal{H}_i^{\text{pred}}$ for each training object, defined as:

$$\mathcal{H}_i^{\text{pred}} = \{H_j^{\text{pred}}\}_{j=1}^{N_{\mathcal{H}^{\text{pred}}}}, \tag{6}$$

where H_j^{pred} denotes the j-th generated grasp for the i-th training object, and $N_{\mathcal{H}^{\text{pred}}}$ is the total number of generated grasps for each training object. We further evaluate all generated grasps in simulation to identify successful ones, and construct a success set $\mathcal{H}_i^{\text{succ}}$ for each training object as:

$$\mathcal{H}_i^{\text{succ}} = \{\, H_j^{\text{pred}} \in \mathcal{H}_i^{\text{pred}} \mid \text{Valid}(H_j^{\text{pred}}) = 1 \,\}, \tag{7}$$

where $\text{Valid}(H_j^{\text{pred}}) = 1$ indicates that the j-th generated grasp H_j^{pred} for training object i successfully passes the physics-based grasp validation. After filtering grasps that succeed in simulation, a dataset of successful grasps generated by the pose generation model can be defined as:

$$\mathcal{D}^{\text{succ}} = \{\, (\mathcal{P}_i^{\text{pred}}, \mathcal{H}_i^{\text{succ}}) \,\}_{i=1}^{N_{\mathcal{D}}}, \tag{8}$$

where $\mathcal{P}_i^{\text{pred}}$ denotes the predicted point cloud of the i-th object, $\mathcal{H}_i^{\text{succ}}$ represents its corresponding set of successful grasps, and $N_{\mathcal{D}}$ is the total number of objects in the dataset.

Instead of the original dataset $\mathcal{D}$, which is usually generated by an analytical method, we adopt the dataset $\mathcal{D}^{\text{succ}}$ which is generated by our trained pose generation model for the subsequent retrieval stage. Since the grasp distribution learned by the trained model differs from that obtained by the analytical method, relying on $\mathcal{D}$ would introduce a distribution mismatch, leading to degraded retrieval quality. In contrast, $\mathcal{D}^{\text{succ}}$ is generated under the same distribution learned by the trained model, which ensures consistency between training and retrieval, resulting in more reliable and higher-quality grasp retrieval in the next module.

Object Similarity Retrieval. After constructing the grasp dataset $\mathcal{D}^{\text{succ}}$, we retrieve high-quality grasps for each test object. Specifically, given a test object i, we first encode its predicted complete point cloud $\mathcal{P}_i^{\text{pred}}$ into a global feature $\boldsymbol{f}_i^{\text{test}}$, and compute the pairwise distance between $\boldsymbol{f}_i^{\text{test}}$ and all training object features $\boldsymbol{f}_j^{\text{train}}$ to form a distance matrix $\mathbf{M}$, whose elements are defined as:

$$\mathbf{M}_{i,j} = \|\boldsymbol{f}_i^{\text{test}} - \boldsymbol{f}_j^{\text{train}}\|_2^2, \tag{9}$$

where $\mathbf{M}_{i,j}$ denotes the squared Euclidean distance between the test object i and the j-th training object in the feature space. A smaller $\mathbf{M}_{i,j}$ indicates higher geometric similarity between the two objects. Based on the distance matrix $\mathbf{M}$, we sort each row in ascending order to obtain the ranking of training objects for each test object:

$$\boldsymbol{\pi}_i = \text{argsort}(\mathbf{M}_{i,:}), \tag{10}$$

where $\boldsymbol{\pi}_i$ denotes the ordered indices of training objects ranked by increasing distance to the test object i. We leverage $\boldsymbol{\pi}_i$ to identify the most relevant training objects for each test object and aggregate their successful grasps. Therefore, we can construct a retrieved set for each test object i:

$$\mathcal{S}_i^{\text{retr}} = \{\, (H_j^{\text{succ}}, \mathcal{P}_j^{\text{pred}}) \,\}_{j=1}^{N_{\mathcal{S}^{\text{retr}}}}, \tag{11}$$

where H_j^{succ} denotes a successful grasp, $\mathcal{P}_j^{\text{pred}}$ is the predicted point cloud corresponding to H_j^{succ}, and $N_{\mathcal{S}^{\text{retr}}}$ determines the number of retrieved grasps.

