2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

RoboPCA: Pose-centered Affordance Learning from Human Demonstrations for Robot Manipulation

Zhanqi Xiao, Ruiping Wang and Xilin Chen

Abstract— **Understanding spatial affordances—comprising the contact regions of object interaction and the corresponding contact poses—is essential for robots to effectively manipulate objects and accomplish diverse tasks. However, existing spatial affordance prediction methods mainly focus on locating the contact regions while delegating the pose to independent pose estimation approaches, which can lead to task failures due to inconsistencies between predicted contact regions and candidate poses. In this work, we propose RoboPCA, a pose-centered affordance prediction framework that jointly predicts task-appropriate contact regions and poses conditioned on instructions. To enable scalable data collection for pose-centered affordance learning, we devise Human2Afford, a data curation pipeline that automatically recovers scene-level 3D information and infers pose-centered affordance annotations from human demonstrations. With Human2Afford, scene depth and the interaction object's mask are extracted to provide 3D context and object localization, while pose-centered affordance annotations are obtained by tracking object points within the contact region and analyzing hand–object interaction patterns to establish a mapping from the 3D hand mesh to the robot end-effector orientation. By integrating geometry–appearance cues through an RGB-D encoder and incorporating mask-enhanced features to emphasize task-relevant object regions into the diffusion-based framework, RoboPCA outperforms baseline methods on image datasets, simulation, and real robots, and exhibits strong generalization across tasks and categories.**

I. INTRODUCTION

Humans excel at manipulating various objects in unstructured environments, including locating task-related interaction regions and selecting appropriate contact poses. This ability is largely ascribed to the profound understanding of spatial affordances [1], [2]. As robots and humans operate in the same workspace, a comprehensive understanding of spatial affordances is crucial for enhancing robotic manipulation capabilities across diverse object categories and tasks.

Affordance had already been extensively studied before its application to robotic manipulation, which is often represented as masks or heatmaps [3], [4], [5]. While such representation can indicate potential interaction regions on objects, it lacks the precise spatial localization required for accurate robot manipulation. To address this, recent works propose representing affordances in terms of contact points and post-contact trajectories [6]. In this representation, the interaction region is denoted by a 2D pixel to specify where the robot should make contact with the object, while the post-contact trajectories are represented by a 2D vector on the image to indicate the direction and movement the robot should execute after making contact. Building on this representation, subsequent works have made many improvements, such as introducing object retrieval or leveraging the reasoning capability of foundation models for generalization across categories [7], [8], [9], [10], [11]. However, contact points and post-contact trajectories alone do not specify the manipulator pose for task execution. Although the final pose could be obtained by filtering the grasp candidates from independent pose estimation approaches [12], [13] based on the predicted contact point, the inconsistency between predicted contact points and grasp candidates may lead to suboptimal or even failed executions. We argue **pose-centered affordances**, represented jointly by the **contact points** and the corresponding **contact poses**, provide a more coherent and expressive formulation for robotic manipulation. By unifying contact localization and pose estimation, this formulation reduces the inconsistency between predicted contact points and grasp candidates and provides a principled basis for generating reliable manipulation strategies.

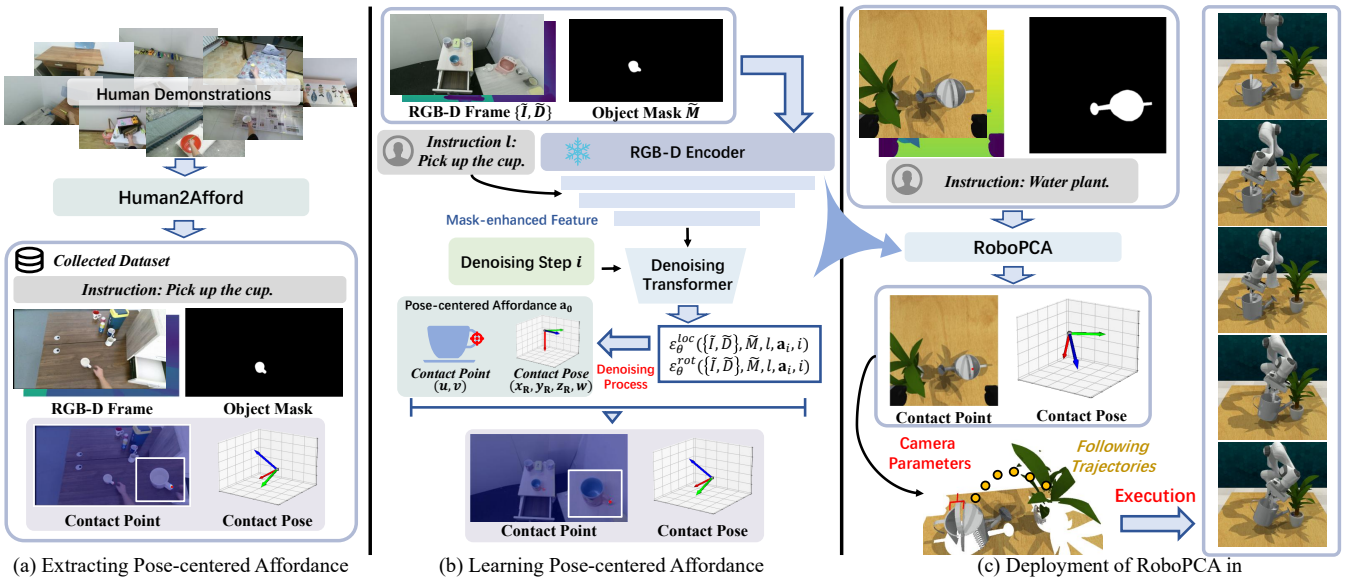
Similar to other tasks in robot learning, learning pose-centered affordances that support the manipulation of arbitrary objects requires large-scale data. While recent works, such as DROID [14], have gathered large-scale robotic demonstrations via human teleoperation, this approach remains difficult to scale to new environments or task demands. In contrast, human demonstrations provide a promising data source, not only due to the scale and diversity of scenes and tasks, but also because they naturally capture dynamics relevant to object interactions. Nevertheless, the absence of 3D information and low-level action labels limits their utility for pose-centered affordance learning, particularly for extracting contact poses.

