

# Learning Granularity-Aware Affordances from Human-Object Interaction for Tool-Based Functional Dexterous Grasping

Fan Yang, Wenrui Chen, Kailun Yang, Haoran Lin, Dongsheng Luo, Conghui Tang, Zhiyong Li, and Yaonan Wang

**Abstract**—To enable robots to use tools, the initial step is teaching robots to employ dexterous gestures for touching specific areas precisely where tasks are performed. Affordance features of objects serve as a bridge in the functional interaction between agents and objects. However, leveraging these affordance cues to help robots achieve functional tool grasping remains unresolved. To address this, we propose a granularity-aware affordance feature extraction method for locating functional affordance areas and predicting dexterous coarse gestures. We study the intrinsic mechanisms of human tool use. On one hand, we use fine-grained affordance features of object-functional finger contact areas to locate functional affordance regions. On the other hand, we use highly activated coarse-grained affordance features in hand-object interaction regions to predict grasp gestures. Additionally, we introduce a model-based post-processing module that transforms affordance localization and gesture prediction into executable robotic actions. This forms GAAF-Dex, a complete framework that learns Granularity-Aware Affordances from human-object interaction to enable tool-based functional grasping with dexterous hands. Unlike fully-supervised methods that require extensive data annotation, we employ a weakly supervised approach to extract relevant cues from exocentric (Exo) images of hand-object interactions to supervise feature extraction in egocentric (Ego) images. To support this approach, we have constructed a small-scale dataset, Functional Affordance Hand-object Interaction Dataset (FAH), which includes nearly 6K images of functional hand-object interaction Exo images and Ego images of 18 commonly used tools performing 6 tasks. Extensive experiments on the dataset demonstrate that our method outperforms state-of-the-art methods, and real-world localization and grasping experiments validate the practical applicability of our approach. The source code and the established dataset are available at <https://github.com/yangfan293/GAAF-DEX>.

**Index Terms**—Visual Affordance, Dexterous Grasping, Dexterous Hand, Tool Manipulation, Hand-Object Interaction.

## I. INTRODUCTION

This work was partially supported by the National Key R&D Program of China under Grant 2022YFB4701400/2022YFB4701404, the National Natural Science Foundation of China under Grant 62273137, 62473139, No. U21A20518, and No. U23A20341, the Hunan Provincial Research and Development Project under Grant 2025QK3019, the Hunan Science Fund for Distinguished Young Scholars under Grant 2024JJ2027, and the Open Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2025B20). (Corresponding author: Wenrui Chen. E-mail: chenwenrui@hnu.edu.cn.)

F. Yang, W. Chen, K. Yang, H. Lin, D. Luo, and C. Tang are with the School of Artificial Intelligence and Robotics, Hunan University, Changsha 410012, China. (E-mail: sysf293@hnu.edu.cn.)

W. Chen, K. Yang, Z. Li, and Y. Wang are also with the National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University, Changsha 410082, China.

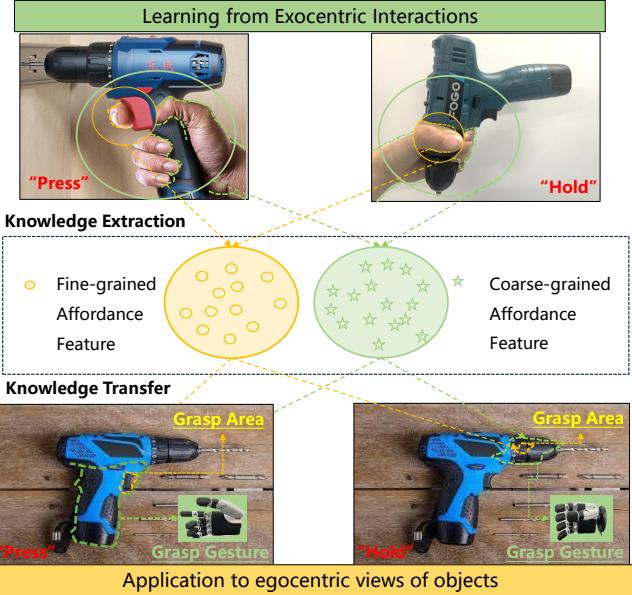


Fig. 1. GAAF-Dex extracts multi-granular affordance features from exocentric images and transfers them to egocentric images, achieving functional grasping through localization, gesture prediction, and post-processing.

ENABLING robots to flexibly utilize tools based on diverse task instructions (e.g., pressing, grasping, or opening) constitutes a cornerstone of human-robot collaboration, with functional grasping [1] serving as a critical initial step. Unlike general grasping, functional grasping imposes stringent requirements, necessitating dynamic identification of task-specific functional regions and generation of corresponding grasping gestures. Specifically, it entails: (1) precise localization of task-specific functional regions (e.g., drill's button or scissor's handle) rather than arbitrary contact points; (2) generation of multi-finger grasps via dexterous hands to meet the complex demands of varied tasks. This work aims to achieve functional grasping through vision-guided approaches, leveraging constraints inherent to objects and tasks.

Conventional vision-based grasping methods primarily focus on 6D pose estimation [2]–[4], capable of determining an object's overall position and orientation, but inadequate for localizing fine-grained functional regions or predicting task-specific grasping gestures. Data-driven approaches attempt to regress hand-object interaction parameters (e.g., contact points and gestures) directly from images [5]–[9] but rely heavily on extensive pixel-level annotations and often utilize human hand models, rendering them ill-suited for practical robotic

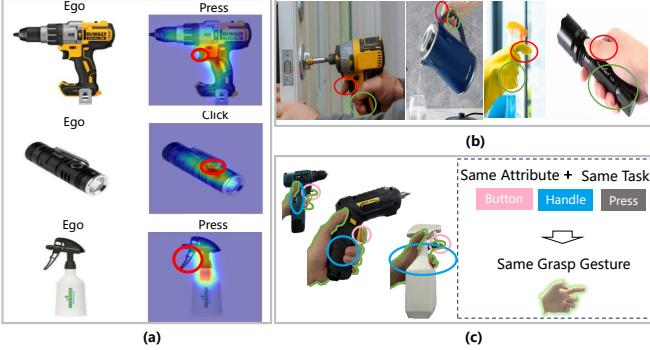


Fig. 2. Research motivations. (a) Functional grasping requires fine-grained affordance localization (red circle), beyond coarse-grained action regions (colored areas). (b) Objects with similar structural attributes employ consistent gestures for identical tasks. (c) Functional fingers (red circle) exhibit spatial separation from other fingers (green circle).

hand applications. Deep reinforcement learning methods train grasping policies in simulation to output end-effector poses and finger configurations [8], [10], [11], yet their dependence on complex simulation setups and parameter tuning limits real-world applicability.

The concept of “affordance” [12] offers a novel perspective to address these shortcomings by characterizing potential physical interactions of object components (e.g., a button for pressing), thereby linking tasks to actions. Recent advances in visual affordance research [13]–[17] have shown promise: object-centric methods can detect, segment, and label “action possibility” regions on objects, while weakly supervised methods [15], [17]–[20] learn affordance regions from exocentric (Exo) hand-object interaction images and transfer them to egocentric (Ego) perspectives, significantly reducing annotation costs. However, these approaches are limited to coarse-grained action region localization, failing to fully exploit multi-granular affordance cues in exocentric images or provide a holistic solution integrating localization and gesture prediction.

To address these limitations, this paper proposes GAAF-Dex, a weakly supervised multi-task framework that leverages affordance cues from exocentric hand-object interaction images to dynamically localize functional regions and generate corresponding coarse-grained gestures based on diverse task instructions, thereby achieving functional grasping, as illustrated in Fig. 1, which depicts the overall pipeline from feature extraction to localization and gesture prediction. This approach tackles three key deficiencies of existing methods: (1) a focus solely on isolated localization tasks, neglecting multi-granular affordance cues in exocentric images; (2) detection limited to coarse regions (as shown in the colored areas in Fig. 2 (a)), whereas functional grasping demands fine-grained localization (see the red circles in Fig. 2 (a)); and (3) reliance of gesture prediction methods on human hand models (e.g., MANO [21]), which perform poorly in robotic hand applications.

GAAF-Dex integrates a vision-driven affordance solution with multi-task learning and a post-processing module to effectively unify perception and action. Initially, a multi-task framework is designed to extract multi-granular affordance features from exocentric hand-object interaction images via weak supervision and transfer them to egocentric perspectives, providing a perceptual foundation for functional grasping.

To address the precise localization demands of functional grasping (e.g., Fig. 2 (a)’s red circle), fine-grained affordance features are extracted, coupled with spatial separation between functional fingers (Fig. 2 (b)’s red circle) and other fingers (Fig. 2 (b)’s green circle), employing spatial analysis and kinematic modeling to achieve functional finger-guided fine-grained localization, accurately targeting key contact points, such as a drill’s button in a “Press” task. Furthermore, inspired by the observation that objects with similar structural attributes (e.g., a drill and a spray bottle’s button-handle design) employ consistent gestures for identical tasks (Fig. 2 (c)), coarse-grained affordance features are extracted to predict task-specific coarse-grained grasping gestures via an affordance-driven prediction network that focuses on high-activation features in exocentric interactions, delivering diverse gestures tailored to robotic hands and overcoming limitations of human hand models. To bridge visual perception and robotic execution, a post-processing module is developed, integrating localization results and gesture predictions to compute a transformation matrix from fingertip to wrist using gesture-derived joint angles and robotic hand models, thereby executing dexterous grasping actions.

Existing datasets, such as those proposed in [5]–[7], [22]–[25], often rely on synthetic images, utilize mesh parameters for gesture representation, or lack multi-view data, rendering them inadequate for the generalization and practicality demands of functional grasping. To address this, the FAH dataset is constructed, comprising approximately 6,000 images covering 18 tools, 6 functions, and 14 gesture labels, requiring only image-level annotations to significantly reduce labeling costs while supporting task scalability and providing a practical research foundation for functional grasping.

The contributions of this work are outlined as follows:

- A weakly supervised multi-task learning framework, GAAF-Dex, is proposed. It integrates fine-grained functional region localization and coarse-grained gesture prediction, leveraging multi-granular affordance features and a model-based post-processing module to bridge perception and control in a complete grasp execution pipeline.
- A dataset named FAH is introduced, containing functional human-object interaction data with both region-level and gesture-level annotations. It enables affordance transfer from exocentric to egocentric views and serves as a benchmark for functional grasp learning.
- The effectiveness and generalizability of GAAF-Dex are demonstrated through extensive experiments on the FAH dataset and real-world, enabling task-conditioned grasping across diverse scenarios and unseen tools.

## II. RELATED WORK

**Visual Affordance Understanding.** Research in vision-based affordance understanding aims to locate areas of objects that are operable. Various methodologies have inferred visual affordances for simple gripper grasps [17], [26]–[28]. Chen *et al.* [29] proposed a framework for detecting 6-DoF task-oriented grasps, processing observed object point clouds to predict diverse grasping poses tailored for distinct tasks. The

studies [30], [31] were conducted to generalize the robot grasping affordance areas beyond labels by incorporating large prediction models.

