

# OmniDexVLG: Learning Dexterous Grasp Generation from Vision Language Model-Guided Grasp Semantics, Taxonomy and Functional Affordance

Lei Zhang<sup>1,2†</sup>, Diwen Zheng<sup>3,2</sup>, Kaixin Bai<sup>1,2</sup>, Zhenshan Bing<sup>3</sup>, Zoltán-Csaba Márton<sup>2</sup>, Zhaopeng Chen<sup>2</sup>, Alois Christian Knoll<sup>3</sup>, Jianwei Zhang<sup>1</sup>

<sup>1</sup> University of Hamburg <sup>2</sup> Agile Robots SE <sup>3</sup> Technical University of Munich

† Corresponding Author: [lei.zhang-1@studium.uni-hamburg.de](mailto:lei.zhang-1@studium.uni-hamburg.de), [zhanglei.cn.de@gmail.com](mailto:zhanglei.cn.de@gmail.com)

**Abstract** — Dexterous grasp generation aims to produce grasp poses that align with task requirements and human-interpretable grasp semantics. However, achieving semantically controllable dexterous grasp synthesis remains highly challenging due to the lack of unified modeling of multiple semantic dimensions, including grasp taxonomy, contact semantics, and functional affordance. To address these limitations, we present OmniDexVLG, a multimodal, semantics-aware grasp generation framework capable of producing structurally diverse and semantically coherent dexterous grasps under joint language and visual guidance. Our approach begins with OmniDexDataGen, a semantic-rich dexterous grasp dataset generation pipeline that integrates grasp-taxonomy-guided configuration sampling, functional-affordance contact point sampling, taxonomy-aware differential force-closure grasp sampling, and physics-based optimization and validation, enabling systematic coverage of diverse grasp types. We further introduce OmniDexReasoner, a multimodal grasp-type semantic reasoning module that leverages multi-agent collaboration, retrieval-augmented generation (RAG), and chain-of-thought (CoT) reasoning to infer grasp-related semantics and generate high-quality annotations that align language instructions with task-specific grasp intent. Building upon these components, we develop a unified Vision-Language-Grasping (VLG) generation model that explicitly incorporates grasp taxonomy, contact structure, and functional affordance semantics, enabling fine-grained control over grasp synthesis from natural language instructions. Extensive experiments in simulation and real-world object grasping and ablation studies demonstrate that our method substantially outperforms state-of-the-art approaches in terms of grasp diversity, contact semantic diversity, functional affordance diversity, and semantic consistency. These results highlight the critical role of multi-dimensional semantic modeling, including grasp taxonomy, contact semantics, and affordance reasoning, in advancing dexterous grasp generation. More details are available on our project website <https://sites.google.com/view/omnidexvlg>.

**Keywords** — LLM/VLM, Multi-Fingered Robotic Hand, Grasp Taxonomy, Generation Model.

## I. INTRODUCTION

Dexterous hands, owing to their high degrees of freedom and complex manipulation capabilities, have demonstrated remarkable potential in fine-grained grasping and coordinated multi-finger tasks. In recent years, advancements in generative models and reinforcement learning have enabled researchers to synthesize reliable dexterous grasp poses from visual observations of objects, substantially improving grasp robustness and generalization across diverse objects and environments [1], [2].

In real-world scenarios, dexterous hands are often required to perform grasping and manipulation tasks with multi-dimensional semantic intent. For example, a task might involve grasping the handle of a kettle using a tripod grasp with the thumb, index, and middle fingers. For humans, such semantic reasoning is naturally achievable through prior knowledge and experience. To model this reasoning process, the concept of grasp taxonomy [3] has emerged as a structured framework that categorizes grasp strategies based on finger, contact point

configurations, opposition types, and force directions, thereby formalizing the diversity of grasp behaviors in both human and robotic hands.

However, existing semantic-guided grasp generation approaches primarily focus on functional grasping, lacking explicit sensitivity to grasp taxonomy and contact semantics [4]–[8]. In the context of objects with multiple functional affordance regions, different task goals often require distinct grasp strategies beyond the simplistic alignment with functional parts. Critical semantic elements such as finger flexion, contact configurations, and force application are highly correlated with the intended grasp type, yet current models struggle to utilize such fine-grained cues, resulting in limited controllability and semantic diversity in generated grasp actions.

Most existing dexterous grasp datasets [1], [9] are constructed through optimization-based methods that guide hand-object distances while satisfying quality metrics like differential force closure (DFC) [10]. While some recent works have introduced multimodal conditioning to enhance semantic grounding, these models largely focus on part-level semantics



Fig. 1: Overview of the OmniDexVLG framework. The framework integrates three core components: OmniDex-DataGen, for functional and grasp taxonomy-aware dexterous grasp dataset generation; OmniDexReasoner, for multi-dimensional semantic understanding using large multimodal models; and OmniDexGraspNet, a 3D vision-language grasp generation model guided by semantic instructions across grasp type, affordance, contact, and finger configuration.

or predefined functional goals. They lack the capacity to model and generalize structured grasp semantics, including contact semantics, grasp taxonomy, and affordance-sensitive representations [1], [2], [9], [11]. This leads to sparse distributions of grasp types in existing datasets [1], [11]. For instance, datasets like DexGraspNet [1] are dominated by simplified grasp types such as the pinch grasp, exposing the limitations of current models in generating diverse contact structures and grasp configurations. Although some studies have introduced manually annotated grasp-type labels [12], how to automatically generate diverse and semantically meaningful grasp types remains an open challenge.

To address these limitations, we propose a novel grasp data generation framework (OmniDexDataGen) that integrates functional affordance cues, contact pattern priors, and finger configuration semantics. This enables the synthesis of grasp samples that are not only physically stable, but also semantically rich and structurally diverse—providing stronger priors for downstream reasoning and generation tasks.

Furthermore, even in existing datasets with semantic labels, the grasp-related semantics are often coarse and limited to object-level part annotations [13] or basic functional descriptors [14]. There remains a lack of a unified semantic understanding model capable of jointly reasoning about grasp taxonomy, contact semantics, and functional affordance. Given the inherent complexity of dexterous hands, semantic reasoning in this context requires modeling the nuanced relationships between hand-object contact topology, force dynamics, and task-driven semantic objectives. The object’s affordance often constrains which grasp types are viable and, in turn, influences the contact structure and pose planning strategy required for successful manipulation.

Motivated by these challenges, we introduce a multimodal semantic understanding framework based on large multimodal

models (LMMs), named OmniDexReasoner, tailored for reasoning over dexterous grasping tasks. Our framework seeks to unify the modeling of grasp type, contact structure, and functional intent under a shared semantic space, enabling fine-grained understanding and generation of natural language grasp instructions for dexterous hands.

To further bridge the semantic gaps in current grasp generation pipelines, especially in representing grasp taxonomy, contact semantics, and affordance, we propose a vision-language grasp generation (VLG) model, OmniDexGraspNet, for semantically grounded grasp generation. By combining multimodal inputs, including language instructions and visual point clouds, with multi-dimensional semantic modeling, our method enables joint reasoning over complex task goals and grasp semantics. This facilitates the generation of grasp poses that are both semantically aligned and physically plausible. Through explicit modeling of structured semantics and controllable generation, our approach not only improves the diversity and realism of generated grasps, but also lays a foundation for more generalized, task-aware robotic manipulation.

Our Contributions are summarized as follows:

- **OmniDexDataGen: Functionality-Aware and Contact-Sensitive Dexterous Grasp Dataset Generation Method with Grasp Taxonomy Diversity.** We propose an optimization-based framework for generating dexterous grasp datasets that are sensitive to grasp taxonomy, hand-object contact points, and functional affordance. Our approach substantially enhances the dataset’s representational richness in terms of contact guidance and grasp semantics. Furthermore, it improves the diversity of grasp types, contact paradigms, and functional affordances.
- **OmniDexReasoner: LMM-Based Functional, Contact, and Taxonomy-Aware Dexterous Grasp Understanding.**

**ing Method.** We propose an LMM-powered framework for dexterous grasping that captures multi-level semantic cues across three core dimensions: functional affordance, contact semantics, and grasp taxonomy. By incorporating a multi-agent collaboration mechanism, retrieval-augmented generation (RAG) and Chain-of-Thought (CoT) reasoning, the proposed method addresses key limitations in current multimodal understanding approaches for dexterous hands, substantially enhancing the comprehension of complex grasping behaviors.

- **OmniDexGraspNet: Semantic-Aware 3D Vision-Language Grasp Pose Generation Model.** We propose a grasp pose generation method that integrates a 3D vision-language model to guide dexterous grasp synthesis using multi-dimensional semantic information and partial object point clouds. The method enables the generation of grasp poses that are sensitive to various semantic dimensions, including functional intent, grasp taxonomy, and contact configuration. It demonstrates strong generalization across objects of different sizes and categories, producing diverse functional grasps with rich semantic and structural variability. Compared to existing grasp generation approaches, our method exhibits superior semantic sensitivity and grasp diversity.

## II. RELATED WORK

### A. Semantic-Aware Dexterous Robotic Grasp Generation

In recent years, dexterous grasp generation has garnered increasing attention with the rise of generative models such as diffusion models [15]–[17] and variational autoencoders [9], and reinforcement learning [18], [19] in the robotics community. Owing to the high DoF and complex manipulation requirements of dexterous hands, grasp synthesis involves rich semantic dimensions, such as grasp types, functional affordance, and contact semantics. Existing approaches have primarily aimed to improve grasp stability and generalization across diverse object categories [1], [11].

To improve semantic grounding, recent works incorporate natural language guidance into grasp generation by leveraging LLMs or VLMs [6], [20]. These methods typically target functional grasping, where language instructions are encoded and fused with visual inputs to generate semantically aligned actions.

The complexity of semantic modeling also varies by end-effector type. For two-finger grippers, the semantics are often limited to object-level affordances [4], [5], whereas dexterous hands require finer modeling of grasp taxonomy, contact semantics, finger configurations, and opposition types. To this end, works [14], [21], [22] exploit hand-object representations as priors for functional grasping generation and transfer.

Moreover, grasp taxonomy and contact semantics have recently received increased attention. For instance, Dexonomy [12] and AnyDexGrasp [23] introduce grasp-type encodings to guide grasp synthesis, while ContactDexNet [9], GrainGrasp [24] and Grasp as You Say [25] leverage contact maps or linguistic references to inform fine-grained contact configurations.

