International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Learning Precise Affordances from Egocentric Videos for Robotic Manipulation

Gen Li[1] · **Nikolaos Tsagkas**[1] · **Jifei Song**[2] · **Ruaridh Mon-Williams**[1] ·
Sethu Vijayakumar[1] · **Kun Shao**[2] ✉ · **Laura Sevilla-Lara**[1] ✉

Received: date / Accepted: date

Abstract Affordance, defined as the potential actions that an object offers, is crucial for robotic manipulation tasks. A deep understanding of affordance can lead to more intelligent AI systems. For example, such knowledge directs an agent to grasp a knife by the handle for cutting and by the blade when passing it to someone. In this paper, we present a streamlined affordance learning system that encompasses data collection, effective model training, and robot deployment. First, we collect training data from egocentric videos in an automatic manner. Different from previous methods that focus only on the object graspable affordance and represent it as coarse heatmaps, we cover both graspable (e. g., object handles) and functional affordances (e. g., knife blades, hammer heads) and extract data with precise segmentation masks. We then propose an effective model, termed Geometry-guided Affordance Transformer (GAT), to train on the collected data. GAT integrates an innovative Depth Feature Injector (DFI) to incorporate 3D shape and geometric priors, enhancing the model's understanding of affordances. To enable affordance-oriented manipulation, we further introduce Aff-Grasp, a framework that combines GAT with a grasp generation model. For comprehensive evaluation, we create an affordance evaluation dataset with pixel-wise annotations, and design real-world tasks for robot experiments. The results show that GAT surpasses the state-of-the-art by 15.9% in mIoU, and Aff-Grasp achieves high success rates of 95.5% in affordance prediction and 77.1% in successful grasping among 179 trials, including evaluations with seen, unseen objects, and cluttered scenes. Project page: https://reagan1311.github.io/affgrasp.

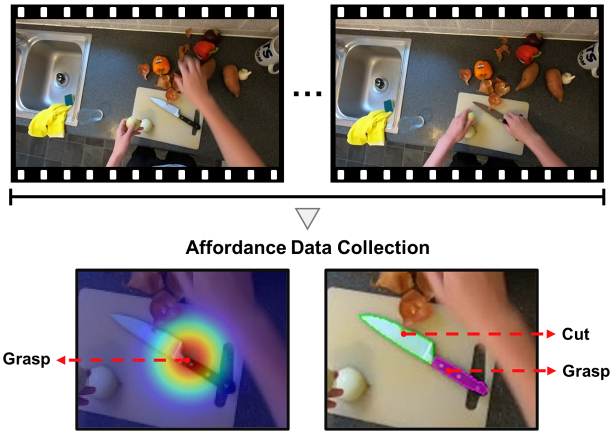
Keywords Affordance learning · Egocentric videos · Human-object interaction · Robotic manipulation

1 Introduction

Humans excel at mastering the use of various tools and selecting the appropriate one based on its function and utility. At a dining table, for instance, we use a knife to cut, a fork to stick, and a spoon to scoop. In addition to recognizing the appearance and semantics of tools, we know how to grasp them and which part to use for the desired function. This capability is driven by affordance, which denotes the potential actions that objects offer [24]. With this understanding, we can effectively manipulate and use a variety of objects, even in unfamiliar environments. Similarly, equipping robots with this actionable knowledge is crucial for intelligent interactions, making it a popular research topic in the fields of computer vision and robotics.

A broad spectrum of work [10, 15, 18, 45, 46, 48, 49] focuses on learning from hand-crafted affordance datasets, which require extensive data collection and costly annotation. In contrast, humans typically acquire knowledge of affordances in a more efficient manner, either through trial-and-error interactions or by observing others. Intuitively, learning through observation is particularly effective, allowing for generalization to objects with similar shapes and appearances. Inspired by this intuition, a number of recent studies [3, 22, 30, 39, 47, 79] have focused on extracting actionable knowledge from videos of humans interacting with objects. These studies reason about object affordances from human videos and represent them as interaction heatmaps. However, two limitations persist in their learning pipeline as illustrated in Figure 1: (1) The focus is only on how humans grasp

✉ Corresponding authors
[1]School of Informatics, University of Edinburgh, Edinburgh, UK
[2]Huawei Noah's Ark Lab, London, UK



(a) Existing work (b) Ours

Fig. 1: Illustration of affordance data collection from videos of human interactions. Existing work [3, 39] collects the graspable affordance in the form of Gaussian heatmaps, whereas we extract both graspable and functional affordances with precise segmentation masks.

objects (graspable affordance), rather than on which part of the tool is being used (functional affordance). (2) Affordances are learned and represented as probabilistic distributions, which are coarse and noisy, making them difficult to apply in real-life situations and susceptible to distractions.

To resolve these limitations, we aim to jointly learn graspable and functional affordances of objects from egocentric videos, focusing on generating precise segmentation maps rather than coarse heatmaps. Specifically, to avoid costly and time-consuming manual annotation, we propose an automated pipeline to collect data without human labor. Given an egocentric video, this pipeline first extracts graspable points on objects from hand-object interactions and functional points from tool-object interactions. Since objects are often occluded by hands or tools during interaction, we identify the pre-contact frame, where contact is about to occur, using an off-the-shelf hand-object detector [64]. We then project the extracted points back to the pre-contact frame through homography [68] or point correspondence. Finally, these graspable and functional points are employed as prompts and fed into the Segment Anything Model (SAM) [33] to obtain part segmentation. We gather training data from two large-scale egocentric video datasets: Epic-kitchens [11] and Ego4d [25].

Although the data are collected in an annotation-free manner, it poses significant challenges for model training. Most training samples have quite low resolution, with cropped areas often comprising only 5% of the original frame. Additionally, factors such as motion blur, occlusion, and homography estimation can make the resulting training samples blurry and noisy. To address these issues, we propose the Geometry-guided Affordance Transformer (GAT), which includes a simple and effective Depth Feature Injector (DFI) to incorporate geometric information during training. DFI allows the model to make predictions in conjunction with 3D geometry, rather than relying solely on blurry appearances. Moreover, we observe that the model yields inferior performance when evaluated on data that significantly differs from the training source domain. To tackle this domain gap, we use the visual foundation model DINOv2 [54] as the image encoder, which has been trained on data from various domains. We keep DINOv2 frozen and employ Low-Rank Adaptation (LoRA) [27] for fine-tuning to enhance the model's generalization ability and prevent overfitting. After training, we combine GAT with a grasp generation model for robotic manipulation, and name this framework Aff-Grasp. Aff-Grasp can adaptively grasp objects based on different task requirements, and identify the functional affordance of objects to complete tasks.

To comprehensively demonstrate the effectiveness of our data collection and model training, we perform evaluations from two perspectives. First, we collect and annotate 721 images from several existing affordance datasets and internet sources, creating a challenging evaluation dataset of great diversity to assess the model's performance. Second, we design a real-world robotic manipulation evaluation with 7 tasks and 34 diverse objects. A task is considered successful if the model makes the correct affordance segmentation, the robot grasps the correct graspable part, and applies the functional part to the target object.

Overall, the contributions of this work can be summarized as follows:

1. **Automated Affordance Data Collection:** We propose an automated pipeline for collecting and annotating affordance data from egocentric human-object interaction videos. Different from previous work, the data are collected with precise segmentation maps for both graspable and functional affordances.
2. **Advanced Affordance Learning Model:** We introduce Geometry-guided Affordance Transformer (GAT) that features an innovative Depth Feature Injector (DFI) to incorporate geometric information. GAT builds on the visual foundation model DINOv2 and employs LoRA for fine-tuning.
3. **Affordance-Oriented Manipulation:** We design a framework, Aff-Grasp, that integrates GAT with a grasp pose generation model to achieve affordance-oriented manipulation. Given a task, Aff-Grasp can localize the most appropriate object (without specifying the object category), grasp the correct part, and utilize its functional part to complete the task.

4. **Comprehensive Vision and Robot Evaluations:** We conduct comprehensive experiments on both static datasets and real robots. A challenging affordance evaluation dataset is created for vision evaluation and affordance-oriented manipulation tasks are designed for robot experiments.

2 Related Work

2.1 Visual Affordance Learning

Affordance learning studies the properties of objects and environments that suggest possible actions. It has garnered considerable attention in the fields of computer vision and robotics. Initial successes in this field were achieved through fully supervised methods [9, 18, 22, 34, 46] that employ convolutional neural networks (CNNs). These methods, however, heavily depend on large-scale, annotated datasets [10, 15, 18, 45, 48, 49], which are costly and time-consuming to produce.

To reduce annotation costs, recent interest has shifted to weakly-supervised approaches, utilizing keypoints [62, 63] and image-level labels [35, 43, 47]. Although these weakly-supervised based methods have made notable progress, they still require substantial manual collection of training data. Distinct from the above methods, we generate the affordance training data in an automatic manner, eliminating the need for manual annotation efforts. Also, we incorporate 3D geometry information to facilitate affordance learning, which has been shown to be beneficial in related affordance literature [46, 55, 69]. Although geometry information like depth maps is not directly available for arbitrary images, recent monocular depth estimation models [5, 31, 57, 58, 77] have become much more reliable through better modeling and data-driven schemes. In this work, we use a state-of-the art depth estimation model, Depth-Anything [77], to produce pseudo depth maps.

2.2 Affordance Learning from Human Videos

An emerging and promising alternative for affordance learning is the automatic extraction of affordance knowledge through the observation of human interactions with objects in natural environments. Given the abundance of human-object interaction video datasets [11–13, 25, 37, 41, 51, 78, 80], recent work has explored how to extract rich affordance-based information from these videos. Liu *el al.* [39] first proposed an automatic pipeline to generate data from egocentric videos for training, and represent affordance as probabilistic distributions in the form of interaction heatmaps. Following this paradigm, VRB [3] further collected trajectory waypoints to estimate the post-contact direction of movement. Robo-ABC [30] created an affordance memory consisting of object images and human contact points, achieving zero-shot generalization via object retrieval and semantic correspondence mapping.

Despite making significant strides in affordance learning, these studies have two limitations. First, the defined affordance is limited to graspable areas of objects, ignoring the important functional parts. Second, the affordance is predicted as coarse heatmaps, which are less accurate when applied in the real world. To address these limitations, we introduce an improved affordance data collection pipeline that localizes both graspable and functional points, and utilizes SAM [33] to produce high-quality masks with point-based prompting.