In contrast, works such as those in [17], [19], [22], [32], [33] explored non-robot-centric perspectives in affordance understanding. Early efforts focused on fully supervised methods that required per-pixel labeling, resulting in high data acquisition costs [22], [34], [35]. To address this, recent studies [15], [19], [20] proposed weakly supervised approaches, leveraging human-object interaction cues from images [19] or videos [17] to supervise affordance learning for object-only views. For example, Locate [19] aggregated features from exocentric images into compact prototypes (human, object parts, and background) to supervise egocentric images, enabling identification of matching object parts. Similarly, Luo *et al.* [17] analyzed hand positions and motions in interaction videos to obtain affordance areas for object-alone images. Zhang *et al.* [32] introduced a bidirectional progressive transformer using video data to achieve joint prediction of hand trajectories and interaction hotspots in first-person scenarios. The research works in [36]–[39] advanced the field of 3D perception. Specifically, the work in [39] leverages unsupervised multi-view stereo (MVS) and neural rendering to enable effective perception of 3D dense and occluded scenes, which can support robotic operations in complex environments. 3D AffordanceNet [36] focused on recognizing 3D affordances of static objects by analyzing their shapes and features, but its reliance on synthetic datasets, lack of dynamic hand-object interactions, and the absence of functional operation tasks limit its applicability to real-world and dynamic applications.

Tool use, however, demands a combination of dexterous manipulation and functional part affordance understanding. To address this, we integrate weakly supervised affordance learning with the generation of dexterous coarse gestures. Our approach not only learns affordance localization from hand-object interactions but also predicts grasping gestures, laying the groundwork for practical tool use.

**Coarse-to-Fine Dexterous Grasping.** Achieving functional tool manipulation with robotic hands necessitates advanced dexterous grasping, which extends beyond basic two or three-finger grippers to involve multi-finger coordination. Previous approaches to achieving precise grasping relied on either model-based methods [40]–[43], which require extensive time for object and hand modeling and suffer from poor generalization, or data-driven methods [1], [44]–[48], which are costly due to the need for extensive labeling of contact points and joints. In contrast, the coarse-to-fine approach used in [1], [5], [6] treated the task of predicting dexterous gestures as a classification problem. After obtaining a specific category of grasp type, fine-tuning was performed, simplifying the high-dimensional data prediction task. GanHand [5] utilized 33 grasp classification types of the MANO model [21] to generate pre-grasp postures, whereas FunctionalGrasp [1] mapped these 33 grasp types of MANO models to the ShadowHand robotic hand model to obtain pre-grasp postures. In contrast, we have designed a classification network for 14 gestures of a low-cost robotic hand, leveraging the consistency of the object’s “task-affordance”. These 14 gestures encompass the daily tool

TABLE I

STATISTICS OF RELATED DATASETS AND THE PROPOSED FAH DATASET.  
INTER-TYPE: INTERACTION TYPE (Ha-O: HAND-OBJECT, Hu-O: HUMAN-OBJECT).  
REAL / SYN.: REAL OR SYNTHETIC DATA. VIEW: PERSPECTIVE. ANNOTATION:  
LEVEL OF ANNOTATION (PIX-LEVEL: PIXEL-LEVEL, IMG-LEVEL: IMAGE-LEVEL).  
HAND POSE: ANNOTATION TYPE (MESH: HAND MESH, ANGLE: JOINT ANGLES).  
AFF. INT.: AFFORDANCE INTERACTION (✓: YES, ×: NO)

Dataset	Year	Inter-Type	Real / Syn.	View	Annotation	Hand Pose	Aff. Int.
ObMan [25]	2019	Ha-O	syn.	Exo	Pix-Level	Mesh	×
YCBAfford [5]	2020	Ha-O	syn.	Exo	Pix-Level	Mesh	×
PAD [24]	2021	Hu-O	real	Exo-Ego	Pix-Level	×	✓
AGD20K [20]	2021	Hu-O	real	Exo-Ego	Img-Level	×	✓
OakInk-Image [7]	2022	Ha-O	real	Exo	Pix-Level	Mesh	✓
AffordPose [6]	2023	Ha-O	syn.	Exo	Pix-Level	Mesh	✓
OakInk2 [23]	2024	Ha-O	real	Exo	Pix-Level	Mesh	✓
FAH (Ours)	2024	Ha-O	real	Exo-Ego	Img-Level	Angle	✓

operation needs of humans.

**Hand-Object Interaction Datasets.** The emergence of relevant datasets has significantly advanced the development of “hand-object” interactions, as shown in Tab. I. OakInk [7] introduced a dataset containing affordances and corresponding gesture labels for 1800 household objects. AffordPose [6] introduced a synthetic dataset for fine-grained hand-object interactions based on specific object part visibility, but its reliance on labor-intensive MANO annotations [21] limits its applicability to real-world scenarios, as it focuses solely on static interactions without addressing functional operations or dynamic interactions critical for robotics applications. AGD20K [17] focused on inferring human intentions from support images of human-object interactions and transferring them to a set of query images. However, it did not consider direct hand-object interactions and lacked gesture annotations. Datasets related to visibility [5]–[7], [22]–[25] faced challenges such as reliance on synthetic data, use of mesh parameters for gestures, and failure to consider human behavior in reasoning about affordance areas. The FAH dataset we constructed provides real paired Exo-Ego view data for learning transferable human-tool interaction knowledge, and includes coarse gesture annotations that extend affordance vision research toward robot-executable manipulation.

### III. PROBLEM FORMULATION

The objective of this study is to address challenges in functional grasping by developing a model, denoted by  $M$ . This model is designed to analyze egocentric RGB images,  $I$ , containing a single object, along with a task description,  $T$ . The model outputs the initial grasping area and a coarse grasping gesture appropriate for the task. Specifically, the model predicts:

$$M(I, T) \rightarrow (P, \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}), \quad (1)$$

where  $P=(x, y, z)$  represents the position where the object should be grasped in the camera coordinate system. The  $z$  coordinate is derived from depth maps, providing depth information about the grasping location. The set  $\{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$  denotes the joint angles for a coarse grasping gesture. Upon obtaining  $P$  and  $\theta_i$ , a post-processing module refines these predictions to determine the precise hand positions and joint angles required for effective functional grasping.

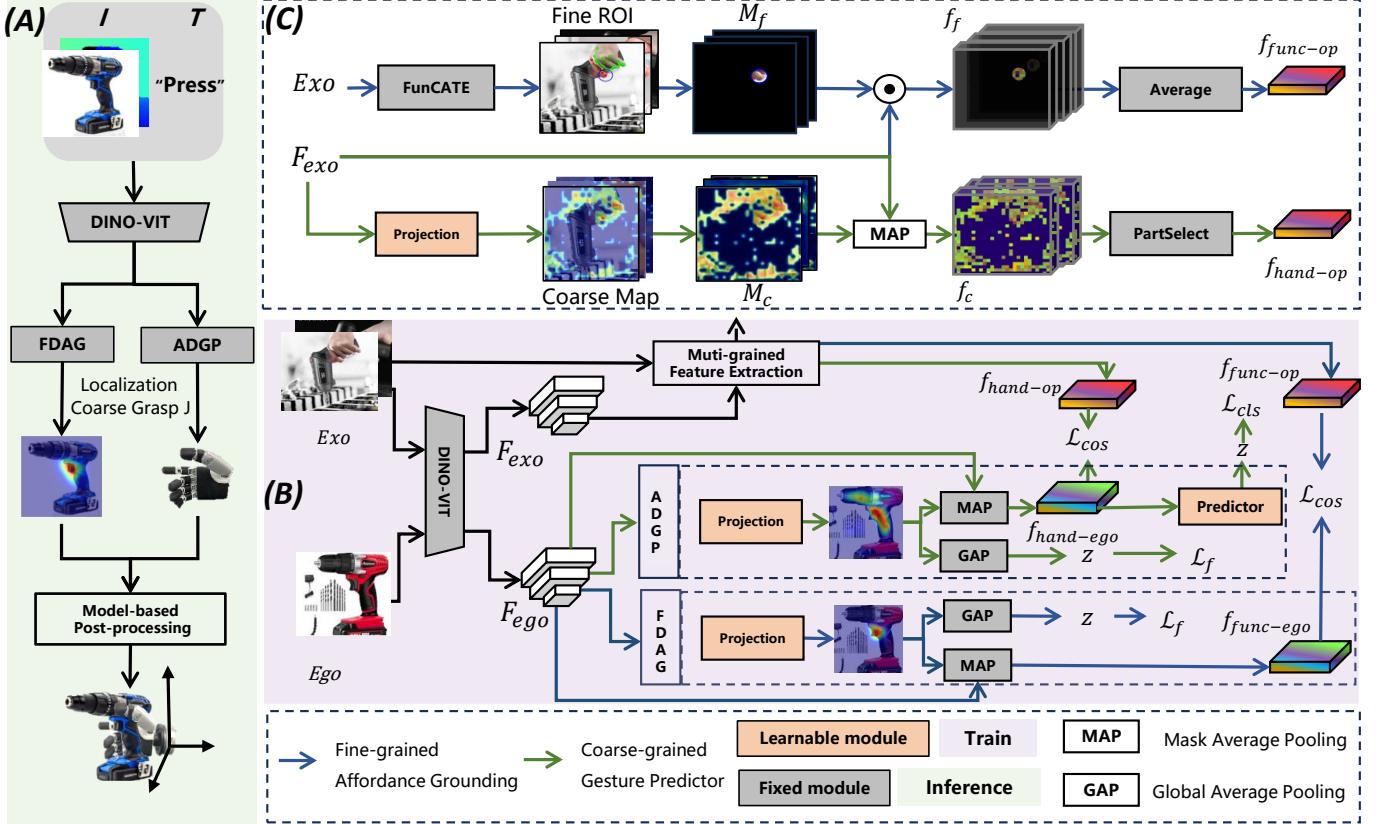


Fig. 3. **Framework of GAAF-Dex.** (A) The inference flow. Given an RGB-D image (where the depth is not involved in training but directly provides according to the affordance location) and a task as input, the Funcfinger-Driven Affordance Grounding (FDAG) module identifies the grasping region's coordinates, and the Affordance-Driven Gesture Predictor (ADGP) module predicts the corresponding coarse gesture. Finally, the Model-based Post-processing Module integrates coordinates and gestures for final execution. (B) The training flow takes  $N$  Exo images and one Ego image as input. The bottom part of the box represents the training procedure for the ADGP and FDAG modules for Ego images. The top part describes the process of extracting coarse-grained affordance features and fine-grained affordance features from Exo images to serve as supervision for the ADGP and FDAG modules. (C) The Multi-grained Feature Extraction (MFE) module extracts coarse- and fine-grained feature prototypes from Exo for supervision.

