RAGG: Retrieval-Augmented Grasp Generation Model

Zhenhua Tang[1], Bin Zhu[2], Yanbin Hao[3], Chong-Wah Ngo[2], Richang Hong[1]*

[1]Hefei University of Technology, Anhui, China
[2]Singapore Management University, Singapore
[3]University of Science and Technology of China, Anhui, China
zhenhuat@foxmail.com, {binzhu, cwngo}@smu.edu.sg, haoyanbin@hotmail.com, hongrc.hfut@gmail.com

Abstract

Intent-based grasp generation inherently involves challenges such as manipulation ambiguity and modality gaps. To address these, we propose a novel Retrieval-Augmented Grasp Generation model (RAGG). Our key insight is that when humans manipulate new objects, they initially mimic the interaction patterns observed in similar objects, then progressively adjust hand-object contact. Consequently, we develop RAGG as a two-stage approach, encompassing retrieval-guided generation and structurally stable grasp refinement. In the first stage, we propose a Retrieval-Augmented Diffusion Model (ReDim), which identifies the most relevant interaction instance from a knowledge base to explicitly guide grasp generation, thereby mitigating ambiguity and bridging modality gaps to ensure semantically correct manipulation. In the second stage, we introduce a Progressive Refinement Network (PRN) with Kolmogorov-Arnold Network (KAN) layers to refine the generated coarse grasp, employing a Structural Similarity Index loss to constrain the spatial relationship between the hand and the object, thus ensuring the stability of the grasp. Extensive experiments on the OakInk and GRAB benchmarks demonstrate that RAGG achieves superior results compared to the state-of-the-art approach, indicating not only better physical feasibility and controllability, but also strong generalization and interpretability for unseen objects.

Introduction

Being able to manipulate objects like humans holds significant implications across various fields, including human-computer interaction (Pollard and Zordan 2005), virtual reality (Höll et al. 2018; Wu et al. 2020), augmented reality (Hürst and Van Wezel 2013), and imitation learning in robotics (Hsiao and Lozano-Perez 2006; Liu et al. 2024a). To understand object affordance, i.e., manipulation intent or grasp type (Corona et al. 2020; Yang et al. 2022) through learning from human experience, recent advances have developed various multimodal hand-object interaction datasets (Taheri et al. 2020; Yang et al. 2022; Jian et al. 2023). These datasets are accompanied by text signals that specify the intent (e.g., hold and use), leading to the definition of a new **intent-based grasp generation** task. Given a 3D object and a manipulation intent, the goal is to generate a 3D hand pose that ensures physically feasible hand-object contact, enables affordance-aware interaction, and achieves robust generalization to unseen objects.

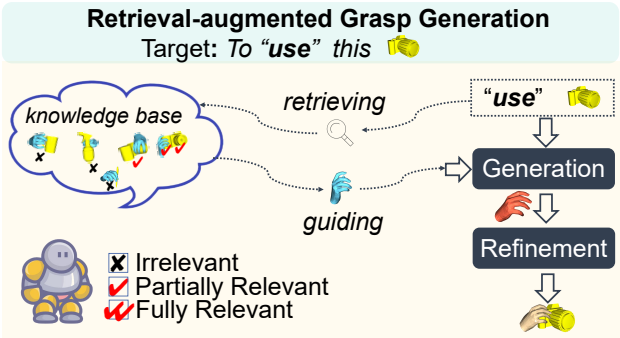

Figure 1: Our RAGG retrieves the most relevant grasp sample based on the object shape and manipulation intent to guide the generation process.

The intent-based grasp generation task inherently involves challenge of manipulation ambiguity, as one intent may result in different poses, complicating model convergence (Jian et al. 2023; Chang and Sun 2024). The gaps between the modalities of text, point cloud, and mesh further exacerbate this challenge in generating semantically correct grasps (Zhu et al. 2019; Liu, Li, and Lin 2023; Zhang et al. 2024). Yang *et al.* (Yang et al. 2022) attempt to address these challenges by employing GrabNet (Taheri et al. 2020). However, this approach is far from satisfactory compared to human-level interaction, as it merely compresses object point cloud, intent text, and hand pose into a latent space for reconstruction. More recently, approaches such as Text2Grasp (Chang and Sun 2024) and DiffH2O (Christen et al. 2024) utilize rich data annotations and textual descriptions as generation prompts to explicitly specify the parts of objects being interacted with, thus improving the model's understanding of manipulation intent and reducing ambiguity. Despite achieving impressive results, their approaches rely heavily on high-quality textual cues and lack the flexibility to handle uncommon objects.

In this paper, we address the above limitations by employing a Retrieval-Augmented Generation (RAG) mechanism (Sheynin et al. 2022; Chen et al. 2022; Blattmann et al. 2022;

*Corresponding author.

Zhang et al. 2023). **Our key insight is that when humans manipulate new objects, they initially mimic their interaction patterns observed in similar objects, then progressively adjust hand-object contact.** By using RAG, we specify a relevant interaction instance as auxiliary information, explicitly guiding the model to mimic the grasping of the referenced instance. This approach mitigates inter-modal barriers by providing a grasping reference to bridge the semantic gap between the object point cloud and the textual intent. Using the reference as a condition further reduces generation uncertainty, alleviating manipulation ambiguity. Additionally, when dealing with unseen objects, our approach offers the flexibility to guide generation by retrieving similar instances from known interaction samples. In contrast, methods like (Chang and Sun 2024; Christen et al. 2024) that rely on textual prompts require re-labeling of unseen objects, such as specifying the new object's category, the contact region, etc. Consequently, we develop a Retrieval-Augmented Grasp Generation model (RAGG), as illustrated in Figure 1, which comprises two stages: retrieval-guided generation and structurally stable grasp refinement.