Retrieval-Augmented Initialization and Sampling. To obtain high-quality initial samples for mobius normalizing flow to generate rotation and translation, given a test object i, we first retrieve the set $\mathcal{S}_i^{\text{retr}}$ and extract the rotation $\mathbf{R}_s^{\text{succ}}$ and translation $\mathbf{t}_s^{\text{succ}}$ from each H_j^{succ}. Together with the grasp type embedding $\mathbf{E}_j$ and the predicted point cloud $\mathcal{P}_j^{\text{pred}}$ of the **training object**, these components are mapped into the latent space of the flow via the transformation:

$$(\mathbf{z}_s^{\text{rot}},\ \mathbf{z}_s^{\text{trans}}) = (f_K^{-1} \circ f_{K-1}^{-1} \circ \cdots \circ f_1^{-1})\big((\mathbf{R}_s^{\text{succ}}, \mathbf{t}_s^{\text{succ}});\ \mathbf{E}_j;\ \mathcal{P}_j^{\text{pred}}\big), \tag{12}$$

where f_k^{-1} denotes the k-th inverse flow layer and $(\mathbf{z}_s^{\text{rot}},\ \mathbf{z}_s^{\text{trans}})$ are the latent variables corresponding to rotation and translation. These latent variables are then used as geometry-aligned initial samples for the **test object** i. Starting from these initial samples, the final global rotation and translation are obtained through the transformation:

$$(\mathbf{R}_s^{\text{pred}},\ \mathbf{t}_s^{\text{pred}}) = (f_K \circ f_{K-1} \circ \cdots \circ f_1)\big((\mathbf{z}_s^{\text{rot}}, \mathbf{z}_s^{\text{trans}});\ \mathbf{E}_i;\ \mathcal{P}_i^{\text{pred}}\big), \tag{13}$$

where f_k represents the k-th forward flow layer conditioned on the predicted complete point cloud $\mathcal{P}_i^{\text{pred}}$ of the **test object** i. We further adopt MLPs to predict the corresponding joint angles $\boldsymbol{\theta}_s$, the pre-grasp pose $\mathbf{h}_p$, and the grasp pose $\mathbf{h}_g$.

3.4 Diversity-Aware Grasp-Type Selection Network

Although GAI improves the quality of pose generation, selecting an appropriate grasp type is equally critical for achieving high grasp success rates and maintaining pose diversity. Current grasp-type classifiers are typically formulated as single-label classification problems, where only the most probable grasp type is selected for each object. This formulation implicitly assumes that each object corresponds to a unique optimal grasp mode, which limits diversity. However, many objects can admit multiple feasible grasp types with similarly high grasp success rates.

To address this issue, we train a Diversity-Aware grasp-type Selection network(DAS). During inference, the input of DAS consists of two components: the global feature of the test object $\boldsymbol{f}_i^{\text{test}}$ and the grasp-type success statistics $\boldsymbol{s}_j$ obtained from $\mathcal{D}^{\text{succ}}$. Here, $\boldsymbol{s}_j \in \mathbb{R}^T$ records the success rates of all grasp types for the retrieved object. These inputs are combined to form a retrieval-enhanced representation $\boldsymbol{f}^{\text{fusion}}$. This representation is processed by a shared backbone to produce a high-level feature, which is then fed to both prediction heads below.

Multi-Label Grasp-Type Classification. The first head is a lightweight MLP that models grasp-type feasibility as a multi-label classification problem [35]. It is supervised using a binary cross-entropy loss with logits. This head estimates whether the success probability of each grasp type exceeds a predefined threshold τ, allowing multiple grasp types to be randomly selected. The choice of threshold τ controls the degree of diversity. A lower τ encourages activating multiple grasp types and thus promotes diversity, whereas a higher τ favors high-confidence predictions and improves grasp success rates.

Single-Label Grasp-Type Classification. The second head performs single-label grasp-type classification and is trained with a cross-entropy loss. It identifies the most probable grasp type and serves as a fallback mechanism. Specifically, when none of the grasp-type probabilities from the first head exceed the threshold τ, the prediction from the second head is selected.

The dual-head design naturally adapts to object difficulty. Geometrically simple objects inherently admit multiple valid grasp types; in such cases, the first head activates several feasible grasp types to encourage diversity. For challenging objects, where success probabilities remain low across all grasp types, the second head selects the most probable grasp mode to ensure a more reliable and stable prediction.