#### IV. METHOD

Given a set of exocentric interaction images and an egocentric image of an object, our core objective is to train two prediction modules, namely the Funcfinger-Driven Affordance Grounding (FDAG) and Affordance-Driven Gesture Predictor (ADGP) modules. We extract affordance region features related to functional fingers and corresponding grasp gesture features from exocentric (Exo) images and transfer these features to egocentric (Ego) images, enabling us to locate the grasp points and gestures of functional fingers in the egocentric images. During the training phase, we utilize image-level affordance labels, whereas in the testing phase, the input is an egocentric image, and the outputs are the object's optimal grasp point  $P$  and the associated coarse grasp gesture  $G$ , as shown in the green background (A) part of Fig. 3.

The training part of our method is illustrated in the purple background (B) part of Fig. 3, and the core idea is as follows: for the input images  $\{I_{exo}, I_{ego}\}$  ( $I_{exo} = \{I_1, I_2, \dots, I_N\}$ ), we first use a network  $\phi$  to extract deep features  $\{\mathcal{F}_{exo}, \mathcal{F}_{ego}\} \in \mathbb{R}^{D \times H \times W}$ . In our case,  $\phi$  is a self-supervised visual transformer (DINO-ViT [49]), which provides excellent part-level features. Subsequently, based on task requirements, we extract fine-grained and coarse-grained visibility cues from Exo images using the Multi-grained Feature Extraction (MFE) module (see Fig. 3 (C)) to supervise

the corresponding features extracted in Ego images by the ADGP and FDAG modules, enabling affordance localization and gesture prediction. Specifically, for functional affordance localization, as guided by the blue dashed line in Fig. 3, we propose a functional finger-driven fine-grained feature extraction method (Sec. IV-A). For coarse gesture prediction, as guided by the green dashed line in Fig. 3, we leverage the Class Activation Mapping (CAM) [50] and the hand-background-object feature prototype selection module from LOCATE [19] to extract coarse-grained features from Exo images (Sec. IV-B). Finally, we design a model-based post-processing module to combine functional areas and coarse grasp gestures, yielding the final end-effector grasp points and coarse-to-fine functional grasp results (Sec. IV-D).

##### A. Fine-Grained Feature Extraction for Affordance Grounding

In this section, we focus on the blue arrow flow in Fig. 3 (B) and (C). First, the Exo images are processed through the FunCATE module in the MFE module (see Sec. IV-A1) to obtain the fine-grained Region of Interest (Fine ROI). The ROI is then used to generate the fine-grained mask  $M_f$  as follows:

$$M_f(x, y) = \begin{cases} 1 & \text{if } \sqrt{(x - x_0)^2 + (y - y_0)^2} \leq r, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

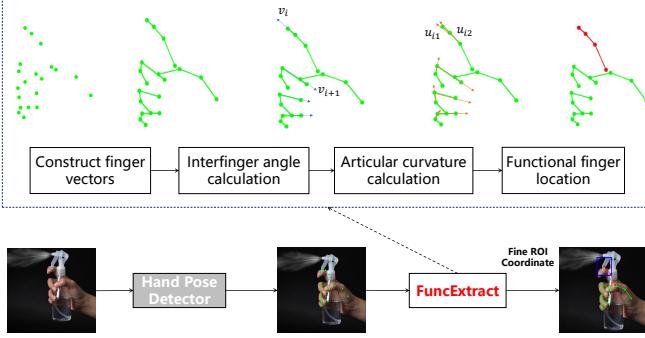


Fig. 4. FunCATE module, which includes a hand pose detector and the FuncExtract module. FuncExtract determines the functional finger by calculating the vector angles between fingers and the joint angles of each finger.

where the mask function  $M_f(x, y)$  defines a circular region centered at  $(x_0, y_0)$  with radius  $r$ , producing a binary mask. Simultaneously, the feature map  $\mathcal{F}_{exo}$  is upsampled as follows:

$$f_{up} = \text{Upsample}(\mathcal{F}_{exo}), \quad (3)$$

where  $f_{up}$  represents the upsampled feature map obtained by resizing  $\mathcal{F}_{exo}$  to match the size of the resized image. This ensures consistency between the feature coordinates and the ROI coordinates. Then,  $f_{up}$  and  $M_f$  are combined via element-wise multiplication to obtain the fine-grained affordance features  $f_f \in \mathbb{R}^{C \times H \times W}$ :

$$f_f = f_{up} \odot M_f. \quad (4)$$

Then, the  $f_{fine}$  from  $N$  images are averaged to generate the fine-grained functional affordance prototype  $f_{func-op}$  for supervision. Meanwhile, the Ego features  $\mathcal{F}_{ego}$  are processed through the FDAG module to obtain  $f_{func-ego}$ , and finally, knowledge transfer from Exo to Ego is performed (see Sec. IV-A2).

1) *FunCATE*: For feature supervision, we perform functional finger-guided cue extraction on Exo images, primarily implemented by the FunCATE module. Specifically, as shown in Fig. 4, we first use the network  $\phi_2$  for gesture recognition on Exo images. In our case,  $\phi_2$  is MediaPipe [51], which has the advantage of accurate landmark detection. It can obtain 21 key points' 2D ( $x, y$ ) coordinates of the human hand.

Then, we apply our proposed functional finger determination algorithm, FuncExtract, to obtain the 2D ( $x, y$ ) coordinates of the functional fingertip. We consider the area with radius  $r$  around these coordinates as our fine ROI.

**FuncExtract:** The FuncExtract module is illustrated in the dotted box at the top of Fig. 4. This module evaluates the spatial looseness between fingers by considering both the spatial alignment (parallelism) and the local curvature (bending angle), and selects the functional finger based on this parameter.

Specifically, when we obtain the 2D coordinates of 21 key points on the human hand, we first vectorize the points of each finger and compute the cosine of the angle between adjacent finger vectors to evaluate the parallelism of the four non-thumb fingers. The cosine of the angle between vectors formed by adjacent finger joints for the  $i$ -th finger (excluding the thumb) angle <sub>$i$</sub>  is calculated as:

$$\text{angle}_i = \frac{v_i \cdot v_{i+1}}{\|v_i\| \|v_{i+1}\|}, \quad i \in \{2, 3, 4, 5\}. \quad (5)$$

Here,  $v_i$  represents the vector of the  $i$ -th finger,  $\cdot$  denotes the dot product of the vectors, and  $\|\cdot\|$  denotes their magnitude.

If the cosine values for all adjacent finger pairs are greater than a predefined threshold  $\tau$ , the four non-thumb fingers are considered parallel, and the thumb is directly identified as the functional finger. Otherwise, we proceed to analyze the joint bending angles of the four non-thumb fingers. The bending angle for each finger is determined by calculating the cosine of the angle between two vectors formed by the adjacent joints of each finger:

$$\begin{aligned} \text{func}_{ID} &= \underset{i \in \{2, 3, 4, 5\}}{\operatorname{argmin}} \left( 1 - \frac{u_{i1} \cdot u_{i2}}{\|u_{i1}\| \|u_{i2}\|} \right), \\ u_{i1} &= p_{i2} - p_{i1}, \\ u_{i2} &= p_{i3} - p_{i2}, \end{aligned} \quad (6)$$

where  $p_{i1}$ ,  $p_{i2}$ , and  $p_{i3}$  represent the coordinates of the first, second, and third joints of the  $i$ -th finger, and  $u_{i1}$ ,  $u_{i2}$  are vectors between the second-to-first and third-to-second joints of the  $i$ -th finger, respectively. The finger with the minimum bending angle is selected as the functional finger,  $\text{func}_{ID}$  is the functional finger identifier, ranging from 2 (index finger) to 5 (little finger).

2) *Functional Part-Level Knowledge Transfer*: Now we focus on the fine-grained feature extraction of Ego and use  $f_{func-op}$  for its supervision. Specifically, in the FDAG module, we first apply the Projection function  $P()$  [19] to the ego image, which utilizes class activation mapping (CAM) techniques [50] to generate a functionally-aware localization map, depicted as follows:

$$P_{\text{func-ego}} = P(\mathcal{F}_{ego} + \text{MLP}(\mathcal{F}_{ego})), \quad (7)$$

where MLP represents a feed-forward layer, and  $P()$  consists of two  $3 \times 3$  convolutional layers, normalization layers, and non-linear activation functions, followed by a  $1 \times 1$  class-aware convolution layer. Each map  $P_c \in \mathbb{R}^{H \times W}$  represents the network activation for the  $c$ -th interaction.

Then, we perform Masked Average Pooling (MAP) between the localization map and  $\mathcal{F}_{ego}$ , aggregating them into an embedding vector. On the other hand, a Global Average Pooling (GAP) layer is applied to the localization map to obtain the task classification scores  $z$ , which are used to compute the cross-entropy loss  $\mathcal{L}_t$  for optimization, depicted as follows:

$$f_{\text{func-ego}} = \text{MAP}(P_{\text{func-ego}}, \mathcal{F}_{ego}), \quad z = \text{GAP}(P_{\text{func-ego}}), \quad (8)$$

where the MAP operation includes a matrix multiplication between the normalized  $P_{\text{func-ego}}$  and  $\mathcal{F}_{ego}$ .

Finally, we use cosine loss  $\mathcal{L}_{\cos}$  and concentration loss  $\mathcal{L}_c$  to ensure the features are correctly extracted while maintaining coherence as follows:

$$\mathcal{L}_{\cos} = \max(1 - \frac{f_{\text{func-op}} \cdot f_{\text{func-ego}}}{\|f_{\text{func-op}}\| \|f_{\text{func-ego}}\|}, 0), \quad (9)$$

$$\mathcal{L}_c = \sum_c \sum_{u,v} \|\langle u, v \rangle - \langle \bar{u}_c, \bar{v}_c \rangle\| \cdot P_{\text{func-ego}} / z_c, \quad (10)$$

$$\bar{u}_c = \sum_{u,v} u \cdot P_{\text{func-ego}} / z_c, \quad \bar{v}_c = \sum_{u,v} v \cdot P_{\text{func-ego}} / z_c, \quad (11)$$

where  $\alpha$  is a margin added to compensate for the domain gap as the two embeddings come from different domains.  $\bar{u}_c$  and  $\bar{v}_c$  represent the center of the  $c$ -th localization map along the  $u$  and  $v$  axes, and  $z_c = \sum_{u,v} P_{\text{func-ego}}$  is a normalization term. The concentration loss forces the high activation regions of the localization maps to be close to the geometric center.

### B. Coarse-Grained Feature Extraction for Gesture Predictor

To achieve gesture prediction, we focus on the green arrow flow in Fig. 3 (B) and (C). For Exo images,  $\mathcal{F}_{\text{exo}}$  is processed through the Projection function in the Multi-grained Feature Extractor to obtain task-specific localization maps, denoted as coarse maps. These coarse maps serve as masks  $M_c$  and are processed with  $\mathcal{F}_{\text{exo}}$  using MAP to generate coarse-grained affordance features  $f_c$ . The features from  $N$  images are concatenated and passed into the PartSelect module [19], which clusters the object, background, and hand features within the hand-object interaction region into  $K$  clusters. Based on the Intersection Over Union (IOU) values of the similarity map with  $\mathcal{F}_{\text{ego}}$  and the saliency map obtained from Ego images processed by DINO-ViT [49], the coarse-grained affordance supervision feature prototype  $P_{\text{hand-op}}$  is derived.