In summary, while recent advancements have explored various aspects of semantic-aware grasp generation, a unified modeling framework that jointly captures grasp types, contact structures, functional intent, and finger configurations remains lacking. The role of semantic information in guiding dexterous grasp synthesis has yet to be fully explored and systematically leveraged.

### B. Dexterous Grasp Taxonomy and Reasoning

Grasp taxonomy [3] serves as a fundamental abstraction for describing human hand-object interaction strategies. It provides a structured categorization of grasp types based on contact point locations, involved anatomical links, opposition types, and force directions. Classical taxonomies, such as those proposed in [3], divide grasps into high-level categories including power, precision, and intermediate grasps, and further into subtypes such as tripod, lateral, and 2-finger pinch. These categories encapsulate nuanced differences in finger articulation, contact configuration, and force application strategies. Notably, recent research continues to identify new grasp types [10], underscoring both the richness of human and robotic grasp behavior and the inherent complexity of modeling such interactions. This ongoing evolution also reflects the challenge of reasoning over grasp taxonomies, which remains an underexplored area, especially for dexterous hands.

Early works employed CNN-based architectures to infer manipulation semantics, including grasp type and object attributes, directly from visual inputs [26]. These models built semantic action representations for reasoning about grasp categories and manipulation intent.

Recently, multimodal large models have demonstrated strong capabilities in semantic understanding and spatial comprehension [27]–[31]. Language-driven frameworks such as SemGrasp [6] and Multi-GraspLLM [20] leverage large language models (e.g., GPT-4) to perform grasp type prediction from natural language instructions. However, current vision-language models (VLMs) or large multimodal models (LMMs) struggle with accurate semantic grounding when interpreting hand-object interactions, often exhibiting semantic hallucinations for similar grasp types.

Consequently, there remains a significant gap in designing models that can reason over grasp taxonomy in a physically grounded, task-aware, and semantically aligned manner. Bridging this gap requires novel methods that integrate structured grasp knowledge with multi-modal embeddings, enabling more controllable and diverse dexterous grasp generation.

### C. Dexterous Hand Pose Data Generation

Existing dexterous grasp datasets [1], [11] are predominantly constructed using optimization-based approaches, which aim to generate stable grasp configurations by minimizing hand-object distances or maximizing grasp quality metrics such as Differentiable Force Closure (DFC) [10].

In terms of grasp type diversity, datasets such as DexGraspNet [1] are heavily biased toward simplified grasp types like the pinch grasp, with limited coverage of more complex or

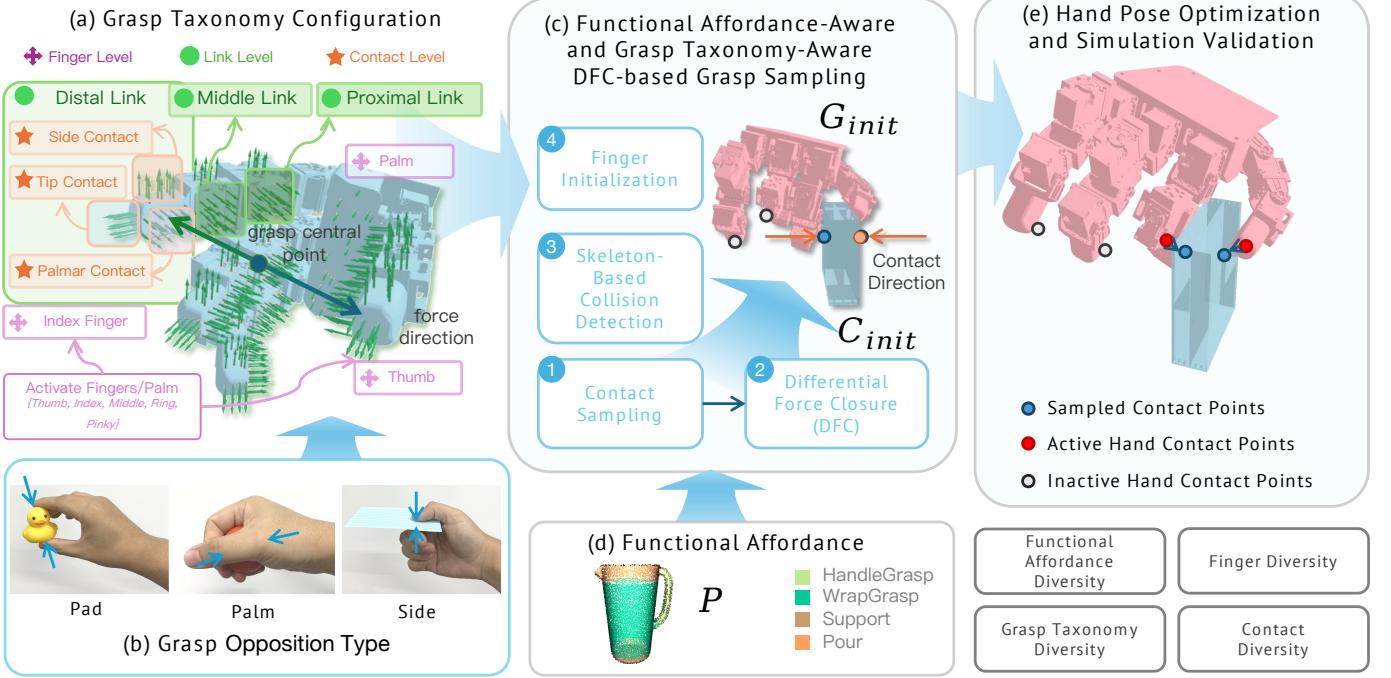


Fig. 2: OmniDexDataGen: Functional affordance-aware, contact and grasp taxonomy-aware dexterous grasp synthesis. Our grasp synthesis framework introduces a contact-level representation for dexterous hand manipulation. Each grasp is described by a multi-level configuration, including active fingers, involved links, and corresponding contact regions. Based on grasp opposition types, grasp-taxonomy-aware differential force closure sampler (Tax-DFCSampler) computes initial grasp poses and functional affordance contact point sampler (AC-Sampler) conducts contact sampling in object-specific affordance zones. Samples are quantitatively evaluated using differential wrench space metrics. High-quality grasps that pass collision filtering are subjected to hand pose optimization stage, further improving their performance. The finalized grasps are validated through physical simulation to ensure their robustness and applicability.

nuanced grasp configurations. This imbalance stems primarily from the lack of explicit modeling of grasp taxonomy and contact configurations during the data generation phase.

Other grasp synthesis methods that aim to cover diverse grasp types typically rely on manually predefined parameters or heuristics specific to each grasp category—such as in [12], [32]. However, these approaches lack the capability to automatically generate grasp configurations conditioned on grasp-type-relevant semantics, limiting their adaptability and scalability to unseen tasks or object categories.

While some recent efforts incorporate multimodal inputs [9], [11], the semantic annotations in these datasets remain relatively sparse and coarse. Most labels are restricted to object part names (e.g., handle, tip) or high-level functional descriptions (e.g., pour, hold) [6], [20], without capturing the full spectrum of semantic dimensions relevant to dexterous grasping.

### III. METHODS

#### A. Problem Statement and Method Overview

Existing dexterous grasp pose generation methods remain limited in their ability to model semantic information, particularly in scenarios requiring fine-grained control. This limitation restricts further improvements in the precision and adaptability of dexterous grasp strategies. Moreover, the scarcity of grasp

datasets with rich, multi-dimensional semantic annotations continues to hinder progress in this research direction.

To address these challenges, we propose a semantic-sensitive dexterous grasp dataset generation method OmniDexDataGen that incorporates multiple levels of semantic information, from functional affordance, contact configuration, to grasp taxonomy, as introduced in Sec. III-B. Building upon this foundation, we develop a semantic reasoning pipeline based on large multimodal models OmniDexReasoner to enhance the understanding of object-grasp relationships and grasp type, as shown in Sec. III-C. We further introduce a 3D vision-language model for semantically guided dexterous grasp pose generation, named OmniDexGraspNet, as detailed in Sec. III-D.

#### B. Functional Affordance-Aware and Contact-Sensitive Dexterous Grasp Dataset Generation Enriched by Grasp Taxonomic Diversity

We propose a dexterous grasp data generation method that is sensitive to both functional affordance and grasp taxonomy, as shown in Fig. 2. This method integrates the functional regions of the object, the contact configurations, and the opposition types associated with various grasp taxonomies. By leveraging an optimization-based strategy, our approach generates high-quality grasp samples with enhanced func-

tional diversity, contact semantic diversity, and grasp taxonomy diversity. The proposed grasp generation pipeline consists of following key parts: Grasp-type-aware configuration sampling, functional affordance-aware contact point sampling (AC-Sampler), grasp taxonomy-aware DFC grasp pose sampling (Tax-DFCSampler), and grasp optimization method.

In the grasp-type-aware configuration sampling stage, representative grasp configurations are automatically sampled according to the dexterous hand grasp taxonomy. In the affordance- and grasp-type-guided pose sampling method, initial grasp poses are sampled within object affordance regions based on task semantics and grasp-type constraints, including opposition type and corresponding contact configuration. Finally, optimization method refines the sampled grasp poses using functionality- and taxonomy-aware hand-object interaction loss and validates them through physical evaluation to ensure reliable data generation.

1) *Grasp Taxonomy-Aware Configuration Sampling*: Originating from grasp taxonomy analysis [3], different grasp types are typically characterized by specific opposition types, force directions, involved fingers in hand-object contact, contacting links of each finger, and the spatial distribution of contact regions. We define a structured representation of a grasp configuration, which includes: the set of activated fingers  $S_f$  participating in the grasp, the specific links  $S_l$  of each finger that are involved in contact, the contact regions  $S_c$  associated with each link and the force direction  $v$  and grasp central point  $c$ .

The goal of grasp configuration sampling is to generate an appropriate grasp configuration  $G = \{S_f, S_l, S_c, v, c\}$  given a specified grasp type  $t_g$ .

Formally, the grasp configuration sampling process is defined as:

$$G = f_{\text{config}}(t_g) \quad (1)$$

where  $G$  denotes the generated grasp configuration, and  $f_{\text{config}}$  is the grasp configuration sampling method conditioned on the specified grasp type.

2) *Functional Affordance-Aware and Grasp Type-Aware Differential Force-Closure Grasp Sampling*: Building on our proposed grasp configuration modeling, we further introduce a Differential Force Closure (DFC)-based grasp pose sampling method (Tax-DFCSampler) that is sensitive to both functional affordance and grasp type. This method is designed to generate initial grasp pose candidates, enhancing the semantic diversity and physical plausibility of the resulting grasp set. The sampling pipeline is detailed in Alg. 1.