2.3 Affordance-Oriented Robotic Manipulation

Recent work [29, 59, 65, 72, 73] has achieved task-oriented grasping with the help of vision-language models (VLMs) and large language models (LLMs) [1, 6, 53, 54, 56, 61]. This involves using LLMs to infer parts to be grasped based on task instructions, and VLMs to precisely locate specific parts. Although this pipeline can produce effective and accurate manipulation, the reasoning process requires extra prompt engineering and is time-consuming. In contrast, less attention has been paid to more efficient affordance-oriented grasping that can derive graspable and functional areas without explicitly specifying corresponding object parts. For example, if tasked with slicing bread, an affordance-oriented system can deduce that the serrated edge of a bread knife is appropriate for slicing, while the handle is the correct part to be grasped. One major barrier is the lack of large-scale affordance datasets, compounded by the difficulty in unifying different datasets due to varying annotation standards. Additionally, a recent study [36] highlighted the insufficient granularity of affordance features extracted from existing foundation models, especially when language instructions are involved. For instance, they would associate “cutting” with the entire knife rather than its blade.

To overcome the data limitations, recent methods such as VRB [3] and Robo-ABC [30] have explored the potential of learning affordances from human videos for robotic manipulation. However, VRB generates only coarse Gaussian heatmaps and necessitates additional policy learning to deploy the affordance model in real robots. Robo-ABC relies on point correspondences that can be noisy and susceptible to background variations, and do not necessarily lead to reliable grasp poses. In contrast, we propose a more effective and robust strategy

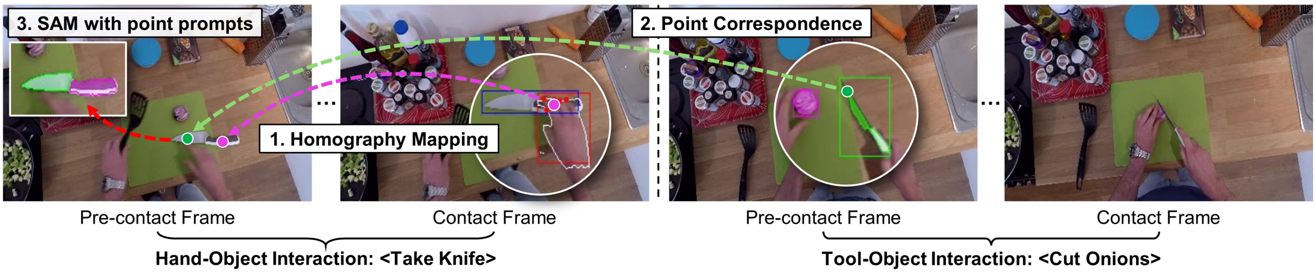


Fig. 2: Illustration of the data collection process from egocentric videos. First, graspable points (depicted in purple) are localized from clips of hand-object interaction and then projected to pre-contact frame by homography. Next, functional points (depicted in green) are identified from tool-object interactions and mapped to the pre-contact frame of hand-object interaction through point correspondence. Lastly, these points are used as prompts for the SAM to obtain affordance masks.

to extract precise affordance masks, and introduce an effective model for affordance learning. When combined with a handful of basic pre-recorded motion primitives (as in [16, 17]) and grasp pose detection models [20, 21, 67], our method enables the utilization of both graspable and functional affordances for robotic manipulation.

3 Method

In this study, our goal is to develop a holistic system that covers data collection, model learning, and robot deployment. To this end, we first develop an automated pipeline to collect images and related affordance annotations from human videos. Next, we propose an effective affordance learning model termed Geometry-guided Affordance Transformer (GAT). GAT is based on a vision transformer [19] using DINOv2 as the encoder [54], and it incorporates a depth feature injector and LoRA layers to facilitate training. Finally, we introduce Aff-Grasp that couples the trained model with an off-the-shelf grasp generation model to achieve affordance-oriented manipulation. Given a task and a cluttered scene, the robot can identify the object with the appropriate affordance, grasp the correct part, and use its functional component on the target object. In Section 3.1, we describe how affordance data are collected from large-scale egocentric videos of human interactions. In Section 3.2, we elaborate on the design of GAT that enables effective affordance learning from collected data. Lastly, in Section 3.3, we explain Aff-Grasp, detailing how it yields affordance-oriented grasp poses for robotic manipulation.

3.1 Data Collection from Egocentric Videos

Given an egocentric video of a human interacting with an object, our aim is to first locate contact points. Human-object interaction videos can generally be categorized into two types: hand-object interaction and tool-object interaction. In hand-object interaction, contact points indicate where the human grasps the object. In tool-object interaction, contact points reveal which part of the tool is used to interact with the target object. These points represent sparse *graspable* and *functional* areas of an object, carrying rich affordance information. As shown in Figure 2, we propose a pipeline to automatically collect these points without manual annotation. The collected points then serve as prompts to produce precise segmentation masks using the Segment Anything Model (SAM) [33].

3.1.1 Graspable Point Localization

Egocentric videos like Epic-Kitchens [11] and Ego4D [25] include timestamped narrations that describe actions and their respective start and end times. Based on these narrations, we first retrieve hand-object interaction clips (associated with actions such as “take” or “hold”) and employ a hand-object detector [64] to generate contact states and hand-object bounding boxes for all frames. Next, we use labeled timestamps to extract the contact frame, which is typically annotated as the start of an action. We conduct an additional hand segmentation in this frame using the hand box as a prompt via EfficientSAM [76]. We then locate the intersection region of the hand mask and object bounding box to sample n contact points $P = \{p_1, p_2, ..., p_n\}$. However, the sampled points are often occluded by hands, and therefore do not accurately represent the graspable affordance area of the object. To collect clean object images free of occlusion, it is necessary to identify the pre-contact frame, i. e., the last frame where the object is fully visible before contact occurs. We utilize the contact states to detect the frame that is closest to the contact frame but without hand-object contact, designating this as the pre-contact frame. Since human motion between adjacent frames is minimal, we follow a similar pipeline to previous studies [30, 39] to project the average position of sampled graspable points to the pre-contact frame by computing a homography transformation, as illustrated by the purple dashed line in Figure 2. Specifically, correspondences between these two frames are detected using the SURF descriptor [4], and the homography is then estimated by sampling at least four pairs of points with the RANSAC [23] algorithm to maximize the number of inliers.

3.1.2 Functional Point Localization

To localize functional points, we first retrieve the relationship between objects and affordances from existing affordance datasets. We then extract related tool-object interaction clips from video narrations based on the object-affordance relationship. For example, most datasets associate affordances “cut” and “grasp” with a knife. After localizing the graspable points from a clip showing a person grasping a knife, we then retrieve the next nearest clip depicting a cutting action with the knife.

However, the tool is often heavily occluded or invisible in the contact frame. Similar to graspable point localization, we need to find the pre-contact frame that shows minimal or no intersection between the tool and the target object. To achieve this, we first employ the same combination of the hand-object detector and EfficientSAM to obtain the bounding box and mask of the hand-held tool. Next, we use an open-vocabulary object segmentation model, GroundedSAM [60], to segment the target object. We then measure the Intersection over Union (IoU) between tool and object bounding boxes in frames prior to the contact frame until the IoU is below a preset threshold. Lastly, we calculate the point distances between masks in this pre-contact frame and extract the point within the tool mask that has the shortest distance to all points in the object mask. An erosion operation is applied to the object mask to ensure that the functional point is inside the object.

Nonetheless, not all object categories have related action clips in the narrations. In such cases, we use the farthest sampling to determine the functional points based on the distance to the grasp points. This simple method also produces accurate functional points, as most tools are designed with graspable and functional parts distributed at opposite ends.

3.1.3 Training Data Generation

After extracting functional points, we first project these points to the pre-contact frame of the hand-object interaction clip where we infer the graspable points. Since the object category remains the same, we compute the point correspondence within object bounding boxes using foundation model features [2], which map the functional point from the tool-object pre-contact frame to the hand-object pre-contact frame (illustrated by the green dashed line in Figure 2). We then label the graspable points as positive and the functional points as negative to obtain the graspable affordance mask. Conversely, the functional affordance mask is generated by reversing the positive and negative labels. Finally, we crop the object images and store them, along with the generated segmentation masks as annotations.

3.2 Geometry-guided Affordance Transformer

The affordance data are collected from egocentric videos without manual labor, but they have two major issues that hinder effective model training. The first issue is the low resolution of the collected images. The object of interests occupy very small areas of the video frames, often resulting in cropped images smaller than 100 pixels in either length or width. The second one is the limited diversity of the training data, characterized by monotonous backgrounds and mostly restricted to indoor scenes. As a result, training a model using typical ImageNet [14] pre-trained representation leads to fairly poor performance (detailed in Section 4.1.3).

To cope with the above issues, we propose an affordance learning architecture called Geometry-guided Affordance Transformer (GAT). The illustration of GAT is presented in Figure 3. It includes a novel Depth Feature Injector (DFI) that integrates geometric priors into image features using pseudo depth maps, a DINOv2 image encoder as feature extractor, and additional LoRA layers for effective fine-tuning.

3.2.1 Depth Feature Injector

We argue that depth maps introduce rich geometric information that can help with foreground-background and part separation. Additionally, they exclude color information, allowing the model to fully focus on the shape information, which is highly relevant to affordances. For instance, graspable parts often consist of shapes like cylinders or spheres, while parts designed for cutting typically feature sharp edges and flat surfaces.