4 Experiments

4.1 Experiment Setting

Dataset. Simulated experiments are conducted in MuJoCo with a Shadow Hand, providing a fully reproducible environment for evaluating dexterous grasping. We use two large-scale datasets, BODex [4] and Dexonomy [3]. BODex contains 1.4 million floating fingertip grasps over 2.4k objects, where each object is normalized and scaled to multiple discrete levels within a fixed range. Single-view partial point clouds are used as input to mimic realistic perception settings, and each sample includes pre-grasp, grasp, and squeeze poses for comprehensive evaluation. Dexonomy is a larger, multi-type dataset with 9.5 million grasps over 10.7k objects, covering 31 grasp types defined by the grasp taxonomy [10]. It also adopts floating scenes and single-view partial point clouds, providing pre-grasp, grasp, and squeeze poses for each sample.

Metrics. We follow the evaluation setting of [3,4] for fair comparison and report a comprehensive set of grasp quality metrics. Grasp Success Rate (GSR) measures the percentage of successful grasps among all attempts, where a grasp is deemed successful if it withstands six external forces in MuJoCo without severe penetration (>1 cm) under the specified object mass and pose tolerance. Object

Success Rate (OSR) quantifies the percentage of objects for which at least one successful grasp is obtained. Contact Distance Consistency (CDC) computes the difference between the maximum and minimum signed contact distances across all fingers, reflecting the uniformity of finger–object contact. Penetration Depth (PD) denotes the maximum penetration distance between the hand and the object for each grasp, while Self-Penetration Depth (SPD) measures the maximum self-intersection distance among different hand links. Diversity (D) is defined as the proportion of variance explained by the first principal component in PCA over grasp translation, rotation, and joint angles.

Baselines. For the BODex dataset, we follow the official benchmark setup of BODex, evaluating four representative supervised learning architectures, including ISAGrasp (ISAG) [7], GraspTTA (GTTA) [16], 3D Diffusion Policy (DP3) [48], and UniDexGrasp (UDG) [46]. These methods span diverse generative paradigms, ranging from naive regression and CVAE-based approaches [32] to diffusion models and normalizing flows [28], thereby providing a comprehensive comparison across different architectural designs. For the Dexonomy dataset, we adopt the diffusion and normalizing flow(nflow) baselines released with the dataset.

Implementation Details. The pose generation models and the grasp-type selection network are trained on a single NVIDIA RTX 3090 GPU. The pose generation model is trained with a batch size of 256, an initial learning rate of 0.001 decayed to 0.0001, and gradient clipping with a threshold of 10. During evaluation, 100 candidate grasps are sampled per object and the top 10 are selected by likelihood. More details are provided in Appendix.

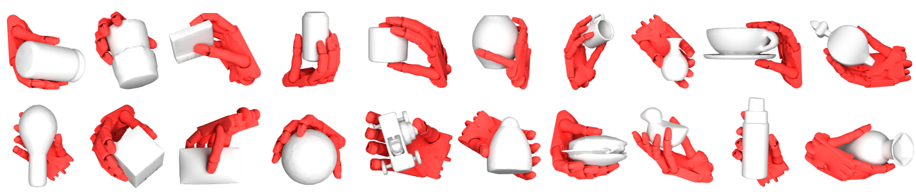

Fig. 3: Visualization of the generated grasp poses of RAGrasp on BODex.

4.2 Comparison with Existing Methods

We first conduct experiments on BODex, a fingertip grasping dataset containing only a single grasp type. As a result, our evaluation on BODex primarily validates the effectiveness of our proposed hand pose initialization strategy(GAI).

Tab. 1 presents the quantitative results on BODex. Our method achieves the highest grasp success rate (GSR) of **86.7%**, corresponding to an **8.2%** relative improvement and a **6.6%** absolute gain over the second-best method. It also attains the lowest PD and matches the best SPD,

Table 1: Quantitative results on BODex.

Method	GSR↑	PD↓	CDC↓	SPD↓	D↓
GTTA [16]	14.8	27.3	29.6	0.15	39.2
ISAG [7]	20.7	17.3	22.4	0.08	38.7
DP3 [48]	62.4	8.91	13.9	0.10	**32.6**
UDG [46]	80.1	3.47	**8.85**	**0.04**	33.4
Ours	**86.7**	**3.38**	9.55	**0.04**	37.7

demonstrating more stable and geometrically consistent grasp configurations. These results highlight the strong performance of our method on BODex. More visualization results are shown in Fig. 3.