In the first stage, we propose a Retrieval-Augmented Diffusion Model (ReDim), which integrates RAG into a diffusion-based generation framework. Specifically, we develop a Joint Retrieval mechanism (Sheynin et al. 2022; Chen et al. 2022; Blattmann et al. 2022; Zhang et al. 2023) that evaluates both object shape and manipulation intent to identify the most relevant interaction instance as a reference. It helps to bridge the semantic gap and reduce ambiguity. However, simply transferring the referenced grasp into target object may destroy the physical feasibility of the interaction. Therefore, we further utilize a Semantic Calibrator Transformer (SCT) to align the referenced grasp with the conditional signals and fuse them with the noise latent during the denoising process. The latent is finally used to predict clean signals from the noise space, resulting in a grasp that is semantically correct in manipulation.

In the second stage, we build upon previous pipelines (Taheri et al. 2020; Yang et al. 2022; Jian et al. 2023; Chang and Sun 2024) that utilize a Fully-Connection Network to refine the generated coarse grasp. However, we empirically found that prior refinements often produce features that loosen the object, leading to greater simulation displacements. To address this, we propose a Progressive Refinement Network (PRN) that improves refinement in two key aspects: (1) by introducing residual Kolmogorov-Arnold Networks (KAN) (Liu et al. 2024b) (ResKAN), which enhance the model's capacity with strong nonlinear representation capabilities (Zhang and Zhang 2024); and (2) by imposing a Structural Similarity Index (SSIM) (Wang et al. 2004) loss to limit structural incorrectness. This strategy encourages the model to leverage the spatial relationship between the hand and the object, thus ensuring the stability of the grasp.

We summarize the main contributions of this work as follows: First, ReDim is a novel intent-based grasp generation model that retrieves the most relevant interaction sample to guide generation, effectively mitigating manipulation ambiguity and modality gaps. Second, PRN contains an innovative architecture and loss function that ensure stable grasping, and it can be seamlessly integrated with other models to enhance their performance. By combining ReDim and PRN, our RAGG achieves superior results on the OakInk and GRAB datasets compared to state-of-the-art approach. Extensive experiments demonstrate its superior physical feasibility, controllability, and strong generalization and interpretability for unseen objects.

Related Work

Based on whether manipulation intent is specified in the conditional signals, we categorize the 3D hand-object interaction generation approaches into general grasp generation and intent-based grasp generation.

General grasp generation. The general grasp generation task explores how hands interact with objects without specifying manipulation intent. Previous regression-based methods (Liu et al. 2019, 2020) directly predict grasp parameters given the object as input, often resulting in either repetitive outcomes or less accurate predictions. Therefore, a line of works employ various generation models to improve the diversity. GanHand (Corona et al. 2020) utilizies a multi-task Generative Adversarial Network (GAN) (Goodfellow et al. 2020) to jointly analyzes the 3D shape/pose of object, predict possible grasp type, and then optimize the 3D hand model MANO (Romero, Tzionas, and Black 2022). GrabNet (Taheri et al. 2020) utilizes Conditional Variational Auto-Encoder (CVAE) (Sohn, Lee, and Yan 2015) to compresses hand pose and object shape features into a low-dimensional space, which is then recovered by sampling from Gaussian noise. And then GOAL(Taheri et al. 2022) expands GrabNet to whole-body grasp task. GF (Karunratanakul et al. 2020) calculates the signed distances to the surfaces of both the hand and the object, enabling the hand conform to the object's surface. Inspired by the success of diffusion model in text-to-image generation tasks, recent works have adapted it to grasp generation. DexDiffuser (Weng et al. 2024) combines a conditional diffusion-based sampler with an evaluator to generate and refine high-quality dexterous grasps from object point clouds. G-HOP (Ye et al. 2024) develops a hand's skeletal distance field to align with the latent signed distance field of the object, creating a coherent and accurate 3D representation. GeneOH Diffusion (Liu and Yi 2024) handles input trajectories with intricate interaction noise by first diffusing them to align with the whitened noise space and then cleaning them using a canonical denoiser.

Moreover, works such as (Brahmbhatt et al. 2019, 2020; Jiang et al. 2021; Li et al. 2022) collect additional contact maps to enhance physical feasibility. GraspTTA (Jiang et al. 2021) introduces a GraspCVAE to generate hand grasps and a ContactNet to predict object contact maps. Contact2Grasp (Li et al. 2022) learns the distribution of contact maps for grasps using a CVAE and then maps these contacts to grasps. ContacGen (Liu et al. 2023) proposes a comprehensive representation that encodes the specific contact parts of both the object and hand, along with the precise touch direction. UGG (Lu et al. 2023) employs a unified framework to jointly generate grasping hands, objects, and their contact information, ensuring diversity and physical feasibility. Despite their

success in generating natural and realistic hand poses for different objects, these methods are limited in their comprehensive understanding of object affordance.

Intent-based grasp generation. To learn how humans manipulate objects with specific intent, recent datasets collect multimodal hand-object interaction data. OakInk (Yang et al. 2022) captures grasps based on various object meshes and action texts, including use, hold, lift-up, hand-out, and receive, defining the intent-based generation task. Affordpose (Jian et al. 2023) goes further by labeling specific part-level affordances on the objects, such as twist, pull, handle-grasp, and the corresponding parts. As a baseline, OakInk (Yang et al. 2022) and AffordPose (Jian et al. 2023) inject the intent representation into GrabNet (Taheri et al. 2020) to guide generation. DiffH2O (Christen et al. 2024) designs a textual descriptor to specify the action of the interaction, as well as the name of the interacting object, reducing the uncertainty. Similarly, Text2Grasp (Chang and Sun 2024) leverages GPT-3 to generate various text prompts for interaction. Moreover, SemGrasp (Li et al. 2024) uses an automatic grasp language annotation methodology based on GPT-4 to augment previous datasets and employs vector quantized variational autoencoder (VQ-VAE) to discretize grasp components into tokens, aligning human grasp poses with language descriptions. Although these methods demonstrate outstanding performance in intent-based grasp generation task, they all rely on high-quality textual prompts for interactions, limiting their versatility with uncommon object categories. In this paper, we enhance the diffusion model with a RAG to explicitly guide generation, mitigating manipulation ambiguity and bridging modality gaps.

Proposed Method

Figure 2 depicts an overview of the proposed RAGG, which consists of two main stages: retrieval-guided generation by ReDim and structurally stable grasp refinement by PRN. This section is organized as follows: we first describe the Joint Retrieval mechanism for identifying the most relevant sample, then introduce ReDim for generating a semantically correct grasp guided by reference and conditional signals, and finally, we apply PRN to refine the initial grasp to ensure physical feasibility.