Specifically, we first obtain pseudo depth maps for each training image with a state-of-the-art depth estimation model Depth-Anything [77]. During model training, the pseudo depth map is first encoded into feature maps using a stem block, which contains several standard 3×3 convolution layers as in the ResNet [26]. These feature

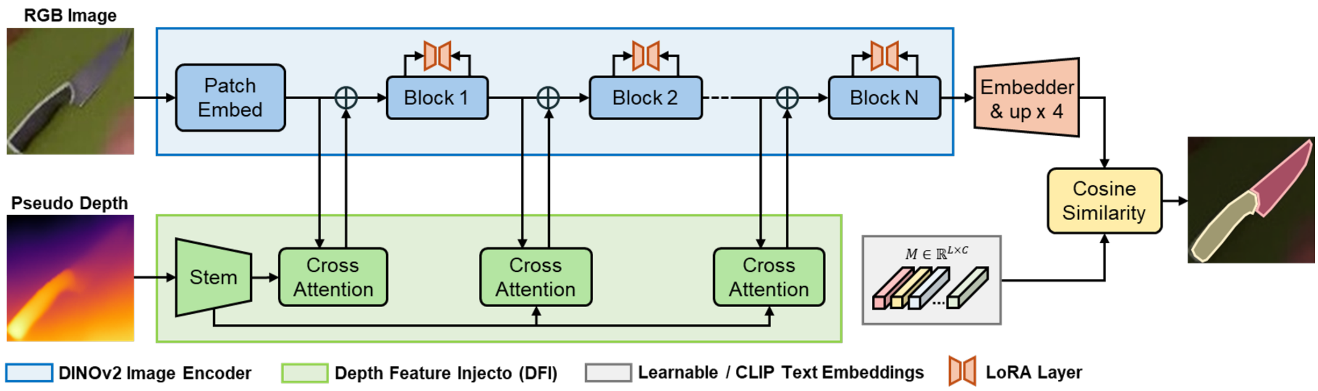


Fig. 3: The architecture of GAT. It consists of a DINOv2 image encoder, a depth feature injector, an embedder, and LoRA layers. The model performs segmentation by computing cosine similarity between upsampled features and learnable or CLIP text embeddings.

maps are then processed by a 1×1 convolution that transforms the channel dimension to match that of the RGB image features.

We divide the whole model into four blocks. At the beginning of each block, the DFI takes the image features $F_i \in \mathbb{R}^{N\times C}$ and depth features $F_d \in \mathbb{R}^{N\times C}$ as input, and outputs updated image features $\hat{F}_i \in \mathbb{R}^{N\times C}$, where N denotes the number of patches. Concretely, DFI contains several cross-attention layers followed by residual connections. In the cross-attention layer, F_i is used as the query, and F_d is adopted as the key and value:

$$Q = \phi_q(F_i),\ \ K = \phi_k(F_d),\ \ V = \phi_v(F_d), \tag{1}$$

$$\hat{F}_i = \beta \cdot \text{softmax}(QK^T/\sqrt{d_k}) \cdot V + F_i, \tag{2}$$

where ϕ is a linear transformation, and d_k is the dimension of the key acting as a scaling factor. Following [8], we set a learnable vector $\beta \in \mathbb{R}^C$, initialized to 0, to balance the output from the cross-attention layer and the image feature. This strategy prevents the image feature from being excessively affected by the depth feature, making the training process more stable.

We observe that DFI constantly brings improvement, even when integrated solely during training (see Section 4.3). This indicates that it can act as a regularization mechanism during training, and can be discarded during inference to speed up the process.

3.2.2 DINOv2 with Low-Rank Adaptation

We notice that directly training from the typical ImageNet pre-trained representation often leads to inferior results. This can be attributed to two primary reasons: First, affordance segmentation focuses on fine-grained object parts, whereas representations trained for image classification emphasize more on global object features [2]. Second, ImageNet pre-trained models exhibit limited diversity, making it challenging to handle data from diverse domains. To address this issue, we employ the self-supervised visual foundation model DINOv2, which has been demonstrated to be highly effective for data-limited affordance learning due to its properties of part-aware representation and part-level correspondence [36]. Furthermore, we introduce LoRA [27] to fine-tune the model without modifying the parameters of the original DINOv2. This strategy helps adaptation across different domains and prevents overfitting. LoRA was originally developed to fine-tune large language models for different downstream tasks. Specifically, it injects trainable rank decomposition matrices to a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ by $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d, k)$. During training, only A and B are trainable, while W_0 remains frozen. This incurs minimal computational cost and memory usage. With LoRA layers, the original forward pass $h = W_0 x$ is modified as:

$$h = W_0 x + \Delta W x = W_0 x + BAx. \tag{3}$$

We apply LoRA to all query, key, and value projection layers, and find that this fine-tuning strategy leads to better transfer learning results.

3.2.3 Classifier and Loss Functions

To speed up inference time for real-world applications, we avoid adding any complex decoder structures. Instead, we process the output feature with an embedder (an MLP), reshape it, and upsample it by a factor of four to increase the resolution $F_{out} \in \mathbb{R}^{C\times \frac{4H}{p} \times \frac{4W}{p}}$, where p is the patch size. Next, we intialize $M \in \mathbb{R}^{L\times C}$ learnable

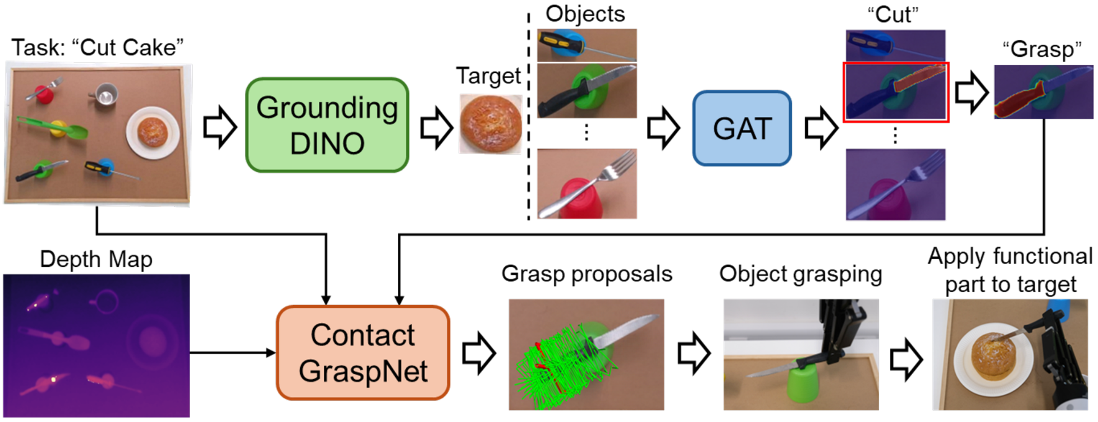


Fig. 4: The framework of Aff-Grasp. It first employs an open-vocabulary detector to locate all objects within the scene, which are then sent to GAT to dertermine if they possess corresponding affordance required for the task. Afterwards, a 6 DoF grasp generation model, utilizing both the object's graspable affordance and the depth map, estimates the potential grasp poses. Finally, the robot executes affordance-specific sequential motion primitives to apply the functional part to the target.

embeddings, where L is the number of affordance categories. We compute the cosine similarity between M and F_{out} to yield the segmentation output, which is then restored to the same size of the input image via bilinear interpolation. Due to the domain gap, we do not add a learnable embedding for the background classification to prevent overfitting. Alternatively, we determine a pixel as background if all its affordance predictions are below a preset threshold τ. Compared to a linear layer and explicit background classifier, the cosine similarity-based segmentation and implicit background prediction are more robust and can effectively improve the performance, as detailed in Section 4.3. In addition, to achieve open vocabulary affordance segmentation, the M can also be replaced with corresponding CLIP text embeddings, as verified in [36].

Since the collected data are highly unbalanced, we utilize a combination of focal loss [38] and dice loss [44] as training objectives:

$$\mathcal{L}_{focal} = -\frac{1}{n}\sum_{i=1}^{n}\big[(1-\hat{y}_i)^{\gamma}\cdot\hat{y}_i\log(y_i) + \hat{y}_i^{\gamma}\cdot(1-\hat{y}_i)\log(1-y_i)\big], \tag{4}$$

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_i^n y_i\hat{y}_i + \epsilon}{\sum_i^n y_i + \sum_i^n \hat{y}_i + \epsilon}, \tag{5}$$

$$\mathcal{L} = \alpha\cdot\mathcal{L}_{focal} + \mathcal{L}_{dice}, \tag{6}$$

where n is the number of valid pixels in output, $\gamma = 2$ is a focusing parameter to balance easy and hard samples, $\epsilon = 1$ is a smoothing factor that prevents division by zero and stabilizes the training, and α is a weighting factor to balance loss values.

3.3 Affordance-Oriented Robotic Manipulation

Our ultimate goal is affordance-oriented robotic manipulation, where given a task and a cluttered scene, the robot can select the object that possesses the related affordance, grasp the correct part, and apply the functional part to the target object to perform desired actions. To achieve this, we propose Aff-Grasp, which integrates GAT to achieve affordance segmentation and transforms the visual affordance to available grasp poses. The framework of Aff-Grasp is shown in Figure 4. Given a task consisting of a verb and a target, such as "cut cake", it first uses an open-vocabulary object detection model [40] to detect the target (cake) and other visible objects. For object other than the target, the input vocabulary is simply set to "objects" for class-agnostic detection. These detected objects are then cropped and sent to GAT to predict affordance areas. The object with the most certain and largest affordance area for the required action (cut) is identified. After that, we extract the graspable affordance area of this object, and generate potential grasp poses within the area, i. e., the knife handle. For grasp pose generation, we select Contact-GraspNet [67] that can produce dense grasp proposals within a specified mask area. Once the object is grasped and lifted, we execute affordance-specific sequential motion primitives to apply the functional affordance area to the target object to complete the task. For the handover task, Aff-Grasp is instructed to find available grasp proposals within the functional affordance area and then pass the graspable part to the human hand.

When CLIP text embeddings are used as classifiers, the action required for a task can be transformed into

Table 1: Statistics on the number of images for each object on the AED.

Total	knife	cup	scissors	hammer	fork	screwdriver	spatula	ladle	pan	shovel	spoon	drill	trowel
721	156	95	78	78	72	59	46	46	31	22	18	10	10

text embeddings for open-vocabulary affordance segmentation [36, 50]. Therefore, unseen affordance vocabularies can also be used at inference, enhancing the model's adaptability and versatility. Overall, Aff-Grasp seamlessly integrates affordance prediction and grasp pose generation to enable robust and flexible robotic manipulation in diverse and cluttered environments.

4 Experiments

In this section, we present experiments from both vision and robot perspectives. Section 4.1 details the vision experiments, where we propose a dataset for evaluation and compare GAT with state-of-the-art models. Section 4.2 describes robotic experiments and compares our grasping framework Aff-Grasp with two related methods that acquire affordance knowledge from human-object interactions. Section 4.3 presents ablation studies that examine the design choices of GAT.