In this work, we aim to enable pose-centered affordance learning from a large amount of unlabeled human demonstrations, tackling two key questions: (1) How can pose-centered affordances be extracted from unlabeled human demonstrations? (2) How to effectively learn pose-centered affordances from collected datasets? To answer the first question, we devise **Human2Afford**, a data curation pipeline that automatically recovers scene-level 3D information and extracts pose-centered affordances from human demonstrations. With each identified pre-contact frame and contact frame pair, Human2Afford first recovers the depth information and extracts the interaction object's mask in the pre-contact frame

*This work is partially supported by Beijing Municipal Natural Science Foundation Nos. L257009, L242025, and Natural Science Foundation of China under contracts Nos. 62495082, 62461160331.

The authors are with the Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China. {zhanqi.xiao}@vipl.ict.ac.cn, {wangruiping,xlchen}@ict.ac.cn

Corresponding author: Ruiping Wang.



(a) Extracting Pose-centered Affordance from Human Demonstrations

(b) Learning Pose-centered Affordance Based on RoboPCA

(c) Deployment of RoboPCA in Manipulation Tasks

Fig. 1. Overview of our pipeline. (a) Pose-centered affordance annotations and complementary scene information are extracted from human demonstrations with Human2Afford for pose-centered affordance learning. (b) RoboPCA builds upon a diffusion framework to predict pose-centered affordances, an RGB-D encoder is used to effectively capture both geometry and appearance cues, and mask-enhanced features are incorporated to emphasize task-relevant object regions. (c) The predicted pose-centered affordances are transformed into 6-DoF poses using camera parameters, guiding the robot to complete the task.

to provide 3D context and object localization to facilitate pose-centered affordance learning. To obtain pose-centered affordance annotations, we first recover the contact pose by analyzing the hand–object interaction patterns to establish a mapping between the estimated 3D hand mesh in the contact frame and the robot end-effector's orientation, and then derive the contact point by tracking object pixels within the contact region in the contact frame back to the pre-contact frame. We collected 10K human–object interaction images with pose-centered affordance annotations and complementary scene information using Human2Afford. Based on the collected dataset, we propose **RoboPCA** for the second question, which is a **P**ose-**C**entered **A**ffordance prediction framework that jointly predicts task-appropriate contact point and corresponding contact pose given instructions. RoboPCA builds upon a diffusion framework to predict pose-centered affordances. To effectively capture both geometry and appearance cues, it leverages a state-of-the-art RGB-D encoder [15] that integrates color and depth information. Additionally, mask-enhanced features are incorporated to emphasize task-relevant object regions, improving the model's ability to localize interaction points and accurately infer corresponding contact poses. The overview of our pipeline is shown in Fig. 1.

We evaluate RoboPCA through extensive experiments on image datasets, simulation, and real-world scenarios. It outperforms baseline methods in contact point prediction precision and manipulation task success rates, achieving improvements of 18.6% on AGD20K [16] evaluated on object categories feasible for robotic manipulation, 38.5% on RLBench [17], and 24.9% in real-world experiments, which demonstrates that RoboPCA not only achieves higher precision in predicting contact points but also provides more reliable and effective guidance for manipulation tasks. We also demonstrate its compatibility with robotic data.

