Task-Oriented Dexterous Grasping via Grasp Taxonomy

Hyunwoo Kim[1], Jeonghwan Choi[1], Dohyun Kim[2], Juntae Hwang[3], Ilsung Jin[4], and Donghan Kim[1*]

***Abstract*— Task-oriented grasping requires selecting grasp strategies that match the intended use of an object, but most existing approaches are limited to parallel grippers and do not fully exploit the contact diversity of dexterous hands. In this work, we propose a task-oriented dexterous grasp generation framework that bridges the gap between task intention and executable multi-finger grasps through a grasp taxonomy and contact representation. We first convert the dexterous hand grasp taxonomy into a consistent 2-point opposition representation using the Contact Web, enabling alignment with task-oriented parallel gripper grasp poses. We then predict a taxonomy-conditioned Contact Semantic Map from the object point cloud, grasp pose, and grasp taxonomy, and use it to guide transformer-based dexterous grasp generation. In real-world experiments on unseen household objects, our approach successfully translates diverse grasp taxonomies into stable and task-consistent dexterous executions. These results demonstrate that a grasp taxonomy is an effective intermediate representation for extending existing task-oriented grasping pipelines to dexterous hands.**

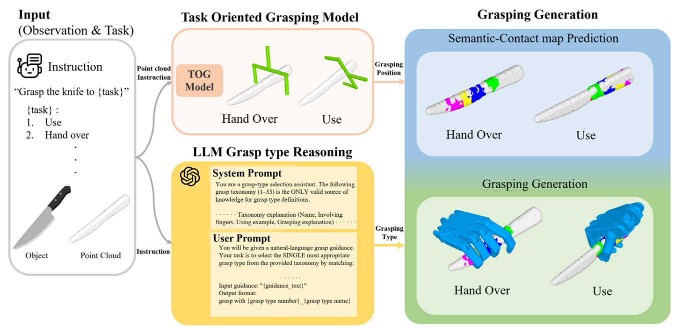

Fig. 1: Overview of the proposed pipeline. Given an instruction and an object point cloud, a task-oriented grasping (TOG) model predicts task-relevant parallel gripper grasp poses, while an LLM performs grasp type reasoning to select an appropriate grasp taxonomy. The semantic contact prediction module estimates a Contact Semantic Map conditioned on the grasp pose and the selected taxonomy. Finally, the grasp generation module synthesizes executable dexterous grasps from the predicted contact distributions.

I. INTRODUCTION

Humans naturally select different grasp strategies for the same object depending on the task. For example, a cup is grasped differently for moving than for pouring. A knife is grasped differently when used for cutting versus when handing it to another person. Such task-dependent grasp selection goes beyond stable object holding; it adjusts grasp position, orientation, and contact structure according to the intended manipulation.

This capability originates from the structure of the human hand, where multiple finger joints provide high degrees of freedom (DoF). This structure enables flexible control of grasp configurations according to the functional use of an object. In contrast, parallel grippers, widely used in robotic manipulation, have limited degrees of freedom and mainly support simple pick-and-place operations. Dexterous robot hands mimic the structure of the human hand to provide high degrees of freedom, enabling more complex object manipulation and interaction.

However, despite this potential, current robot hand research still lacks diverse object interaction strategies based on different grasp types. As a result, the capabilities of high-DoF robot hands remain underutilized.

*: Corresponding author. [1]Department of Electronic Engineering (AgeTech-Service Convergence Major), College of Electronic Information, Kyung Hee University, Yongin, Republic of Korea. [2]Department of Electronic Engineering, College of Electronic Information, Kyung Hee University, Yongin, Republic of Korea. [3]Department of Convergence Security, Kyung Hee University, Yongin 17104, Republic of Korea. [4]Department of Artificial Intelligence, Kyung Hee University, Yongin 17104, Republic of Korea. E-mail: hyunwoo.kim@khu.ac.kr, jjeonghi@khu.ac.kr, doran3887@naver.com, s6tae@khu.ac.kr, ilseong827@khu.ac.kr, donghani@khu.ac.kr

Recent large-scale learning approaches such as DexGraspNet [1] and UniDexGrasp [2], along with dataset construction efforts such as Dexonomy [3] that incorporate grasp taxonomy, enable diverse dexterous grasp generation and significantly improve the generalization of high-DoF grasp synthesis. Language-conditioned approaches also attempt to generate grasps that reflect user intent, suggesting the possibility of semantic alignment. However, these studies mainly focus on geometric constraints or conditional generation based on grasp types. They do not explicitly model how task context is structured through grasp taxonomy and translated into contact configurations.

To address this limitation, we propose a framework that extends task-oriented grasping [4]–[7] to dexterous hands. Our key idea is to use a grasp taxonomy as an intermediate representation between task intention and hand–object contact structure.

We first reformulate the dexterous grasp taxonomy [8] as a 2-point opposition structure using the Contact Web [9]. This representation establishes a correspondence with the parallel gripper grasp poses. Given a grasp pose and a grasp taxonomy, the model predicts a Contact Semantic Map [10] that represents finger contact distributions on the object surface. A transformer-based grasp generation module then produces executable dexterous grasp poses from the predicted contact information. We summarize our main contributions as follows:

- **Taxonomy-to-opposition grasp representation.** We introduce a Contact Web-based approach that translates the dexterous grasp taxonomy into a consistent 2-point opposition representation, enabling alignment between dexterous grasps and parallel gripper grasp poses.