4.1 Vision Experiments

4.1.1 Evaluation Dataset

To evaluate the effectiveness of GAT, we require a diverse and challenging affordance dataset that has consistent object and affordance categories with the collected training data. After carefully inspecting existing datasets, we found that most of them are not compatible with our evaluation requirements. Many datasets either have a small number of categories [10, 49] or are collected in the lab environment with limited diversity [34, 46]. Some datasets contain a large number of images but have coarse keypoint-based annotations [22, 43] or small resolutions [32]. Therefore, we create an Affordance Evaluation Dataset (AED) by manually annotating 721 images collected from several existing affordance datasets and internet resources. AED contains 13 object categories and 8 affordance classes. In Figure 5, we present examples from existing affordance datasets alongside AED, highlighting the necessity for a new evaluation dataset. The statistics regarding the number of images per object category are listed in Table 1.

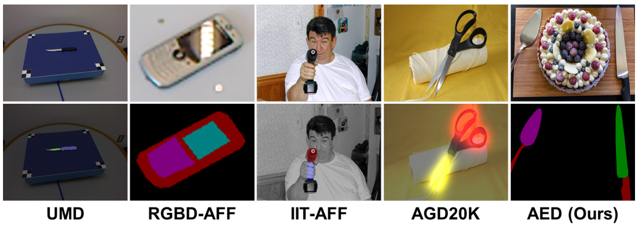


Fig. 5: Examples from existing affordance datasets. UMD [46] is collected in fixed lab environment with coarse annotations. RGBD-AFF [32] has very low resolution and clean background. IIT-AFF [49] includes humans and occluded parts are also annotated. AGD20K [43] is annotated with keypoints and transformed to coarse heatmaps with a Gaussian kernel. In contrast, AED contains natural images with pixel-wise annotations.

4.1.2 Implementation Details

All experiments are conducted on two GeForce RTX 3090 GPUs using the Adamw [42] optimizer, with a learning rate of 1e−3 and batch size 8 for 15 epochs. DINOv2-base is used as the feature extractor. For data collection from egocentric videos, we utilize narratives to filter irrelevant objects and only collect the 13 objects contained in the evaluation dataset. Collected images are first resized to 476 × 476 and then randomly cropped to 448 × 448. Both horizontal and vertical flipping are used for data augmentation. Three metrics, mean intersection-over-union (mIoU), F1-score (F1), and accuracy (Acc), are adopted for evaluation.

4.1.3 Quantitative and Qualitative Comparisons

Table 2 shows the results of different state-of-the-art segmentation approaches on the proposed AED. They use either pre-trained ImageNet backbones to extract feature maps, or obtains representations from visual foundation models like CLIP [56] and DINOv2 [54]. Thus, we divide these models into two sections based on the pre-training strategies. For ImageNet pre-trained models, we employ classical CNN segmentation models such as DeepLabV3+ [7] and PSPNet [81], as well as a transformer-based segmentation model SegFormer [75]. For visual foundation-based models, we choose ZegCLIP [82], DINOv2 [54], ViT-Adapter [8], and OOAL [36]

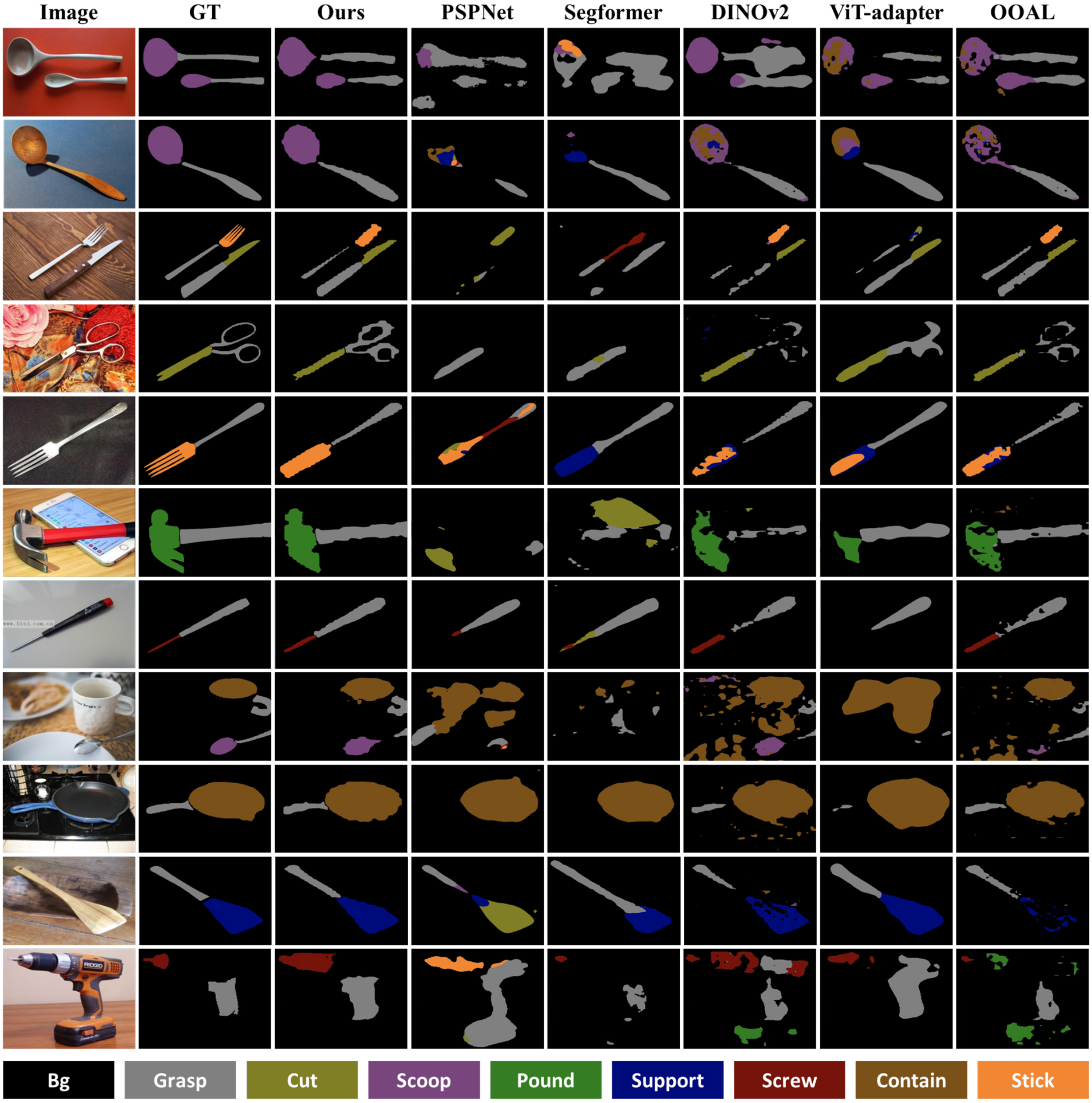


Fig. 6: Qualitative comparison between our approach and other segmentation models on the AED.

Table 2: Quantitative comparison on the AED.

Pre-train	Method	mIoU	F1	Acc
ImageNet	DeepLabV3+ [7]	13.46	22.27	23.05
	PSPNet [81]	16.90	27.32	26.46
	SegFormer [75]	23.72	36.86	37.19
Foundation Models	ZegCLIP [82]	18.33	26.41	25.55
	DINOv2 [54]	46.16	62.49	63.61
	ViT-Adapter [8]	48.36	64.66	65.80
	OOAL [36]	52.72	68.70	65.79
	GAT (Ours)	68.62	81.09	83.51

to compare with GAT, as they represent the state-of-the-art in leveraging visual foundation models for semantic or affordance segmentation tasks. We notice that methods using pre-trained ImageNet backbones generally produce much inferior results compared to those based on foundation models, confirming the huge domain gap between training and evaluation sources. Among the foundation model based approaches, our model GAT significantly outperforms the second best counterpart, OOAL, achieving higher performance in mIoU, F1-score, and accuracy by 15.9%, 12.39%, and 17.72%, respectively.

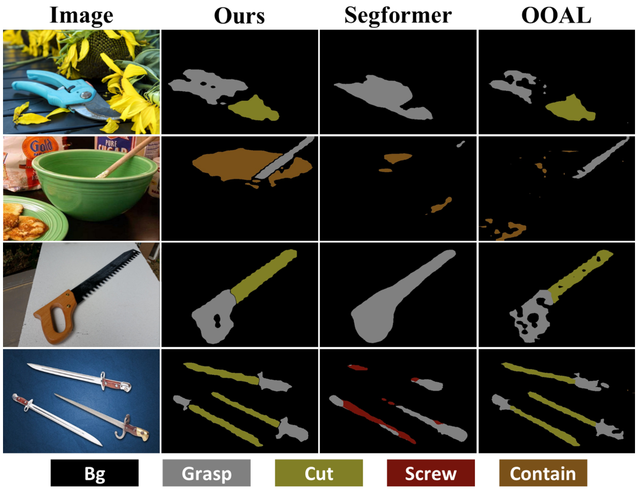
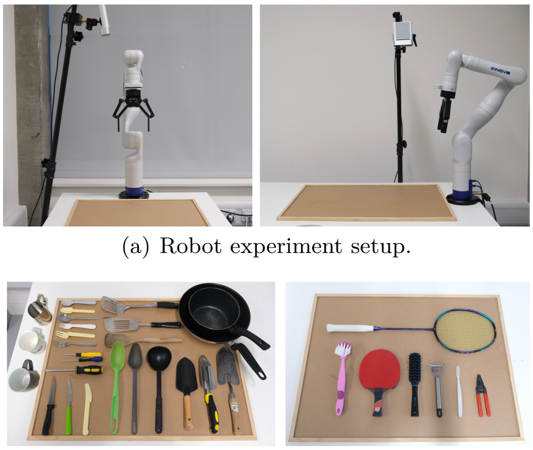


Fig. 7: Qualitative comparison on unseen objects.

Figure 6 depicts the qualitative comparison between our methods and other models. We find that models like PSPNet and Segformer often yield incomplete or incorrect affordance predictions, which may result from the low diversity in the pre-trained ImageNet representation. On the other hand, most models based on DINOv2 can coarsely generate correct affordance prediction map, but often suffer from incomplete part activation and noisy segmentation around object boundaries. In contrast, the results from GAT are part-focused, exhibit well-preserved boundary segmentation, and are capable of handling complex objects like drills. Notably, our results stand out from other counterparts when dealing with images containing multiple objects, as shown in the 1st, 3rd, 6th, and 8th examples in Figure 6.