II. RELATED WORK

A. Visual Affordance Learning

Affordance focuses on determining where and how to interact with diverse objects based on the visual inputs and instructions. One line of work learns affordances from annotated datasets [18], [19], which is often prohibitively expensive. To alleviate the labeling cost, some studies learn affordances by exploring effective interaction in simulated environments [20], [21]; however, acquiring diverse virtual assets also incurs substantial cost. Others leverage large-scale models to automatically annotate contact regions [4], [5], but these annotations are typically performed on single-frame images or static assets and thus fail to capture the dynamic information about how to interact with objects.

In contrast, human demonstrations have gained increasing attention as a more general source for affordance learning, as they naturally encode interaction dynamics and are abundantly available online, offering rich scene and task diversity. Building upon this formulation proposed by [6], which represents affordance as contact points and post-contact trajectories by a 2D pixel and a 2D vector on the image, object retrieval has been employed to enable cross-category transfer [7], [8], and knowledge from large-scale models has been leveraged to enhance the robustness of contact point prediction [9], [10], [11]. With respect to post-contact trajectories, recent studies have extended the 2D vector representation into point-based trajectories to capture curved or non-linear motions that cannot be expressed in a single vector, thereby offering a more faithful description of complex interaction dynamics [22], [23]. However, these

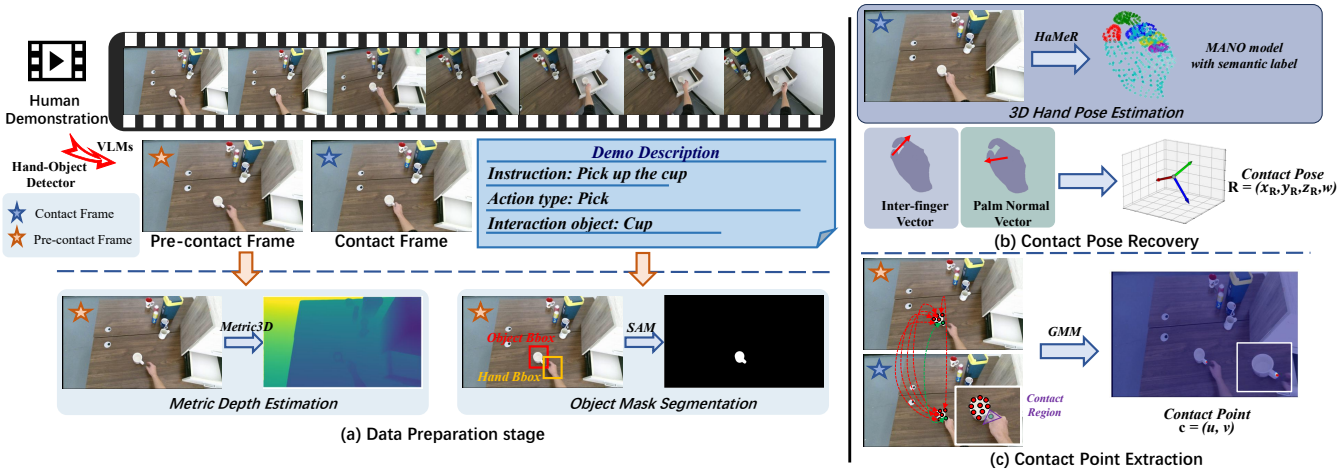


Fig. 2. Overview of Human2Afford. (a) Given a human demonstration, we identify the demo description and extract key frames using a hand–object detector and VLMs. Depth and the interaction object mask are then obtained via metric depth estimation and segmentation. (b) Using the 3D hand mesh from a hand pose estimator, we extract the contact pose based on the inter-finger vector and palm normal. (c) Object points are tracked from the pre-contact to the contact frame, and points within the inter-finger contact region are modeled with GMM to extract the contact point.

approaches require an independent pose estimation module to generate the final contact pose by filtering, which could lead to suboptimal or failed execution due to inconsistencies between predicted contact points and grasp candidates. In contrast, our model jointly predicts contact points and corresponding poses to provide a more consistent representation for robot manipulation.

B. Robot Learning from Humans

Human demonstrations offer a natural and abundant source of supervision for robotic learning, as they inherently capture the dynamics of interaction. With the availability of large-scale video datasets [24], [25], [26], recent studies have explored utilizing human videos to facilitate robot learning. One line of approaches learns visual representations from human videos and leverages the pre-trained visual encoders to facilitate policy network training [27], [28], [29], [30]. Another line of research focuses on learning reward functions from human videos [31], [32], [33]. Additionally, some works leverage the motion attributes extracted from human videos, such as wrist trajectories or 3D hand poses, to guide policy learning or to serve as intermediate supervision for manipulation tasks [34], [35], [36]. Building upon these directions, affordance provides an explicit and interpretable representation that can naturally transfer from human to robot embodiment. In this work, our model provides a principled basis for generating reliable, context-aware manipulation strategies through learning pose-centered affordances that jointly predict contact points and contact poses directly from human demonstrations.