- **Taxonomy-conditioned Contact Semantic Map.** We propose a semantic contact representation that predicts finger-level contact distributions conditioned on a grasp pose and grasp taxonomy, linking semantic grasp types with hand–object contact structure.
- **Task-oriented dexterous grasp generation framework.** We develop a unified pipeline that integrates grasp type reasoning and contact-aware grasp synthesis to generate executable dexterous grasps.

II. RELATED WORK

A. Task-Oriented Grasping

Recent studies have made significant progress in task-oriented grasping by inferring grasp poses that satisfy task requirements and human intentions [4]–[7]. However, these approaches mainly focus on parallel grippers, which limits their ability to represent complex human-like interactions. To overcome this limitation, we extend the parallel gripper-based pipeline to high-DoF dexterous hands. Our framework leverages reliable grasp poses predicted by parallel gripper models to constrain the search space of dexterous manipulation, while enabling task-aware multi-contact interactions.

B. Dexterous Grasp Generation

Recent research on dexterous grasp generation has evolved in diverse directions, largely driven by advances in model architectures. Methods such as UniDexGrasp [2], DexGraspNet [1], DexGraspNet 2.0 [11], and ContactDexNet [12], alongside CVAE-based approaches like GraspCVAE [13], have significantly improved geometric contact modeling for dexterous grasp synthesis. These methods leverage large-scale synthetic data and generative models to learn diverse grasp configurations.

Transformer-based approaches have also gained traction; for instance, Dexterous Grasp Transformer (DGTR) [14] and Grasp as You Say [10] directly generate grasp poses from explicit language instructions. Furthermore, large-scale datasets such as Dexonomy [3] enable diverse dexterous grasp learning by providing rich taxonomy-level annotations.

Despite these advances, existing methods mainly focus on object geometry or textual conditions and do not explicitly address how a comprehensive task context is structured into specific grasp types.

C. Grasp Taxonomy & Representation

Early studies categorized human grasp behaviors from anatomical and functional perspectives [15], [16]. Subsequent studies further analyzed grasp usage patterns in daily manipulation tasks and highlighted the relationship between grasp types and object interactions [17]. Insights from human grasping have also inspired the design and control of robotic hands [18]. The Feix taxonomy [8] introduced the concepts of the Virtual Finger and opposition groups for describing grasp types, while the Contact Web [9] models multi-finger contact structures. However, directly associating such multi-contact representations with the two-point action space of parallel grippers remains challenging. We address this gap by reformulating multi-finger contact structures into a consistent two-point opposition mechanism (OP_1, OP_2).

III. METHOD

A. Overview

The goal of this work is to generate a dexterous hand grasp pose $\mathcal{G}^{dex}$ that satisfies the task intention, given an object point cloud $\mathcal{O}$, a task-oriented parallel gripper grasp pose $\mathcal{G}^{parallel}$, and a grasp taxonomy τ inferred by an LLM. While task-oriented grasping methods typically represent grasp poses for parallel grippers, dexterous hands involve complex multi-finger contact structures, making direct extension from two-point parallel grasps difficult.

We address this challenge with a three-stage framework. First, we reformulate dexterous grasp contacts as a Contact Web-based two-point opposition representation to establish geometric correspondence with parallel grasps. Next, we predict a taxonomy-conditioned Contact Semantic Map that models finger-level contact distributions on the object surface. Finally, a generative model produces executable dexterous grasp poses from the predicted contact representation.

B. Transferring Dexterous Hand Taxonomy to Parallel Gripper Pose

Dexterous grasps involve complex multi-finger contact patterns that vary across grasp types. This variability makes it difficult to directly relate dexterous grasps to the two-contact structure of parallel grippers. To address this challenge, we introduce the **Contact Web** concept and reformulate multi-finger contact structures into a consistent **two-point opposition** (OP_1, OP_2) representation.

1) Extracting Contact Points and the Contact Web. Each finger may contact the object at multiple locations; we represent these contacts with a single representative point for each finger. For finger i, we compute the shortest distance between the distal link vertices and the object point cloud. Points within a threshold distance ϵ are treated as contact candidates. A distance-weighted average over the candidates yields the representative contact point $\mathbf{c}_i$. The set of all representative contacts forms the **Contact Web** $\mathcal{W} = \{\mathbf{c}_i\}_{i=1}^{n}$ [9].

2) From Multi-Contact to Two-Point Opposition. If exactly two contact points are present, they can be directly used as a two-point opposition structure. However, most dexterous grasps involve three or more contacts. In such cases, simply selecting two points may fail to preserve the structural meaning of the grasp. To address this issue, we leverage the **opposition concept in the Feix taxonomy** [8]. In this taxonomy, a grasp is defined as an opposition between two functional groups, referred to as Virtual Fingers (VF_1 and VF_2). Thus, a grasp can be interpreted as the interaction between a reference group (VF_1) and its opposing group (VF_2).