In addition, we perform a further qualitative comparison on unseen objects to explore the models' generalization ability. As displayed in Figure 7, novel objects such as shears, saw, bowl, and sword are used. It is apparent that Segformer cannot make accurate affordance predictions for these objects. OOAL demonstrates acceptable potential on unseen objects but often produce less confident and inconsistent results. In comparison, GAT shows excellent performance on these out-of-distribution objects with much more complete segmentation maps.



(a) Robot experiment setup.

(b) Experimental objects.

Fig. 8: (a) Experimental setup. (b) Seen (left) and unseen (right) objects used in the experiments.

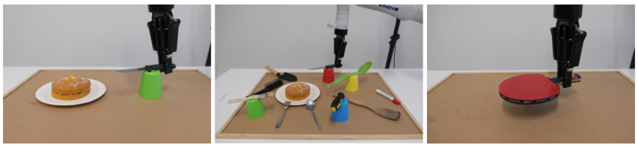


Fig. 9: Illustration of accuracy, robustness, and generalization evaluations. The accuracy evaluation requires the model to recognize the affordance of a single object and execute related task. The robustness evaluation involves accurately selecting a object in a cluttered scene to perform a specified affordance task. The generalization evaluation accesses if the model can reason about the graspable area of unseen objects.

4.2 Robot Experiments

4.2.1 Implementation Details

To evaluate the effectiveness of learned visual affordances, we deploy Aff-Grasp in a 7 DoF Kinova Gen3 robot arm. The arm is equipped with a Robotiq 2F-85 parallel jaw gripper, and a calibrated Azure Kinect RGB-D camera is mounted next to the robot to capture the scene of the workspace (see Figure 8(a)). To enable open-vocabulary affordance recognition, we utilize CLIP text embeddings as the classifier, and discard the DFI to speed up inference time. Real-world experiments are conducted with 34 diverse objects (shown in Figure 8(b)) to evaluate three essential properties: accuracy, robustness, and generalization. We provide illustrations for these experiments in Figure 9, and introduce them in detail as follows:

1. *Accuracy evaluation:* Given a single seen object on the workspace, we evaluate whether the model can recognize correct affordances of the object and perform related affordance task. This evaluation is performed with 24 objects, each of which is repositioned 3 times during the experiment.
2. *Robustness evaluation:* Given multiple seen and unseen objects in a cluttered scene and an affordance task, we evaluate the model's ability to identify which object should be selected to perform the specific task. This requires the model to make robust predictions in the presence of distractors. The evaluation

Table 3: Success rates for accuracy evaluation.

Models	Correct Affordance	Successful Grasp	Successful Interaction
LOCATE [35]	42/72 (58.3%)	33/72 (45.8%)	n/a
Robo-ABC [30]	62/72 (86.1%)	44/72 (61.1%)	n/a
Aff-Grasp (Ours)	70/72 (97.2%)	57/72 (80.6%)	47/72 (65.3%)

Table 4: Success rates for robustness evaluation.

	Cut	Stir	Scoop	Screw	Pour	Stick	Handover	Total
Correct Affordance	8/9	9/9	9/9	8/9	9/9	9/9	17/18	69/72 (95.8%)
Successful Grasp	7/9	6/9	7/9	6/9	7/9	6/9	13/18	53/72 (73.6%)
Successful Interaction	7/9	6/9	5/9	5/9	6/9	5/9	11/18	46/72 (63.9%)

Table 5: Success rates for generalization evaluation and inference time for affordance prediction components.

Models	Correct Affordance	Successful Grasp	Inference Time (s)
LOCATE [35]	20/35 (57.1%)	15/35 (42.9%)	0.0047
Robo-ABC [30]	24/35 (68.6%)	21/35 (60.0%)	12.92
Aff-Grasp (Ours)	32/35 (91.4%)	28/35 (80.0%)	0.0063

is conducted across 7 affordance tasks. Each task is tested with 3 diverse objects, except for the handover task, which is tested with 6 objects, each possessing different functional affordances. Every object is repositioned 3 times during the experiments.

3. *Generalization evaluation:* Given novel objects not encountered during training, we evaluate if the model can still recognize the correct graspable areas. This evaluation assess if the model can generalize the graspable affordance prediction to novel objects, which is a crucial factor in robotic manipulation. It is conducted with 7 novel objects, each repositioned 5 times.

Success rate is adopted as metric and reported from three aspects: correct affordance prediction, successful grasp, and successful interaction. For experiments in cluttered scenes, we assume that only one object is available to complete the target task. We do not perform manipulation policy learning, as it is beyond the focus of this work. Instead, we design motion primitives for each affordance and assume that the operating direction of the tool is known.

4.2.2 Comparison Methods

We choose two relevant methods, LOCATE [35] and Robo-ABC [30], that learn affordances in a similar manner for comparison.

- LOCATE: It is a state-of-the-art affordance grounding model that learns affordances from human-object interaction images using action labels as weak supervision. The method builds on DINO-ViT to identify object parts by clustering visual features from interaction regions of exocentric images, and then transfers the discovered parts to egocentric images for affordance grounding.
- Robo-ABC: It extracts object images and contact points from egocentric videos and store these as an affordance memory. During inference, it first retrieves the most similar objects to the target and then utilizes semantic correspondence from the diffusion model to map the contact point to the current object.

4.2.3 Quantitative and Qualitative Comparisons

The results for the accuracy evaluation are shown in Table 3. It is clear that success rates of Aff-Grasp significantly exceed those of its competitors, with 11.1% higher affordance prediction rate and 19.5% higher in successful grasping compared to Robo-ABC. Also, it is worth mentioning that our success rate for affordance prediction is measured on both graspable and functional affordances, whereas other two methods are measured solely on the graspable affordance. We observe that LOCATE struggles to make accurate predictions in real-world scenarios. Robo-ABC has relatively accurate affordance predictions, but it generates grasp proposals based on point correspondences, which do not represent the most confident

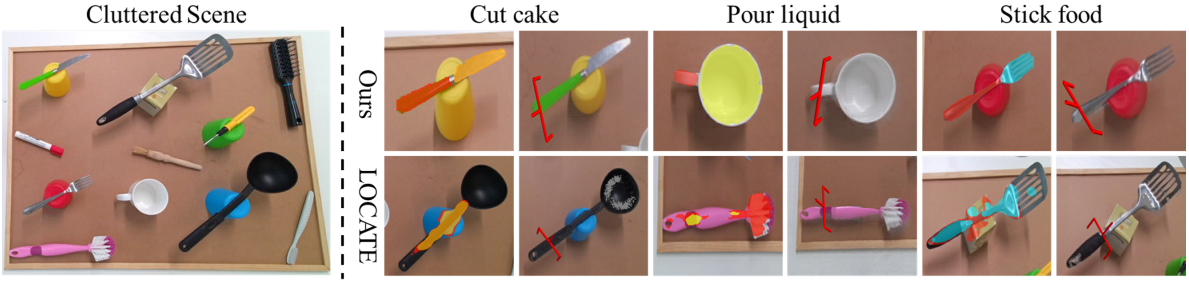


Fig. 10: Qualitative comparison of affordance prediction and final grasp pose for 3D point clouds in the cluttered scene. LOCATE fails to identify related objects for desired tasks, whereas Aff-Grasp can select the correct object with accurate affordance segmentation and is not affected by cluttered scenes.

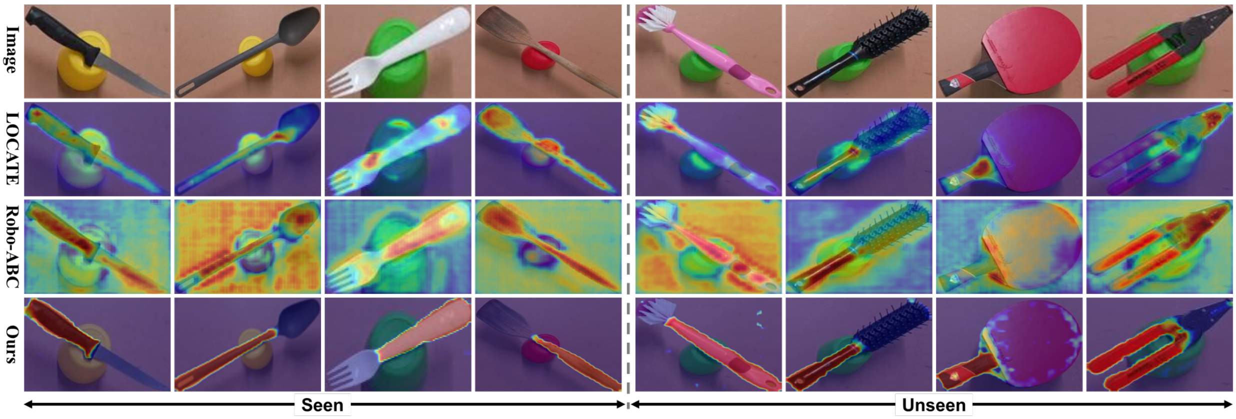
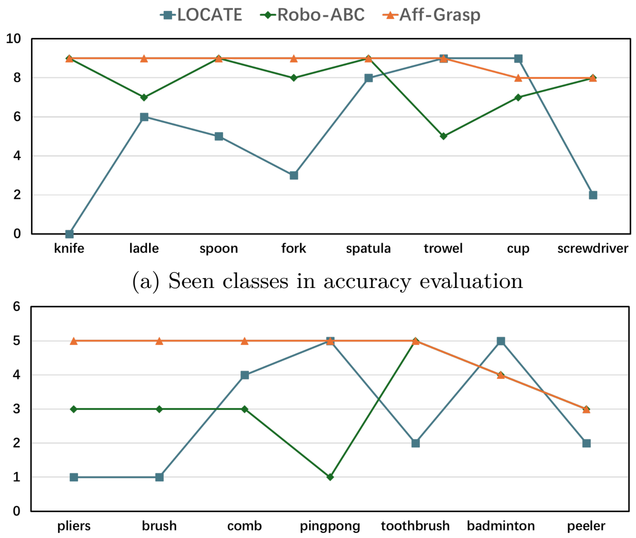


Fig. 11: Qualitative comparison of graspable affordance predictions on seen and unseen object categories.

grasp. Consequently, even though Robo-ABC frequently makes correct affordance predictions, 29% of its proposed grasping points do not lead to successful grasps.