C. Diffusion Models in Robotics

Diffusion models have emerged as a powerful paradigm for modeling complex data distributions through iterative denoising, and have also demonstrated strong potential in robotic learning as effective frameworks for policy learning [37], [38], [39], [40], [41]. Diffusion Policy [38] provides a general framework for generating robot action trajectories via a conditional denoising diffusion process. Building on this, subsequent works have explored enhancing trajectory generation by incorporating reward signals to guide the denoising process [37], while other approaches adopt a more factorized policy learning framework that unifies action keypose prediction and trajectory diffusion generation for learning robot manipulation from demonstrations [40], [41]. However, these approaches often suffer from limited generalization ability, making it difficult to transfer across diverse objects and environments. Instead of predicting full trajectories, our model uses diffusion models for pose-centered affordance learning. Conditioned on 3D scenes' information and instructions, it jointly generates contact points and poses that generalize across diverse tasks and object types.

III. METHOD

A. Problem Formulation

We aim to learn a pose-centered affordance prediction model $\mathbf{a} = \pi(\{\tilde{I}, \tilde{D}\}, l, \tilde{M})$ from human demonstrations, where $\{\tilde{I}, \tilde{D}\}$ denotes an RGB-D frame (with image $\tilde{I}$ and depth $\tilde{D}$), l represents the language instruction and $\tilde{M}$ is the target object's mask relevant to the instruction. The depth frame could be obtained either from a depth sensor or a metric-depth estimation foundation model [42], and the target object mask can be easily acquired through integrating an open-vocabulary object detection method [43] and segmentation foundation models [44]. We formulate the pose-centered affordance representation $\mathbf{a}$ in terms of contact points $\mathbf{c}$, represented as 2D points in pixel space, and end-effector orientations $\mathbf{R}$, represented as quaternions in 3D space relative to the camera coordinate frame. Specifically, $\mathbf{a} = \{\mathbf{c}, \mathbf{R}\}$, where $\mathbf{c} = (u, v)$ and $\mathbf{R} = (w, x_\mathbf{R}, y_\mathbf{R}, z_\mathbf{R})$, with $(w, x_\mathbf{R}, y_\mathbf{R}, z_\mathbf{R})$ denoting the quaternion components. Using the camera intrinsic parameters $\mathbf{K} = (f_x, f_y, c_x, c_y)$ and the depth value $\tilde{D}_{u,v}$ at pixel (u, v), the 3D contact position $\mathbf{p} = (x, y, z)$ can be computed as:

$$z = \tilde{D}_{u,v}, \quad x = (u - c_x) \cdot \frac{z}{f_x}, \quad y = (v - c_y) \cdot \frac{z}{f_y}. \tag{1}$$

The final 6-DoF pose, which the robot uses to interact with the target object under the given instruction, is then represented as $\tau = (\mathbf{p}, \mathbf{R}) = (x, y, z, w, x_{\mathbf{R}}, y_{\mathbf{R}}, z_{\mathbf{R}})$. To enhance clarity in subsequent discussions, we will use $\mathbf{R}$ to directly represent the contact pose.

B. Pose-centered Affordance Extraction from Human Videos

Human videos are often captured with moving monocular cameras, where the camera pose and depth per frame are unknown. The absence of low-level action labels and 3D scene information impedes robots from acquiring manipulation skills directly from human demonstrations. To enable learning pose-centered affordances from human videos, we devise **Human2Afford** that automatically recovers scene-level 3D information and extracts pose-centered affordances from human demonstrations, as shown in Fig. 2. Here, we introduce the key components in our data curation pipeline.

Data Preparation. Given a video clip V consisting of T frames i.e. $V = \{I_1, \ldots, I_T\}$, depicting human–object interactions (e.g., a person picking up a cup), we first prompt a vision-language model (VLM), specifically Gemini-2.0-Flash [45], to identify the action depicted and the category of the interacting object. To extract pose-centric affordances from human videos, we first identify the contact frame where the interaction occurs and the pre-contact frame where the target object remains unoccluded. We utilize a widely-adopted hand-object interaction detector [46] to identify the interaction state in each frame and obtain the corresponding hand bounding boxes. Since significant changes in camera pose between the pre-contact frame and the contact frame can hinder 3D pose recovery due to the lack of camera parameter labels, we further leverage Gemini to examine frames around the critical interval of interaction state changes, filtering out false positives and selecting contact and pre-contact frames $\{I_c, I_p\}$ that are temporally close to satisfy the viewpoint consistency requirement. We then employ a metric-depth foundation model [42] to recover the depth $\hat{D}$ of I_p, and use GroundingDINO [43] and SAM2 [44], guided by hand bounding box priors, to obtain the interacting object's mask in I_c, which is then projected back onto I_p, denoted as $\hat{M}$.