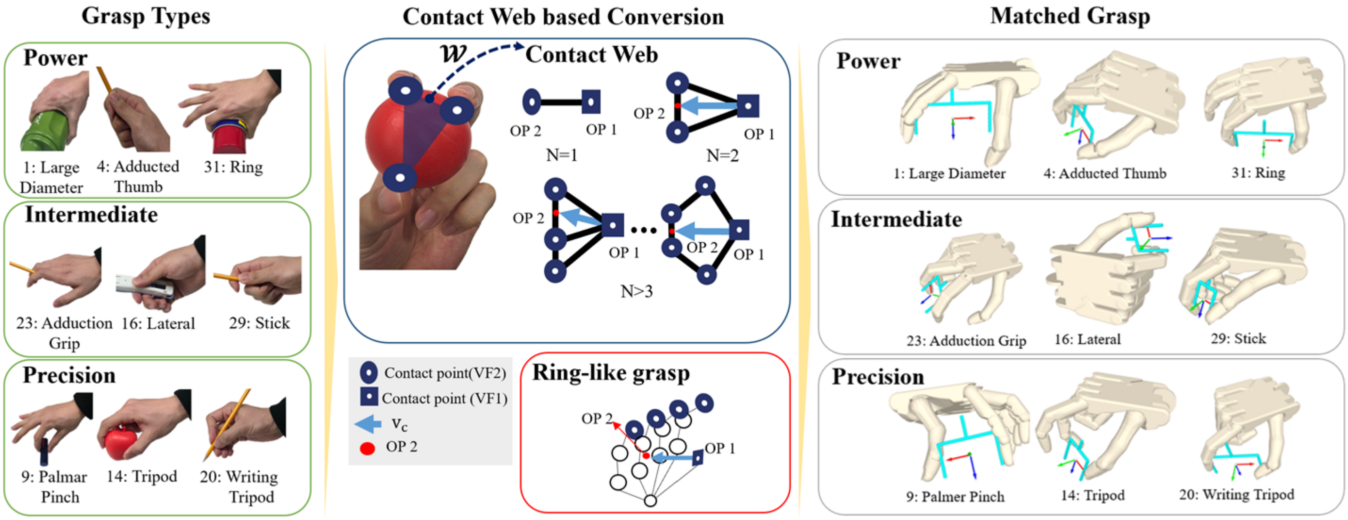

Fig. 2: Illustration of the proposed conversion from dexterous grasp taxonomy to a parallel gripper representation. **Left:** Representative grasp types from the Feix taxonomy grouped into **Power**, **Intermediate**, and **Precision** categories. **Middle:** Contact Web–based conversion that reduces multi-finger contacts into a **two-point opposition structure** (OP_1, OP_2). For typical grasps, OP_1 is selected from the VF1 group and OP_2 is determined from the contact web along the contact vector $\mathbf{v}_c$. For **ring-like grasps**, where palm contacts form a wrap configuration, OP_2 is defined as the centroid of finger contact points and OP_1 is selected from the opposing palm region. **Right:** Example dexterous grasp configurations aligned with the converted parallel gripper pose for each taxonomy.

First, we define the representative contact point of the fingers belonging to VF_1 in the Contact Web as the **Opposition Point 1** ($\mathrm{OP}_1 \in R^3$), which serves as the reference point of the grasp. Once OP_1 is determined, we define a contact vector $\mathbf{v}_c$ along the surface normal direction at OP_1. This vector represents the direction in which the grasp applies force to the object.

When the Contact Web contains three or more contact points, we construct a triangle fan $\mathcal{T}$ centered at OP_1 by connecting the remaining contact points:

$$\mathcal{T} = \bigcup_k \triangle(\mathrm{OP}_1, \mathbf{c}_k, \mathbf{c}_{k+1}). \tag{1}$$

A triangle fan connects multiple points to a shared center (OP_1), forming triangles that radiate outward. This structure approximates the local contact surface around OP_1. If exactly three contact points exist, a single triangle is sufficient. When four or more contacts are present, the fan-triangulated surface provides a consistent local geometry around OP_1.

We then project the contact vector $\mathbf{v}_c$ onto this triangulated surface and search for the intersection point starting from OP_1:

$$\mathrm{OP}_2 = \mathrm{Intersection}(\mathrm{OP}_1 + \lambda \mathbf{v}_c, \mathcal{T}), \quad \lambda > 0. \tag{2}$$

The intersection occurs on the edges of the triangles forming the fan rather than inside the triangular faces. As a result, OP_2 lies on one of the line segments of the Contact Web.

3) Ring-Like Grasp Opposition. Wrap grasps of the power grasp–palm opposition type form a ring-like structure where the palm and multiple fingers surround the object. Since VF_1 corresponds to the palm, the palm–object contact is widely distributed and cannot be represented by a single contact point. Therefore, the Contact Web-based opposition definition is not applied.

Instead, we define OP_2 as the average of the distal contact points of the fingers assigned to the VF_2 group. The corresponding representative point on the palm side is then defined as OP_1.

4) Approach Axis Definition. The approach direction is defined according to the orientation properties of the Feix grasp taxonomy. For intermediate grasps, the wrist direction is used as the approach axis. For other grasp types, the palm normal direction is used as the approach axis.

The closing axis is defined as the direction from OP_1 to OP_2. Together with the approach axis, this direction defines an orthogonal coordinate frame for the parallel gripper grasp.