We further evaluate our method on Dexonomy under the **type-conditioned** setting to validate the effectiveness of both the proposed hand pose initialization strategy and the grasp-type selection network. The results are shown in Tab. 2.

Table 2: Quantitative results on Dexonomy under type-conditioned settings.

Method	GSR↑	OSR↑	CDC↓	PD↓	D↓
nflow	63.9	91.3	13.9	8.6	25.7
Ours(div)	65.6	91.3	14.2	8.5	**25.5**
Ours(succ)	**71.5**	**91.4**	**13.4**	**8.3**	26.2

When our method favors diversity, which means adopting a lower threshold in DAS, it encourages the activation of more feasible grasp types, and our method achieves the best diversity score of **25.5**. In contrast, when favoring grasp success, which means adopting a higher threshold, the grasp success rate reaches **71.5%**. This corresponds to an **11.9%** relative improvement and a **7.6%** absolute gain while still maintaining competitive diversity.

We also evaluate our method on Dexonomy under the **type-unconditioned** setting. Tab. 3 reports detailed success rates of 31 grasp types on Dexonomy. Compared with the baselines provided by [3], our approach achieves the best average success rate of **39.1%**, representing a **13.0%** relative improvement and a **4.5%** absolute gain over the second-best method. Improvements are observed across a broad range of grasp families, including both power and precision grasps. For representative large-scale grasps such as Large Diameter and Prismatic 4 Fin-

Table 3: Type-unconditioned grasp success rates of 31 grasp types on Dexonomy. Diffusion and nflow are baselines provided by Dexonomy.

Method	Avg. Success↑	Large Diameter	Small Diameter	Medium Wrap	Adducted Thumb	Light Tool	Prismatic 4 Finger	Prismatic 3 Finger
diffusion	24.6	42.1	6.7	14.6	21.7	2.3	39.6	34.9
nflow	34.6	55.5	9.2	20.7	30.2	**4.2**	48.8	45.4
ours	**39.1**	**57.8**	**11.0**	**23.3**	**33.3**	4.1	**57.7**	**53.9**

Method	Prismatic 2 Finger	Palmar Pinch	Power Disk	Power Sphere	Precision Disk	Precision Sphere	Tripod	Fixed Hook
diffusion	27.0	33.4	16.2	25.0	23.5	34.3	19.2	6.5
nflow	40.7	48.0	22.5	40.5	34.9	46.9	32.9	9.3
ours	**44.9**	**52.3**	**26.6**	**48.8**	**43.8**	**56.8**	**39.5**	**9.6**

Method	Lateral	Index Finger Extension	Extension Type	Writing Tripod	Parallel Extension	Adduction Grip	Tip Pinch	Lateral Tripod
diffusion	14.8	13.4	43.9	25.1	42.2	19.6	26.9	8.8
nflow	28.4	15.8	61.2	40.2	52.0	33.1	41.6	17.0
ours	**34.0**	**18.0**	**66.9**	**46.1**	**54.8**	**41.1**	**46.9**	**22.7**

Method	Sphere 4 Finger	Quadpod	Sphere 3 Finger	Stick	Palmar	Ring	Ventral	Inferior Pincer
diffusion	39.7	29.3	28.9	12.0	24.9	35.4	15.7	35.0
nflow	51.3	42.4	38.2	**16.5**	30.8	47.8	19.1	46.3
ours	**53.5**	**51.8**	**45.3**	16.1	**33.7**	**50.1**	**20.5**	**47.4**