Contact Pose Recovery. Under the assumption that there is no significant change in camera pose between I_p and the I_c, we first leverage a 3D hand pose estimator [47] to estimate the hand pose in I_c. With the hand mesh given by [47], we propose a heuristic method that establishes a mapping from the recovered human hand pose to the robot end-effector orientation. Since human-object interactions predominantly involve the thumb, index, and middle fingers, we identify the finger pair that primarily applies force to the object by jointly analyzing the inter-finger mesh distances and their spatial relationships with the object mask $\hat{M}$. Based on the inter-finger vector of the selected finger pair $\hat{\mathbf{v}}_{fp}$ and the averaged normal vector of the palm region $\hat{\mathbf{n}}_{palm}$, we recover the robot contact pose $\mathbf{R}_c$ from the human hand pose, here $\mathbf{R}_c$ is aligned with the orientation $\mathbf{R}$ in the I_p, denoted as:

$$\mathbf{R} = \mathbf{R}_c = \mathscr{F}(\hat{\mathbf{v}}_{fp}, \hat{\mathbf{n}}_{palm}). \tag{2}$$

Contact Point Extraction. To extract the contact point c, we employ SpaTracker [48], an off-the-shelf dynamic point tracker, to automatically track the interaction object from I_p to I_c, providing temporally consistent object localization even under occlusion. We then analyze the overlap between the object points and the hand mesh to identify points within the area formed by the thumb, index, and middle fingers in I_c. These points are back-projected to I_p, denotes as $\{\mathbf{c}_i\}^N$. Treating $p(\mathbf{c})$ as the distribution over $\{\mathbf{c}_i\}^N$, we fit a Gaussian mixture model (GMM) with parameters (μ_k, σ_k) by maximizing the likelihood over all $\mathbf{c}_i$:

$$p(\mathbf{c}) = \underset{\mu_1,\ldots,\mu_K,\sigma_1,\ldots\sigma_K}{argmax} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathscr{N}(\mathbf{c}_i \mid \mu_k, \sigma_k). \tag{3}$$

The final contact point in I_p is represented as the average of the Gaussian means:

$$\mathbf{c} = \frac{1}{K} \sum_{k=1}^{K} \mu_k \tag{4}$$

C. Pose-centered Affordance Learning

RoboPCA is trained as a conditional diffusion probabilistic model to infer pose-centered affordance $\mathbf{a} = \{\mathbf{c}, \mathbf{R}\}$ given the RGB-D frame of the scene $\{\tilde{I}, \tilde{D}\}$, target object mask $\hat{M}$, and a language instruction l through iterative denoising process. We represent rotations $\mathbf{R}$ using the 6D rotation representation of [49] to avoid the discontinuities of the quaternion representation. A variance schedule $\{\beta_i \in (0,1)\}_{i=1}^N$ is associated with the diffusion process, which defines how much noise is added at each diffusion step. Given a sample $\mathbf{a}_0$, the noise version of $\mathbf{a}_0$ at step i can be written as $\mathbf{a}_i = \sqrt{\bar{\alpha}_i}\mathbf{a}_0 + \sqrt{1-\bar{\alpha}_i}\varepsilon$, where $\varepsilon \sim \mathscr{N}(0,1)$ is a sample noise from a Gaussian distribution with the same dimension as $\mathbf{a}_0$, $\bar{\alpha}_i = \prod_{j=1}^{i} \alpha_i$ and $\alpha_i = 1 - \beta_i$.

RoboPCA models a learned gradient of the denoising process with a denoising transformer $\hat{\varepsilon} = \varepsilon_\theta(\mathbf{a}_i; i, \{\tilde{I}, \tilde{D}\}, l, \tilde{M})$ that takes the noisy pose-centered affordance $\mathbf{a}_i$, diffusion step i, and conditioning information from the RGB-D frame of current scene $\{\tilde{I}, \tilde{D}\}$, object mask $\tilde{M}$ and language instruction l as input, to predict the noise component $\hat{\varepsilon}$.

At each diffusion step i, we convert the visual observation of the scene $\{\tilde{I}, \tilde{D}\}$ and noised pose-centered affordance estimate $\mathbf{a}_i$ to a set of tokens, where each token is represented as a latent embedding and a position in pixel coordinates. Since human demonstrations lack camera pose annotations, we incorporate geometric information, which is crucial for affordance understanding, by leveraging a state-of-the-art RGB-D encoder [15] to encode each RGB-D frame. Moreover, with the object mask $\tilde{M}$, we deploy the same encoder to the masked RGB-D frame to enhance the model's perception of task-relevant object regions. The features from both the full and masked frames are then concatenated to obtain the mask-enhanced features of the scene. The noisy estimate $\mathbf{a}_i$ of the clean pose-centered affordance $\mathbf{a}_0$ is transferred to a latent embedding vector with an MLP, and the language task instruction is mapped to