The position and orientation of the parallel gripper grasp are defined as $\mathbf{p}^{parallel}$ and $\mathbf{R}^{parallel}$, respectively:

$$\mathbf{p}^{parallel} = \frac{\mathrm{OP}_1 + \mathrm{OP}_2}{2}. \tag{3}$$

The orientation matrix is constructed from the approach axis $\mathbf{a}_{app}$ and the closing axis $\mathbf{a}_{close}$. The lateral axis $\mathbf{a}_{lat}$ is obtained by their cross product:

$$\mathbf{R}^{parallel} = [\mathbf{a}_{lat}\ \mathbf{a}_{close}\ \mathbf{a}_{app}], \quad \mathbf{a}_{lat} = \mathbf{a}_{app} \times \mathbf{a}_{close}. \tag{4}$$

Finally, the parallel gripper grasp pose is expressed as:

$$\mathcal{G}^{parallel} = (\mathbf{p}^{parallel}, \mathbf{R}^{parallel}). \tag{5}$$

Through this formulation, the multi-contact structure of a dexterous grasp is reformulated into a geometrically consistent two-point opposition representation, enabling direct correspondence with parallel gripper grasps. This parallel

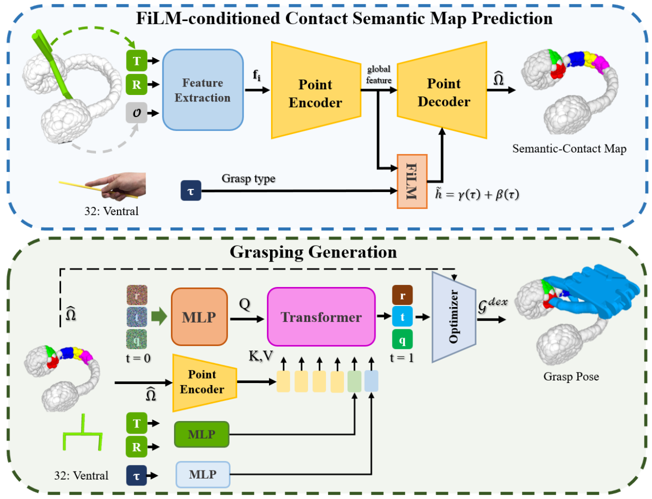

Fig. 3: Overview of the proposed Contact Semantic-based dexterous grasp generation framework. The pipeline consists of two stages: **(top) FiLM-conditioned Contact Semantic Map Prediction**, which estimates contact distributions based on the object point cloud, grasp pose, and grasp taxonomy; and **(bottom) Grasp Generation**, which utilizes a transformer-based module for dexterous grasp synthesis followed by test-time optimization.

grasp pose provides the geometric condition for predicting the Contact Semantic Map in the next stage.

C. FiLM-Conditioned Contact Semantic Map Prediction

We propose a **FiLM-conditioned Contact Semantic Map Prediction Model** that predicts the contact distribution of a dexterous hand given an object point cloud, a grasp pose, and a grasp taxonomy label. The Contact Semantic Map represents finger-level contact distributions on the object surface by assigning a semantic contact label to each point in the object point cloud.

Feature Representation. The contact structure is determined by the relative geometry between the hand and the object rather than the absolute position of the object in the world frame. Therefore, we transform the object point cloud into a **hand-centric coordinate frame** defined by the grasp pose.

Given the grasp center $\mathbf{p}$ and rotation matrix $\mathbf{R}$, each point $\mathbf{x}_i$ and its surface normal $\mathbf{n}_i$ are transformed into the hand frame as:

$$\tilde{\mathbf{x}}_i = \mathbf{R}^\top(\mathbf{x}_i - \mathbf{p}), \qquad \tilde{\mathbf{n}}_i = \mathbf{R}^\top \mathbf{n}_i. \tag{6}$$

To encode the geometric relationship between the object surface and the grasp direction, we define **alignment features** for each point. We first compute the **grasp-center distance** and then measure the alignment between the point position and the grasp approach direction, as well as the alignment between the surface normal and the approach direction:

$$d_i = \|\tilde{\mathbf{x}}_i\|, \qquad a_i^{\text{pos}} = \frac{\tilde{\mathbf{x}}_i}{\|\tilde{\mathbf{x}}_i\|} \cdot \hat{\mathbf{d}}, \qquad a_i^{\text{surf}} = \tilde{\mathbf{n}}_i \cdot \hat{\mathbf{d}}. \tag{7}$$

Here, $\hat{\mathbf{d}}$ denotes the unit vector of the grasp approach direction. Finally, each point is represented by the following hand-centric geometric feature:

$$\mathbf{f}_i = \left[\tilde{\mathbf{x}}_i,\ \tilde{\mathbf{n}}_i,\ d_i,\ a_i^{\text{pos}},\ a_i^{\text{surf}}\right]. \tag{8}$$

Here, $\tilde{\mathbf{x}}_i$ denotes the hand-centric coordinates, $\tilde{\mathbf{n}}_i$ represents the surface normal expressed in the hand frame, and d_i is the grasp-center distance. The terms a_i^{pos} and a_i^{surf} represent the position–approach alignment and surface–approach alignment, respectively. These features capture both the local surface geometry and the spatial relationship between the grasp and the object. The resulting point features are used as input to a PointNet++ encoder [19].

Taxonomy-Conditioned FiLM Generator. Grasp taxonomy determines the functional grasp type and influences the resulting finger contact patterns. Even with the same object geometry and grasp pose, the contact distribution can vary depending on the selected taxonomy. To incorporate this information, we apply FiLM conditioning [20] to modulate intermediate network features.

The taxonomy label τ is first converted into an embedding vector. From this embedding, the network generates modulation parameters $\boldsymbol{\gamma}(\tau)$ and $\boldsymbol{\beta}(\tau)$ for each layer feature $\mathbf{h}$. The modulated feature $\tilde{\mathbf{h}}$ is computed as:

$$\tilde{\mathbf{h}} = \boldsymbol{\gamma}(\tau) \odot \mathbf{h} + \boldsymbol{\beta}(\tau), \tag{9}$$

where $\boldsymbol{\gamma}(\tau)$ and $\boldsymbol{\beta}(\tau)$ control the scaling and shifting of the feature representation. This conditioning allows the network to predict Contact Semantic Maps that reflect both the grasp pose and the grasp taxonomy.