The objective of Tax-DFCSampler is to, given a specific grasp configuration, incorporate object-level functional affordance information to generate a set of physically plausible contact points and estimate their corresponding initial grasp poses. The process can be formally defined as:

$$\{G_{\text{init}}, C_{\text{init}}\} = f_{\text{Tax-DFCSampler}}(M_{\text{obj}}, G) \quad (2)$$

where,  $G_{\text{init}}$  is the initial grasp pose.  $C_{\text{init}}$  denotes the corresponding set of contact points.  $M_{\text{obj}}$  is the object mesh annotated with an affordance map.  $G$  is the grasp configuration and  $f_{\text{Tax-DFCSampler}}$  is the grasp sampling function.

To enhance sensitivity to functional affordance, functional affordance contact point sampling method (AC-Sampler) first samples two primary contact points  $C$  from the affordance map of the object surface. These candidate contacts are then validated using a DFC estimator  $f_{\text{DFC}}$  [10] to ensure that they meet force-closure constraints.

$$f_{\text{DFC}} = \|Gc\|^2$$

$$G = \begin{bmatrix} I_3 & \cdots & I_3 \\ [\psi_1]_{\times} & \cdots & [\psi_n]_{\times} \end{bmatrix} \quad (3)$$

$$[\psi_k]_{\times} = \begin{bmatrix} 0 & -\psi_k^{(z)} & \psi_k^{(y)} \\ \psi_k^{(z)} & 0 & -\psi_k^{(x)} \\ -\psi_k^{(y)} & \psi_k^{(x)} & 0 \end{bmatrix}$$

where,  $\Psi = \{\psi_1, \dots, \psi_n\}$  denotes the set of contact point candidates. term  $c \in \mathbb{R}^{n \times 3}$  represents the normals of object surface at the contact points in  $\Psi$ , and  $n$  indicates the number of contact points. Specifically, an oversampling strategy is employed, allowing the DFC estimator to filter out physically infeasible candidates and retain a batch-sized set of valid contact points. The mid-point between  $C_m$  is selected as the grasp central point for further initial hand pose alignment, and random permutations are added to increase diversity in the resulting grasp poses.

Next, finger joint poses are initialized based on sampled the hand configuration according and the grasp type. By default, the initial pose is defined such that the index finger and thumb are positioned in opposition, enabling a neutral pre-grasp configuration commonly used for precision or pad opposition grasps. However, for grasp types that involve side contact regions, the initial hand open pose is specifically adapted to place the thumb in a side opposition configuration. Then, to generate contact points for all activated fingers, all fingertip contact points of all activated fingers are projected along the contact force direction onto the object surface to obtain a full set of candidate contact points. To prevent overfitting to a single structure, the non-activated fingers are randomly perturbed. Following this, we perform parallel hand-object collision checking using the kinematic skeleton of the dexterous hand. Specifically, the generated grasp candidates are split into multiple chunks for efficient parallel evaluation. For each pose, the number of collisions between hand links and the object is computed. Only those candidates with a collision count below a predefined threshold are retained as valid initial grasp poses  $G_{\text{init}}$ .

3) *Grasp Optimization and Simulation Validation*: After constructing the grasp configuration that incorporates functional affordance, grasp taxonomy, and contact sensitivity, and obtaining the initial grasp pose through DFC-based sampling, we further refine the candidate poses via an optimization procedure. During optimization, only the activated fingers and wrist pose are updated, while the non-activated fingers remain fixed to their original configuration. The overall objective function comprises the following three key components: Semantic hand-object interaction loss, Differential Force Closure (DFC) estimation loss and Penetration and joint-limit regularization.

The construction of the loss function proceeds as follows. For each activated finger, we randomly sample contact

---

**Algorithm 1** Functionality- and Grasp-Type-Aware Grasp Pose Sampling (AC-Sampler + Tax-DFCSampler)

---

**Require:**  $M_{\text{obj}}$ : Object mesh with functional affordance map  
 $G$ : Grasp configuration from taxonomy  
 $f_{\text{DFC}}$ : Differential Force Closure estimator  
 $N$ : Number of required grasp candidates  
 $\tau$ : Collision threshold

**Ensure:**  $G_{\text{init}}$ : Set of valid initial grasp poses

```

1:  $C_{\text{valid}} \leftarrow \emptyset$ 
2: while  $|C_{\text{valid}}| < N$  do
3:    $C_m \leftarrow \text{SamplePrimaryContacts}(M_{\text{obj}}, G)$ 
4:   if  $f_{\text{DFC}}(C_m)$  is valid then
5:      $C_{\text{valid}} \leftarrow C_{\text{valid}} \cup \{C_m\}$ 
6:   end if
7: end while
8:  $G_{\text{init}} \leftarrow \emptyset$ 
9: for each  $C$  in  $C_{\text{valid}}$  do
10:   $C'_m \leftarrow \text{RandomSwap}(C_m)$ 
11:   $d \leftarrow \text{GetForceDirection}(G)$ 
12:   $F_{\text{act}} \leftarrow \text{GetActivatedFingers}(G, d)$ 
13:   $p_{\text{center}} \leftarrow \text{EstimateGraspCenter}(C'_m, d)$ 
14:   $G_{\text{pose}} \leftarrow \text{ConstructHandPose}(p_{\text{center}}, d, G, \text{addNoise}=\text{True})$ 
15:  RandomizeIdleFingers( $G_{\text{pose}}$ )
16:   $n_{\text{col}} \leftarrow \text{CheckCollision}(G_{\text{pose}}, M_{\text{obj}})$ 
17:  if  $n_{\text{col}} < \tau$  then
18:     $G_{\text{init}} \leftarrow G_{\text{init}} \cup \{(G_{\text{pose}}, C'_m)\}$ 
19:  end if
20: end for
21: return  $G_{\text{init}}$ 
```

---

points from its designated contact region as specified by the grasp configuration. Corresponding target contact points are sampled from the object surface. The contact alignment is then optimized by computing the Signed Distance Function (SDF) between the hand mesh and object surface, minimizing interpenetration while encouraging semantically valid contact. The DFC estimator is employed as an additional optimization objective to promote high-quality, force-closure grasps. During the entire optimization process, joint angle updates are constrained by anatomical joint limits, ensuring biomechanically feasible and physically plausible hand postures.

After optimization, each refined grasp candidate is validated within a physics simulation environment. Specifically, we adopt a joint impedance control strategy to enable compliant yet stable dexterous grasping, simulating realistic contact forces and evaluating the stability of the grasp during execution. For each joint  $i$ , the control torque  $\tau_i$  is computed as:

$$\tau_i = k_i (q_i^{\text{target}} - q_i) - d_i \dot{q}_i + g_i \quad (4)$$

where,  $\tau_i$  denotes the torque applied to joint  $i$ ,  $q_i$  is the current joint position,  $q_i^{\text{target}}$  represents target joint position,  $\dot{q}_i$  denotes joint velocity,  $k_i$ ,  $d_i$  and  $g_i$  are the stiffness, damping coefficients, and external compensation terms, respectively. Only those grasp poses that successfully complete the simulated grasp—without slippage or failure—are retained in the final dataset.

### C. LMM-Grounded Dexterous Grasping Multi-Dimensional Semantics Understanding Method

To enrich the generated grasp dataset with multi-dimensional semantic annotations, we propose OmniDexReasoner, a multimodal large model (LMM)-based semantic reasoning framework for dexterous grasping, which is sensitive to functional affordance, contact semantics, and grasp taxonomy, as shown in Fig. 3. By integrating multi-agent collaboration, Chain-of-Thought (CoT) reasoning, and Retrieval-Augmented Generation (RAG) strategies, the proposed framework addresses key gaps in current research on dexterous hand semantic understanding. Our semantic reasoning pipeline consists of three key modules: hand-object interaction understanding model, grasp taxonomy reasoning model and dexterous grasp multi-dimensional semantic generation. This framework enhances the expressiveness of semantic annotations in dexterous grasp datasets and provides high-quality semantic priors to support downstream multimodal models.

1) *Hand–Object Interaction Understanding Model*: To understand the interactive information between a robotic hand and an object, we propose a model for hand-object interaction understanding. This model leverages multimodal information, including object attributes, robotic hand configurations, hand-object contact data, and grasping scene images, to automatically predict the semantic relationships between the hand and the object. This process can be modeled in a probabilistic framework as:

$$P(A|I) = P(A|H, O, C, V) \quad (5)$$

where  $A$  represents the semantic outputs of the model, and  $I$  denotes the multimodal input.  $H$  denotes information about the robotic hand, such as the types of available fingers, available finger links, joint configurations, and maximum graspable size.  $O$  represents object-related information, including the object's name, affordance annotations.  $C$  captures the contact information between the hand and the object, including the contact semantic map  $M_c$  and corresponding grasp affordance estimation result  $a^*$ .  $V$  refers to multi-view images of the grasping scene. Output  $A$  contains of final functional affordance estimation classification result  $a_{\text{sem}}$ , textual object descriptions, contact fingers and links, finger flexion configurations, palm contact status, and the contact semantic map.

To model the contact information, we compute the spatial distance between various links and fingers of the hand and surface points on the object, similar to [9]. This allows us to construct a Contact Semantic Map  $M_c$ , assigns semantic contact labels to the object surface points. The map is defined as:

$$M_c(p_i) = (l_j, f_k), \quad \text{if } d(p_i, L_{jk}) < \delta \quad (6)$$

where,  $p_i \in \mathcal{P}$  denotes a surface point sampled from the object.  $l_j$  denotes the  $j$ -th link.  $f_k$  denotes the  $k$ -th finger.  $L_{jk}$  is the 3D model of the  $j$ -th link on the  $k$ -th finger.  $d(p_i, L_{jk})$  is the minimum Euclidean distance between  $p_i$  and  $L_{jk}$ .  $\delta$  is a contact threshold. Each contact point on the object surface is thus labeled with the corresponding hand link and finger involved in the contact.