For Ego images, the ADGF module first performs the same operation as the FDAG module, extracting coarse-grained hand-object interaction features  $f_{\text{hand-ego}}$  through the Projection layer.  $f_{\text{hand-ego}}$  is then passed through a GAP layer to obtain the task-related class label. The normalized  $f_{\text{hand-ego}}$  is combined with  $\mathcal{F}_{\text{ego}}$  via MAP to obtain the supervised coarse-grained affordance features.

Finally, as discussed in Sec. I, these coarse-grained affordance features originate from the most active regions of hand-object interactions and include affordance-guided coarse gestures. We add a coarse gesture predictor to  $f_{\text{hand-ego}}$ . Specifically, we use a Fully Connected (FC) classification network on  $f_{\text{hand-ego}}$  to predict the grasp type  $C$ , classifying it into one of the 14 grasp types that best suit the target object. This network is trained using the cross-entropy loss  $\mathcal{L}_{\text{class}}$ . The predicted grasp type  $C$  is associated with a representative hand configuration  $H_C$ , consisting of the joint angles of the five fingers and the abduction angle of the thumb.

### C. Training Supervision

In summary, during the training phase, the total loss consists of the following four parts:

$$\mathcal{L} = \mathcal{L}_{\cos} + \lambda_c \mathcal{L}_c + \mathcal{L}_{\text{class}} + \mathcal{L}_t, \quad (12)$$

where  $\lambda_c$  is the weight balancing these four terms,  $\mathcal{L}_{\cos}$  is defined in Eq. 9 as the cosine similarity loss between the exo coarse and fine-grained affordance feature prototypes and the ego coarse and fine-grained affordance features;  $\mathcal{L}_c$ , defined in Eq. 10, is the clustering loss;  $\mathcal{L}_t$  represents the cross-entropy loss for task classification of the Exo coarse-grained feature prototype and the Ego coarse-grained and fine-grained functional affordance feature prototypes, as described in Sec. IV-A2;  $\mathcal{L}_{\text{class}}$  is the cross-entropy loss for ego gesture type prediction, as defined in Sec. IV-B.

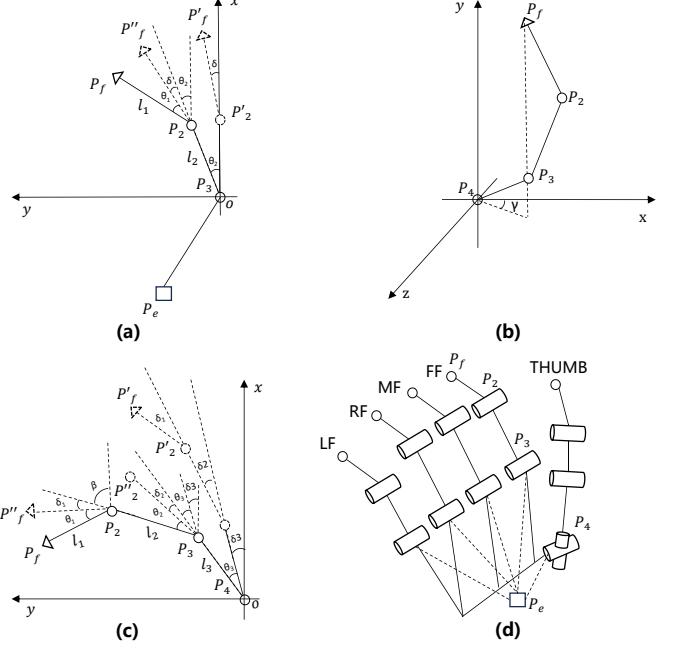


Fig. 5. Diagrams illustrating fingertip-to-end coordinate transformations based on a model. (a) shows the coordinate transformation of the flexion and extension joints of the index finger, (b) shows the coordinate transformation of the abduction and adduction joints of the thumb, (c) shows the coordinate transformation of the flexion and extension joints of the thumb, and (d) provides an overall reference for Inspire hand joints.

### D. Model-based Post-processing Module

In this module, we first extract the top  $N$  brightest RGB-D pixels from the functional affordance grounding predicted by FDAG, corresponding to the pixels with the highest probability values in the heatmap. Let the 3D coordinates of these pixels be  $P_i = [x_i, y_i, z_i]^T$  ( $i=1, 2, \dots, N$ ), where  $z_i$  is obtained from the depth camera. Then, the contact point is determined by calculating the centroid of these pixels, given by  $P_{\text{centroid}} = \frac{1}{N} \sum_{i=1}^N P_i$ . Finally, the contact point coordinates  $P_{\text{centroid}}$  are converted to the global coordinate system using the hand-eye calibration matrix, yielding the final functional fingertip contact point  $P_{wf} = [x_{wf}, y_{wf}, z_{wf}]^T$ .

Then, based on the proportional relationship and joint angles of the robotic hand model's finger joints, we transform the fingertip coordinates  $P_{wf}$  to obtain the wrist end coordinates  $P_{we}$  in the global coordinate system. Specifically, as shown in Fig. 5 (d), outside the thumb, the other four fingers of the Inspire hand have the same structure, a motor drives the two finger joints to flex and stretch. We take the index finger as an example as shown in Fig. 5 (a), where  $P_2, P_3$  represent the node of the first and second finger joint rotation axes, respectively.  $P'_i$  represents the position of the joint node when the drive motor is in the zero position. Here,  $P'_f - P'_2$  is at an  $\delta$  angle to the X-axis.  $P''_f$  represents the hypothetical position of the fingertip if only the second phalanx moves. We establish a hand coordinate system with  $P_3$  as the origin,  $O$ .  $P_3 - P'_2$  is the positive direction of the x-axis, and the z-axis coincides with the rotation axis of  $P_3$ .

The coordinate  $P_f = [x_{hf}, y_{hf}, z_{hf}]^T$  of the fingertip in the

hand coordinate system can be obtained as follows:

$$P_f = R(\theta_2) \begin{bmatrix} l_2 \\ 0 \\ 0 \end{bmatrix} + R(\theta_1 + \theta_2 + \delta) \begin{bmatrix} l_1 \\ 0 \\ 0 \end{bmatrix}, \quad (13)$$

where the  $l_1$  and  $l_2$  represent the first and second direct lengths, respectively.  $\theta_1$  and  $\theta_2$  come from the linear transformation of the index finger angle of our predicted coarse gestures.

$$R(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The thumb is driven by a motor for flexion and extension of the three joints, as shown in Fig. 5 (c). The calculation method of the fingertip to the end in flexion and extension is similar to that of the index finger, but the difference is that the thumb also has a lateral swing movement, as shown in Fig. 5 (b). The mapping of the fingertip  $P_f = [x_{hf}, y_{hf}, z_{hf}]^T$  to the end in the lateral swing process is as follows:

$$P_f = \begin{bmatrix} \cos \gamma & 0 & \sin \gamma \\ 0 & 1 & 0 \\ -\sin \gamma & 0 & \cos \gamma \end{bmatrix} P_f^o, \quad (14)$$

where the  $\gamma$  is the abduction-adduction angle,  $P_f^o$  denotes the thumb fingertip coordinate in the hand coordinate system, obtained from the flexion and extension angles using a forward kinematics formulation, similar to that in Equation 13.

The end coordinate  $P_{we}$  in the world coordinate system can be obtained by the following equation:

$$P_{we} = R_{wf}(P_e - P_f) + P_{wf}, \quad (15)$$

where  $P_e = [x_{he}, y_{he}, z_{he}]^T$  is the wrist end coordinate in the hand coordinate system, which is directly obtained from the mechanical structure.  $R_{wf}$  represents the rotation matrix of the object correctly grasped by the hand. Since this method focuses on the functional area and does not involve rotation, we assume that it is a known quantity.

Finally, to quickly and stably achieve coarse-to-fine gesture adjustment, we adopt the Functionally Integrated Adaptive Force-Feedback Manipulation (FAFM) algorithm [52] to refine the coarse gesture angles. During this process, continuous force feedback is received, and the adjustment stops when the rate of change of the force derivative reaches zero, indicating a stable grasp.

## V. ESTABLISHED DATASET

To advance research in dexterous functional manipulation, we introduce the FAH dataset, specifically designed for complex functional grasping tasks. FAH features diverse human-tool interaction examples that reflect common scenarios in both domestic and industrial settings. The dataset contains nearly 6K images, including 5616 training images (3951 exocentric and 1555 egocentric) and 232 egocentric test images.

Based on the Finger-to-Function (F2F) knowledge graph [52], FAH includes 18 commonly used tools (e.g., "Screwdriver", "Plug", "Kettle", "Drill") and 6 task types ("Press", "Click", "Hold", "Open", "Clamp", "Grip"), as defined in Table II.

TABLE II  
DEFINITIONS OF SIX TASKS IN THE FAH DATASET WITH EXAMPLES OF TARGET OBJECTS AND CORRESPONDING CONTACT PARTS FOR EACH TASK.

Task	Definition	Example (Object/Part)
Press	Applies a gesture to a tool's button with sustained force to maintain functionality for subsequent operations.	Drill/Button
Click	Applies a gesture to a tool's switch with brief force to trigger functionality for subsequent operations.	Mouse/Switch
Hold	Five-finger grasp to ensure stability, facilitating subsequent actions.	Bottle/Body
Open	Objects that are detached or twisted for special functionality.	Bottle/Lid
Clamp	Two-finger opposing force on a single region for precise control.	Plug/Body
Grip	Multi-finger balance across discontinuous regions.	Scissors/Handle

### A. Image Collection

Exocentric images were collected from three main sources: the AGD20K dataset [20], high-resolution product images from e-commerce platforms, and publicly available images retrieved using object-related keywords. To supplement underrepresented interaction types, we recruited 10 volunteers to photograph themselves using tools in natural hand-tool interactions, covering cases like "Clamp Knife", "Click Kettle", "Click Mouse", "Hold Drill", and "Open Valve".

A key design decision was to focus on single-hand exocentric images, which are essential for learning precise functional interaction cues. Multi-hand images were excluded to avoid interference in fine-grained affordance feature extraction. For egocentric images—where no human-object interaction is present—we directly selected standalone tool images from the same high-quality sources. Examples are shown in Fig. 6 (a).

Fig. 6 (b1), (b2), and (b3) illustrate the instance distributions across the exocentric train set, egocentric train set, and egocentric test set. The distributions are highly consistent, with the "Hold" task being the most common in all sets (43% in the exocentric set), while "Clamp" and "Grip" are less frequent (6% and 11% respectively). Each task-tool pair in the exocentric train set contains at least 100 images. The most common pair, "Press Spraybottle", includes 349 images. These statistics reflect a balanced and comprehensive dataset design that captures both common and nuanced interactions.