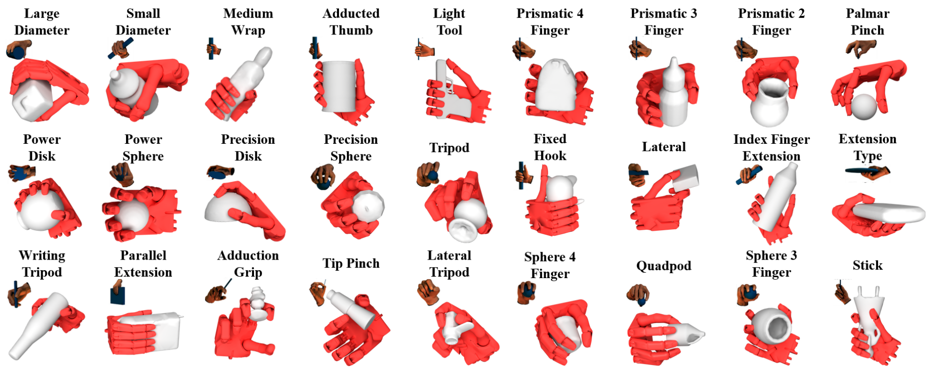

Fig. 4: Visualization of the generated grasp poses of RAGrasp on Dexonomy.

ger, the success rate increases to 57.8% and 57.7%, respectively. Clear gains are also achieved on fine-grained and contact-sensitive grasps, for example Precision Sphere reaching 56.8% and Writing Tripod reaching 46.1%. More visualization results are shown in Fig. 4.

Table 4: Ablation for GAI on Dexonomy. *Base* denotes normalizing flow baseline without GAI. *Comp* refers to the point cloud completion module of GAI. *Ret* represents the object retrieval module of GAI. *Data* represents the data collection module of GAI. *Base+Comp+Ret* denotes the setting where retrieval is performed using ground-truth poses from the dataset instead of model generated poses.

Configuration	Avg. Success↑	Large Diameter	Small Diameter	Medium Wrap	Adducted Thumb	Light Tool	Prismatic 4 Finger	Prismatic 3 Finger
Base	34.2	53.4	9.2	20.7	30.2	**4.2**	48.3	45.4
Base+Comp	36.0	55.5	9.7	22.1	30.8	3.4	52.6	49.4
Base+Comp+Ret	30.4	47.3	8.6	18.6	25.2	2.9	46.8	43.2
Base+Comp+Data+Ret	**39.1**	**57.8**	**11.0**	**23.3**	**33.3**	4.1	**57.7**	**53.9**

Configuration	Prismatic 2 Finger	Palmar Pinch	Power Disk	Power Sphere	Precision Disk	Precision Sphere	Tripod	Fixed Hook
Base	40.7	45.6	22.5	40.5	34.9	46.9	32.9	9.3
Base+Comp	41.7	48.6	24.6	42.3	38.8	50.4	35.3	9.0
Base+Comp+Ret	36.0	43.9	21.6	34.5	28.1	42.3	30.2	7.0
Base+Comp+Data+Ret	**44.9**	**52.3**	**26.6**	**48.8**	**43.8**	**56.8**	**39.5**	**9.6**

Configuration	Lateral	Index Finger Extension	Extension Type	Writing Tripod	Parallel Extension	Adduction Grip	Tip Pinch	Lateral Tripod
Base	28.4	15.8	61.7	40.2	50.2	33.1	41.6	17.0
Base+Comp	28.7	16.6	64.2	42.8	53.1	35.8	43.9	18.7
Base+Comp+Ret	24.7	14.8	51.5	35.4	46.5	32.3	40.4	15.0
Base+Comp+Data+Ret	**34.0**	**18.0**	**66.9**	**46.1**	**54.8**	**41.1**	**46.9**	**22.7**

Configuration	Sphere 4 Finger	Quadpod	Sphere 3 Finger	Stick	Palmar	Ring	Ventral	Inferior Pincer
Base	48.5	42.4	38.2	**16.5**	30.8	47.1	19.1	45.5
Base+Comp	51.4	45.9	40.7	15.2	33.1	47.7	19.4	45.7
Base+Comp+Ret	39.2	38.1	30.6	13.2	28.3	39.4	17.1	40.7
Base+Comp+Data+Ret	**53.5**	**51.8**	**45.3**	16.1	**33.7**	**50.1**	**20.5**	**47.4**

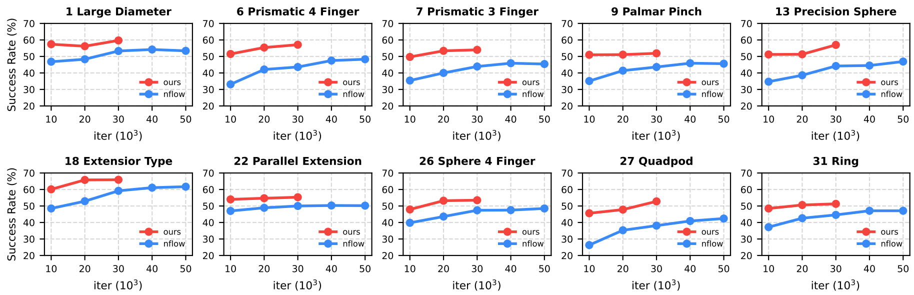

Fig. 5: Success rates over training iterations across 10 representative grasp types.