## Multi-Dimensional Functional Action, Contact and Grasp Taxonomy Semantic-Description Generation

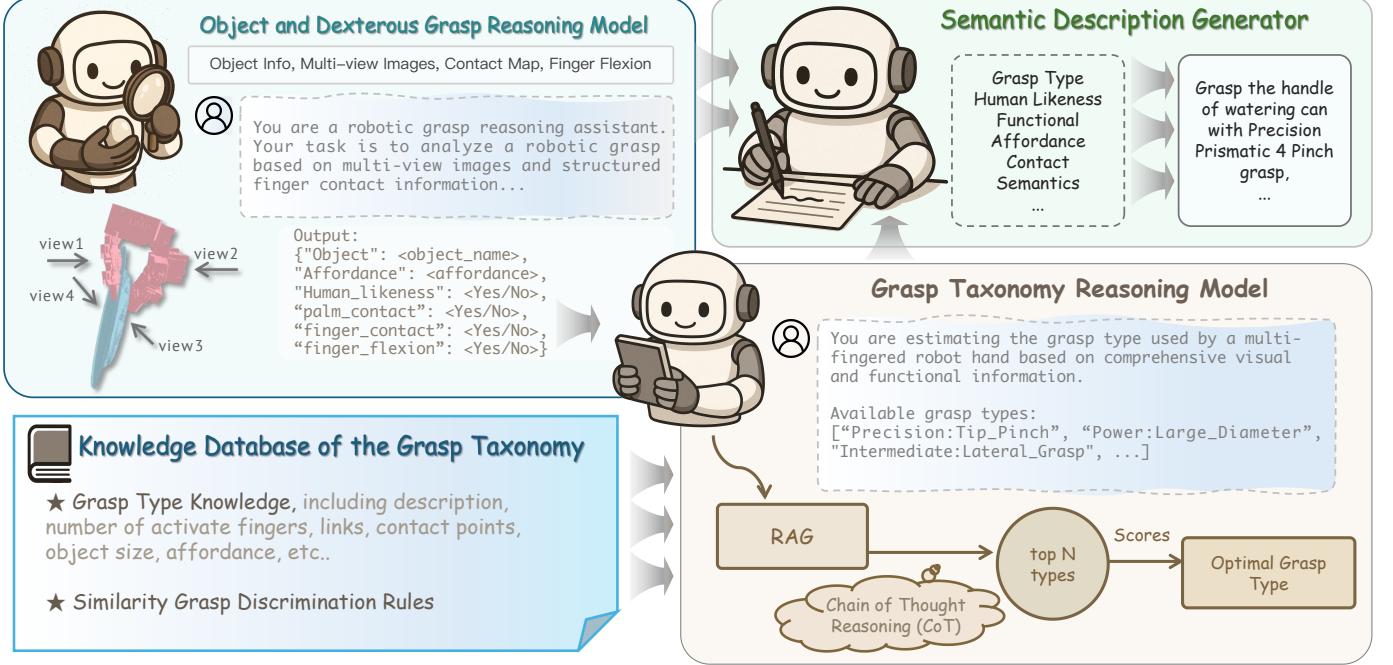


Fig. 3: OmniDexReasoner: LMM-Based dexterous grasping multi-dimensional semantic reasoning, including functional affordance, contact semantic information, finger configuration, grasp taxonomy and human-consistency.

Within the agent, the Contact Semantic Map is further analyzed to identify the involved fingers and links, and whether the palm is in contact with the object. Palm contact is a key indicator of grasp type, such as power grasp or precision grasp. However, the degree of force exerted by the palm requires further estimation using visual cues from the grasping scene.

To infer object affordances from hand-object interactions, we employ a voting-based functional grasp affordance classification method. This method integrates the Contact Semantic Map  $M_c$  with an affordance label map  $M_a$  defined over the object surface. For each contact point  $p_i$ , we count the associated affordance label and determine the most frequent one via voting:

$$a^* = \arg \max_{a \in \mathcal{A}} \sum_{p_i \in \mathcal{P}} \mathbb{I}[M_c(p_i) \neq \emptyset] \cdot \mathbb{I}[M_a(p_i) = a] \quad (7)$$

where,  $a^*$  is the predicted functional affordance.  $\mathcal{A}$  is the set of predefined affordance classes (e.g., HandleGrasp, WrapGrasp, Press, Pour, Cut, Stab, Pull, Push, Open, Twist, Hammer, Pry, Support, Lift, Lever, None).  $\mathbb{I}[\cdot]$  is the indicator function.  $M_c(p_i) \neq \emptyset$  implies point  $p_i$  is in contact with the hand.  $M_a(p_i)$  is the affordance label for point. This process selects the dominant affordance type based on the most frequent contact-affordance label co-occurrence.

In addition, we incorporate multi-view images of the grasping scene as part of the input, denoted as the set  $\{V\}$ . These images contain rich visual information about the physical contact between different parts of the hand and the object, as well as the affordance-relevant features of the interaction. They also reveal the pose of each finger on the robotic hand. The degree of finger flexion jointly reflects the object's size,

the contact force distribution, and the potential grasp strategy. In general, larger objects tend to require power grasps, while smaller or more intricate objects are more likely to involve precision grasps.

Based on all the aforementioned inputs, the agent is tasked with predicting the semantic attributes of the hand-object interaction, including contact affordances and finger joint configurations, evaluating human-consistency, and then validating the contact understanding derived from conventional methods. More implementation details are introduced in the supplementary materials.

**2) Grasp Taxonomy Understanding with Multimodal Reasoning:** The grasp type is influenced by a combination of complex factors, including hand-object contact information, force direction, finger joint configurations, and object affordances. However, current Large Multimodal Models (LMMs) still suffer from significant hallucination when interpreting physical interactions in the real world, making them unreliable for accurate grasp type prediction. To address this challenge, we propose a novel grasp taxonomy semantic understanding model aimed at reasoning about grasp types  $T$  from multimodal input signals  $I = \{H, O, C, A, V\}$ , where,  $H$ ,  $O$ ,  $C$ ,  $A$ , and  $V$  represent hand configuration, object properties, hand-object contact information, affordance cues, and multi-view scene images, respectively.

We formulate grasp taxonomy classification as a two-level hierarchical task from coarse-level classification to fine-level classification. Coarse-level classification predicts whether the grasp is Power, Precision, or Intermediate. Fine-level classification: Further classify into specific taxonomy subtypes (e.g., pinch, tripod, hook, etc.).

Grasp taxonomy classification is challenging because of high retrieval and reasoning complexities. The grasp taxonomy space is large and fine-grained, where many grasp types differ only subtly in visual or semantic attributes. This complexity increases the classification difficulty. Moreover, in different task scenarios, the same object may afford different grasp types depending on human preferences or task-specific goals. This context-dependent variability makes it difficult to classify grasps using physical information alone. Therefore, it is essential to incorporate both affordance information and semantic reasoning mechanisms to assist grasp type classification. Secondly, task-driven preferences often lead to different grasping strategies even under similar contact conditions, placing high demands on the model’s reasoning capabilities. Meanwhile, LMMs frequently generate hallucinated or inconsistent predictions when interpreting physical scenes. To enhance the robustness and reliability of grasp reasoning, external knowledge, semantic context, and multimodal cues must be integrated into the inference pipeline.

To address the challenges mentioned, we incorporate Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) into our grasp taxonomy reasoning framework to enhance the performance of grasp taxonomy reasoning.

The Retrieval-augmented mechanism is introduced to complement the LMM with an external structured knowledge database about the grasp taxonomy. The database is designed to retrieve relevant semantic and functional information prior to grasp classification. Specifically, the domain-specific knowledge bases consist of a grasp type database and rules about similar grasp discrimination. The grasp type database includes detailed descriptions of all defined grasp types and corresponding semantic knowledge about fingers, links, contacts, and affordance configurations, as well as the use-case examples. This improves the model’s ability to ground predictions in contextually relevant knowledge. Formally, the model becomes:

$$T = f_{\text{GTR}}(I, \text{Retrieve}(q)) \quad (8)$$

where,  $f_{\text{GTR}}$  is the multimodal model that integrates both inputs and retrievals for final Grasp Taxonomy Reasoning prediction.  $q$  is the query constructed from  $I$ .  $\text{Retrieve}(q)$  denotes the retrieved context from the knowledge database.

Chain-of-Thought (CoT) prompting is also incorporated to perform intermediate reasoning steps before arriving at a final grasp type prediction. This enhances interpretability and logical consistency in inference. The reasoning chain is structured as: grasp scene, affordance inference, contact type identification, grasp taxonomy prediction.

*3) Dexterous Grasp Multi-Dimensional Semantic Description Model:* To enhance the semantic interpretability of dexterous grasp behaviors and integrate multi-modal and multi-level semantic information, we propose a Dexterous Grasp Descriptor Generator. This module is designed to produce natural language descriptions that characterize grasp actions by jointly reasoning over key semantic components embedded in the grasping process.

The generator takes as input a combination of semantic signals derived from the grasping interaction, including the functional affordance of the object, the contact semantics

between the hand and the object, and the associated grasp taxonomy label. By modeling the interdependence between these factors, the generator is capable of composing diverse and contextually appropriate textual descriptions that reflect the nature of the grasp. For instance, “Two fingertips pinch the narrow handle, performing a precision grasp adapted for fine manipulation tasks.”

This descriptor generation module enhances the interpretability of the grasping process by establishing a semantic bridge between physical interaction and linguistic representation. As well, it enables the language-guided grasping pose generation network training, where a natural language interface is essential for aligning robotic actions with human intent.

#### D. Vision Language Dexterous Hand Grasping Pose Action Generation Model

To enable effective perception and response to multi-dimensional semantic cues in dexterous grasping tasks, we propose a language-conditioned Vision-Language Grasping Generation (VLG) model that generates semantically aligned dexterous grasp poses  $\mathbf{g} \in \mathbb{R}^d$  based on partial point cloud observations  $P \in \mathbb{R}^{N \times 3}$  of the target object and language instruction  $\mathcal{L}$ , as shown in Fig. 4. The model is capable of modeling and leveraging three key semantic dimensions: functional affordance, contact semantics, and grasp taxonomy, allowing it to produce grasp actions that are highly consistent with the task-specific semantic intent.

The proposed network comprises the key components: Multimodal Conditioning Embedding and Grasp Pose Diffusion Generator. The input language instruction and point cloud are embedded into a shared semantic space via a Vision-Language Model (VLM):

$$\mathbf{c} = f_{\text{VLM}}(\mathcal{L}, P) \quad (9)$$

where  $f_{\text{VLM}}$  denotes a pretrained transformer-based model (e.g., LLaMA-1B) with cross-attention between the language and visual tokens. The output  $c \in \mathbb{R}^{d_c}$  serves as a semantic condition for pose generation.