Joint Retrieval

Basically, there are two steps for Joint Retrieval. The first step is to establish a knowledge base from massive hand-object interaction samples. The second step is to retrieve the appropriate instance from the knowledge base as a reference.

Knowledge base. For the hand-object interaction sample $\{H_i, O_i, I_i\}$ in the training set, where $H_i = \left\{\theta \in R^{1\times 96}, P \in R^{1\times 3}\right\}$ indicates the 6D rotation of 16 hand joints and the root position, $O_i \in R^{N\times 3}$ indicates the 3D point cloud of object and I_i is the manipulation intent, it will take extremely high retrieval costs for calculating the similarities across all entities. To support the retrieval process, we establish a lightweight knowledge base from all the training data, containing diverse interaction samples. Specifically, for each data sample, we first extract its point cloud features of the object $f_i^{obj} \in R^{4096}$ using the Basis Point Set (BPS) model (Prokudin, Lassner, and Romero 2019), which efficiently represents the shape of the object. We then translate the intention into one-hot embedding $f_i^{int} \in R^1$ and combine it with the BPS feature to achieve an object-intent joint feature $F_i = \left\{f_i^{obj}, f_i^{int}\right\} \in R^{4097}$.

Next, we use the K-means algorithm to cluster the samples into K clusters by calculating the cosine distance between the object-intent joint feature of each sample and the cluster centers. Finally, for each cluster we retain only the L samples closest to the center, resulting in a knowledge base containing $K \times L$ samples. Each entity $\left\{H_i^{KB}, F_i^{KB}\right\}$ in the knowledge base consists of the hand pose H_i^{KB} and the object-intent joint feature F_i^{KB}.

Retrieval reference. To find the appropriate grasp samples, we compute the object-intent joint feature $\left\{f^{obj}, f^{int}\right\}$ of the generated conditions in the same way (by BPS and one-hot embedding). As shown in left-top of Figure 2, the similarity score s_i between given conditions and the i_{th} entity in knowledge base is calculated as below:

$$s_i = \left\langle \left\{f^{obj}, f^{int}\right\}, F_i^{KB} \right\rangle, \tag{1}$$

where $\langle ., . \rangle$ denotes the cosine similarity between the two feature vectors. This feature enables the representation of both the object shape and manipulation semantics in interaction. The similarity decreases when the objects are not similar in shape or the intention of the interaction differs, and vice versa. We then sort all elements by the score s_i in Equ. 1 and select v most relevant entities from the knowledge, constituting an grasp set M for target grasp. Formally, we have:

$$M = \{H_i \mid H_i = \{P_i, \theta_i\}, i \in Top_v(s_1, s_2, \ldots, s_{K\times L})\}. \tag{2}$$

The set contains the most probable grasping gestures under the current conditions, helping to bridge modality gaps and reduce ambiguity. In this paper, we set v to 1. To simplify the notation, we denote the given object as o, the intent as i, and the retrieved grasp as m. And they are then fed into the denoiser of ReDim to explicitly guide the generation.

Retrieval-Augmented Diffusion Model

In this section, we introduce how to reference the retrieved information via our proposed ReDim. Figure 2 (a) depicts the overall architecture of ReDim, a versatile generation model that can be parameterized as two Markov chains: 1) a diffuse phase that gradually perturbs data to noise, and 2) a reverse phase that reconstructs the uncontaminated data using a denoiser.

Diffuse phase. In the diffuse phase, we first sample a timestep $t \sim U(0, T)$, where T is the maximum number of timesteps. Then the real grasp data y_0 is diffused to the corrupted pose y_t by adding t-step independent Gaussian noise $\epsilon \sim U(0, 1)$. Following DDPMs (Ho, Jain, and Abbeel 2020), this process can be formulated as:

$$q(y_t|y_0) := \sqrt{\bar{\alpha}} y_0 + \varepsilon\sqrt{1-\bar{\alpha}_t}, \tag{3}$$