D. Dexterous Grasp Pose Generation

We represent a dexterous grasp pose as $\mathcal{G}^{dex} = (\mathbf{t}, \mathbf{r}, \mathbf{q})$, where $\mathbf{t} \in R^3$ denotes the grasp position, $\mathbf{q} \in R^{22}$ represents the joint angles of the hand, and $\mathbf{r} \in S^3$ denotes the wrist rotation expressed as a unit quaternion.

Since the rotation component lies on the unit sphere S^3 rather than in Euclidean space, the model predicts a conditional vector field $v_\theta(\mathcal{G}_t^{dex}, t \mid c)$ from an intermediate grasp state $\mathcal{G}_t^{dex}$. After each update, the rotation component is normalized to satisfy $\|\mathbf{r}\|_2 = 1$.

The condition $c = (\hat{\Omega}, \tau, \mathcal{G}^{parallel})$ comprises the contact semantic map fused with the object point cloud, the grasp taxonomy label, and the parallel gripper grasp pose.

During training, the model predicts the final grasp pose $\hat{\mathcal{G}}_1^{dex}$ from an intermediate grasp state $\mathcal{G}_t^{dex}$ at $t \in [0, 1)$ using a one-step update:

$$\hat{\mathcal{G}}_1^{dex} = \mathcal{G}_t^{dex} + (1-t)\, v_\theta\big(\mathcal{G}_t^{dex}, t \mid c\big), \quad t \in [0, 1). \tag{10}$$

The training objective consists of joint parameter alignment, fingertip position alignment, and Chamfer distance alignment:

$$\mathcal{L}_{\text{grasp}} = \lambda_{\text{para}}\mathcal{L}_{\text{para}} + \lambda_{\text{tip}}\mathcal{L}_{\text{tip}} + \lambda_{\text{chamfer}}\mathcal{L}_{\text{chamfer}}. \tag{11}$$

The detailed formulation of each loss term follows existing dexterous grasp optimization frameworks.

During inference, the predicted grasp $\hat{\mathcal{G}}_1^{dex}$ reflects the object geometry and task intention, but small penetrations or imperfect contacts may remain. To refine the predicted grasp, we apply test-time adaptation (TTA) and directly optimize the normalized grasp parameters:

$$\mathcal{L}_{\text{TTA}} = \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{cmap}}\mathcal{L}_{\text{cmap}} + \lambda_{\text{tip}}\mathcal{L}_{\text{tip}} + \lambda_{\text{aff}}\mathcal{L}_{\text{aff}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}}. \tag{12}$$

We follow the loss formulation of prior optimization-based grasping methods [10], [21]. Since our method uses a finger-level contact semantic map instead of a single hand-level contact map, the contact map loss and the affordance distance loss are reformulated to incorporate semantic label consistency.

Contact Map Loss ($\mathcal{L}_{\text{cmap}}$). Unlike prior works that define a single hand-level contact map, we use a finger-level contact semantic map. Let $\mathcal{O} = \{\mathbf{o}_j\}$ denote the object point cloud and let $s_j \in \{1, \ldots, K\}$ be the semantic contact label of point $\mathbf{o}_j$, where K is the number of fingers. The object contact region assigned to finger k is defined as:

$$\mathcal{O}_k = \{\, \mathbf{o}_j \in \mathcal{O} \mid s_j = k \,\}. \tag{13}$$

Let $\mathcal{H}_k$ denote the surface point set of the k-th finger. The contact map loss minimizes the distance between each finger surface and its corresponding semantic contact region:

$$\mathcal{L}_{\text{cmap}} = \sum_{k=1}^{K} \frac{1}{|\mathcal{O}_k|} \sum_{\mathbf{o}\in\mathcal{O}_k} \min_{\mathbf{h}\in\mathcal{H}_k} \|\mathbf{o} - \mathbf{h}\|_2^2. \tag{14}$$

Affordance Contact Loss ($\mathcal{L}_{\text{aff}}$). The affordance contact loss constrains hand contact candidates to lie close to the task-relevant object regions. Unlike prior works that consider a single affordance area, our formulation conditions the distance constraint on the contact semantic label.

For a hand contact candidate $\mathbf{p}_i$ belonging to finger k, the distance to the corresponding semantic affordance region is defined as:

$$d_k(\mathbf{p}_i) = \min_{\mathbf{o}\in\{\mathbf{o}_j\in\mathcal{O}_k | a_j > \tau_2\}} \|\mathbf{p}_i - \mathbf{o}\|_2. \tag{15}$$

The affordance contact loss is defined as:

$$\mathcal{L}_{\text{aff}} = \sum_i \mathbf{I}\big((d_k(\mathbf{p}_i^c) < \tau_1) \vee (d_k(\mathbf{p}_i^r) < \tau_1)\big) \cdot d_k(\mathbf{p}_i^r), \tag{16}$$

where $\mathbf{p}_i^c$ and $\mathbf{p}_i^r$ denote the contact candidate of the initial coarse hand and the refined hand, respectively.

E. LLM-based Grasp Type Selection

We use a large language model (LLM) to select an appropriate grasp taxonomy from natural language task instructions. The 33 grasp types defined in the Feix taxonomy are provided as part of the system prompt, and the model is constrained to select exactly one grasp type from this set.