### B. Data Annotation

**Image-level Annotation:** We annotated each training image with task and object labels. Exocentric images were labeled based on observed human-object interactions, while egocentric images—without interactions—were assigned task labels by mapping from tool categories, and object labels were directly annotated. Two annotators labeled independently, with a third resolving disagreements. The task-tool relationship is many-to-many, totaling 23 combinations. For example, the task "Hold" applies to tools like "Knife", "Flashlight", "Cup", and "Door Handle", while a single tool may map to multiple affordances, such as "Knife" with "Hold" and "Clamp", and "Flashlight" with "Hold" and "Click".

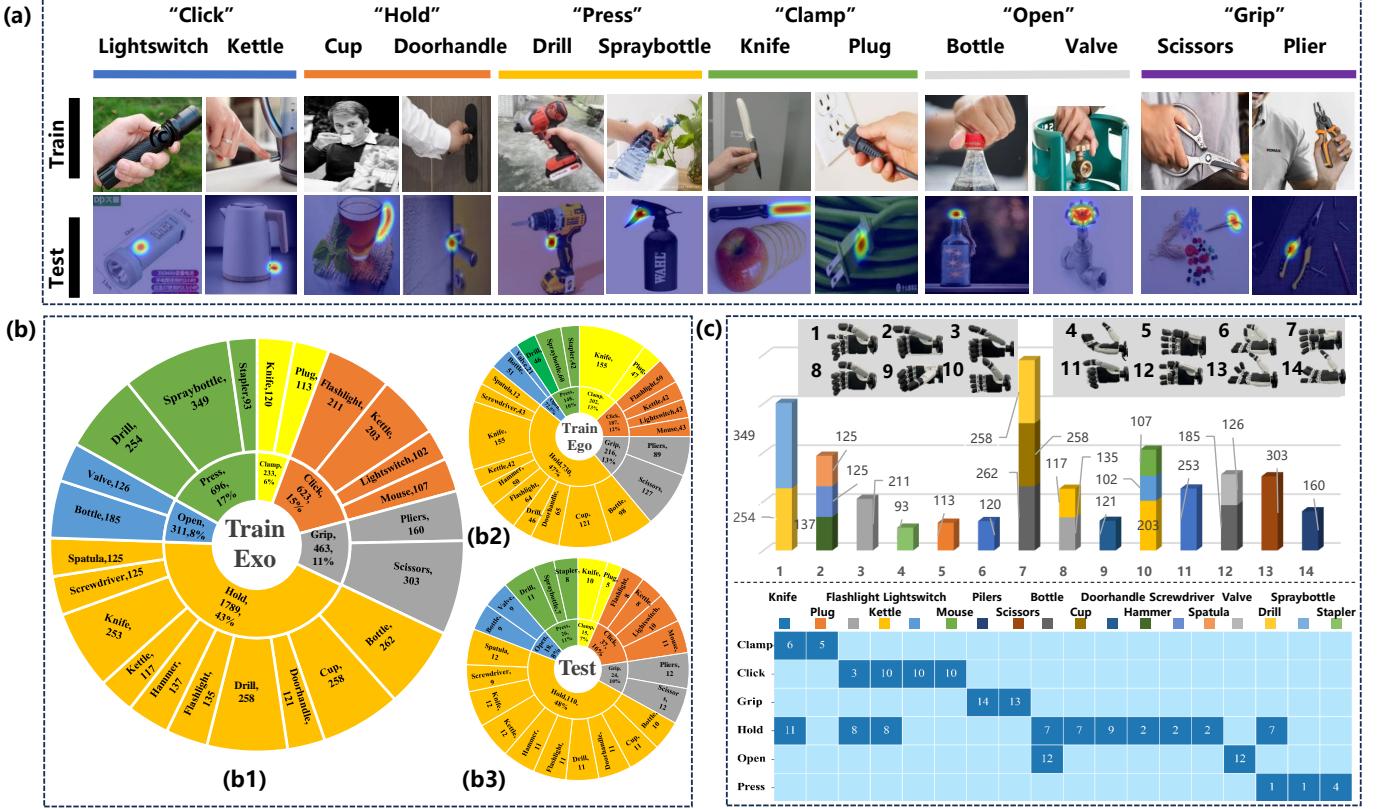


Fig. 6. Properties of the FAH dataset. (a) Examples from the dataset. (b) Instance count distribution of 18 tools and 6 tasks in the training and test sets. (c) Distribution properties of 14 coarse gestures: instance counts and their distribution across object categories (top); confusion matrix between affordance and object categories (bottom, horizontal axis: object categories, vertical axis: affordance (task) categories, table numbers: corresponding gestures).

**Affordance Annotation:** For test images, we adopted a heatmap-based annotation approach similar to AGD20K. Annotators focused on marking the contact areas of “functional fingers”. Three volunteers used the LabelMe tool to draw polygons over expected interaction regions based on typical “Task Tool” usage. The annotations were averaged and smoothed using a Gaussian blur.

**Coarse Gesture Annotation:** Following F2F [52], we assigned one of 14 coarse grasping gestures to each “Task Tool” pair in the FAH dataset. Their distributions across object categories are shown at the top of Fig. 6 (c), whereas the bottom shows a confusion matrix mapping gestures to task and object categories. For each gesture, we recorded five-finger flexion angles and thumb abduction angles. Importantly, our method is robot-agnostic and can be adapted to different robotic hands by adjusting gesture parameters accordingly.

## VI. EXPERIMENTS

We evaluate our approach on three levels: (1) qualitative and quantitative assessment of affordance localization based on functional fingers; (2) validation of affordance-based coarse gesture prediction; and (3) real-world dexterous grasping experiments with fixed rotation to verify localization, gesture prediction, and overall grasp success.

### A. Setups

**Implementation Details.** We use the DINO-ViT-S [49] pretrained on ImageNet [53] (unsupervised) with a patch size

of 16 to extract deep features. Each training iteration inputs one egocentric and  $N=3$  exocentric images. Images are resized to 512×512, randomly cropped to 448×448, and horizontally flipped. We train with SGD ( $\text{lr}=1e-3$ , weight decay= $5e-4$ , batch size=16). The loss weight  $\lambda_c$  is set to 0.07, and the margin  $\alpha$  to 0.5. In the first epoch,  $L_{\cos}$  is disabled to avoid supervision from inaccurate initial localizations.

**Metrics.** For affordance grounding, we adopt Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS) following prior work [19], [20].

For gesture prediction, the accuracy for tool  $j$  in task  $t$ ,  $A_{j,t} = \frac{C_{j,t}}{N_{j,t}}$ , where  $C_{j,t}$  is the number of correct predictions and  $N_{j,t}$  is the total number of samples. The overall average accuracy is  $AA = \frac{\sum_{t,j} C_{j,t}}{\sum_{t,j} N_{j,t}}$ .

For dexterous grasping, we measure the grasp success rate as defined in [48]: a grasp is successful if the hand holds the object stably for at least ten seconds and correctly performs the intended action on the tool’s functional area.

### B. Results of Functional Affordance Grounding

In this section, we present both qualitative and quantitative results to demonstrate the effectiveness and efficiency of our proposed method on the FAH test set. Our baselines include two weakly supervised methods, Cross-view-AG [20] and LOCATE [19], and two fully supervised methods, PSPNet [54] and DeepLabv3 [55].

**Qualitative Analysis.** We present the visibility grounding visualizations of two weakly supervised baseline methods, our

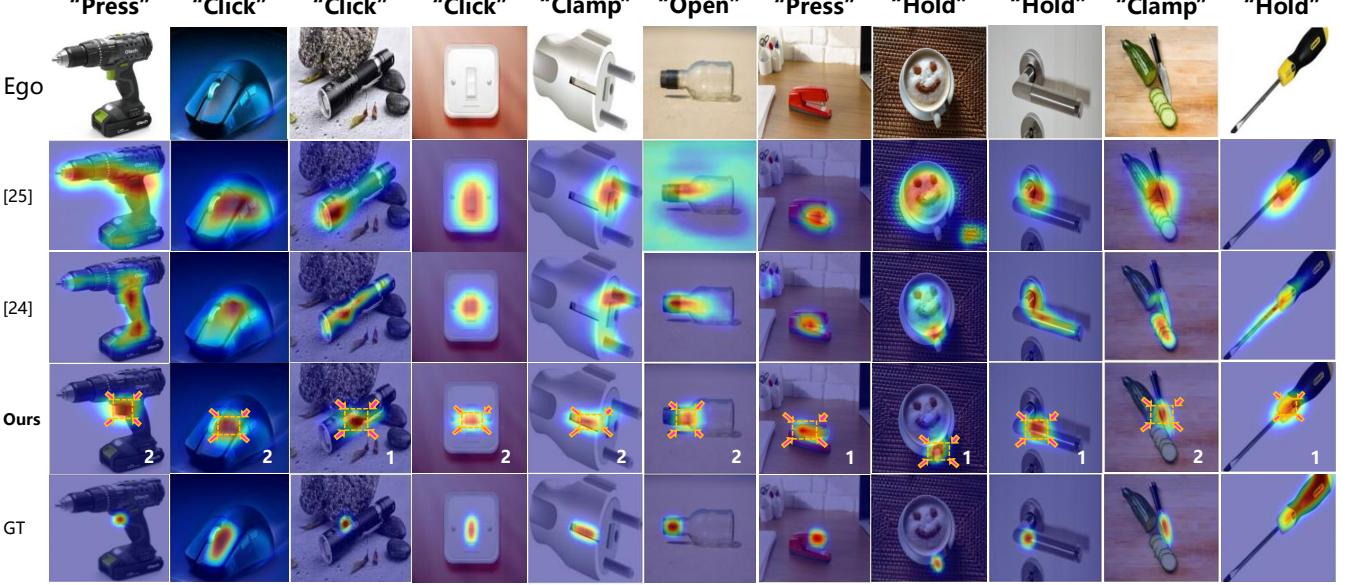


Fig. 7. Qualitative comparison of our method with LOCATE [19] and Cross-view-AG [20] on the FAH test set. The digits “1” and “2” in the fourth row of each image represent the functional finger indices, as calculated by Sec. IV-A1, where “1” denotes the thumb and “2” denotes the index finger.