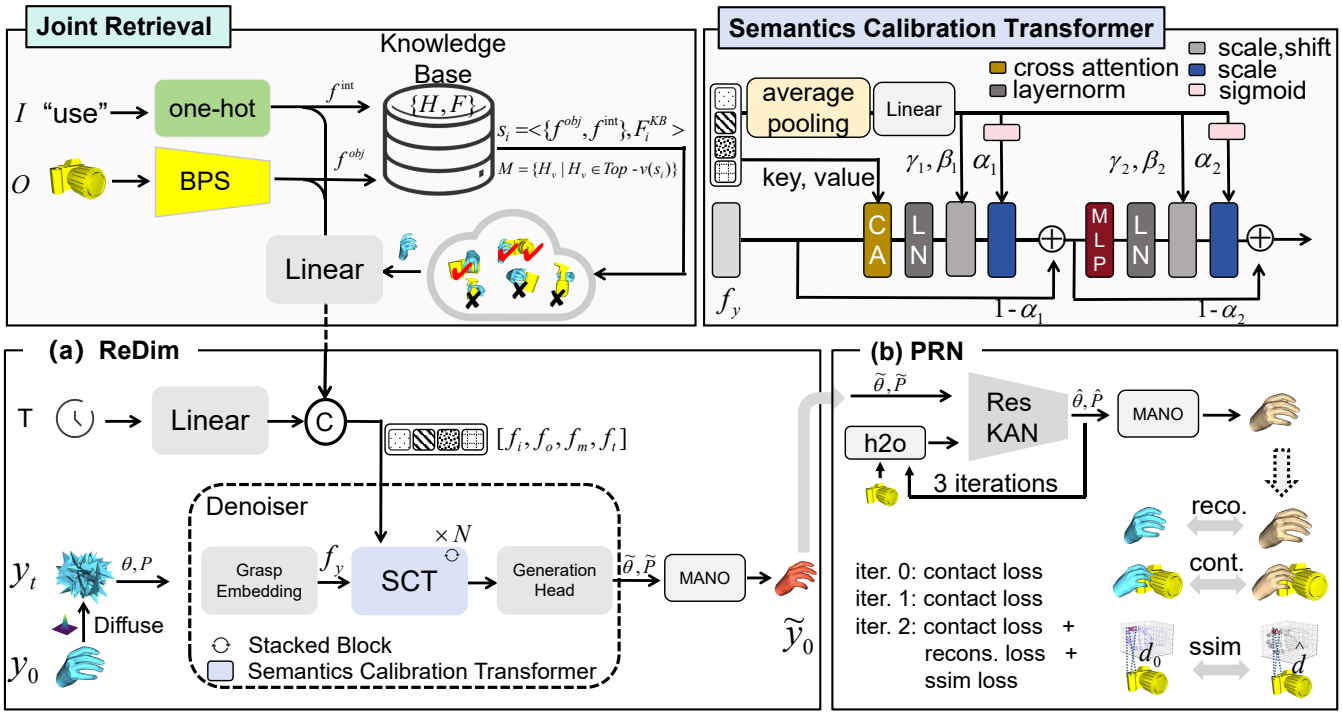

Figure 2: Overview of the proposed RAGG, comprising two stages: (a) retrieval-guided generation by ReDim and (b) structurally stable grasp refinement by PRN. (a) ReDim perturbs real grasp poses into noise, then integrates the Joint Retrieval mechanism and Semantics Calibration Transformer (SCT) into the denoiser, generating a semantically correct grasp. (b) PRN refines the generated coarse grasp using a residual KolmogorovArnold Network (ResKAN), optimized with multiple loss functions over three iterations.

where $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$ and $\alpha_t := 1 - \beta_t$, β_t is the cosine noise variance schedule. When T is large enough, the distribution of $q(y_T)$ is nearly an isotropic Gaussian distribution.

Denoise phase. Subsequently, y_t is sent to a denoiser $\mathfrak{D}$ conditioned on given conditional signals (i.e., object o and intent i), timestep t and also the reference grasp set m to reconstruct the hand pose $\tilde{y}_0$ without noise:

$$\tilde{y}_0 = \mathfrak{D}(y_t, i, o, t, m), \tag{4}$$

The denoiser consists of a grasp embeding, multiple stacked Semantics Calibration Transformer (SCT) blocks for context (i.e., object, intent, referenced grasp, and input noise) aggregation and a generation head.

Grasp embedding. Taking the noisy grasp pose $\{\theta, P\}$ decoupled from y_t as input, ReDim firstly projects the grasp pose to high-dimensional feature by a grasp embedding layer. This layer applies two independent linear transformations to the rotation and root position, followed by a GELU activation function. The resulting features are then concatenated and passed through another linear layer with an activation function. This process produces a feature vector f_y with a shape of $1 \times C$.

Semantics Calibration Transformer blocks. The Semantics Calibration Transformer (SCT) aims to align and selectively absorb information from all conditional signals, including objects, intents, and referenced grasps, for generation. To achieve this, we customize the advanced multimodal fusion transformer proposed in (Peebles and Xie 2023). As shown in the top-right of Figure 2, the one-hot feature of intent, BPS feature of the object, referenced grasp, and time step pass through four linear layers to obtain representations of the same dimension, denoted respectively as $f_i, f_o, f_m, f_t \in \mathbb{R}^{1\times c}$. After concatenation, they form the key $\mathbf{K}$ and value $\mathbf{V}$ vectors using two additional linear layers for cross attention, while the query $\mathbf{Q}$ is obtained by a linear transformation of the noisy grasp f_y. Moreover, we adopt the adaptive normalization layers (Peebles and Xie 2023) to regress the dimension-wise scale and shift parameters γ and β from the average of the features $[f_i, f_o, f_m, f_t]$. Formally, we have:

$$\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2 = \mathit{Linear}(\mathit{Avg}([f_i, f_o, f_m, f_t])), \tag{5}$$

where $[\cdot,\cdot]$ denotes concatenation, and Avg denotes the compression of the four features into a single feature using average pooling. The learnt parameters α_1 and α_2 are introduced to weight the residual connection. For simplicity, we omit the activation layer. Each SCT block is constructed using cross attention (CA), layer normalization (LN), and a multilayer perceptron (MLP) in sequence, with weighted skip connections:

$$\begin{aligned} H &= \gamma_1 * LN(CA(Q,K,V)) + \beta_1, \\ H_y &= \alpha_1 * H + (1-\alpha_1) * f_y, \\ Z &= \gamma_2 * LN(MLP(H_y)) + \beta_2, \\ Z_y &= \alpha_2 * Z + (1-\alpha_2) * H_y, \end{aligned} \tag{6}$$

where H, H_y and Z are the hidden features. The output Z_y serves as the input to the next block until the final block.

Generation head. Instead of predicting noise as formulated by DDPM (Ho, Jain, and Abbeel 2020), we follow MDM (Tevet et al. 2022) and predict the clean signal by Equ. 4. A linear layer is established on top of the SCT blocks to generate the target grasp pose $\left\{\tilde{\theta}, \tilde{P}\right\}$, which then passes through the MANO layer (Romero, Tzionas, and Black 2022) to obtain the hand mesh $\tilde{y}_0 \in R^{778\times 3}$. Thus, we can easily apply the standard reconstruction loss:

$$\mathcal{L}_{rec.} = \lambda_1 \left\|\tilde{y}_0 - y_0\right\|^2 + \lambda_2 \left\|\tilde{e}_0 - e_0\right\|^2 , \quad (7)$$

where $\tilde{e}_0$ and e_0 denote the edges between pairs of vertices of the MANO grasp, with $\lambda_1 = 34.825$ and $\lambda_2 = 29.85$. The contact losses are given by:

$$\mathcal{L}_{h2o} = \lambda_3 \left\|h2o(\tilde{y}_0, O) - h2o(y_0, O)\right\|^2 , \quad (8)$$

$$\mathcal{L}_{o2h} = \lambda_4 \left\|o2h(\tilde{y}_0, O) - o2h(y_0, O)\right\|^2 , \quad (9)$$

where $h2o$ denotes the signed distance from every vertex on the MANO hand to the object mesh, and $o2h$ denotes the signed distance from every vertex on the object mesh to the hand mesh, with $\lambda_3 = 34.825$ and $\lambda_4 = 29.85$.