For aligning the vision-language model with point cloud encoder, the PointNet-based encoder, projector are trained following the training strategy from PointLLM [33], [34]. During first training stage, the projector is trained using the multimodal alignment dataset with point clouds, images and natural language annotations [33], [34]. The point cloud encoder extracts local geometric features  $f_{\text{pcd}}$  from partial object point clouds. The extracted features are passed through a projector that maps them into a shared multimodal embedding space to enable semantic alignment with language features from language tokenizer and encoder.

The multimodal feature fusion module captures complex cross-modal relationships between visual geometry and linguistic semantics, enabling the model to understand what to grasp, how to grasp.

Grasp pose diffusion transformer generator is introduced to generate semantically aligned grasp poses from semantic condition  $c$ . The grasp pose is represented as a latent variable  $z_\theta$ , generated by gradually denoising a Gaussian-initialized latent  $\mathbf{z}_T \sim \mathcal{N}(0, I)$  through a conditional denoising network:

### 3D Vision Language Grasp Pose Generation with Functional, Contact and Grasp Taxonomy Awareness

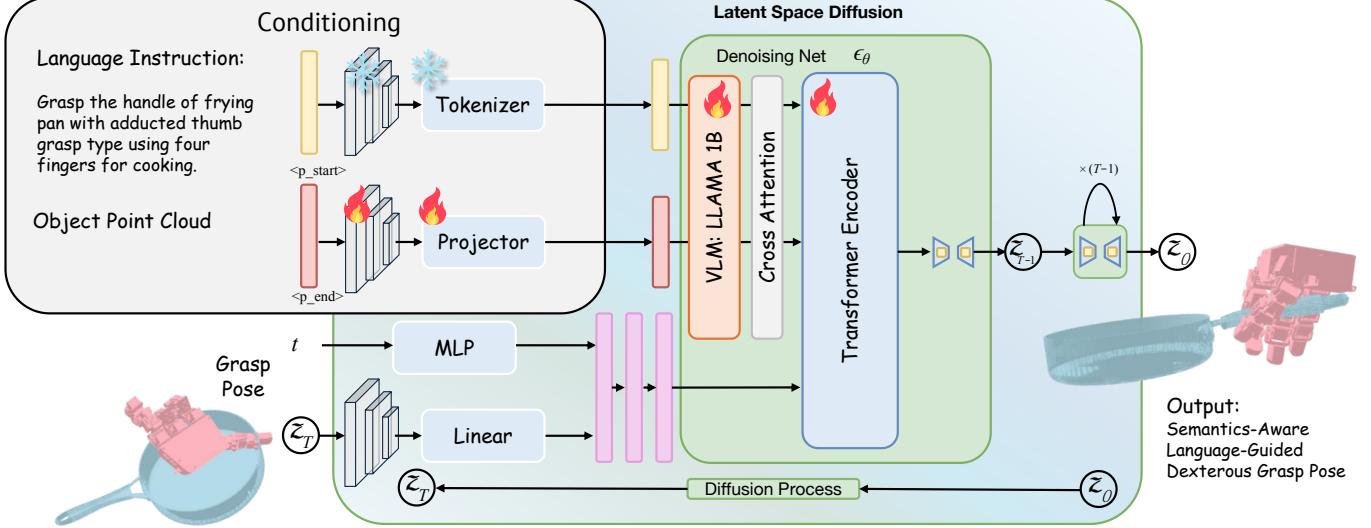


Fig. 4: OmniDexGraspNet: Vision Language Dexterous Hand Grasping Pose Generation Model. The textual task description and the partial point cloud of the object are encoded into their respective latent features using a text encoder and a point cloud encoder. Simultaneously, the initial grasp pose is mapped into the latent space via a multilayer perceptron (MLP). These multimodal latent features are then fused and processed by a diffusion transformer encoder, which is trained to predict and remove noise in the grasp pose representation. This denoising process refines the grasp configuration to produce a physically plausible and task-relevant hand pose following language instruction. The input text instructions include key semantic elements such as the grasp affordance location, the intended grasp type, contact configuration, and the high-level task objective.

$$\mathbf{z}_0 = \mathbf{z}_T - \sum_{t=1}^T \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t) \quad (10)$$

where  $\epsilon_\theta$  is the noise prediction network , conditioned on timestep  $t$  and semantic features  $\mathbf{c}$ .

The training pipeline consists of two main stages. After training projector for point cloud understanding, pose generator is trained with generated dexterous grasp dataset annotated with rich semantic labels based on proposed OmniDexDataGen and OmniDexReasoner. The OmniDexGraspNet is trained to generate high-quality grasp poses that are sensitive to multiple semantic conditions and adaptable to different language instructions.

## IV. EXPERIMENTS

### A. Experimental Setup

Real world experiment setup is shown in Fig. 5. We validate our proposed grasping approach on a real robotic platform composed of a Diana robotic arm and a LEAP Hand dexterous manipulator. During execution, the system first generates semantically-informed grasp poses using our method. Grasp trajectory planning is then performed via a constraint-based optimization algorithm, implemented using the PyRoki motion planning library [35]. Finally, the robot executes the planned grasp through impedance control, ensuring compliant and stable interaction during the grasping process.

To conduct controlled simulation experiments on grasp dataset generation, semantic reasoning, and action generation, we utilize both the LEAP hand and the DLR-HIT II hand



Fig. 5: Real world experimental setup with 7-axis Diana robotic arm, LEAP hand and RealSense D415 3D camera.

in Isaac Gym simulation environments. These settings allow us to evaluate the generalizability and performance of the proposed dataset generation method OmniDexDataGen and semantic reasoning method OmniDexReasoner across different hand hardware. The simulation experiments include the following components: Grasp Dataset Generation Evaluation,

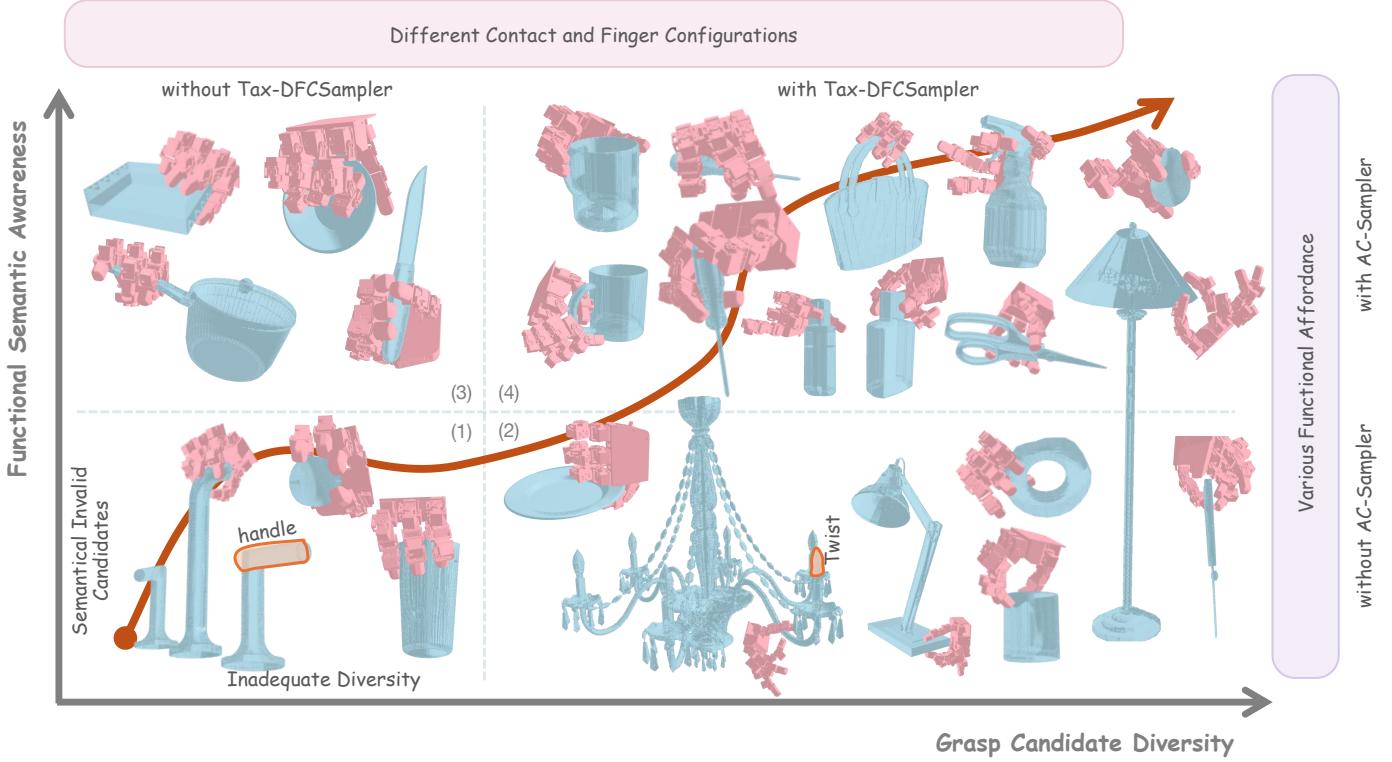


Fig. 6: Grasp generation results using state-of-the-art and different variants of our method. (1) DexGraspNet [1], (2) OGG + Tax-DFCSampler, (3) OGG + AC-Sampler, (4) Ours: OGG + AC-Sampler + Tax-DFCSampler.

Multi-Dimensional Semantic Understanding Module Evaluation, Semantic-Conditioned Grasp Action Generation Evaluation. OmniDexGraspNet is evaluated in simulation and real world using LEAP hand.

### B. Experiments of Grasp Dataset Generation Method

1) Qualitative Experiment: The qualitative evaluation of OmniDexDataGen includes comparisons against state-of-the-art grasp dataset generation methods [1], as well as an ablation study across several variant configurations of our approach. The evaluated baselines and ablations are summarized as follows:

- BASELINE 1: Optimization-based dexterous grasp generation (OGG), DexGraspNet [1], which employs global contact point sampling without any semantic awareness and guidance.
- BASELINE 2: OGG + Grasp taxonomy-aware grasp sampling (Tax-DFCSampler)
- BASELINE 3: OGG + Affordance-aware contact point sampling (AC-Sampler).
- Our OmniDexGen with both AC-Sampler and Tax-DFCSampler.

The qualitative results are summarized in Fig. 6 with leap hand and Fig. 7 with DLR-HIT II hand. Overall, our method demonstrates superior performance across multiple aspects of grasp generation, including controllability over functional affordance, diversity of contact configurations, variation in finger joint configurations, and grasp taxonomy coverage.



Fig. 7: Grasp generation results with DLR-HIT II hand.