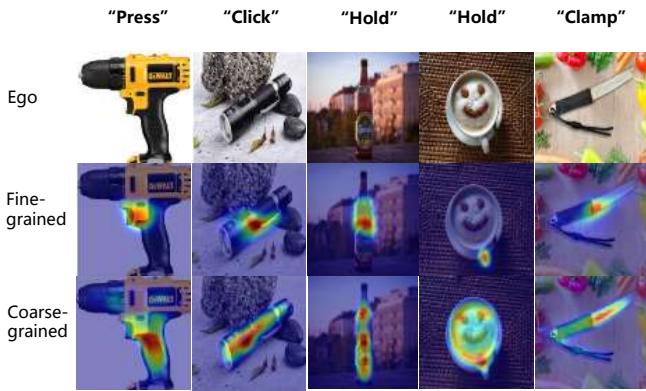


Fig. 8. Visualization of fine-grained and coarse-grained affordance feature regions. The second row shows fine-grained affordance regions used to localize functional contact areas between fingers and tools, while the third row depicts coarse-grained affordance regions for predicting rough grasp gestures.

method, and the Ground Truth (GT). As shown in Fig. 7, our visibility localization is more concentrated in the areas where functional fingers should contact, compared to the two baseline methods. We highlighted our method’s predictions with pink dashed boxes, which are essential for dexterous manipulation oriented toward functional usage. When the robot performs the corresponding actions in the “*Task Tool*” scenarios, stricter functional area localization is required. Particularly for tools like drills and flashlights that have specific buttons, corresponding to the first and third images in the fourth row of Fig. 7, our method accurately localizes to smaller button areas. For tools without buttons, our method also successfully localizes to the areas consistent with the human functional fingers. For instance, in the “*Clamp Plug*” task, the localization is on the right side of the plug’s head, which is the area the index finger needs to contact.

In Fig. 8, we compare fine-grained and coarse-grained affordance feature extraction regions. The second row highlights the fine-grained regions utilized for precise hand-object contact

TABLE III  
COMPARISON TO STATE-OF-THE-ART WEAKLY SUPERVISED METHODS AND FULLY SUPERVISED METHODS (\*) ON THE FAH TEST SET. THE **BEST** AND **SECOND-BEST** RESULTS ARE HIGHLIGHTED IN **BOLD** AND **UNDERLINED**, RESPECTIVELY. THE INFERENCE TIME IS EVALUATED ON A 4060Ti GPU. ( $\uparrow/\downarrow$  MEANS HIGHER/LOWER IS BETTER).

Model	KLD ( $\downarrow$ )	SIM ( $\uparrow$ )	NSS ( $\uparrow$ )	Time (s) ( $\downarrow$ )
PSPNet [54]*	6.876	0.186	0.467	0.2160
DeepLabv3 [55]*	2.226	0.184	0.252	0.0540
Cross-view-AG [20]	1.695	0.269	1.124	<u>0.0226</u>
LOCATE [19]	<u>1.537</u>	<b>0.317</b>	<u>1.131</u>	<b>0.0221</b>
Ours	<b>1.458</b>	0.311	<b>1.316</b>	0.0228

localization, while the third row illustrates the larger coarse-grained regions designed for gesture prediction. The comparison demonstrates that our DAAF-Dex network effectively extracts affordance features tailored for distinct functionalities, namely contact region localization and coarse gesture prediction. For instance, in the “*Press Drill*” task (first column), the button region (pressed by the index finger) in the second row represents the fine-grained feature extraction region, whereas the handle region in the third row serves as the coarse-grained feature extraction region for a full-hand grasp.

**Quantitative results.** We present the performance of the latest methods from related tasks, which involve weakly supervised object localization. As shown in Tab. III, our method demonstrates significant improvements over competing methods across most metrics. Specifically, our approach achieves a 5.1% improvement in KLD and a 16.3% improvement in NSS over the state-of-the-art grounding method LOCATE [19]. Our SIM score of 0.311 is slightly lower than LOCATE’s score of 0.06. This minor reduction is because SIM focuses more on the similarity of overlapping regions rather than their size. While LOCATE also performs part-level detection, it heavily relies on pre-trained DINO-ViT [49] for extracting part-level features, which are then clustered into background, human, and object categories, often neglecting the decoupling of fine-

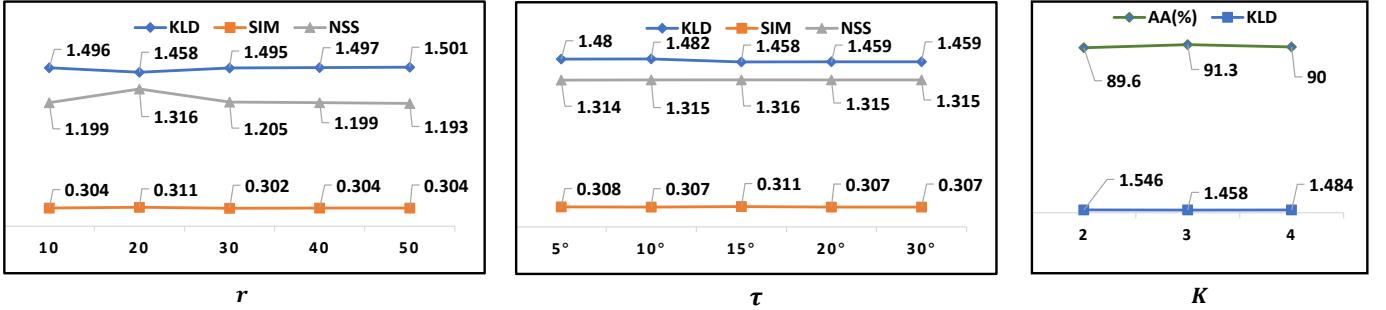


Fig. 9. Hyper-parameter study. We investigate the influence of  $r$  in the FunCATE module, the threshold  $\tau$  in the FuncExtract module, and the number of clusters  $K$  in the PartSelect module, respectively.

TABLE IV  
ACCURACY OF PREDICTED GRASPING GESTURES FOR 6 TASKS AND 18 TOOLS. (FLASHLIGHT: FL, HAMMER: HM, KETTLE: KT, SPATULA: SP, SCISSORS: SC, CUP: CP, DOORHANDLE: DH, BOTTLE: BT, KNIFE: KN, SCREWDRIVER: SD, DRILL: DR, STAPLER: ST, SPRAYBOTTLE: SB, LIGHTSWITCH: LS, MOUSE: MS, PLUG: PG, PLIERS: PL, VALVE: VL, AVERAGE PRECISION: AP)

	FL.	HM.	KT.	SP.	SC.	CP.	DH.	BT.	KN.	SD.	DR.	ST.	SB.	LS.	MS.	PG.	PL.	VL.	AA
Hold	81.82	90.91	58.33	50	-	100	90.91	100	100	100	100	-	-	-	-	-	-	86.37	
Press	-	-	-	-	-	-	-	-	-	-	100	87.5	100	-	-	-	-	96.15	
Click	75	-	100	-	-	-	-	-	-	-	-	-	-	100	100	-	-	94.59	
Clamp	-	-	-	-	-	-	-	-	100	-	-	-	-	-	80	-	-	93.33	
Grip	-	-	-	-	-	100	-	-	-	-	-	-	-	-	-	91.67	-	95.83	
Open	-	-	-	-	-	-	-	-	100	-	-	-	-	-	-	-	100	100	
AA	78.95	90.91	75	50	100	100	90.91	100	100	100	100	87.5	100	100	100	80	91.67	100	91.3

grained, object-related part features. In contrast, our approach, by localizing functional fingers during hand-object interactions, extracts more detailed functional part-level features of objects, enhancing the precision of our localization and facilitating a deeper and more effective transfer of knowledge, leading to superior performance in other metrics.

Additionally, we compare the inference time on a single image, as shown in the Time column of Tab. III. All three methods demonstrate excellent inference speed, with times around  $0.02s$ . Although our method is  $0.0002s$  slower than the fastest method, LOCATE [19], this slight difference is negligible in practical applications. Moreover, unlike LOCATE [19] and Cross-view-AG [20], which only perform the localization task, our method also achieves coarse gesture prediction.

To further validate the performance of our method, we compared it with two fully supervised segmentation methods, PSPNet [54] and DeepLabv3 [55]. Since these methods rely on pixel-level annotations, we trained them using 1,665 pixel-annotated Ego images from the FAH training set. The results demonstrate that our method significantly outperforms these fully supervised methods in both KLD and NSS metrics. Moreover, the inference time of our method is only  $0.0228s$ , which is much faster than that of PSPNet [54] ( $0.216s$ ) and DeepLabv3 [55] ( $0.054s$ ). These results indicate that fully supervised methods perform poorly on small and imbalanced datasets, while our weakly supervised approach achieves superior performance with reduced annotation costs and faster inference efficiency.

**Hyperparameter Analysis.** We further investigate the impact of the parameter  $r$  in the FunCATE module (Fig. 9, left), the threshold  $\tau$  in the FuncExtract module (Fig. 9, middle), and the number of clusters  $K$  in the PartSelect module (Fig. 9, right). It can be observed that the threshold  $\tau$  has no significant impact on the results. Parameters  $r$  and  $K$  are respectively used

to extract fine-grained and coarse-grained interaction features from exocentric data, aiding in affordance localization and gesture prediction. Their final results align with the principles of our algorithm design: an overly large  $r$  captures excessive background noise, while an overly small  $r$  fails to fully capture tool button features due to finger occlusion. When  $K = 3$ , gesture prediction accuracy reaches its highest, as more precise clustering based on the three semantic features-human, object, and background-effectively captures object features.

### C. Result of Coarse Gesture Predictor

Table IV presents the grasping gesture prediction accuracy for six tasks and 18 tools, analyzed from the following three perspectives:

**Overall Accuracy:** The average accuracy across all task-tool combinations is 91.3%, indicating high prediction reliability. Tasks such as “Hold Cup”, “Hold Bottle”, and “Open Bottle” achieve 100% accuracy, while “Hold Spatula” is the lowest at 50%, likely due to indistinct handle features and limited training data (125 samples).

**Average Accuracy per Task:** All six tasks achieve over 85% average accuracy. Notably, the “Hold” task, which spans 11 tools, reaches 86.37%, demonstrating the model’s ability to extract shared features for dexterous manipulation.

**Average Accuracy per Tool:** Tool-wise accuracy varies. “Cup” and “Bottle” achieve 100%, while “Spatula” and “Kettle” yield 50% and 75%, respectively. These lower scores are attributed to subtle and less distinctive interaction features, which challenge the gesture prediction module.