Progressive Refinement Network

To further improve the physical feasibility of interaction, we input the coarse hand pose $\left\{\tilde{\theta}, \tilde{P}\right\}$ and the signed distance from the hand to the object into the Progressive Refinement Network (PRN) shown in Figure 2 (b). It consists of three residual blocks, each comprising two Kolmogorov-Arnold Network (KAN) layers, referred to as ResKAN, which enhance the models nonlinear representation. The detailed architecture of ResKAN is provided in the appendix. However, optimizing the refinement network using vertex-independent loss functions (i.e., Equ. 7 and Equ. 8) may cause all fingers to move away from the object, thereby increasing displacement. Hence, we propose using a Structural Similarity Index (SSIM) loss to maintain hand structure while adjusting the pose for stable grasping.

Specifically, we compute SSIM for each of the 17 palm parts based on the MANO model. Each part has a distance set, $\hat{d}$, calculated from the distances between its palm points and the object, along with the ground-truth distance set, d_0. Using their means, (μ_1, μ_2), variances, (σ_1^2, σ_2^2), and covariance, σ_{12}, we obtain the SSIM score:

$$SSIM(\hat{d}, d_0) = -\frac{(2\mu_1\mu_2 + C_1)(2\sigma_{12} + C_2)}{(\mu_1^2 + \mu_2^2 + C_1)(\sigma_1^2 + \sigma_2^2 + C_2)} , \quad (10)$$

where C_1 and C_2 are set to 0.0001 and 0.0009, respectively, to avoid numerical instability as the denominator approaches zero. We then sum the SSIM loss across each local region. This loss term constrains both local and global hand structure dynamics, minimizing when the spatial relationship between the hand and object is correct, thereby ensuring grasping stability. Finally, in line with previous works (Taheri et al. 2020; Yang et al. 2022), we optimize the network by applying contact, reconstruction, and SSIM losses over three iterations.

Experimental Results

We comprehensively evaluate the proposed RAGG and compare it with the state-of-the-art technique, GrabNet (Taheri et al. 2020), which has been expanded for the intent-based grasp generation task by (Yang et al. 2022).

Datasets and Evaluation Metrics

OakInk is a large-scale knowledge repository for understanding hand-object interactions, which contains 1800 object models of 32 categories. Building on the previous protocol (Yang et al. 2022), we use 9 categories of objects (i.e., bottle, camera, cylinder bottle, eyeglasses, game controller, lotion pump, mug, pen, and trigger sprayer) with 2 intents (i.e., use and hold) for training and testing, making our setup more challenging. To evaluate the generalization ability , we further select 3 unseen object categories: bowl, headphone, and knife for testing.

GRAB contains real human grasps for 51 objects from 10 different subjects. In this dataset, we select objects on which subject S1 performs pass manipulation to test the out-of-domain performance. These include two object categories identical to those in OakInK (camera and mug), as well as two unseen object categories (wineglass and toothpaste).

Evaluation metrics. Following previous works, we assess the physical feasibility of generated grasps by 1) penetration depth, 2) solid intersection volume (Yang et al. 2021), and 3) simulation displacement (Hasson et al. 2019). Additionally, we measure controllability by 4) downscaling the generated grasps using t-SNE based on the intent and calculating the center distance of different clusters. We also conduct a perceptual evaluation by 5) asking five volunteers to judge whether the generated hand poses on target objects demonstrate the given intents, and then count the success rate.

Performance Comparison on OakInk

Table 1 summarizes the performance comparisons in terms of physical feasibility and controllability. Overall, compared to the competitor GrabNet (Yang et al. 2022), RAGG achieves lower average penetration depth (i.e., 0.506 cm vs. 0.5133 cm, and 0.410 cm vs. 0.541 cm), intersection volume (i.e., 3.667 cm^3 vs. 5.165 cm^3, and 2.973 cm^3 vs. 4.867 cm^3), and simulation displacement (i.e., 1.274 cm vs. 1.964 cm, and 1.598 cm vs. 1.604 cm) under both manipulation intents, highlighting its advantages in physical feasibility. For the hold intent, RAGG achieves smaller penetration depths, insertion values, and simulation displacements on 7 out of 9, 9 out of 9, and 7 out of 9 objects, respectively. For the use intent, these values were achieved on 8 out of 9, 9 out of 9, and 6 out of 9 objects, respectively. Moreover, RAGG achieves much higher average center distances for generated grasps, demonstrating a better understanding of manipulation intent and the ability to generate distinguishable grasps. The higher success rates of human perception further demonstrate the advantage of RAGG in intent-based grasp generation.