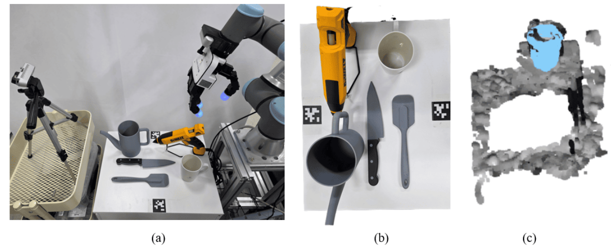

Fig. 4: Experimental setup and data acquisition process. (a) Real-world experiment platform with a UR3 manipulator, Allegro Hand V5, and an RGB-D camera. (b) Unseen household objects used in the experiments: glue gun, mug, pitcher, knife, and spatula. (c) Scene reconstruction using multi-view depth observations. The segmented target object is highlighted in sky blue.

The input grasp guidance contains cues about object geometry, task intention, and contact requirements. The LLM compares these cues with the taxonomy definitions and selects the most appropriate grasp type. The selected taxonomy label is then used as a conditioning input for the subsequent dexterous grasp generation process.

IV. Experiment

A. Experimental Setup

We evaluate the proposed framework through both geometric analysis and real-world robot experiments. The experimental platform consists of a UR3 manipulator, an Allegro Hand V5, and an Intel RealSense D435i depth camera.

The 3D object shapes are reconstructed from multi-view depth observations. Object segmentation is performed using Grounding DINO [22] and SAM [23], followed by ICP-based point cloud registration to obtain the final object point cloud. The experimental setup and scene construction process are illustrated in Fig. 4.

For the real-world experiments, we select five unseen household objects and define two task-oriented manipulation scenarios for each. Parallel gripper grasp poses are generated using Contact-GraspNet [24] and GraspGPT [4]. The top two poses, ranked by grasp quality score , are utilized in our evaluation. The task-relevant grasp taxonomy and its functionally opposite taxonomy are extracted using GPT-4o.

The training data used in this study is derived from the Dexonomy dataset [3]. This dataset is used to train both the Contact Semantic Map prediction model and the dexterous grasp generation model. It is also employed in the geometric analysis of the taxonomy-to-parallel conversion, where we measure the angular and positional deviations between the original dexterous grasps and the converted parallel grasps.

B. Dexterous Hand to Parallel Conversion Analysis

Fig. 4 illustrates the distribution of taxonomy-wise conversion errors using box plots. Most precision and pinch-type grasps (e.g., Tip Pinch, Precision Sphere, and Inferior Pincer) exhibit a median angular deviation of approximately 3°–6° and a center deviation of about 0.5–1.0 cm. This indicates that grasp types with clear multi-finger opposition structures

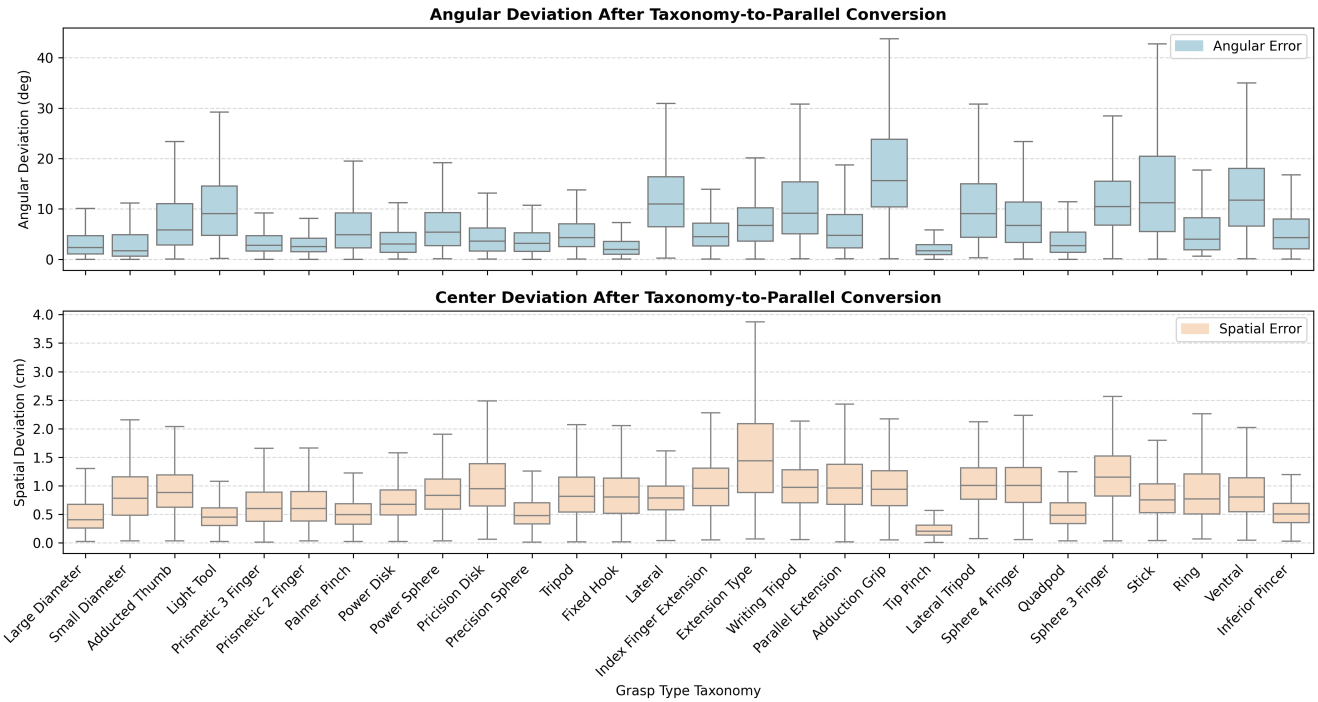

Fig. 5: Angular deviation (top) and **Center deviation (bottom)** after taxonomy-to-parallel conversion across grasp types. Each box plot shows the geometric deviation between the converted parallel gripper pose and the original dexterous grasp. Angular deviation measures the difference between the estimated closing axis and the original opposition direction, while center deviation measures the Euclidean distance between the two grasp centers.