In addition, we note that Aff-Grasp is capable of recognizing the correct affordance in cluttered scenes. As presented in Table 4 for the robustness evaluation, Aff-Grasp achieves a high success rate in affordance prediction, accurately predicting affordances 95% of the time, even in the presence of multiple seen and unseen objects acting as distractors. Table 5 reports results from the generalization evaluation and the inference time for the affordance prediction component of the models. It can be observed that our method is efficient and significantly more accurate in predicting the correct graspable areas for unseen objects, leading to a much higher success rate in grasping. While LOCATE also has a fast inference speed, it fails to accurately infer graspable affordances. In contrast, Robo-ABC's performance on unseen objects is considerably reduced and suffers from a much longer inference time. Although it does not require additional training, its retrieval and correspondence mapping processes are quite time-consuming and computationally expensive, making it less suitable for real-world applications. In Figure 12, we show success rates of individual classes for accuracy and generalization evaluations. It is evident that our results are accurate and stable over all categories, while the results of LOCATE and Robo-ABC show frequent fluctuations.

Qualitative comparisons are illustrated in Figure 10 and 11. In Figure 10, we display affordance segmentation maps and grasp poses for point clouds in a cluttered scene, where both seen and unseen objects serve as interferences (robustness evaluation). It is noted that LOCATE is unable to localize the correct object to execute the specified affordance task, while our model successfully identifies the matching object and predicts accurate segmentation maps. In Figure 11, we present raw predictions of graspable affordance from each model for seen and unseen objects. It can be observed that LOCATE often produces incomplete and wrong predictions. Robo-ABC occasionally makes prediction within the right object part area, but also produces high activation for the background or the entire object. In comparison, Aff-Grasp consistently makes precise segmentation predictions for both seen and unseen objects and is not affected by the background. Furthermore, we display affordance and



(a) Seen classes in accuracy evaluation

(b) Unseen classes in generalization evaluation

Fig. 12: Success rates of correct affordance predictions for each individual object from the accuracy and generalization evaluations. The total numbers of trials are 9 and 5, respectively.

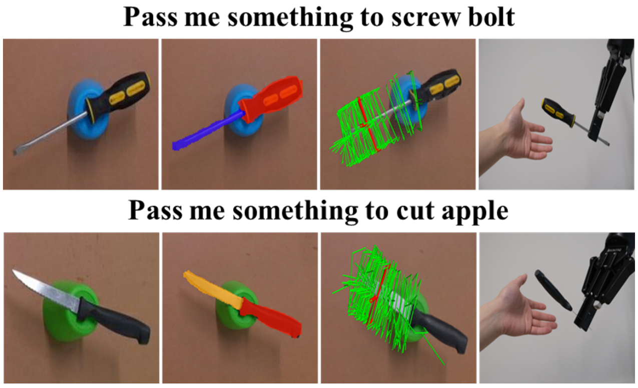

Pass me something to screw bolt
Pass me something to cut apple

Fig. 13: The Aff-Grasp framework can perform the handover task by generating grasp poses within the functional parts of objects, and orienting the graspable parts towards the human hand. Green indicates all potential grasps, while red marks the final selected grasp.

grasp pose predictions for the handover task in Figure 13. When the robot is asked to pass something to the subject for a task, Aff-Grasp generates grasp proposals based on the functional affordance mask and directs graspable parts towards the subject's hand.

4.3 Ablation Study

To explore the impact of each component in our model, we perform ablation experiments on the embedder, loss function, classifiers, designed modules, and hyper-parameters.

Table 6: Ablation results of embedder, loss functions, classifiers, and proposed modules. The baseline model is a DeiT III model with a linear layer and binary cross entropy loss. "w/o bg" means that there is no background classifier. "DFI-training only" denotes that the DFI is only used during training, and discarded at inference.

Methods	mIoU	F1	Accuracy
Baseline - DeiT III	31.02	44.55	35.85
w/ DINOv2	45.45	61.78	70.86
w/ embedder	48.83	65.10	71.07
w/ embedder & up×4	51.41	64.26	67.27
w/ focal loss	50.70	66.97	70.12
w/ focal & dice loss	53.12	69.13	74.55
linear layer w/o bg	54.96	70.50	71.97
cosine similarity	55.52	71.01	71.54
cosine similarity w/o bg	56.70	72.00	71.22
+ DFI-training only	60.15	74.92	79.87
+ DFI	64.66	78.35	79.74
+ LoRA	68.62	81.09	83.51

The ablation results are summarized in Table 6. We first set up a baseline model, which employs a frozen DeiT III [71] backbone that is fully supervised on ImageNet-1k. We then add a simple linear layer for patch-wise classification and utilize binary cross-entropy as the loss function. Based on this baseline, we first explore the impact of the embedder and loss functions. We find that a larger feature map followed by an embedder is beneficial, and the combination of focal loss and dice loss also brings improvements. Then, we analyze the results under different classification schemes, including the linear layer, cosine similarity, and whether to learn a background classifier. It is clear that implicit background prediction leads to better performance. Given the large gap between training and evaluation data, learning a background classifier can easily result in overfitting. Also, employing cosine similarity as the classifier can better utilize the inherent features of DINOv2, producing better results than a linear classifier. Lastly, we investigate the influence of DFI and LoRA. Notably, DFI improves performance significantly by a large margin, with 7.96% and 6.35% increases in mIoU and F1 score. In particular, DFI can also be used solely in training and discarded at inference, thereby improving results without extra computational cost. Additionally, integrating LoRA layers to fine-tune the foundation features is also helpful, leading to a 3.96% improvement in mIoU with marginal additional parameters. In Figure 14, we show the qualitative ablation results to visually examine the effects of DFI and LoRA. The segmentation results indicate that DFI is particularly effective at locating tiny and slender parts,

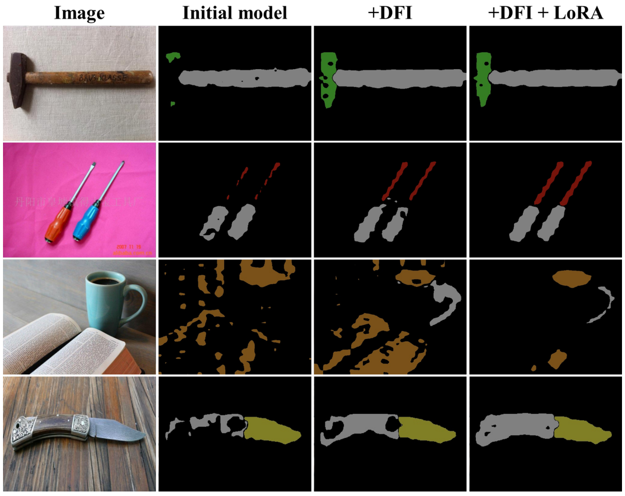


Fig. 14: Visualization of qualitative improvements with DFI and LoRA.

Table 7: Ablation study on different depth representations in DFI.

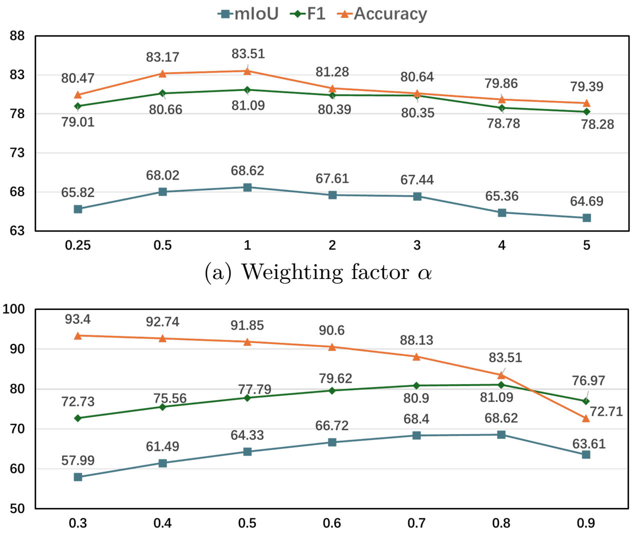
Depth map	mIoU	F1	Accuracy
Color depth (jet)	61.82	76.29	78.34
Color depth (inferno)	62.38	76.57	77.31
Color depth (viridis)	63.92	77.81	78.95
Grayscale depth	64.66	78.35	79.74

while LoRA further enhances performance with refined boundaries and more complete segmentation maps.

To further understand the effectiveness of DFI module, we perform experiments using different depth maps as input. We observe that DFI is more effective with depth representations that have low contrast. As listed in Table 7, the jet colormap, known for high-contrast visual effect, yields the worst results in DFI. In comparison, the less expressive grayscale depth achieves the best performance among other colored counterparts. We speculate that grayscale input focuses more on the geometric information, whereas color depth may introduce noise to some extent. In Table 8, we show the impact of different classification embeddings. The learnable embeddings yield the best results, but lose the ability to reason about unseen affordances. While performance degrades slightly when using CLIP text embeddings as classifiers, this approach retains the ability for open-vocabulary affordance segmentation. Therefore, we use learnable embeddings for vision evaluation and CLIP-L/14 text embeddings for robot experiments.

Table 8: Ablation study on different classification embeddings: learnable or CLIP text embeddings.

Embeddings	mIoU	F1	Accuracy
CLIP-B/32	66.47	79.70	79.31
CLIP-B/16	66.04	79.37	79.15
CLIP-L/14	66.91	80.02	81.09
Learnable embeds	68.62	81.09	83.51



(a) Weighting factor α

(b) Threshold τ for background classification

Fig. 15: Ablation study on hyper-parameters.

Finally, we conduct experiments with different hyerparameter settings, focusing on the threshold τ for background classification and the weighting factor α in the loss function. As presented in Figure 15, the model obtains the highest performance in mIoU and F1-score with a weighting factor α of 1. For the background classification threshold τ, a smaller value leads to higher accuracy, as only confident predictions are counted as foreground. In this case, only mIoU and F1 score can truly reflect the performance. We thus choose 0.8 as the default threshold.