Obj.	Method	Hold			Use			Control.	
		Pen. ↓	Ins. ↓	Dis. ↓	Pen. ↓	Ins. ↓	Dis. ↓	Cen. ↑	Hum. ↑
Bot.	GrabNet	1.140 ± 0.909	15.486 ± 22.390	2.441 ± 2.708	1.070 ± 0.849	10.727 ± 15.844	**2.476 ± 2.902**	4.296	56.341
	RAGG	**0.855 ± 0.760**	**8.423 ± 12.920**	**1.939 ± 2.338**	**0.841 ± 0.694**	**7.054 ± 15.138**	3.275 ± 3.460	**37.155**	**75.362**
Cam.	GrabNet	1.440 ± 0.800	17.167 ± 23.346	2.274 ± 1.732	1.641 ± 0.922	21.897 ± **25.595**	3.130 ± **1.996**	10.949	60.268
	RAGG	**1.283 ± 0.655**	**11.617 ± 12.150**	**1.703 ± 0.973**	**1.180 ± 0.660**	**19.185** ± 28.028	**3.122** ± 2.173	**28.928**	**88.839**
Cyl.	GrabNet	0.702 ± 0.810	8.395 ± 16.976	2.419 ± 2.783	0.634 ± 0.757	6.924 ± **14.602**	2.429 ± 2.731	6.110	45.847
	RAGG	**0.697 ± 0.744**	**6.819 ± 11.159**	**1.421 ± 1.112**	**0.619 ± 0.721**	**5.908** ± 15.136	**2.309 ± 2.523**	**42.045**	**58.866**
Eye.	GrabNet	**0.074 ± 0.075**	0.573 ± 0.971	**4.120 ± 1.211**	0.047 ± **0.063**	0.391 ± 0.817	**4.097 ± 1.342**	8.828	64.063
	RAGG	0.078 ± 0.093	**0.213 ± 0.298**	4.966 ± 1.416	**0.033** ± 0.064	**0.076 ± 0.236**	5.182 ± 2.188	**12.702**	**93.750**
Gam.	GrabNet	1.280 ± **0.367**	7.944 ± **2.759**	0.476 ± 0.148	1.318 ± **0.290**	12.225 ± 5.128	1.286 ± 1.273	2.615	50.000
	RAGG	**1.058** ± 0.554	**3.899** ± 3.315	**0.464 ± 0.064**	**0.503** ± 0.376	**3.023 ± 2.181**	**0.438 ± 0.163**	**14.041**	**75.000**
Lot.	GrabNet	1.042 ± 0.598	3.749 ± 3.528	1.419 ± 0.469	**0.691 ± 0.476**	4.898 ± 4.874	1.118 ± 0.191	**23.474**	50.000
	RAGG	**0.227 ± 0.198**	**0.481 ± 0.622**	**1.282 ± 0.313**	0.973 ± 0.603	**3.537 ± 3.573**	**1.105 ± 0.063**	21.841	**50.000**
Mug	GrabNet	**0.383** ± 0.318	2.857 ± 3.897	1.725 ± 2.095	0.429 ± 0.370	2.852 ± 3.341	1.172 ± 0.699	**69.602**	94.883
	RAGG	0.411 ± **0.301**	**2.405 ± 2.747**	**1.001 ± 0.702**	**0.299 ± 0.248**	**1.616 ± 1.910**	**1.047 ± 0.475**	68.367	**100.000**
Pen	GrabNet	0.347 ± 0.289	1.101 ± 1.427	**2.393 ± 2.900**	0.454 ± 0.221	1.386 ± 1.282	**0.969 ± 1.003**	10.34	77.404
	RAGG	**0.322 ± 0.249**	**1.085 ± 1.089**	4.026 ± 3.798	**0.235 ± 0.212**	**0.195 ± 0.269**	2.547 ± 2.464	**30.992**	**96.154**
Spa.	GrabNet	1.432 ± 0.864	21.535 ± 25.470	2.489 ± 2.754	1.500 ± 0.848	19.425 ± 17.828	2.397 ± 2.297	1.03	52.273
	RAGG	**1.276 ± 0.769**	**15.472 ± 13.634**	**1.298 ± 0.747**	**0.863 ± 0.367**	**9.099 ± 8.158**	**1.284 ± 0.997**	**5.646**	**95.455**
Ave.	GrabNet	0.5133 ± 0.584	5.165 ± 14.011	1.964 ± 2.593	0.541 ± 0.595	4.867 ± 12.049	1.604 ± **1.775**	33.094	82.463
	RAGG	**0.506 ± 0.497**	**3.667 ± 7.080**	**1.274 ± 1.318**	**0.410 ± 0.478**	**2.973 ± 7.839**	**1.598** ± 1.898	**61.395**	**93.440**

Table 1: Performance comparisons in terms of physical feasibility and controllability on OakInk dataset. "Hold" and "Use" indicate the conditional manipulation intent. The best result in each column are marked in **bold**. ↓ indicates that the lower value, the better generation. And ↑ indicates that the higher value, the better generation.

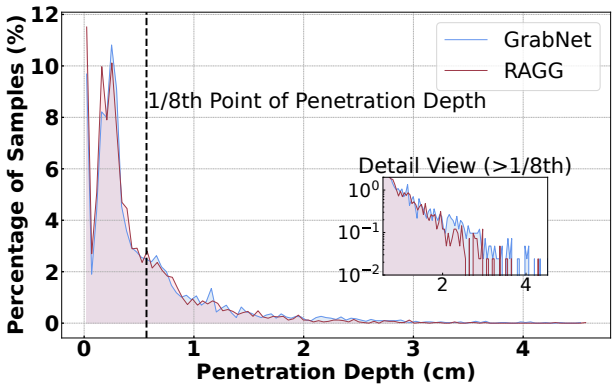

Figure 3: Generation result distributions on OakInk. The horizontal axe represents penetration depth, while the vertical axe shows the proportion of grasps per interval.

In Figure 3, we present the result distributions of RAGG and GrabNet (Yang et al. 2022). As shown in Figure 3, RAGG produces a higher percentage of samples with lower penetration depths and significantly fewer samples in the region of larger penetration depths compared to GrabNet. A similar pattern is observed in the appendix, where RAGG generates a greater proportion of samples with smaller simulation displacements. These results confirm the superiority of RAGG in achieving lower average penetration depth, stable grasping, and better distribution across error ranges.

Performance Comparison on Unseen Object

In Table 2, we further compare the quantitative results for unseen object categories on OakInk. The results indicate that RAGG achieves lower intersection volumes across all three categories, including bowl, headphone, and knife, and exhibits smaller penetration depths and displacements in the bowl and headphone categories, suggesting higher grasp quality. Additionally, RAGG consistently achieves greater center distances compared to GrabNet (Yang et al. 2022).