Fig. 3. Qualitative results on AGD20K. The label ★ ★ ★ ★ ★ indicate the predicted contact points of different methods.

language tokens using a pre-trained CLIP language encoder. We fuse tokens of scene and $\mathbf{a}_i$ by applying relative self-attentions among all tokens, and additionally fuse language tokens using cross-attentions to get the final conditional tokens. We use the rotary positional embeddings [50] to encode relative positional information in attention layers. The conditional tokens is then fed into MLPs to predict the noise $\varepsilon_\theta^{loc}(\{\tilde{I},\tilde{D}\},\tilde{M},l,\mathbf{a}_i,i)$ and $\varepsilon_\theta^{rot}(\{\tilde{I},\tilde{D}\},\tilde{M},l,\mathbf{a}_i,i)$ added to $\mathbf{a}_0$'s contact point $\mathbf{c}_0$ and contact pose $\mathbf{R}_0$.

During training stage, we randomly sample a diffusion step i and add noise $\varepsilon=(\varepsilon^{loc},\varepsilon^{rot})$ to ground truth pose-centered affordance $\mathbf{a}_0=(\mathbf{c}_0,\mathbf{R}_0)$. We use L1 loss for the prediction of ε. The objective loss of the denoising transformer $\mathscr{L}_\theta$ can be expressed as (5), where ω_1 and ω_2 are hyperparameters controlling the relative weights of $\mathscr{L}_\theta^{loc}$ and $\mathscr{L}_\theta^{rot}$.

$$\begin{aligned}\mathscr{L}_\theta^{loc}&=\|\varepsilon_\theta^{loc}(\{\tilde{I},\tilde{D}\},\tilde{M},l,\mathbf{a}_i,i)-\varepsilon^{loc}\|,\\ \mathscr{L}_\theta^{rot}&=\|\varepsilon_\theta^{rot}(\{\tilde{I},\tilde{D}\},\tilde{M},l,\mathbf{a}_i,i)-\varepsilon^{rot}\|,\\ \mathscr{L}_\theta&=\omega_1\cdot\mathscr{L}_\theta^{loc}+\omega_2\cdot\mathscr{L}_\theta^{rot}.\end{aligned}\tag{5}$$

During inference, a sample $\mathbf{a}_N\sim\mathscr{N}(0,1)$ is first drawn. The predicted contact point $\hat{\mathbf{c}}$ and contact pose $\hat{\mathbf{R}}$ are obtained through progressively denoising the sample N times with ε_θ, following (6), where $\mathbf{z}\sim\mathscr{N}(0,1)$ is a random variable of appropriate dimension. Scaled-linear and square cosine scheduler are used for $\mathbf{c}$ and $\mathbf{R}$ seperately to achieves better performance [40].

$$\mathbf{a}_{i-1}=\frac{1}{\sqrt{\alpha_i}}\left(\mathbf{a}_i-\frac{\beta_i}{\sqrt{1-\bar{\alpha}_i}}\varepsilon_\theta(\{\tilde{I},\tilde{D}\},\tilde{M},l,\mathbf{a}_i,i)\right)+\frac{1-\bar{\alpha}_{i+1}}{1-\bar{\alpha}_i}\beta_i\mathbf{z}\tag{6}$$

IV. EXPERIMENTS

To verify the effectiveness of the RoboPCA, we conduct extensive experiments in three perspectives: image-based affordance localization reasoning (Sec. IV-A), zero-shot manipulation in simulation (Sec. IV-B)) and in real-world settings (Sec. IV-C) as shown in Fig. 4 and further conduct an ablation study on the effectiveness of mask-enhanced features, joint pose-centered learning, and compatibility with robot data for further validation (Sec. IV-D).

We compare our method against four baselines, namely VRB [6], RAM [8], MOKA [9], and RoboPoint [10], which represent training-based, retrieve-and-transfer and VLM-based approaches respectively. As none of the baseline methods predict the manipulation pose directly, we utilize AnyGrasp [12] to produce grasp proposals and filter the final pose based on the contact point predicted by baselines as in [8]. A rule-based mechanism is incorporated to avoid collisions with objects for our method accordingly. For manipulation tasks, prescribed post-contact waypoints are provided to facilitate task completion. The depth and object masks used for evaluation are obtained using the same pipeline as stated in Sec. III-B, if unavailable.

TABLE I

EVALUATION RESULTS ON AGD20K DATASET.

Models	SR ↑	NSS ↑	DTM ↓
RAM [8]	0.1824	0.1892	0.0689
RoboPoint [10]	0.2138	0.2188	0.0508
VRB [6]	0.2846	0.2557	0.2069
MOKA [9]	0.3711	0.3390	**0.0331**
RoboPCA (ours)	**0.4403**	**0.4083**	0.0445

A. Image-Based Affordance Localization Reasoning

To evaluate the precision of contact point localization, we conduct experiments on AGD20K [16] datasets. Following [7], we select all the objects that are feasible for robotic manipulation and supplemented instances of drawer and cupboard categories following the same labeling procedure for comprehensive evaluation. Three metrics are reported for evaluation, including **Success Rate (SR)**, **Normalized Scanpath Saliency (NSS)**, and **Distance to Mask (DTM)**.