can be reliably reformulated into the proposed two-point opposition representation.

In contrast, grasp types such as Adduction Grip, Stick, and Ventral show relatively larger angular deviations and higher variance. This variability likely arises because these grasps depend strongly on the geometric scale of the object, such as its diameter or length. When the object size changes, the relative configuration of contact points varies significantly, increasing the estimation error of the opposition axis.

Despite this variation, the center deviation remains within 1.5 cm for most grasp taxonomies, suggesting that the proposed conversion method preserves the contact center in a geometrically stable manner.

C. Ablation Study on Contact Semantic Map Prediction

We quantitatively evaluate how accurately the proposed contact semantic map prediction model reconstructs finger-level contact structures under different grasp conditions. The full model achieves an mIoU of 0.462, an mF1-score of 0.631, and a recall of 0.916.

Removing the grasp-center distance feature reduces the recall to 0.788. Removing the alignment features causes a sharp drop in mIoU to 0.080. Furthermore, removing hand-centric canonicalization leads to a severe performance collapse, reducing mIoU to 0.036.

These results indicate that grasp-centric coordinate normalization and direction-aware geometric encoding play a critical role in learning grasp-dependent contact topology.

TABLE I: Ablation study on semantic finger contact prediction. We remove key geometric components in Sec. III-C. Metrics: mIoU, mPrecision, mRecall, and mF1.

Method	mIoU	mPrecision	mRecall	mF1
Full model	**0.462**	0.482	**0.916**	**0.631**
w/o Distance Feature	0.435	**0.495**	0.788	0.605
w/o Alignment Feature	0.080	0.125	0.249	0.147
w/o Hand-centric Rep.	0.036	0.067	0.183	0.069

D. Real World Experiment

To evaluate the proposed system in real-world settings, we conduct task-oriented grasping experiments on five unseen household objects. For each object, two manipulation tasks are defined. For each task, we use the top two parallel grasp poses (Top-1 and Top-2) generated by GraspGPT.

To analyze the role of grasp taxonomy, we additionally evaluate a condition where a functionally opposite taxonomy is applied to the same grasp pose. Each condition is repeated 15 times and evaluated as **Success (Succ)**, **Slip (Slip)**, or **Failure (Fail)**.

Success indicates that the robot successfully grasped and lifted the object while maintaining a stable grasp for more than 5 seconds. **Slip** indicates that the object was initially lifted but slipped or the grasp configuration changed during lifting. **Failure** indicates that grasp pose generation failed or that no valid contact was formed during execution.

All results are reported as percentages averaged over repeated trials for each object–task pair.

(1) Influence of Grasp Pose Quality: Using the Top-

TABLE II: Real-world grasping results under different grasp pose ranking and taxonomy conditions.

Object	Task	Success Rate (%)								
		Top-1 Pose			Top-2 Pose			Top-1 + WrongTax		
		Succ	Slip	Fail	Suc	Slip	Fail	Suc	Slip	Fail
Knife	Hand over	**66.7**	20.0	13.3	**40.0**	33.3	26.7	**53.3**	13.3	33.3
	Cut	**66.7**	0.0	33.3	26.7	6.7	**66.7**	**40.0**	20.0	**40.0**
Glue gun	Use	**86.7**	0.0	13.3	**66.7**	26.7	6.7	20.0	13.3	**66.7**
	Pick	20.0	**40.0**	**40.0**	33.3	**53.3**	13.3	26.7	33.3	**40.0**
Mug	Drink	**93.3**	0.0	6.7	26.7	**66.7**	6.7	**40.0**	**40.0**	20.0
	Pour	**66.7**	20.0	13.3	**80.0**	13.3	6.7	0.0	0.0	**100.0**
Pitcher	Pour	**66.7**	13.3	20.0	**40.0**	20.0	**40.0**	0.0	**86.7**	13.3
	Pick	33.3	**40.0**	26.7	**53.3**	13.3	33.3	**73.3**	26.7	0.0
Spatula	Stir	**53.3**	26.7	20.0	**60.0**	26.7	13.3	**46.7**	40.0	13.3
	Hand over	40.0	**46.7**	13.3	**53.3**	20.0	26.7	**53.3**	20.0	26.7

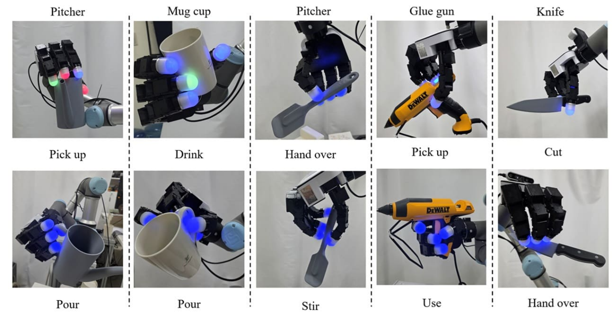

Fig. 6: Successful task-oriented dexterous grasp examples across different grasp taxonomies.