Obj.	Method	Physical Feasibility			Control.
		Pen. ↓	Ins. ↓	Dis. ↓	Cen. ↑
Bow.	GrabNet	0.314	6.102	**1.073 ± 1.113**	6.097
	RAGG	**0.312**	**4.958**	1.086 ± 1.179	**30.476**
Hea.	GrabNet	0.506	5.082	3.430 ± 2.128	5.431
	RAGG	**0.461**	**4.415**	**2.904 ± 1.657**	**12.081**
Kni.	GrabNet	**0.226**	1.023	3.961 ± 2.375	1.741
	RAGG	0.261	**0.949**	**3.756 ± 2.211**	**39.979**
Cam.	GrabNet	0.788	3.985	2.016 ± 2.167	17.299
	RAGG	**0.499**	**2.060**	**1.861 ± 1.707**	**20.195**
Mug	GrabNet	0.328	2.969	1.012 ± 0.553	4.520
	RAGG	**0.273**	**2.210**	**0.867 ± 0.365**	**15.608**
Too.	GrabNet	0.747	3.438	3.989 ± 4.139	19.609
	RAGG	**0.638**	**2.271**	**3.688 ± 2.937**	**39.763**
Win.	GrabNet	0.153	1.213	**2.779 ± 1.881**	5.777
	RAGG	**0.098**	**0.159**	4.680 ± 2.873	**30.511**

Table 2: Performance comparisons of unseen object categories on OakInk and out-of-domain objects on GRAB.

To further verify the generalization to out-of-domain objects, we test the performance on the GRAB dataset. Table 2 shows the comparisons between our RAGG and GrabNet (Yang et al. 2022), where RAGG achieves lower penetration and intersection values in all objects, indicating its superior generation quality. RAGG consistently demonstrates superior controllability for intent-based grasp generation, with higher center distance values across all object categories.

	Component				Physical Feasibility			Control.
	Base	OriR.	JR	PRN	Pen. ↓	Ins. ↓	Dis. ↓	Cen. ↑
#1	✓				0.672	11.177	1.457	33.479
#2	✓	✓			0.527	5.016	1.784	33.094
#3	✓			✓	0.489	4.277	1.571	32.892
#4	✓				0.554	4.553	1.255	57.767
#5	✓	✓			0.477	3.940	1.725	57.357
#6	✓		✓		0.594	5.270	1.102	61.508
#7	✓	✓	✓		0.468	3.793	1.787	57.449
#8	✓			✓	0.464	3.480	1.462	58.310
#9	✓		✓	✓	0.458	3.320	1.436	61.359

Table 3: Performance contribution of each component in the proposed RAGG on OakInk dataset. "Base" refers to using only the generation models. "OriR." denotes the original refinement network used by GrabNet (Yang et al. 2022).

Ablation Study

We conduct ablation studies to assess the impact of different design components on the OakInk dataset. Table 3 details the contribution of each component towards the overall performance. Without incorporating the Joint Retrieval mechanism and refinement network, RAGG achieves smaller penetration depth and displacement distance compared to GrabNet (i.e., #4 vs. #1), highlighting the advantage of the diffusion model. By introducing the Joint Retrieval mechanism to form our ReDiM, it better understands manipulation semantic, achieving comparable results in physical feasibility to the baseline of RAGG while significantly enhancing controllability by 61.508 (i.e., #6 vs. #4). The model is further improved by using refinement, which significantly reduced the penetration values (i.e., #7 vs. #6). It can be observed that using the Joint Retrieval mechanism to first mimic semantically similar grasps helps the refinement model with further adjustments (i.e., #7 vs. #5). However, employing the original refinement (Yang et al. 2022) causes the hand to release the object, reducing penetration but increasing displacement for both GrabNet and RAGG (i.e., #2 vs. #1 and #7 vs. #6). Alternatively, our PRN achieves better refinement quality, reducing penetration depth by 0.010 cm, intersection volume by 0.473 cm^3, and displacement by 0.351 cm (i.e., #9 vs. #7). Additionally, PRN can be conveniently applied to GrabNet, consistently enhancing its physical feasibility in terms of both penetration and displacement (i.e., #3 vs. #2).

Qualitative Analysis

In this section, we validate our RAGG through generated grasp visualization and correlation visualization between the referenced sample and the generated grasp.

Result Visualization of Seen Object Categories on OakInk. Figure 4 showcases intent-based grasp generation by our RAGG and GrabNet. Four examples are randomly selected from the sprayer and camera categories in OakInk. For each object, both hold and use manipulation intents are applied, with generated results shown from two different views. Overall, RAGG demonstrates superior generation quality across all four samples compared to GrabNet. Specifically, under the hold intent, RAGG consistently shows better physical feasibility, while GrabNet often produces results with significant interpenetration. Under the intent to use, RAGG exhibits greater controllability, such as correctly generating a grasping gesture to press the shutter rather than simply holding the camera.

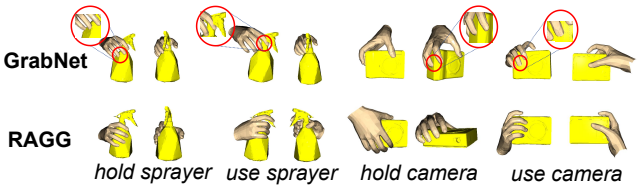

Figure 4: Examples of generated grasps from seen object categories by GrabNet and RAGG.

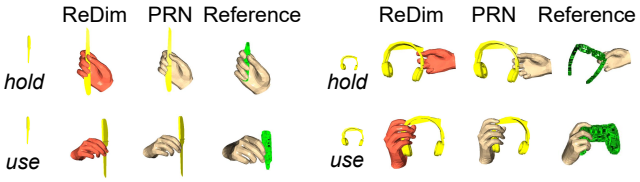

Figure 5: Examples of the generated grasp by RAGG and the corresponding reference.

Correlation Visualization Between the Referenced Sample and the Generated Grasp. To demonstrate the effectiveness of RAG, we present the results of RAGG (ReDim and PRN) alongside the corresponding references in Figure 5. The hand poses generated by ReDim closely align with the referenced grasps, highlighting the interpretability and effectiveness of RAG. For instance, when tasked with using an unseen object like a headphone, ReDim mimics the pose of the "use game controller" instance, albeit with slight penetration. This initial grasp is then refined by the PRN to achieve a more physically feasible interaction.