In contrast, BASELINE 1 performs poorly across all these dimensions, lacking both semantic awareness and structural diversity. Lack of link- and contact region-aware contact sampling also leads to spatial discontinuity of sampled contact points and unstable pose optimization, especially when applied to large-scale objects, as shown in Fig. 6 (1).

BASELINE 2, while incorporating improvements over the original version, does not explicitly model functional affordance sampling. As a result, it often generates grasps that are semantically misaligned—for instance, grasping the main body of a refrigerator rather than its handle. Our approach with AC-Sampler supports adaptive grasp generation across objects of varying sizes. Representative samples generated by our approach are illustrated in the Fig. 6 (2).

BASELINE 3 removes the taxonomy-aware DFC sampler, as shown in Fig. 6 (3), leading to limited variation in contact regions and finger configurations. Consequently, the generated grasps are heavily biased towards multi-finger pinch types, lacking diversity across the grasp taxonomy spectrum.

TABLE I: Comparison experimental results of proposed functional affordance-aware and grasp type-aware dexterous grasp dataset generation method with state-of-the-art and different variants.

Model	Grasp Diversity				Contact Surface Coverage		Affordance Diversity
	KL $\downarrow$	STD <sub>translation</sub> $\uparrow$	STD <sub>orientation</sub> $\uparrow$	STD <sub>joint</sub> $\uparrow$	Average Hausdorff Distance (CM) $\downarrow$	KL $\downarrow$	
DexGraspNet [1]	0.259	7.834	0.713	0.921	36.634		0.288
Ours wo Affordance Contact Sampler	0.083	9.274	1.057	1.328	35.498		0.109
Ours wo grasp taxonomy-aware DFC sampler	0.209	9.987	1.182	1.053	26.173		0.214
Ours: OGG + AC-Sampler + Tax-DFCSampler	<b>0.047</b>	<b>11.709</b>	<b>1.413</b>	<b>1.623</b>	<b>25.518</b>		<b>0.055</b>

These qualitative observations highlight the importance of jointly modeling functional intent, contact semantics, and grasp type priors in achieving semantically grounded and physically diverse dexterous grasp generation, as shown in Fig. 6 (4).

2) *Evaluation Metrics for Quantitative Experiments:* To comprehensively evaluate the quality and semantic diversity of the generated grasp candidates, we adopt a set of metrics that quantify grasp pose variability, contact surface coverage, and functional affordance diversity. Specifically, we use Standard deviations (STDs) of hand position, orientation and joint angles and Kullback–Leibler (KL) divergence to assess pose-level diversity, average Hausdorff distance to measure diversity in contact semantics, and an affordance-level KL divergence to evaluate semantic diversity in grasp functionality.

Using KL divergence calculation, we estimate the diversity of generated categorical data—such as grasp types or functional affordances—by comparing the model’s output distribution to a uniform reference distribution, which represents the ideal diversity. Given a total of  $N$  categories, the ideal distribution is defined as:

$$P_{\text{ideal}} = \left[ \frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N} \right] \quad (11)$$

Let  $P_{\text{gen}} = [p_1, p_2, \dots, p_N]$  be the empirical distribution over categories produced by the model. The Kullback–Leibler (KL) divergence from the uniform distribution is computed as:

$$\text{KL}(P_{\text{gen}} \| P_{\text{ideal}}) = \sum_{i=1}^N p_i \log \left( \frac{p_i}{1/N} \right) \quad (12)$$

A lower KL divergence indicates that the generated distribution is closer to uniform, implying higher diversity across categories. This evaluation approach is applicable to both grasp type diversity and functional affordance distribution diversity.

Specifically, average hausdorff distance (AHD) is calculated using contact semantic maps of generated grasp candidates to quantify contact surface coverage. Given a set of  $N$  generated grasps in each object, we denote the contact semantic maps as point sets  $\{C_1, C_2, \dots, C_N\}$ . The pairwise average Hausdorff distance between each unique pair  $C_i, C_j$  is computed as:

$$\text{AHD}(C_i, C_j) = \frac{1}{|C_i|} \sum_{x \in C_i} \min_{y \in C_j} \|x - y\| + \frac{1}{|C_j|} \sum_{y \in C_j} \min_{x \in C_i} \|y - x\| \quad (13)$$

The final contact surface coverage score is defined as the average AHD across all unique grasp pairs:

$$\text{Score}_{\text{AHD}} = \frac{2}{N(N-1)} \sum_{i < j} \text{AHD}(C_i, C_j) \quad (14)$$

A higher AHD indicates a wider spatial distribution of contact points across the object surface, reflecting the model’s ability to explore a broader range of feasible grasping regions, rather than its sensitivity to functional semantics.

3) *Quantitative Experiments and Results:* In the quantitative experiments for grasp generation, we evaluate the generated grasp candidates across above key metrics. The quantitative results are summarized in Tab. I.

Quantitative results in Tab. I demonstrate that the proposed combination of the AC-Sampler and Tax-DFCSampler improves the diversity of the generated grasp dataset. Specifically, the full model achieves the best performance across all grasp diversity metrics, including translation, rotation, and articulation, as indicated by the highest standard deviations and the lowest KL divergence. Moreover, the average Hausdorff distance shows that this combination also leads to better coverage of contact semantic patterns. The KL divergence of affordance distributions is further reduced, indicating that the generated samples exhibit richer functional affordance variations compared to all baselines and ablations.

4) *Analysis and Discussion:* By introducing functional affordance-driven contact sampling, our method substantially enhances the semantic alignment between generated grasp poses and the intended functional regions of the object. Furthermore, the integration of grasp taxonomy-aware DFC sampling enables the generation of grasps across a wide range of grasp type, each characterized by distinct finger combinations, link involvement, and contact configurations.

Our method achieves superior performance in terms of grasp type diversity, functional affordance expressiveness, and contact semantic diversity.

5) *Limitation of OmniDexGen.:* Despite the capability to generate diverse grasp types, the visual and structural similarity between many grasp types poses challenges for straightforward classification. Existing off-the-shelf classifiers struggle to distinguish between subtle variations. Then, OmniDexReasoner is proposed to automatically infer the semantic category of generated grasps, enabling better annotation and downstream evaluation.

### C. Experiment of LMM-Based Dexterous Grasping Semantics Understanding Method

1) *Qualitative Experiments:* In the qualitative experiments, we demonstrate the capability of the proposed semantic reasoning module to interpret multiple grasp pose candidates across multiple semantic dimensions, as illustrated in Fig. 8. Inference results along three core semantic dimensions: contact semantics, grasp taxonomy, and functional affordance. The

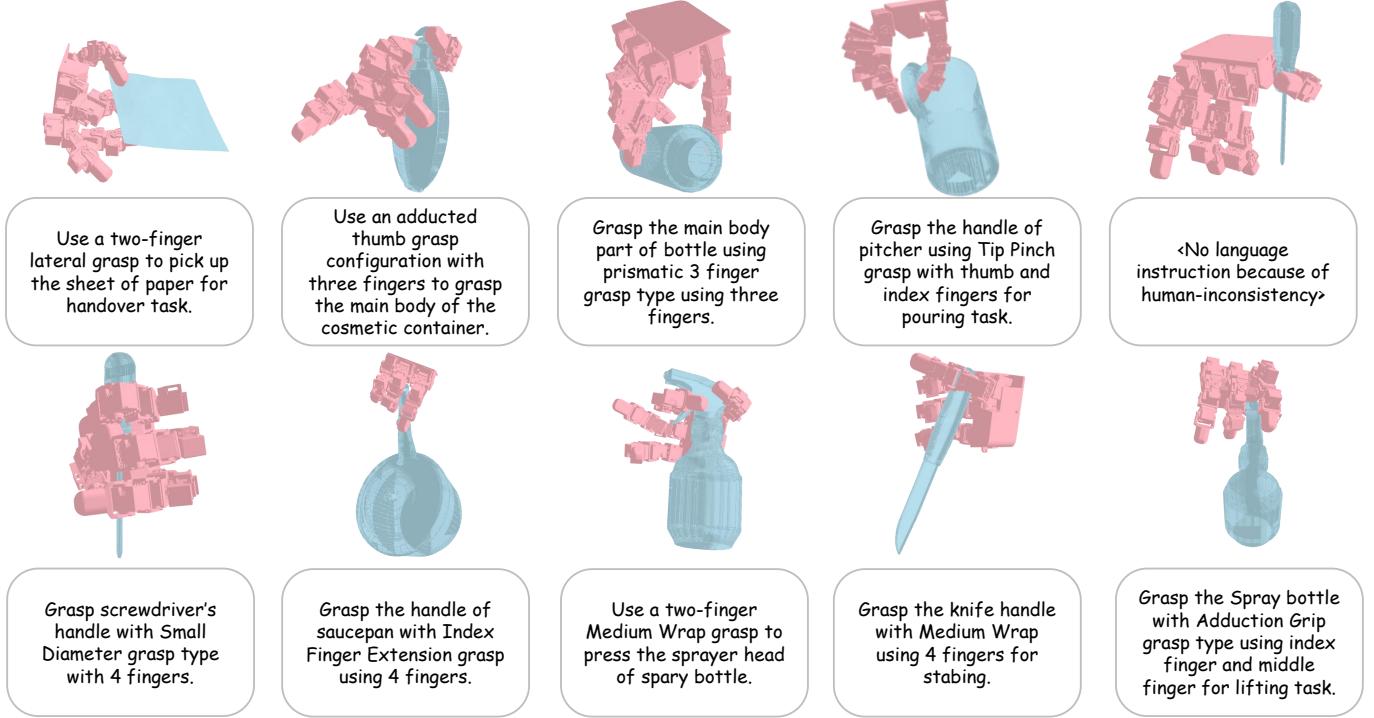


Fig. 8: Results of dexterous grasp semantic understanding model OmniDexReasoner.

TABLE II: Quantitative results of OmniDexReasoner compared with state-of-the-art baselines and its ablated variants.  $ACC_{coarse}$ : classification accuracy of coarse grasp types (power/intermediate/precision).  $ACC_{fine}$ : classification accuracy of fine grasp types.  $ACC_{human}$ : classification accuracy of human-like grasp candidates.  $ACC_{finger}$ : classification accuracy of finger contact status.  $ACC_{func}$ : classification accuracy of functional affordance.