### D. Performance on Common Everyday Tools

We conducted experimental validation on the FAH dataset in real-world scenarios, including both unseen scenes of seen

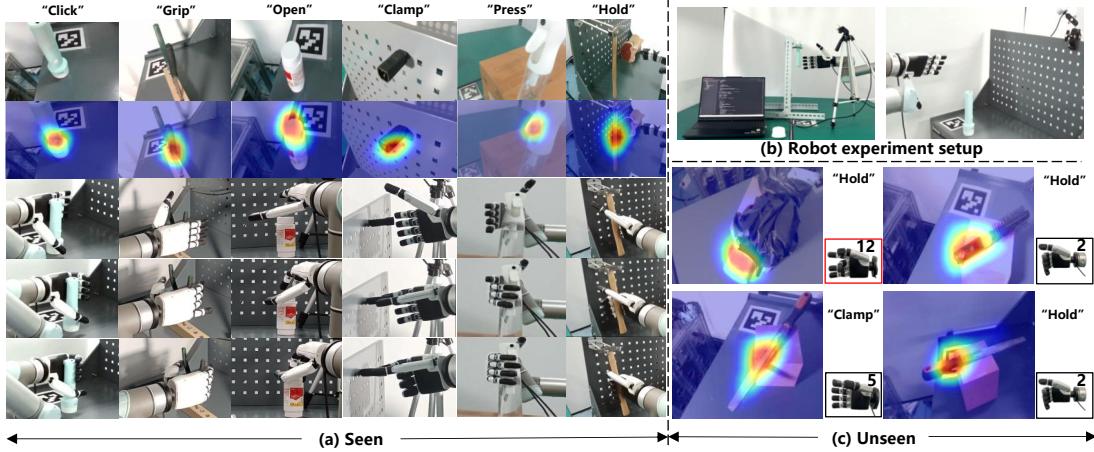


Fig. 10. Experiments in real-world scenarios: (a) Seen categories (rows 1-5: Ego image from camera view, affordance localization, approach based on localization, coarse grasping, fine grasping); (b) Hardware setup with a tool rack (left) and natural placement (right); (c) Four representative unseen categories.

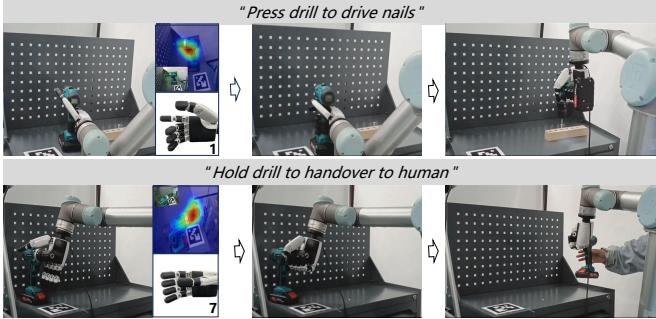


Fig. 11. Our method can dynamically adapt to subsequent different functional manipulations (the green stereoscopic frame is obtained using the 6D pose algorithm [56]).

categories and unseen categories. The hardware setup for the experimental scenarios is shown in Fig. 10 (b): the left side depicts a scene with tools suspended on a tool rack, while the right side shows a more natural placement scenario. Both setups include an Inspire Hand, a UR5 industrial robotic arm, an Intel RealSense D435i camera, a tool holder, and a control computer. The Inspire Hand, a cost-effective anthropomorphic manipulator, features six degrees of freedom: two for the thumb and one for each of the other fingers. Each degree of freedom is driven by a linear motor.

We first demonstrate the complete process—from localization to pre-grasping to functional grasping-on seen categories (unseen instances) in the FAH dataset. As shown in Fig. 10 (a), for different tool instances across our six defined tasks, our algorithm accurately localizes functional regions and predicts corresponding coarse gestures, achieving functional grasping via a post-processing module.

Our method also exhibits generalization on unseen categories, as shown in Fig. 10 (c). For the affordance prediction task, all four unseen task-tool combinations successfully localize functional regions, such as the handle for “Hold Umbrella” and the grip for “Hold Comb.” For the gesture prediction task, except for “Hold Umbrella,” the model accurately predicts reasonable gestures for different task-tool combinations. However, “Hold Umbrella” is incorrectly predicted as a “Clamp”-related gesture (highlighted by the red box in Fig. 10 (c)),

indicating that our gesture prediction network design requires further improvement. For the same tool “Rasp,” our algorithm successfully predicts distinct localizations and gestures based on varying tasks, as shown in the second row of Fig. 10 (c).

Furthermore, we showcase our method’s dynamic adaptability for subsequent functional operations. As shown in Fig. 11, for the same tool, different affordance instructions guide distinct localizations and gestures: “Press Drill” can facilitate subsequent “nailing,” while “Hold Drill” can support “handing to a person.”

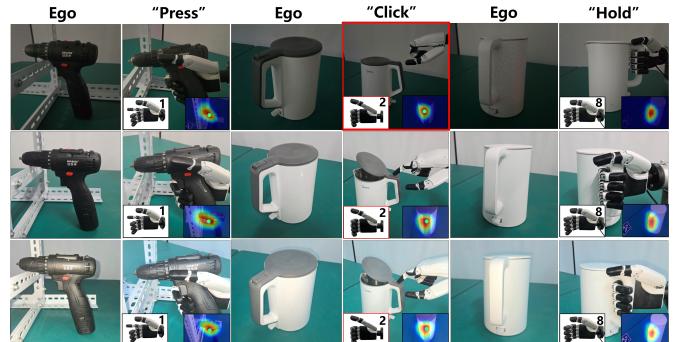


Fig. 12. Experimental results under different lighting conditions for various tools and tasks: Row 1 - dim; Row 2 - normal; Row 3 - bright. In images from even-numbered columns, the bottom left corner shows the predicted coarse hand gesture category, while the bottom right corner indicates affordance localization. Green dots represent functional finger contact points. Red boxes highlight prediction or grasping failure cases.

Secondly, our method demonstrated robustness to lighting variations. As shown in Fig. 12, whether for the complex button-pressing task “Press Drill” (columns 1 and 2) or the simpler “Hold Kettle” task (columns 5 and 6), our approach consistently predicted correct coarse grasping gestures and affordance localizations under dim, normal, and bright lighting conditions. For the task “Click Kettle,” the functional affordance contact region—the kettle’s switch—was correctly localized in all three lighting conditions (see the bottom-right corner of column 4). However, the coarse gesture category was incorrectly predicted as “type2” instead of the correct “type10” in all lighting conditions. Remarkably, despite the

TABLE V

THE SUCCESS RATE OF THE REPRESENTATIVE “*Task Tool*” ACROSS 15 TRIALS IN REAL-WORLD EXPERIMENTS. (POSITIONING SUCCESS RATE: POS., COARSE GESTURE PREDICTION SUCCESS RATE: CG., FUNCTIONAL GRASP SUCCESS RATE: FG., TASK COMPLETION TIME: TCT IN SECONDS.)

	Press DR.	Hold DR.	Hold KT.	Click FL.	Hold HM.	Press SB.
Pos.	66.67	13.33	86.67	66.67	93.33	93.33
CG.	46.67	93.33	100	66.67	93.33	46.67
FG.	26.67	40	86.67	66.67	73.33	46.67
TCT	14	13	13	13	12	12

incorrect coarse gesture prediction, the kettle lid was successfully opened under normal and bright lighting conditions. This success is attributed to our functional finger determination module and the model-based post-processing module, which allowed the functional finger (index finger) to accurately interact with the switch even under the incorrect “*type2*” gesture. This demonstrates the framework’s tolerance for process errors during functional operations.

We also recorded the success rates of localization, coarse gesture prediction, and functional grasping, as well as task completion times, for 6 representative “*Task-Tool*” combinations across 15 real-world experiments. As shown in Tab. V, except for “*Hold Drill*”, all other localization success rates exceeded 50%. The localization success rate for “*Hold Drill*” was only 13%, attributed to limitations of the backbone model DINO-ViT [49]. This model provides part-level features but struggles to effectively extract features from the drill head, which lacks part-level characteristics.

Regarding coarse gesture prediction success rates, “*Hold*” tasks exhibited high success rates exceeding 93.33%, while other tasks showed relatively lower rates. For functional grasping success rates, we observed that, despite occasional errors in localization or coarse gesture prediction, functional grasping could still be completed. For instance, although the localization success rate for “*Hold Drill*” was only 13.33%, the grasping task could still be successfully completed as precise localization is less critical for grasping the drill. Conversely, for “*Task Tool*” combinations with high localization and coarse gesture prediction success rates, occasional lower functional grasping success rates were observed. For example, the “*Press Drill*” task requires precise pressing of the button with the functional finger, posing significant challenges for selecting the end-effector grasping point. Although our model-based coordinate transformation method achieved some success, error propagation prevented precise localization.

Lastly, we recorded the average Task Completion Times (TCT) for six “*Task Tool*” combinations, from model inference to functional grasping completion, as shown in the last column of Tab. V. The task completion times ranged from 12s to 14s, demonstrating relatively stable efficiency across different tasks. The small variation of 1~2s was primarily caused by force feedback-driven adjustments during the transition from coarse to fine-grained grasping. These consistently stable task completion times across diverse task-tool combinations highlight the robustness of our method in adapting to various task scenarios.

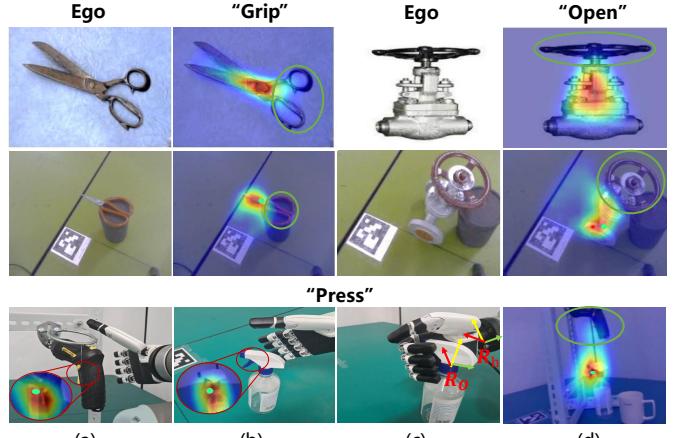


Fig. 13. Presentation of failed cases. The green circle represents the area that should be located, and green dots represent functional finger contact points based on the affordance grounding.

## VII. CONCLUSION AND DISCUSSION

In this work, we propose a weakly supervised method to learn affordance cues from exocentric images of hand-object interactions, which are used to supervise corresponding features in Ego images containing only objects. This enables the localization of functional grasping areas and coarse grasp gestures. Additionally, a model-based post-processing module refines these localizations and gestures to determine wrist-end grasp points and adjust grasps from coarse to fine, ensuring functional grasping conditions are met.

Despite the effectiveness of our method in perception-to-control functional grasping, challenges remain. Fig. 13 highlights failure cases from the data set (first row) and real-world scenarios (second row), showing similar errors. For “*Open Valve*”, localization was below the valve. These errors likely arise from functional finger features in Exo training images that overlap with incorrect regions, suggesting the need to optimize feature selection for complex tool manipulation tasks. In the third row of Fig. 13, failure cases in the “*Press*” task across different tools and scenarios are presented. In (a) and (b), the affordance grounding in the RGB images was generally accurate, but depth extraction failed due to background inclusion. For example, in (a), the extracted depth corresponds to the tool rack, as indicated by the green dots. To address this issue, we plan to improve the localization capability in 3D environments. (c) illustrates a failure caused by the inconsistency between the initial rotation of the hand  $R_h$  and the rotation of the object  $R_o$ . We aim to solve this problem by incorporating rotational affordance. (d) highlights the challenge of object recognition in complex scenes. In a multi-object scenario, we intended to grasp the “*Drill*” within the green box but mistakenly localized on the “*Spraybottle*.” To address this issue, we plan to leverage the multimodal alignment capability of Vision-Language Models (VLMs) to align features from natural language task instructions with those of target objects in the image, enhancing object identification and localization in complex scenes.