As shown in Tab. I, RoboPCA achieves a high success rate of 44.03%, which is 18.6% higher than the second-best method, MOKA, which leverages the strong reasoning capabilities of VLMs. This demonstrates the effectiveness of our method in localizing the appropriate contact point across categories, as shown in Fig. 3. Besides, RoboPCA also achieves higher NSS and similarly low DTM compared to MOKA, indicating that the predicted contact points are closer to the centers of the ground-truth masks.

B. Zero-shot Generalization across Tasks in Simulation

To evaluate the zero-shot generalization of RoboPCA across tasks, we evaluate the task success rates on RLBench [17]. 10 representative tasks from RLBench are selected as the evaluation set, including common object grasping, manipulation of task-relevant object regions, and articulated objects. Before evaluation, the manipulation regions of target objects are ensured to be fully visible by adjusting the camera viewpoint. Each task is evaluated with 25 episodes scored either 0 or 100, indicating failure or success in task execution. We report the average success rate for each task. The entire evaluation follows the testing pipeline of the modular approach as in [9], the task is first decomposed into a sequence of subtasks aligned with the robot's meta-skills, based on the given task instruction and visual input. Each subtask is then executed by its corresponding module to generate action trajectories. We replace the contact-relevant meta-skill modules with affordance

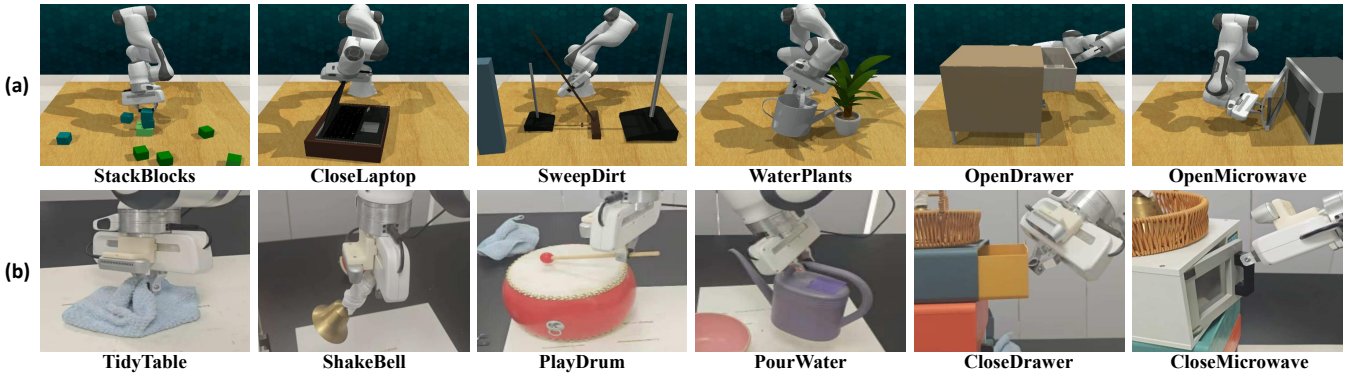


Fig. 4. Examples of task settings in simulation and the real world. We evaluate our method on 10 tasks in simulation (a; only 6 shown), and 9 tasks in the real world (b; only 6 shown) across various object categories to validate its effectiveness.

models to evaluate whether the 6-DoF pose predicted by the affordance model could support successful task execution.

As shown in Tab. II, our method outperforms all baselines on most of the tasks, yielding an average success rate of 64.8%. Compared to our method, RoboPoint and VRB suffer from providing precise contact points for manipulation, leading to suboptimal performance on tasks, such as WaterPlants, which need to locate the contact point precisely on the handle of the watering can. RAM struggles to infer affordances for objects with diverse attributes (e.g., appearance), as its retrieval mechanism relies on object-level similarity within the affordance memory. Although MOKA achieves the second-best performance by leveraging the strong reasoning capabilities of VLMs, its performance is limited by AnyGrasp's capability to generate grasp proposals in cluttered environments and under varying viewpoints. Moreover, the inherent randomness in VLMs' outputs can lead to failures in tasks that require highly consistent action patterns, such as StackBlocks. In contrast, our method achieves higher consistency between the predicted contact points and contact poses by jointly learning pose-centered affordances. Furthermore, by incorporating geometric features, it supports cross-category transfer and generalizes to objects with diverse shapes and appearances beyond the training categories.