5 Conclusion

In this paper, we present a streamlined affordance learning system that integrates data collection, model training, and robot deployment. Specifically, we first collect training samples with segmentation masks as annotations from videos of humans interacting with common objects. To effectively train on the collected data, we introduce an affordance learning model named Geometry-guided Affordance Transformer (GAT). GAT features a depth feature injector that incorporates geometric and shape information, which is relevant and beneficial for affordance understanding. Building on GAT, we develop a framework, Aff-Grasp, that facilitates affordance-oriented manipulation. Aff-Grasp enables robots to select the desired object and grasp the correct part without explicitly

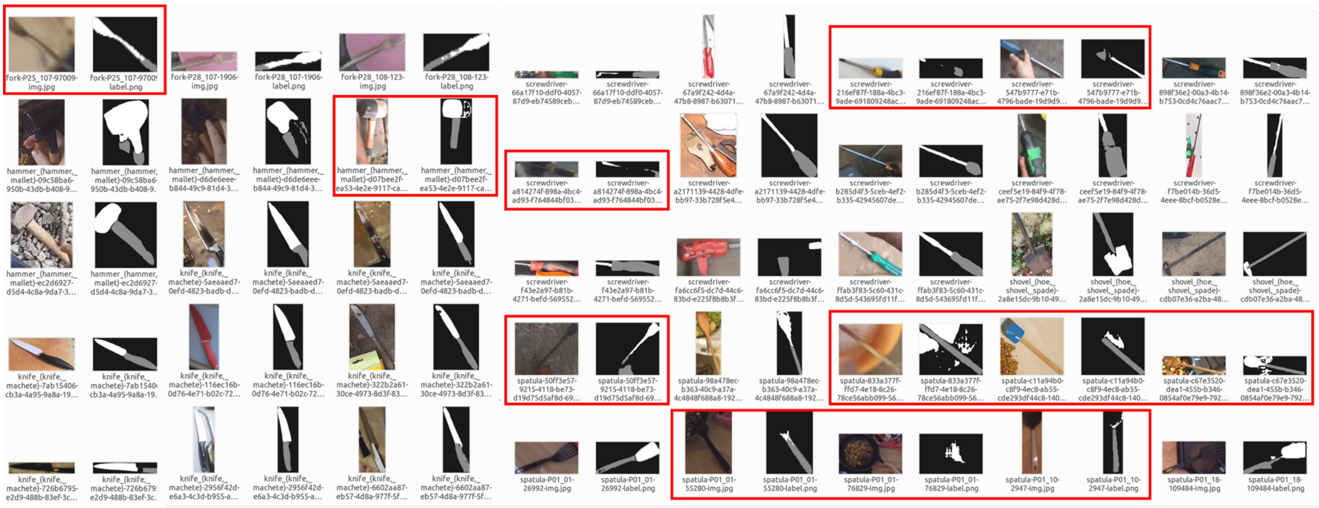

fork-P25_107-97009-img.jpg
fork-P25_107-9700(label.png
fork-P28_107-1906-img.jpg
fork-P28_107-1906-label.png
fork-P28_108-123-img.jpg
fork-P28_108-123-label.png
hammer_(hammer,_mallet)-09c58ba6-950b-43db-b408-9...
hammer_(hammer, mallet)-09c58ba6-950b-43db-b408-9.
hammer_(hammer,_mallet)-d6de6eee-b844-49c9-81d4-3...
hammer_(hammer,_mallet)-d6de6eee-b844-49c9-81d4-3...
hammer_(hammer,_mallet)-d07bee2f-ea53-4e2e-9117-ca...
hammer_(hammer,_mallet)-d07bee2f-ea53-4e2e-9117-ca...
hammer_(hammer,_mallet)-ec2d6927-d5d4-4c8a-9da7-3...
hammer_(hammer, mallet)-ec2d6927-d5d4-4c8a-9da7-3..
knife_(knife,_machete)-5aeaaed7-0efd-4823-badb-d...
knife_(knife,_machete)-5aeaaed7-0efd-4823-badb-d...
knife_(knife,_machete)-5aeaaed7-0efd-4823-badb-d...
knife_(knife,_machete)-5aeaaed7-0efd-4823-badb-d...
knife_(knife,_machete)-7ab15406-cb3a-4a95-9a8a-19...
knife_(knife,_machete)-7ab1540(cb3a-4a95-9a8a-19.
knife_(knife,_machete)-116ec16b-0d76-4e71-b02c-72...
knife_(knife,_machete)-116ec16b-0d76-4e71-b02c-72...
knife_(knife,_machete)-322b2a61-30ce-4973-8d3f-83...
knife_(knife,_machete)-322b2a61-30ce-4973-8d3f-83...
knife_(knife,_machete)-726b6795-e2d9-488b-83ef-3c...
knife_(knife,_machete)-726b679! e2d9-488b-83ef-3c.
knife_(knife,_machete)-2956f42d-e6a3-4c3d-b955-a...
knife_(knife,_machete)-2956f42d-e6a3-4c3d-b955-a...
knife_(knife,_machete)-6602aa87-eb57-4d8a-977f-5f...
knife_(knife,_machete)-6602aa87-eb57-4d8a-977f-5f...
screwdriver-66a17f10-ddf0-4057-87d9-eb74589ceb...
screwdriver-66a17f10-ddf0-4057-87d9-eb74589ceb...
screwdriver-67a9f242-4d4a-47b8-8987-b63071...
screwdriver-67a9f242-4d4a-47b8-8987-b63071...
screwdriver-216ef87f-188a-4bc3-9ade-691809248ac...
screwdriver-216ef87f-188a-4bc3-9ade-691809248ac...
screwdriver-547b9777-e71b-4796-bade-19d9d9...
screwdriver-547b9777-e71b-4796-bade-19d9d9...
screwdriver-898f36e2-00a3-4b14-b753-0cd4c76aac7...
screwdriver-898f36e2-00a3-4b14-b753-0cd4c76aac7...
screwdriver-a814274f-898a-4bc4-ad93-f764844bf03...
screwdriver-a814274f-898a-4bc4-ad93-f764844bf03...
screwdriver-a2171139-4428-4dfe-bb97-33b728f5e4...
screwdriver-a2171139-4428-4dfe-bb97-33b728f5e4...
screwdriver-b285d4f3-5ceb-4ef2-b335-42945607de...
screwdriver-b285d4f3-5ceb-4ef2-b335-42945607de...
screwdriver-ceef5e19-84f9-4f78-ae75-2f7e98d428d...
screwdriver-ceef5e19-84f9-4f78-ae75-2f7e98d428d...
screwdriver-f7be014b-36d5-4eee-8bcf-b0528e...
screwdriver-f7be014b-36d5-4eee-8bcf-b0528e...
screwdriver-f43e2a97-b81b-4271-befd-569552...
screwdriver-f43e2a97-b81b-4271-befd-569552...
screwdriver-fa6cc6f5-dc7d-44c6-83bd-e225f8b8b3f...
screwdriver-fa6cc6f5-dc7d-44c6-83bd-e225f8b8b3f...
screwdriver-ffab3f83-5c60-431c-8d5d-543695fd11f...
screwdriver-ffab3f83-5c60-431c-8d5d-543695fd11f...
shovel_(hoe,_shovel,_spade)-2a8e15dc-9b10-49...
shovel_(hoe,_shovel,_spade)-2a8e15dc-9b10-49...
shovel_(hoe,_shovel,_spade)-cdb07e36-a2ba-48...
shovel_(hoe,_shovel,_spade)-cdb07e36-a2ba-48...
spatula-50ff3e57-9215-4118-be73-d19d75d5af8d-69...
spatula-50ff3e57-9215-4118-be73-d19d75d5af8d-69...
spatula-98a478ec-b363-40c9-a37a-4c4848f688a8-192...
spatula-98a478ec-b363-40c9-a37a-4c4848f688a8-192...
spatula-833a377f-ffd7-4e18-8c26-78ce56abb099-56...
spatula-833a377f-ffd7-4e18-8c26-78ce56abb099-56...
spatula-c11a94b0-c8f9-4ec8-ab55-cde293df44c8-140...
spatula-c11a94b0-c8f9-4ec8-ab55-cde293df44c8-140...
spatula-c67e3520-dea1-455b-b346-0854af0e79e9-792...
spatula-c67e3520-dea1-455b-b346-0854af0e79e9-792...
spatula-P01_01-26992-img.jpg
spatula-P01_01-26992-label.png
spatula-P01_01-55280-img.jpg
spatula-P01_01-55280-label.png
spatula-P01_01-76829-img.jpg
spatula-P01_01-76829-label.png
spatula-P01_10-2947-img.jpg
spatula-P01_10-2947-label.png
spatula-P01_18-109484-img.jpg
spatula-P01_18-109484-label.png

Fig. 16: Screenshot of the collected data. Noisy annotations are highlighted with red bounding boxes.

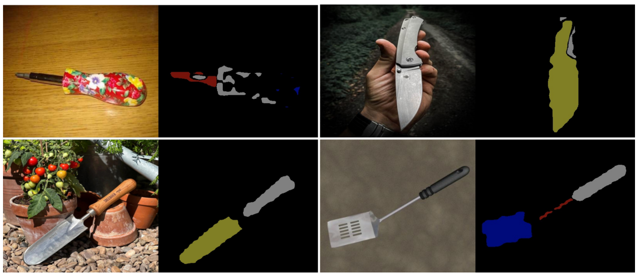


Fig. 17: Failure cases. The model fails to recognize objects with complex texture and confuses parts with similar shapes and appearances.

specifying the object category. To demonstrate the effectiveness of our data collection process and the proposed model, we perform evaluations from both vision and robot perspectives. Extensive experiments show consistent and robust performance, demonstrating the effectiveness of the entire system from data collection to model training and robot deployment.

5.1 Limitations

Despite achieving good performance in static datasets and real-world scenes, it is worth noting that there are several limitations in this work.

Quality of Collected Data: The quality of the collected data is affected by a variety of factors. On the one hand, occlusion, motion blur, poor lighting conditions, inaccurate narrations, and unpredictable subject behavior from the video data can lead to noisy results. On the other hand, the hand-object detector and the open-vocabulary object detection model can produce incorrect predictions, further affecting the usability of the data. To mitigate these issues, we first add some constraints to reduce the error rate, such as setting high thresholds to filter out uncertain predictions. We then visualize all data samples and manually remove those with completely wrong annotations. Figure 16 displays a screenshot of the collected data. Notably, the proposed data collection pipeline is not perfect and the annotations of many samples are noisy and incomplete. Nevertheless, we retain these noisy data to assess the model's performance in this challenging situation.