In summary, as one of the earliest works to integrate affordance perception with practical dexterous grasping, our method holds significant real-world value. We present a task-

oriented perception-action framework with important applications in various domains. It can enable assistive robots to handle surgical tools in healthcare, support industrial robots in assembly tasks, and facilitate domestic robots in unstructured environments. Our modular, hardware-agnostic approach is adaptable to various robotic platforms and can be enhanced with multimodal data, making it applicable across industries such as agriculture, logistics, and space exploration.

## REFERENCES

- [1] Y. Zhang *et al.*, “FunctionalGrasp: Learning functional grasp for robots via semantic hand-object representation,” *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 3094–3101, 2023.
- [2] M. Kovic, D. Krägic, and J. Bohg, “Learning task-oriented grasping from human activity datasets,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3352–3359, 2020.
- [3] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, “PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes,” in *Proc. RSS*, 2018.
- [4] S. Yu, D. Zhai, Y. Guan, and Y. Xia, “Category-level 6-D object pose estimation with shape deformation for robotic grasp detection,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 1857–1871, 2025.
- [5] E. Corona, A. Pumarola, G. Alenyà, F. Moreno-Noguer, and G. Rogez, “GanHand: Predicting human grasp affordances in multi-object scenes,” in *Proc. CVPR*, 2020, pp. 5030–5040.
- [6] J. Jian, X. Liu, M. Li, R. Hu, and J. Liu, “AffordPose: A large-scale dataset of hand-object interactions with affordance-driven hand pose,” in *Proc. ICCV*, 2023, pp. 14 667–14 678.
- [7] L. Yang *et al.*, “OakInk: A large-scale knowledge repository for understanding hand-object interaction,” in *Proc. CVPR*, 2022, pp. 20 921–20 930.
- [8] P. Mandikal and K. Grauman, “Learning dexterous grasping with object-centric visual affordances,” in *Proc. ICRA*, 2021, pp. 6169–6176.
- [9] H. Jiang, S. Liu, J. Wang, and X. Wang, “Hand-object contact consistency reasoning for human grasps generation,” in *Proc. ICCV*, 2021, pp. 11 087–11 096.
- [10] B. Wu *et al.*, “Generative attention learning: A “general” framework for high-performance multi-fingered grasping in clutter,” *Autonomous Robots*, vol. 44, no. 6, pp. 971–990, 2020.
- [11] H. Li, Y. Zhang, Y. Li, and H. He, “Learning task-oriented dexterous grasping from human knowledge,” in *Proc. ICRA*, 2021, pp. 6192–6198.
- [12] J. J. Gibson, “The theory of affordances,” *Hilldale, USA*, vol. 1, no. 2, pp. 67–82, 1977.
- [13] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele, “Functional object class detection based on learned affordance cues,” in *Proc. ICVS*, 2008, pp. 435–444.
- [14] H. Kjellström, J. Romero, and D. Krägic, “Visual object-action recognition: Inferring object affordances from human demonstration,” *Computer Vision and Image Understanding*, vol. 115, no. 1, pp. 81–90, 2011.
- [15] T. Nagarajan, C. Feichtenhofer, and K. Grauman, “Grounded human-object interaction hotspots from video,” in *Proc. ICCV*, 2019, pp. 8687–8696.
- [16] J. Chen, D. Gao, K. Q. Lin, and M. Z. Shou, “Affordance grounding from demonstration video to target image,” in *Proc. CVPR*, 2023, pp. 6799–6808.
- [17] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “Learning visual affordance grounding from demonstration videos,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 16 857–16 871, 2024.
- [18] Y. Li, T. Nagarajan, B. Xiong, and K. Grauman, “Ego-Exo: Transferring visual representations from third-person to first-person videos,” in *Proc. CVPR*, 2021, pp. 6943–6953.
- [19] G. Li, V. Jampani, D. Sun, and L. Sevilla-Lara, “LOCATE: Localize and transfer object parts for weakly supervised affordance grounding,” in *Proc. CVPR*, 2023, pp. 10 922–10 931.
- [20] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “Learning affordance grounding from exocentric images,” in *Proc. CVPR*, 2022, pp. 2242–2251.
- [21] T. Feix, J. Romero, H.-B. Schmidtmayer, A. M. Dollar, and D. Krägic, “The grasp taxonomy of human grasp types,” *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2016.
- [22] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “Grounded affordance from exocentric view,” *International Journal of Computer Vision*, vol. 132, no. 6, pp. 1945–1969, 2024.
- [23] X. Zhan *et al.*, “OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion,” in *Proc. CVPR*, 2024, pp. 445–456.
- [24] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “One-shot affordance detection,” *arXiv preprint arXiv:2106.14747*, 2021.
- [25] Y. Hasson *et al.*, “Learning joint reconstruction of hands and manipulated objects,” in *Proc. CVPR*, 2019, pp. 11 807–11 816.
- [26] F.-J. Chu, R. Xu, and P. A. Vela, “Learning affordance segmentation for real-world robotic manipulation via synthetic images,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1140–1147, 2019.
- [27] P. Ardón, É. Pairet, R. P. A. Petrick, S. Ramamoorthy, and K. S. Lohan, “Learning grasp affordance reasoning through semantic relations,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4571–4578, 2019.
- [28] R. Xu, F.-J. Chu, C. Tang, W. Liu, and P. A. Vela, “An affordance keypoint detection network for robot manipulation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2870–2877, 2021.
- [29] W. Chen, H. Liang, Z. Chen, F. Sun, and J. Zhang, “Learning 6-DoF task-oriented grasp detection via implicit estimation and visual affordance,” in *Proc. IROS*, 2022, pp. 762–769.
- [30] Y. Song, P. Sun, Y. Ren, Y. Zheng, and Y. Zhang, “Learning 6-DoF fine-grained grasp detection based on part affordance grounding,” *arXiv preprint arXiv:2301.11564*, 2023.
- [31] T. Nguyen *et al.*, “Language-conditioned affordance-pose detection in 3D point clouds,” *arXiv preprint arXiv:2309.10911*, 2023.
- [32] Z. Zhang, H. Luo, W. Zhai, Y. Cao, and Y. Kang, “Bidirectional progressive transformer for interaction intention anticipation,” in *Proc. ECCV*, vol. 15117, 2024, pp. 57–75.
- [33] A. Delitzas, A. Takmaz, F. Tombari, R. W. Sumner, M. Pollefeys, and F. Engelmann, “SceneFun3D: Fine-grained functionality and affordance understanding in 3D scenes,” in *Proc. CVPR*, 2024, pp. 14 531–14 542.
- [34] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “Leverage interactive affinity for affordance learning,” in *Proc. CVPR*, 2023, pp. 6809–6819.
- [35] C.-Y. Chuang, J. Li, A. Torralba, and S. Fidler, “Learning to act properly: Predicting and explaining affordances from images,” in *Proc. CVPR*, 2018, pp. 975–983.
- [36] S. Deng, X. Xu, C. Wu, K. Chen, and K. Jia, “3D affordancenet: A benchmark for visual object affordance understanding,” in *Proc. CVPR*, 2021, pp. 1778–1787.
- [37] W. Tong, M. Zhang, G. Zhu, X. Xu, and E. Q. Wu, “Robust depth estimation based on parallax attention for aerial scene perception,” *IEEE Transactions on Industrial Informatics*, vol. 20, no. 9, pp. 10 761–10 769, 2024.
- [38] W. Tong *et al.*, “Edge-assisted epipolar transformer for industrial scene reconstruction,” *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 701–711, 2025.
- [39] K. W. Tong, Y. Cai, Y.-W. Jie, Y. Duan, Y. Hou, and E. Q. Wu, “Neural rendering and flow-assisted unsupervised multi-view stereo for real-time monocular tracking and scene perception,” *IEEE Transactions on Automation Science and Engineering*, 2025.
- [40] A. Rodriguez, M. T. Mason, and S. Ferry, “From caging to grasping,” *The International Journal of Robotics Research*, vol. 31, no. 7, pp. 886–900, 2012.
- [41] C. Rosales, R. Suárez, M. Gabiccini, and A. Bicchi, “On the synthesis of feasible and prehensile robotic grasps,” in *Proc. ICRA*, 2012, pp. 559–556.
- [42] S. El-Khoury, R. De Souza, and A. Billard, “On computing task-oriented grasps,” *Robotics and Autonomous Systems*, vol. 66, pp. 145–158, 2015.
- [43] R. M. Murray, Z. Li, and S. S. Sastry, *A mathematical introduction to robotic manipulation*. CRC press, 2017.
- [44] W. Shang, F. Song, Z. Zhao, H. Gao, S. Cong, and Z. Li, “Deep learning method for grasping novel objects using dexterous hands,” *IEEE Transactions on Cybernetics*, vol. 52, no. 5, pp. 2750–2762, 2022.
- [45] V. Mayer, Q. Feng, J. Deng, Y. Shi, Z. Chen, and A. Knoll, “FFHNet: Generating multi-fingered robotic grasps for unknown objects in real-time,” in *Proc. ICRA*, 2022, pp. 762–769.
- [46] W. Wei *et al.*, “DVGG: Deep variational grasp generation for dexterous manipulation,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1659–1666, 2022.
- [47] H. Wang, H. Zhang, L. Li, Z. Kan, and Y. Song, “Task-driven reinforcement learning with action primitives for long-horizon manipulation skills,” *IEEE Transactions on Cybernetics*, vol. 54, no. 8, pp. 4513–4526, 2024.
- [48] T. Zhu, R. Wu, J. Hang, X. Lin, and Y. Sun, “Toward human-like grasp: Functional grasp by dexterous robotic hand via object-hand semantic

- representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 521–12 534, 2023.
- [49] M. Caron *et al.*, “Emerging properties in self-supervised vision transformers,” in *Proc. ICCV*, 2021, pp. 9630–9640.
- [50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. CVPR*, 2016, pp. 2921–2929.
- [51] C. Lugaressi *et al.*, “MediaPipe: A framework for building perception pipelines,” *arXiv preprint arXiv:1906.08172*, 2019.
- [52] F. Yang *et al.*, “Task-oriented tool manipulation with robotic dexterous hands: A knowledge graph approach from fingers to functionality,” *IEEE Transactions on Cybernetics*, vol. 55, no. 1, pp. 395–408, 2025.
- [53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. CVPR*, 2009, pp. 248–255.
- [54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. CVPR*, 2017, pp. 6230–6239.
- [55] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.
- [56] B. Wen, W. Yang, J. Kautz, and S. Birchfield, “FoundationPose: Unified 6D pose estimation and tracking of novel objects,” in *Proc. CVPR*, 2024, pp. 17 868–17 879.