Conclusion

We have presented RAGG, a novel Retrieval-Augmented Grasp Generation model for intent-based grasp generation. RAGG comprises two main stages: retrieval-guided generation using a Retrieval-Augmented Diffusion Model (ReDim) and structurally stable grasp refinement through a Progressive Refinement Network (PRN). ReDim effectively bridges the semantic gaps between different modalities by retrieving the most relevant interaction instance from a knowledge base and integrating it using a Semantics Calibration Transformer (SCT). PRN further refines the generated grasp by incorporating a residual Kolmogorov-Arnold Network (ResKAN) and applying a Structural Similarity Index loss to ensure a stable grasp. Our extensive experiments on the OakInk and GRAB datasets demonstrate that RAGG outperforms state-of-the-art method, achieving superior results in terms of physical feasibility and controllability, while also demonstrating generalization to unseen objects.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under grants No. 61932009 and No. U23B2031.

References

Blattmann, A.; Rombach, R.; Oktay, K.; Müller, J.; and Ommer, B. 2022. Retrieval-augmented diffusion models. *NeurIPS*.

Brahmbhatt, S.; Handa, A.; Hays, J.; and Fox, D. 2019. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *IROS*.

Brahmbhatt, S.; Tang, C.; Twigg, C. D.; Kemp, C. C.; and Hays, J. 2020. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*.

Chang, X.; and Sun, Y. 2024. Text2Grasp: Grasp synthesis by text prompts of object grasping parts. *arXiv preprint arXiv:2404.15189*.

Chen, W.; Hu, H.; Saharia, C.; and Cohen, W. W. 2022. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

Christen, S.; Hampali, S.; Sener, F.; Remelli, E.; Hodan, T.; Sauser, E.; Ma, S.; and Tekin, B. 2024. DiffH2O: Diffusion-Based Synthesis of Hand-Object Interactions from Textual Descriptions. *arXiv preprint arXiv:2403.17827*.

Corona, E.; Pumarola, A.; Alenya, G.; Moreno-Noguer, F.; and Rogez, G. 2020. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *CVPR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*.

Höll, M.; Oberweger, M.; Arth, C.; and Lepetit, V. 2018. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *VR*.

Hsiao, K.; and Lozano-Perez, T. 2006. Imitation learning of whole-body grasps. In *IROS*.

Hürst, W.; and Van Wezel, C. 2013. Gesture-based interaction via finger tracking for mobile augmented reality. *Multimedia Tools and Applications*, 62: 233–258.

Jian, J.; Liu, X.; Li, M.; Hu, R.; and Liu, J. 2023. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *ICCV*.

Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*.

Karunratanakul, K.; Yang, J.; Zhang, Y.; Black, M. J.; Muandet, K.; and Tang, S. 2020. Grasping field: Learning implicit representations for human grasps. In *3DV*.

Li, H.; Lin, X.; Zhou, Y.; Li, X.; Huo, Y.; Chen, J.; and Ye, Q. 2022. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *arXiv preprint arXiv:2210.09245*.

Li, K.; Wang, J.; Yang, L.; Lu, C.; and Dai, B. 2024. SemGrasp: Semantic Grasp Generation via Language Aligned Discretization. *arXiv preprint arXiv:2404.03590*.

Liu, M.; Pan, Z.; Xu, K.; Ganguly, K.; and Manocha, D. 2019. Generating grasp poses for a high-dof gripper using neural networks. In *IROS*.

Liu, M.; Pan, Z.; Xu, K.; Ganguly, K.; and Manocha, D. 2020. Deep differentiable grasp planner for high-dof grippers. *arXiv preprint arXiv:2002.01530*.

Liu, S.; Zhou, Y.; Yang, J.; Gupta, S.; and Wang, S. 2023. ContactGen: Generative Contact Modeling for Grasp Generation. In *ICCV*.

Liu, X.; and Yi, L. 2024. GeneOH Diffusion: Towards Generalizable Hand-Object Interaction Denoising via Denoising Diffusion. *arXiv preprint arXiv:2402.14810*.

Liu, Y.; Chen, W.; Bai, Y.; Liang, X.; Li, G.; Gao, W.; and Lin, L. 2024a. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

Liu, Y.; Li, G.; and Lin, L. 2023. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11624–11641.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024b. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Lu, J.; Kang, H.; Li, H.; Liu, B.; Yang, Y.; Huang, Q.; and Hua, G. 2023. UGG: Unified Generative Grasping. *arXiv preprint arXiv:2311.16917*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*.

Pollard, N. S.; and Zordan, V. B. 2005. Physically based grasping control from example. In *SIGGRAPH*.

Prokudin, S.; Lassner, C.; and Romero, J. 2019. Efficient learning on point clouds with basis point sets. In *ICCV*.

Romero, J.; Tzionas, D.; and Black, M. J. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*.

Sheynin, S.; Ashual, O.; Polyak, A.; Singer, U.; Gafni, O.; Nachmani, E.; and Taigman, Y. 2022. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Taheri, O.; Choutas, V.; Black, M. J.; and Tzionas, D. 2022. GOAL: Generating 4D whole-body motion for hand-object grasping. In *CVPR*.

Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Weng, Z.; Lu, H.; Kragic, D.; and Lundell, J. 2024. DexDiffuser: Generating Dexterous Grasps with Diffusion Models. *arXiv preprint arXiv:2402.02989*.

Wu, M.-Y.; Ting, P.-W.; Tang, Y.-H.; Chou, E.-T.; and Fu, L.-C. 2020. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *Journal of Visual Communication and Image Representation*, 70: 102802.

Yang, L.; Li, K.; Zhan, X.; Wu, F.; Xu, A.; Liu, L.; and Lu, C. 2022. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Li, J.; and Lu, C. 2021. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*.

Ye, Y.; Gupta, A.; Kitani, K.; and Tulsiani, S. 2024. G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis. In *CVPR*.

Zhang, F.; and Zhang, X. 2024. GraphKAN: Enhancing Feature Extraction with Graph Kolmogorov Arnold Networks. *arXiv preprint arXiv:2406.13597*.

Zhang, H.; Zhu, B.; Cao, Y.; and Hao, Y. 2024. Hand1000: Generating realistic hands from text with only 1,000 images. *arXiv preprint arXiv:2408.15461*.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023. Remodiffuse: Retrieval-augmented motion diffusion model. In *ICCV*.

Zhu, B.; Ngo, C.-W.; Chen, J.; and Hao, Y. 2019. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *CVPR*.