Model Weakness: The model prediction can be susceptible to complex texture. As shown in Figure 17, the model fails to make correct predictions when the target objects have complex textures or packaging. Also, the model sometimes confuses object parts with similar materials and shapes. For example, the head of a trowel is incorrectly recognized as having a "cutting" affordance.

Grasp and Interaction Success Rate: Due to issues such as depth measurement errors, partial point clouds, unreliable grasp poses, and robot self-collision, correct affordance prediction does not guarantee a successful grasp, and a successful grasp does not always result in effective tool-object interaction. Since the focus of this paper is primarily on visual affordance prediction, we did not fine-tune the grasp generation model, nor did we perform policy learning to improve the grasp and interaction success rate.

5.2 Discussion

Affordance vs. Part. One may argue that parts are more direct and explicit instructions than affordances, as actions or verbs are often more abstract than semantics or nouns. A spectrum of recent work [28, 52, 59, 70] also utilize open-vocabulary part segmentation models [66, 74] and large language models [1, 53] to specify the desired grasping parts for robots. However, understanding object affordances holds great significance for embodied intelligence. Firstly, human's instructions are typically high-level and abstract. For example, we would instruct a robot to "cut an apple for me", rather than specifying "grasp the knife handle and cut the apple with the knife blade". Therefore, affordance understanding helps in the interpretation of natural instructions from humans. Second, reasoning about object parts from task instructions using large language models is time-consuming. A direct understanding of affordances can streamline the process by allowing robots to infer actionable areas from high-level instructions without extensive part-based prompting and reasoning. Thus, affordance-based approaches contribute to more intuitive and efficient interactions between robots and their environments, aligning more closely with how humans naturally communicate and perform tasks.

Video Datasets. Although this work collects affordance data from egocentric videos, we observe that the same pipeline can also be applied to exocentric human-object interaction videos. This flexibility highlights the robustness and adaptability of our approach in different visualization perspectives. Egocentric videos provide a first-person viewpoint, which is highly beneficial for capturing the user's direct interactions with objects, allowing for a more intimate and precise understanding of affordances. On the other hand, exocentric videos, which capture interactions from a third-person perspective, can offer a comprehensive view of the context in which interactions occur.

Additionally, video datasets collected in simple or laboratory environments [41, 78, 80] are preferable for ensuring high accuracy and usability of the training data. These controlled settings typically offer good lighting, background uniformity, and clear object boundaries, providing consistent and reliable data.

Potential Applications. Our method can accurately infer affordances of various common tools, making it highly suitable for use in manufacturing settings. In such environments, robots need to select, grasp, and use different tools for tasks such as assembly, maintenance, or inspection. By adapting to different shapes and orientations, our method enhances operational efficiency and reduces the need for human intervention.

Moreover, our approach improves affordance understanding, fostering more intuitive interactions between humans and robots. This advancement makes robots better collaborators in shared environments, especially in collaborative processes involving human-object handovers. For instance, in assembly lines where humans and robots work together, our method enables robots to interact with human workers by correctly grasping and passing tools or components. This not only improves workflow efficiency but also ensures safer and more coordinated collaboration.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. ECCVW What is Motion For (2022)
3. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from human videos as a versatile representation for robotics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13,778–13,790 (2023)
4. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding **110**(3), 346–359 (2008)
5. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
7. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818 (2018)
8. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 (2022)
9. Chuang, C.Y., Li, J., Torralba, A., Fidler, S.: Learning to act properly: Predicting and explaining affordances from images. arXiv preprint arXiv:1712.07576 (2017)
10. Chuang, C.Y., Li, J., Torralba, A., Fidler, S.: Learning to act properly: Predicting and explaining affordances from images. In: CVPR (2018)
11. Damen, D., Doughty, H., Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: The epic-kitchens dataset: Collection, challenges and baselines. IEEE Transactions on Pattern Analysis & Machine Intelligence (01), 1–1 (2020)

12. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV) (2018)
13. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision pp. 1–23 (2022)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
15. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: CVPR (2021)
16. Di Palo, N., Johns, E.: Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. arXiv preprint arXiv:2402.13181 (2024)
17. Di Palo, N., Johns, E.: On the effectiveness of retrieval, alignment, and replay in manipulation. IEEE Robotics and Automation Letters (2024)
18. Do, T.T., Nguyen, A., Reid, I.: Affordancenet: An end-to-end deep learning approach for object affordance detection. ICRA (2018)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
20. Fang, H.S., Wang, C., Fang, H., Gou, M., Liu, J., Yan, H., Liu, W., Xie, Y., Lu, C.: Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. IEEE Transactions on Robotics (2023)
21. Fang, H.S., Wang, C., Gou, M., Lu, C.: Graspnet-1billion: A large-scale benchmark for general object grasping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11,444–11,453 (2020)
22. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2Vec: Reasoning Object Affordances from Online Videos. CVPR (2018)
23. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
24. Gibson, J.J.: The ecological approach to visual perception. (1979)
25. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. arXiv preprint arXiv:2110.07058 (2021)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
28. Huang, H., Lin, F., Hu, Y., Wang, S., Gao, Y.: Copa: General robotic manipulation through spatial constraints of parts with foundation models. arXiv preprint arXiv:2403.08248 (2024)
29. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models. In: CoRL (2023)
30. Ju, Y., Hu, K., Zhang, G., Zhang, G., Jiang, M., Xu, H.: Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. arXiv preprint arXiv:2401.07487 (2024)
31. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
32. Khalifa, Z., Shah, S.A.A.: A large scale multi-view rgbd visual affordance learning dataset. In: 2023 IEEE International Conference on Image Processing (ICIP), pp. 1325–1329. IEEE (2023)
33. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
34. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International journal of robotics research **32**(8), 951–970 (2013)
35. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: Locate: Localize and transfer object parts for weakly supervised affordance grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10,922–10,931 (2023)
36. Li, G., Sun, D., Sevilla-Lara, L., Jampani, V.: One-shot open affordance learning with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3086–3096 (2024)
37. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)

38. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988 (2017)
39. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: CVPR (2022)
40. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
41. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21,013–21,022 (2022)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
43. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning affordance grounding from exocentric images. CVPR (2022)
44. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV), pp. 565–571. Ieee (2016)
45. Mur-Labadia, L., Martinez-Cantin, R., Guerrero, J.J.: Bayesian deep learning for affordance segmentation in images. arXiv preprint arXiv:2303.00871 (2023)
46. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. ICRA (2015)
47. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8688–8697 (2019)
48. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Detecting object affordances with convolutional neural networks. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2765–2770. IEEE (2016)
49. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Object-based affordances detection with convolutional neural networks and dense conditional random fields. In: IROS (2017)
50. Nguyen, T., Vu, M.N., Vuong, A., Nguyen, D., Vo, T., Le, N., Nguyen, A.: Open-vocabulary affordance detection in 3d point clouds. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5692–5698. IEEE (2023)
51. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9226–9235 (2019)
52. van Oort, T., Miller, D., Browne, W.N., Marticorena, N., Haviland, J., Suenderhauf, N.: Open-vocabulary part-based grasping. arXiv preprint arXiv:2406.05951 (2024)
53. OpenAI: Gpt-4v(ision) system card (2023). URL https://cdn.openai.com/papers/GPTV%20System%20Card.pdf. Accessed: [date]
54. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
55. Qian, S., Chen, W., Bai, M., Zhou, X., Tu, Z., Li, L.E.: Affordancellm: Grounding affordance from vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 7587–7597 (2024)
56. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748–8763. PMLR (2021)
57. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ICCV (2021)
58. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022)
59. Rashid, A., Sharma, S., Kim, C.M., Kerr, J., Chen, L.Y., Kanazawa, A., Goldberg, K.: Language embedded radiance fields for zero-shot task-oriented grasping. In: 7th Annual Conference on Robot Learning (2023)
60. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
61. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10,684–10,695 (2022)

62. Sawatzky, J., Gall, J.: Adaptive binarization for weakly supervised affordance segmentation. In: ICCVW (2017)
63. Sawatzky, J., Srikantha, A., Gall, J.: Weakly supervised affordance detection. CVPR (2017)
64. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9869–9878 (2020)
65. Shen, W., Yang, G., Yu, A., Wong, J., Kaelbling, L.P., Isola, P.: Distilled feature fields enable few-shot language-guided manipulation. In: CoRL (2023)
66. Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., Yan, Z.: Going denser with open-vocabulary part segmentation. arXiv preprint arXiv:2305.11173 (2023)
67. Sundermeyer, M., Mousavian, A., Triebel, R., Fox, D.: Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. 2021 IEEE International Conference on Robotics and Automation (ICRA) (2021)
68. Szeliski, R., et al.: Image alignment and stitching: A tutorial. Foundations and Trends® in Computer Graphics and Vision **2**(1), 1–104 (2007)
69. Thermos, S., Potamianos, G., Daras, P.: Joint object affordance reasoning and segmentation in rgb-d videos. IEEE Access **9**, 89,699–89,713 (2021)
70. Tong, E., Opipari, A., Lewis, S., Zeng, Z., Jenkins, O.C.: Oval-prompt: Open-vocabulary affordance localization for robot manipulation through llm affordance-grounding. arXiv preprint arXiv:2404.11000 (2024)
71. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: European Conference on Computer Vision, pp. 516–533. Springer (2022)
72. Tsagkas, N., Rome, J., Ramamoorthy, S., Mac Aodha, O., Lu, C.X.: Click to grasp: Zero-shot precise manipulation via visual diffusion descriptors. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2024)
73. Wang, Y., Li, Z., Zhang, M., Driggs-Campbell, K., Wu, J., Fei-Fei, L., Li, Y.: D^{3}fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. arXiv:2309.16118 (2023)
74. Wei, M., Yue, X., Zhang, W., Kong, S., Liu, X., Pang, J.: Ov-parts: Towards open-vocabulary part segmentation. arXiv preprint arXiv:2310.05107 (2023)
75. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12,077–12,090 (2021)
76. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16,111–16,121 (2024)
77. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10,371–10,381 (2024)
78. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: OakInk: A large-scale knowledge repository for understanding hand-object interaction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
79. Yoshida, T., Kurita, S., Nishimura, T., Mori, S.: Text-driven affordance learning from egocentric vision. arXiv preprint arXiv:2404.02523 (2024)
80. Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., Lu, C.: Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 445–456 (2024)
81. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890 (2017)
82. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11,175–11,185 (2023)