Model	Grasp Taxonomy			Contact Semantic	Functional Affordance
	$ACC_{coarse} \uparrow$	$ACC_{fine} \uparrow$	$ACC_{human} \uparrow$	$ACC_{contact} \uparrow$	$ACC_{func} \uparrow$
<b>Effectiveness of Multi-Agent Collaboration</b>					
$\pi_{single}$	0.52	0.23	0.23	0.36	0.24
$\pi_{multi}$	<b>0.89</b>	<b>0.76</b>	<b>0.59</b>	<b>0.68</b>	<b>0.63</b>
<b>Ablation Study of Key Semantic Modalities.</b>					
$\pi_{multi}$ WO VISION	0.84	0.75	0.34	0.61	0.58
$\pi_{multi}$ WO FINGER INFO	0.88	0.51	0.58	0.69	0.61
$\pi_{multi}$ WO CONTACT INFO	0.47	0.45	0.59	0.23	0.66
$\pi_{multi}$ WO AFFORDANCE INFO	0.74	0.70	0.57	0.66	0.22
<b>Effectiveness of Chain-of-Thought and Retrieval-Augmented Generation.</b>					
$\pi_{multi}$ + CoT	0.95	0.80	0.65	0.75	0.63
$\pi_{multi}$ + RAG	0.90	0.77	0.60	0.71	0.67
$\pi_{multi}$ + CoT + RAG (Ours)	<b>0.97</b>	<b>0.83</b>	<b>0.68</b>	<b>0.77</b>	<b>0.70</b>
<b>Comparison experiments using different LMMs.</b>					
QWEN7B	0.73	0.33	0.49	0.39	0.49
QWEN72B (Ours)	<b>0.97</b>	<b>0.83</b>	<b>0.68</b>	<b>0.77</b>	<b>0.70</b>
GPT4O-MINI	0.95	0.85	0.73	0.79	0.74
GPT4O	<b>0.97</b>	<b>0.92</b>	<b>0.84</b>	<b>0.82</b>	<b>0.78</b>

visualization results show that our model can generate structurally coherent and semantically consistent grasp descriptions, exhibiting a strong ability to capture the interdependence among different semantic dimensions.

2) *Quantitative Experiments*: To quantitatively evaluate the reasoning accuracy, we construct a manually annotated benchmark dataset consisting of 100 grasp poses. Each pose is labeled with ground truth annotations with contact configuration, coarse and fine grasp types, and functional affordance.

This annotated set serves as the Ground Truth (GT) for evaluation. Based on this GT, we compute multidimensional classification accuracies, including coarse and fine grasp taxonomy classification accuracies ( $ACC_{coarse}$  and  $ACC_{fine}$ ), human-consistency accuracy  $ACC_{human}$ , contact semantics classification accuracy  $ACC_{contact}$ , and functional affordance classification accuracy  $ACC_{func}$ .

The experiment results of variants using different agent configurations, key semantic modalities, various LMMs are shown

in Tab. II. The experiments with different agent configurations and key semantic modalities are executed with Qwen 72B [36]. Comparison experiments using different LLMs are finished with optimal configuration using Qwen 7B, Qwen 72B [36], GPT4o-mini and GPT4o [37].

*3) Ablation Study Results and Analysis:* Results demonstrate that the multi-agent variant substantially outperforms its single-agent counterpart, achieving notable improvements in  $ACC_{coarse}$  (0.89),  $ACC_{contact}$  (0.68), and  $ACC_{func}$  (0.63). This highlights the advantage of collaborative reasoning in multi-agent settings for complex embodied tasks.

Ablation studies based on Qwen72B further reveal the critical role of multi-modal information. Removing visual inputs severely degrades  $ACC_{human}$ , while the absence of finger contact cues mainly impacts  $ACC_{fine}$ . Contact semantics are essential for  $ACC_{contact}$ , and affordance-related inputs are key to functional reasoning, as evidenced by the sharp drop in  $ACC_{func}$  when removed. These findings underscore the necessity of integrating diverse semantic modalities for robust performance.

Moreover, incorporating chain-of-thought (CoT) and retrieval-augmented generation (RAG) mechanisms leads to consistent improvements across all metrics. The combination of CoT and RAG yields the best results, suggesting that multi-step reasoning and external knowledge retrieval are complementary in enhancing fine-grained grasp understanding and affordance inference.

Finally, cross-model comparisons show that GPT-4o achieves state-of-the-art performance, outperforming Qwen72B and GPT-4o-mini across all evaluation dimensions. This confirms the scalability and generalization capability of the proposed reasoning framework when paired with advanced large language models.

#### D. Experiments of 3D Vision-Language Dexterous Grasp Pose Generation Network

*1) Simulation Comparison Experiments:* We conduct comparative grasping experiments to further evaluate the effectiveness of our grasp generation framework OmniDexGraspNet. In each trial, the model is provided with a partial point cloud of the target object along with a corresponding language instruction that specifies the intended grasp semantics. The estimated grasp poses are validated in the simulation environment by determining whether the object can be held stably. In addition to grasp success rate, we also record the resulting contact configuration, grasp taxonomy, and functional affordance of each successful grasp to assess the semantic consistency and diversity of the generated behaviors. Comparisons are made between our proposed method and existing state-of-the-art baselines with 100 grasping trials. The results are summarized in Tab. III.

As illustrated in the inference examples (Fig. 11), BERT-based GraspGPT [4] fails to identify appropriate grasp candidates within the designated regions when presented with complex semantic instructions.

*2) Real-World Experiments:* We further evaluate our OmniDexGraspNet in real world platform and compare with

TABLE III: Semantic information-guided grasping generation experiments results in simulation and real world. GSR: Grasp success rate,  $ACC_{FC}$ : Finger contact semantic accuracy,  $ACC_{GT}$ : Grasp taxonomy semantic accuracy,  $ACC_{FA}$ : Functional grasp affordance semantic accuracy.

Model	Simulation				Real-World			
	$GSR^\dagger$	$ACC_{FC}^\dagger$	$ACC_{GT}^\dagger$	$ACC_{FA}^\dagger$	$GSR^\dagger$	$ACC_{FC}^\dagger$	$ACC_{GT}^\dagger$	$ACC_{FA}^\dagger$
GRASP GPT [4]	0.72	—	—	0.67	0.56	—	—	0.34
Ours	<b>0.80</b>	<b>0.85</b>	<b>0.63</b>	<b>0.72</b>	<b>0.71</b>	<b>0.52</b>	<b>0.50</b>	<b>0.54</b>

TABLE IV: Grasp success rate across different finger configurations.

Method	2-Finger	3-Finger	4-Finger	Overall
Ours	64.98%	76.23%	78.83%	73.34%

state-of-the-art methods, including GraspGPT [4]. The grasp generation results and corresponding semantic distribution are shown in Fig. 9 and the quantitative results are summarized in Tab. III.

OmniDexGraspNet enables grasp generation guided by diverse semantic inputs, including grasp type, functional affordance, contact semantics, and finger configurations. However, in real-world experiments, we observe a degradation in contact accuracy and grasp taxonomy semantic accuracy. This performance drop is primarily attributed to the dynamic nature of contact during physical execution, where finger-object interactions often evolve over time. Consequently, such changes can alter the grasp type originally intended by the semantic prompt.

We also conduct experiments across different numbers of fingers and measured the corresponding grasp success rates, as shown in Tab. IV. The results show a clear trend: grasp success rate increases progressively with the number of fingers.

*3) Failure Examples:* Failure cases are primarily attributed to suboptimal force closure in the generated grasps, or to object slippage during execution. As illustrated in the Fig. 12, these failures often occur when the generated grasp lacks sufficient stability or fails to fully constrain the object, leading to unsuccessful grasp execution.

## V. CONCLUSION AND FUTURE WORKS

In this work, we addressed the limited semantic modeling capability in existing dexterous grasp synthesis methods and introduced OmniDexVLG, a multimodal semantic-guided framework that explicitly incorporates grasp taxonomy, contact semantics, and functional affordance into the grasp generation process. This enables the synthesis of more semantically consistent and structurally diverse dexterous grasp poses under natural language instructions.

To support fine-grained semantic supervision, we proposed OmniDexDataGen, a comprehensive grasp data generation pipeline that integrates taxonomy-driven topological configuration sampling, functional-affordance-aware contact point sampling, taxonomy-aware differential force closure grasp sampling, and physics-based optimization. This framework produces dexterous grasp samples that cover a broad range of semantic categories and contact structures.



Fig. 9: Grasp taxonomy, contact semantic and functional affordance distribution of dexterous grasp generation.

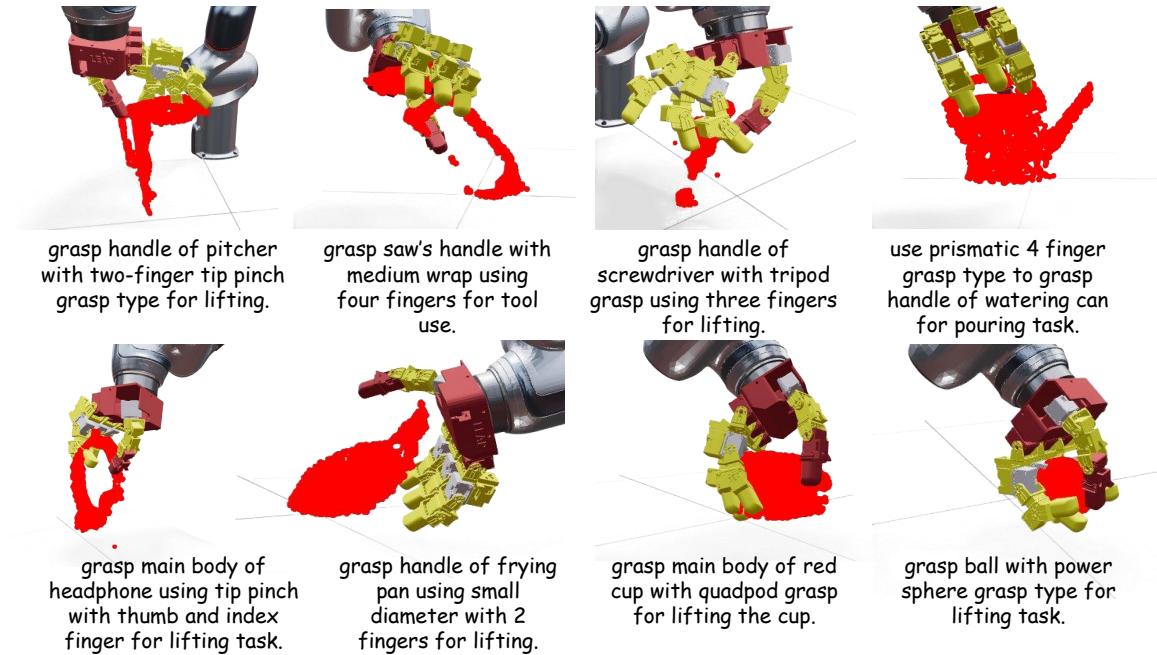


Fig. 10: Real-world inference results using proposed OmniDexGraspNet.