1 grasp pose results in an average success rate of 59.3%, whereas using the Top-2 pose reduces the success rate to 48.0%. At the same time, the slip rate increases from 20.7% to 28.0%. This indicates that grasp pose quality directly affects contact formation and grasp stability. Lower-ranked poses often produce unstable contact locations, which increases the likelihood of slipping during the lifting phase.

(2) Influence of Grasp Taxonomy: When the same grasp pose is used but the taxonomy is replaced with a functionally opposite grasp type, the success rate drops significantly from 59.3% to 35.3%. This result suggests that even with an identical grasp pose, the selected grasp taxonomy alters the resulting finger contact configuration and significantly influences grasp stability. This effect is particularly pronounced for tool-like objects, where task-inconsistent grasp types often lead to slippage or failure.

(3) Slip Failure Analysis: Many failure cases manifest as slippage rather than complete grasp failure. This indicates that the initial contact formation succeeds, but the grasp lacks sufficient stability during lifting. Slippage occurs more frequently for elongated or cylindrical objects, especially when the contact configuration is not aligned with the object's center of mass. These observations highlight that task-oriented grasping requires not only an appropriate grasp pose but also a contact configuration that reflects the intended task.

V. CONCLUSIONS

In this work, we proposed a framework extending task-oriented grasping to dexterous hands using a grasp taxonomy and the Contact Web. While real-world experiments show good generalization, some limitations remain. First, evaluating the LLM-based taxonomy selection is inherently challenging due to the multimodal nature of valid grasps. Second, our physical evaluation focuses on initial stability rather than complex downstream task completion. Third, relying on geometric cues without explicit physical reasoning occasionally causes slippage. Finally, performance heavily depends on upstream grasp predictions. Future work will integrate functional task metrics, physical constraints like friction cones, and end-to-end architectures to address these issues.

VI. ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program(IITP-2023-RS-2023-00266615) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation), This work was supported by the BK21 plus program ”AgeTech-Service Convergence Major” through the National Research Foundation (NRF) funded by the Ministry of Education of Korea[5120200313836], Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022- 00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)), This work was supported by the Technology Innovation Program(RS-2024-00469147, RS-2023-KT225094, RS-2025-02702977, RS-2024-00444344, RS-2024-00507746, RS-2026-25548755) funded By the Ministry of Trade Industry Energy(MOTIE, Korea), This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) and Korea Smart Farm RD Foundation(KosFarm) through Smart Farm Innovation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA) and Ministry of Science and ICT(MSIT), Rural Development Administration(RDA)(2540000882)

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] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen *et al.*, “Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4737–4746.

[3] J. Chen, Y. Ke, L. Peng, and H. Wang, “Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy,” *Robotics: Science and Systems*, 2025.

[4] C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, “Graspgpt: 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] Y. Ju, K. Hu, G. Zhang, G. Zhang, M. Jiang, and H. Xu, “Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation,” in *European Conference on Computer Vision*. Springer, 2024, pp. 222–239.

[7] D. Huang, W. Dong, C. Tang, and H. Zhang, “Hgdiffuser: efficient task-oriented grasp generation via human-guided grasp diffusion models,” in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 19 538–19 545.
[8] T. Feix, J. Romero, H.-B. Schmiedmayer, 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.
[9] S. B. Kang and K. Ikeuchi, “Grasp recognition using the contact web,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1. IEEE, 1992, pp. 194–201.
[10] 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. 46 881–46 907, 2024.
[11] J. Zhang, H. Liu, D. Li, X. Yu, H. Geng, Y. Ding, J. Chen, and H. Wang, “Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes,” in *8th Annual Conference on Robot Learning*, 2024.
[12] 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,” in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 8495–8502.
[13] H. Jiang, S. Liu, J. Wang, and X. Wang, “Hand-object contact consistency reasoning for human grasps generation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 107–11 116.
[14] G.-H. Xu, Y.-L. Wei, D. Zheng, X.-M. Wu, and W.-S. Zheng, “Dexterous grasp transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 933–17 942.
[15] J. R. Napier, “The prehensile movements of the human hand,” *The Journal of Bone & Joint Surgery British Volume*, vol. 38, no. 4, pp. 902–913, 1956.
[16] M. R. Cutkosky *et al.*, “On grasp choice, grasp models, and the design of hands for manufacturing tasks.” *IEEE Transactions on robotics and automation*, vol. 5, no. 3, pp. 269–279, 1989.
[17] I. M. Bullock, J. Z. Zheng, S. De La Rosa, C. Guertler, and A. M. Dollar, “Grasp frequency and usage in daily household and machine shop tasks,” *IEEE Transactions on Haptics*, 2013.
[18] A. M. Dollar and R. D. Howe, “The human hand as an inspiration for robot hand development,” *IEEE Transactions on Robotics*, 2010.
[19] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in neural information processing systems*, vol. 30, 2017.
[20] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
[21] Y.-L. Wei, M. Lin, Y. Lin, J.-J. Jiang, X.-M. Wu, L.-A. Zeng, and W.-S. Zheng, “Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 11 818–11 828.
[22] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” in *European conference on computer vision*. Springer, 2024, pp. 38–55.
[23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.
[24] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, “Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes,” in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.