4.3 Ablation Study

Ablation for GAI. We analyze the contribution of our proposed GAI on Dexonomy. Tab. 4 reports the ablation results across 31 grasp types. The baseline model achieves relatively limited success rates. Introducing the completion module further improves geometric perception. The performance drop of *Base+Comp+Ret* reveals a clear domain shift between dataset grasps and model-generated grasps. This mismatch leads to less effective retrieval priors and consequently degrades the grasp generation performance. The result highlights the necessity of the proposed data collection module, which constructs the retrieval database using model-generated grasps to maintain distribution consistency. These results demonstrate that GAI plays a critical role in improving grasp success rates.

Performance Analysis. Fig. 5 illustrates the success rate as a function of training iterations across 10 representative grasp types. The horizontal axis denotes the number of training iterations measured in thousands, and the vertical axis represents the grasp success rate in percentage. Both methods improve as training progresses, but our model improves more rapidly and consistently achieves higher success rates across different grasp types. High success rates are already observed within the first 30000 iterations. This indicates that the proposed GAI provides high-quality latent priors that accelerate optimization and enable competitive performance with substantially fewer training iterations.

Ablation for DAS. We further analyze the effect of the threshold in the DAS.

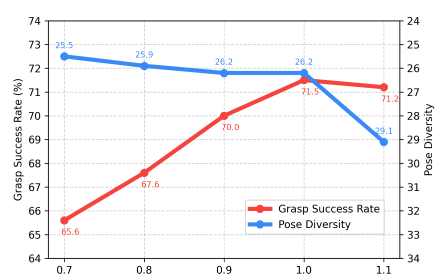

Fig. 6: Ablation on threshold in DAS

Activation occurs when the predicted success probability of a grasp type is greater than or equal to the threshold. When the threshold is set to 1.1, all outputs of the first head fall below the threshold, effectively deactivating the multi-label branch and reducing the model to rely solely on the second head for single-label prediction. In this case, the model selects only the most probable grasp type, behaving as a pure success-oriented classifier, which leads to limited grasp diversity, as illustrated

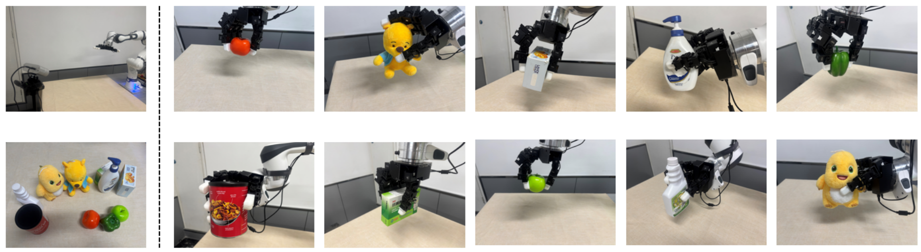

Fig. 7: Visualization of real world experiments.

in Fig. 6. As the threshold decreases, more feasible grasp types are activated, increasing grasp diversity while exhibiting a tendency toward reduced success rates. Notably, when the threshold is set to 1.0, the model maintains a high success rate while preserving competitive diversity, achieving a favorable balance between grasp feasibility and variation. This behavior naturally corresponds to scenarios where an object can be confidently grasped in multiple ways, allowing the model to activate several plausible grasp types and randomly select one among them instead of collapsing to a single dominant mode.

4.4 Real World Experiments

We conducted real-world grasping experiments to verify the practical effectiveness of our method. Ten objects are evaluated, with each object tested in three independent trials, achieving an overall success rate of 70%, as illustrated in Fig. 7. The model is trained on the leap hand [31] dataset provided by BODex. Our experimental setup consists of a leap hand mounted on a Franka Research 3 robotic arm, with an Orbbec Femto Bolt depth camera for perception. As the training data are generated under floating-object settings without environmental constraints, some predicted grasps may be infeasible in tabletop scenarios. Therefore, we filter out grasp poses that may result in collisions between the hand, the object, or the table, with the assistance of a motion planner to ensure collision-free feasibility during execution. In practice, the robotic arm is first moved to the predicted pose of the dexterous hand root, after which the leap hand joints are commanded to the predicted angles. Ten videos and additional implementation details are provided in Appendix.