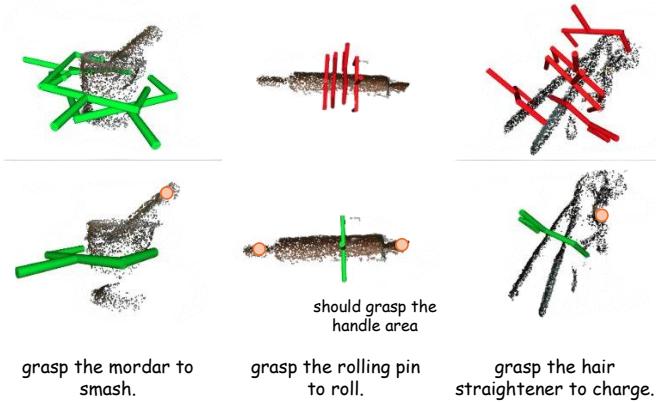


Fig. 11: Inference results of GraspGPT [4]. The inference grasp poses are aligned based on the dexterous hand finger configuration for grasp validation.



Fig. 12: Example of failures. (a) Invalid force closure. (b) Object slippage.

Furthermore, we developed OmniDexReasoner, a dexterous grasp semantic inference module that leverages large multimodal model reasoning together with RAG and CoT mechanisms to decode latent grasp intentions embedded in language and task context, enabling automatic and reliable semantic annotation.

We conducted extensive ablation studies to evaluate the contribution of each proposed component. The functional-affordance-aware contact sampler effectively prevents semantically invalid grasps—particularly on large-scale objects—and enhances both the validity and diversity of functional affordance distributions. The taxonomy-aware DFC sampler enriches grasp diversity across semantic dimensions by encouraging varied grasp types and contact configurations. Within OmniDexReasoner, RAG improves contextual grounding across multi-agent reasoning, while CoT enhances discrimination among semantically similar grasp types. Together, these mechanisms substantially improve grasp taxonomy inference accuracy and stability.

Comprehensive simulation and real-world experiments demonstrate that OmniDexVLG outperforms existing approaches in terms of grasp diversity, contact distribution quality, semantic consistency, and generalization across functional tasks, validating the effectiveness of the proposed multi-semantic modeling paradigm for dexterous grasp generation.

In future work, we plan to extend our framework to support task-chained language instructions and continuous manipulation trajectories, enabling more complex and semantically grounded robotic manipulation capabilities.

## REFERENCES

- [1] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, “Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation,” *arXiv preprint arXiv:2210.02697*, 2022.
- [2] A. Murali, B. Sundaralingam, Y.-W. Chao, J. Yamada, W. Yuan, M. Carlson, F. Ramos, S. Birchfield, D. Fox, and C. Eppner, “Graspgen: A diffusion-based framework for 6-dof grasping with on-generator training,” *arXiv preprint arXiv:2507.13097*, 2025. [Online]. Available: <https://arxiv.org/abs/2507.13097>
- [3] T. Feix, J. Romero, H.-B. Schmidtmayer, A. M. Dollar, and D. Kragic, “The grasp taxonomy of human grasp types,” *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

- [4] C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, “Graspgrpt: Leveraging semantic knowledge from a large language model for task-oriented grasping,” *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7551–7558, 2023.
- [5] C. Tang, D. Huang, W. Dong, R. Xu, and H. Zhang, “Foundationgrasp: Generalizable task-oriented grasping with foundation models,” *IEEE Transactions on Automation Science and Engineering*, 2025.
- [6] K. Li, J. Wang, L. Yang, C. Lu, and B. Dai, “Semgrasp: Semantic grasp generation via language aligned discretization,” in *European Conference on Computer Vision*. Springer, 2024, pp. 109–127.
- [7] L. Huang, H. Zhang, Z. Wu, S. Christen, and J. Song, “Fungrasp: functional grasping for diverse dexterous hands,” *IEEE Robotics and Automation Letters*, 2025.
- [8] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, “Dexterous functional grasping,” *arXiv preprint arXiv:2312.02975*, 2023.
- [9] L. Zhang, K. Bai, G. Huang, Z. Bing, Z. Chen, A. Knoll, and J. Zhang, “Contactdexnet: Multi-fingered robotic hand grasping in cluttered environments through hand-object contact semantic mapping,” *arXiv preprint arXiv:2404.08844*, 2024.
- [10] T. Liu, Z. Liu, Z. Jiao, Y. Zhu, and S.-C. Zhu, “Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator,” *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 470–477, 2021.
- [11] Y. Zhong, Q. Jiang, J. Yu, and Y. Ma, “Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22584–22594.
- [12] J. Chen, Y. Ke, L. Peng, and H. Wang, “Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy,” *arXiv preprint arXiv:2504.18829*, 2025.
- [13] Y. Song, P. Sun, P. Jin, Y. Ren, Y. Zheng, Z. Li, X. Chu, Y. Zhang, T. Li, and J. Gu, “Learning 6-dof fine-grained grasp detection based on part affordance grounding,” *IEEE Transactions on Automation Science and Engineering*, 2025.
- [14] Y. Zhang, J. Hang, T. Zhu, X. Lin, R. Wu, W. Peng, D. Tian, and Y. Sun, “Functionalgrasp: Learning functional grasp for robots via semantic hand-object representation,” *IEEE Robotics and Automation Letters*, 2023.
- [15] Z. Weng, H. Lu, D. Krägic, and J. Lundell, “Dexdiffuser: Generating dexterous grasps with diffusion models,” *IEEE Robotics and Automation Letters*, 2024.
- [16] R. M. Liu, M. Li, K. Shaw, and D. Pathak, “Ifg: Internet-scale guidance for functional grasping generation,” *arXiv preprint arXiv:2511.09558*, 2025.
- [17] Y. Wang, L. Zhang, Y. Tu, H. Zhang, K. Bai, Z. Chen, and J. Zhang, “Tooleenet: Tool affordance 6d pose estimation,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 10519–10526.
- [18] T. G. W. Lum, M. Matak, V. Makoviychuk, A. Handa, A. Allshire, T. Hermans, N. D. Ratliff, and K. Van Wyk, “Dextrah-g: Pixels-to-action dexterous arm-hand grasping with geometric fabrics,” *arXiv preprint arXiv:2407.02274*, 2024.
- [19] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, “General in-hand object rotation with vision and touch,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2549–2564.
- [20] H. Li, W. Mao, W. Deng, C. Meng, H. Fan, T. Wang, Y. Osumu, P. Tan, H. Wang, and X. Deng, “Multi-graspllm: A multimodal llm for multi-hand semantic guided grasp generation,” *arXiv preprint arXiv:2412.08468*, 2024.
- [21] S. Wang, Y. Yang, Y. Luo, D. Li, W. Wei, Y. Zhang, P. Hu, Y. Fu, H. Duan, J. Sun *et al.*, “Scaleadfg: Affordance-based dexterous functional grasping via scalable dataset,” *arXiv preprint arXiv:2511.09602*, 2025.
- [22] R. Wu, T. Zhu, X. Lin, and Y. Sun, “Cross-category functional grasp transfer,” *IEEE Robotics and Automation Letters*, 2024.
- [23] H.-S. Fang, H. Yan, Z. Tang, H. Fang, C. Wang, and C. Lu, “Anydex-grasp: General dexterous grasping for different hands with human-level learning efficiency,” *arXiv preprint arXiv:2502.16420*, 2025.
- [24] F. Zhao, D. Tsetserukou, and Q. Liu, “Graingrasp: Dexterous grasp generation with fine-grained contact guidance,” in *2024 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2024, pp. 6470–6476.
- [25] Y.-L. Wei, J.-J. Jiang, C. Xing, X.-T. Tan, X.-M. Wu, H. Li, M. Cutkosky, and W.-S. Zheng, “Grasp as you say: Language-guided dexterous grasp generation,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 46881–46907, 2024.
- [26] M. Cai, K. M. Kitani, and Y. Sato, “Understanding hand-object manipulation with grasp types and object attributes.” in *Robotics: science and systems*, vol. 3, 2016.
- [27] M. Ni, L. Zhang, Z. Chen, K. Bai, Z. Chen, J. Zhang, and W. Zuo, “Don’t let your robot be harmful: Responsible robotic manipulation via safety-as-policy,” *IEEE Robotics and Automation Letters*, 2025.
- [28] Y. Jin, D. Li, J. Shi, P. Hao, F. Sun, J. Zhang, B. Fang *et al.*, “Robotgpt: Robot manipulation learning from chatgpt,” *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2543–2550, 2024.
- [29] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An *et al.*, “Robobrain: A unified brain model for robotic manipulation from abstract to concrete,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1724–1734.
- [30] D. Li, Y. Jin, Y. Sun, Y. A. H. Yu, J. Shi, X. Hao, P. Hao, H. Liu, X. Li *et al.*, “What foundation models can bring for robot learning in manipulation: A survey,” *The International Journal of Robotics Research*, p. 02783649251390579, 2024.
- [31] Z. Song, G. Ouyang, M. Li, Y. Ji, C. Wang, Z. Xu, Z. Zhang, X. Zhang, Q. Jiang, Z. Chen *et al.*, “Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models,” *arXiv preprint arXiv:2505.16517*, 2025.
- [32] H. Duan, Y. Li, D. Li, W. Wei, Y. Huang, and P. Wang, “Learning realistic and reasonable grasps for anthropomorphic hand in cluttered scenes,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 1893–1899.
- [33] R. Xu, S. Yang, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, “Pointilm-v2: Empowering large language models to better understand point clouds,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [34] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, “Pointilm: Empowering large language models to understand point clouds,” in *European Conference on Computer Vision*. Springer, 2024, pp. 131–147.
- [35] C. M. Kim, B. Yi, H. Choi, Y. Ma, K. Goldberg, and A. Kanazawa, “Pyroki: A modular toolkit for robot kinematic optimization,” *arXiv preprint arXiv:2505.03728*, 2025.
- [36] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, “Qwen2. 5-vl technical report,” *arXiv preprint arXiv:2502.13923*, 2025.
- [37] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.