5 Conclusion

We presented RAGrasp, a retrieval-augmented framework for dexterous grasp generation from partial observations. The framework introduces a Geometry-Aware hand pose Initialization strategy (GAI), which leverages a data collection module and object similarity retrieval to provide geometry-aligned initial samples that facilitate grasp pose generation. In addition, we propose a Diversity-Aware grasp-type Selection network (DAS), a dual-head classifier that improves grasp diversity while allowing flexible diversity adjustment. Extensive evaluations on BODex and Dexonomy demonstrate state-of-the-art grasp success rates.

Acknowledgement

This work was supported in part by the National Key Research and Development Project under Grant 2024YFB4708100, the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM504, National Natural Science Foundation of China under Grants 62572384, U24A20325, 62125305, 62503384, and Key Research and Development Plan of Shaanxi Province under Grant 2024PT-ZCK-80.

References

1. Ahn, D., Kang, J., Lee, S., Min, J., Kim, M., Jang, W., Cho, H., Paul, S., Kim, S., Cha, E., et al.: A noise is worth diffusion guidance. arXiv preprint arXiv:2412.03895 (2024)
2. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
3. Chen, J., Ke, Y., Peng, L., Wang, H.: Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. arXiv preprint arXiv:2504.18829 (2025)
4. Chen, J., Ke, Y., Wang, H.: Bodex: Scalable and efficient robotic dexterous grasp synthesis using bilevel optimization. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 01–08. IEEE (2025)
5. Chen, S.X., Vaxman, Y., Ben Baruch, E., Asulin, D., Moreshet, A., Lien, K.C., Sra, M., Sen, P.: Tino-edit: Timestep and noise optimization for robust diffusion-based image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6337–6346 (2024)
6. Chen, S., Bohg, J., Liu, C.K.: Springgrasp: Synthesizing compliant, dexterous grasps under shape uncertainty. arXiv preprint arXiv:2404.13532 (2024)
7. Chen, Z.Q., Van Wyk, K., Chao, Y.W., Yang, W., Mousavian, A., Gupta, A., Fox, D.: Learning robust real-world dexterous grasping policies via implicit shape augmentation. arXiv preprint arXiv:2210.13638 (2022)
8. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The grasp taxonomy of human grasp types. IEEE Transactions on human-machine systems **46**(1), 66–77 (2015)
11. Ferrari, C., Canny, J., et al.: Planning optimal grasps. In: Proceedings., 1992 IEEE International Conference on Robotics and Automation, 1992. vol. 3, pp. 2290–2295. IEEE (1992)
12. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9380–9389 (2024)
13. He, J., Li, D., Yu, X., Qi, Z., Zhang, W., Chen, J., Zhang, Z., Zhang, Z., Yi, L., Wang, H.: Dexvlg: Dexterous vision-language-grasp model at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14248–14258 (2025)

14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16750–16761 (2023)
16. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11107–11116 (2021)
17. Li, A.H., Culbertson, P., Burdick, J.W., Ames, A.D.: Frogger: Fast robust grasp generation via the min-weight metric. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6809–6816. IEEE (2023)
18. Li, J.: Sd2actor: Continuous state decomposition via diffusion embeddings for robotic manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13751–13760 (2025)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
20. Liu, M., Pan, Z., Xu, K., Ganguly, K., Manocha, D.: Deep differentiable grasp planner for high-dof grippers. arXiv preprint arXiv:2002.01530 (2020)
21. Liu, T., Liu, Z., Jiao, Z., Zhu, Y., Zhu, S.C.: 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)
22. Liu, Y., Liu, H., Yin, Y., Wang, Y., Chen, B., Wang, H.: Delving into discrete normalizing flows on so (3) manifold for probabilistic rotation modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21264–21273 (2023)
23. Lu, J., Kang, H., Li, H., Liu, B., Yang, Y., Huang, Q., Hua, G.: Ugg: Unified generative grasping. In: European Conference on Computer Vision. pp. 414–433. Springer (2024)
24. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
25. Miller, A.T., Allen, P.K.: Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine **11**(4), 110–122 (2004)
26. Murray, R.M., Li, Z., Sastry, S.S.: A mathematical introduction to robotic manipulation. CRC press (2017)
27. Qi, Z., Bai, L., Xiong, H., Xie, Z.: Not all noises are created equally: Diffusion noise selection and optimization. arXiv preprint arXiv:2407.14041 (2024)
28. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
30. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
31. Shaw, K., Agarwal, A., Pathak, D.: Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. arXiv preprint arXiv:2309.06440 (2023)

32. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)
33. Song, X., Li, Y., Zhang, Y., Liu, Y., Jiang, L.: An overview of learning-based dexterous grasping: recent advances and future directions. Artificial Intelligence Review **58**(10), 300 (2025)
34. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
35. Tarekegn, A.N., Ullah, M., Cheikh, F.A.: Deep learning for multi-label learning: A comprehensive survey. arXiv preprint arXiv:2401.16549 (2024)
36. Tian, J., Wang, L., Zhou, S., Wang, S., Li, J., Sun, H., Tang, W.: Pdfactor: Learning tri-perspective view policy diffusion field for multi-task robotic manipulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15757–15767 (2025)
37. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 5026–5033. IEEE (2012)
38. Turpin, D., Wang, L., Heiden, E., Chen, Y.C., Macklin, M., Tsogkas, S., Dickinson, S., Garg, A.: Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In: European Conference on Computer Vision. pp. 201–221. Springer (2022)
39. Turpin, D., Zhong, T., Zhang, S., Zhu, G., Liu, J., Singh, R., Heiden, E., Macklin, M., Tsogkas, S., Dickinson, S., et al.: Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation. arXiv preprint arXiv:2306.08132 (2023)
40. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
41. Wang, R., Zhang, J., Chen, J., Xu, Y., Li, P., Liu, T., Wang, H.: Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. arXiv preprint arXiv:2210.02697 (2022)
42. Wang, S., Wang, L., Zhou, S., Tian, J., Li, J., Sun, H., Tang, W.: Flowram: Grounding flow matching policy with region-aware mamba framework for robotic manipulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12176–12186 (2025)
43. Wang, W., Wei, F., Zhou, L., Chen, X., Luo, L., Yi, X., Zhang, Y., Liang, Y., Xu, C., Lu, Y., et al.: Unigrasptransformer: Simplified policy distillation for scalable dexterous robotic grasping. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12199–12208 (2025)
44. Wei, Y.L., Jiang, J.J., Xing, C., Tan, X.T., Wu, X.M., Li, H., Cutkosky, M., Zheng, W.S.: Grasp as you say: Language-guided dexterous grasp generation. Advances in Neural Information Processing Systems **37**, 46881–46907 (2024)
45. Xu, G.H., Wei, Y.L., Zheng, D., Wu, X.M., Zheng, W.S.: Dexterous grasp transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17933–17942 (2024)
46. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., et al.: Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4737–4746 (2023)

47. Yan, H., Li, Z., Luo, K., Lu, L., Tan, P.: Symmcompletion: High-fidelity and high-consistency point cloud completion with symmetry guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9094–9102 (2025)
48. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. arXiv preprint arXiv:2403.03954 (2024)
49. Zhang, J., Liu, H., Li, D., Yu, X., Geng, H., Ding, Y., Chen, J., Wang, H.: Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In: 8th Annual Conference on Robot Learning (2024)
50. Zhong, Y., Huang, X., Li, R., Zhang, C., Chen, Z., Guan, T., Zeng, F., Lui, K.N., Ye, Y., Liang, Y., et al.: Dexgraspvla: A vision-language-action framework towards general dexterous grasping. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 18836–18844 (2026)
51. Zhong, Y., Jiang, Q., Yu, J., Ma, Y.: Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22584–22594 (2025)
52. Zhou, Z., Shao, S., Bai, L., Zhang, S., Xu, Z., Han, B., Xie, Z.: Golden noise for diffusion models: A learning framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17688–17697 (2025)
53. Zhu, Y., Zhong, Y., Yang, Z., Cong, P., Yu, J., Zhu, X., Ma, Y.: Evolvinggrasp: Evolutionary grasp generation via efficient preference alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11665–11674 (2025)