C. Real-World Experiments

In the real-world setting, we carefully design 9 tasks involving interactions with diverse household objects, encompassing articulated objects (e.g., drawers), objects with function-specific regions (e.g., drumsticks), and deformable objects (e.g., cloth). We employ a Franka Emika robotic arm with a parallel gripper and utilize an on-hand RealSense D435i camera to capture the RGB-D image of the scene. We conduct 10 trials for each task, under different spatial arrangements and with varying object instances, to validate the generalization capability across tasks and categories.

As shown in Tab. III, RoboPCA achieves an 83.3% success rate on average among all 9 tasks, 24.9% higher than the second-best method, RAM, which yields conclusions consistent with the experimental results on the simulation platform. Although the strong real-world grasp generation performance of AnyGrasp enables baseline methods to accomplish some pick-and-place tasks even with inaccurate contact point predictions, they struggle to support tasks that require precise prediction of contact point within task-specific regions (e.g., PlayDrum). Some qualitative results are provided in Fig. 5 to further illustrate the results.

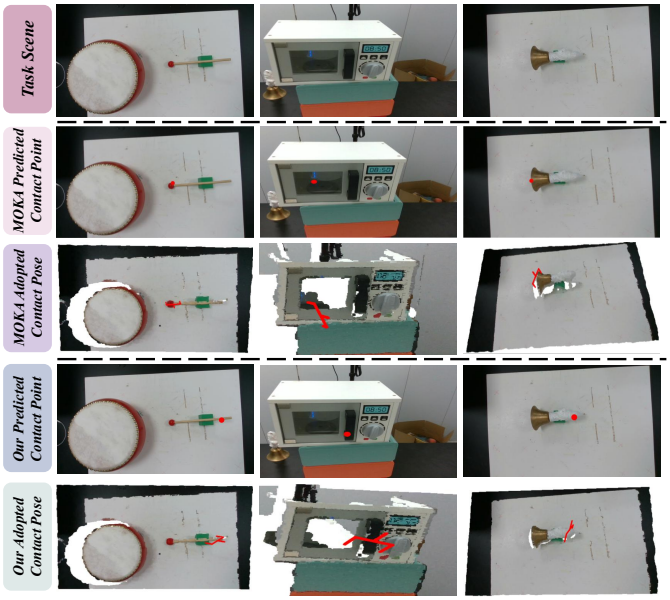


Fig. 5. Qualitative comparison of our model's predicted contact points and poses with MOKA in real-world settings.

D. Ablation Study

As in Tab. IV, we conduct detailed ablation experiments on a subset of manipulation tasks on RLBench, which consist of articulated object and function-specific region manipulation. These tasks are selected as they are more representative of evaluating how effectively the predicted pose-centered affordances support tasks that require precise manipulation of function-specific object regions. Each task is evaluated with 25 episodes under the same settings as in Sec. IV-B.

Effectiveness of mask-enhanced features. We compare our model trained with and without mask-enhanced features. For the model without mask-enhanced features, the RGB-D encoder encodes tokens from the original RGB-D frame directly to represent the scene, which allows us to evaluate the effectiveness of mask-enhanced features on the model's

TABLE II

SUCCESS RATE (%) ON RLBENCH MULTI-TASK SETTING.

Models	Pick Cup	Stack Blocks	Take Umbrella	Water Plants	Sweep Dirt	Close Laptop	Open Drawer	Close Drawer	Open Microwave	Close Microwave	Avg.
RoboPoint [10]	84	4	24	12	44	36	24	56	4	72	36.0
VRB [6]	84	60	32	4	48	44	20	84	8	64	44.8
RAM [8]	84	40	40	36	64	44	16	60	8	60	45.2
MOKA [9]	84	36	32	20	48	**52**	28	**88**	8	72	46.8
RoboPCA (ours)	**88**	**72**	**88**	**44**	**68**	**52**	**36**	84	**32**	**84**	**64.8**

TABLE III

SUCCESS RATE (%) ON 9 TASKS IN REAL-WORLD SETTING.

Models	Put Into	Tidy Table	Pour Water	Shake Bell	Play Drum	Open Drawer	Close Drawer	Open Microwave	Close Microwave	Avg.
MOKA [9]	90	**100**	**80**	50	60	50	70	10	80	65.6
RAM [8]	**100**	80	60	50	30	**70**	70	**60**	80	66.7
RoboPCA (ours)	**100**	90	70	**80**	**100**	**70**	**80**	**60**	**100**	**83.3**

TABLE IV

ABLATION RESULTS ON 5 REPRESENTATIVE TASKS EVALUATED ON SUCCESS RATE (%). AT01: SWEEPDIRT, AT02: OPENDRAWER, AT03: CLOSEDRAWER, AT04: OPENMICROWAVE, AT05: CLOSEMICROWAVE.

Models	AT01	AT02	AT03	AT04	AT05	Avg.
RoboPCA (ours)	**68**	**36**	84	32	84	60.8
w/ robot data	60	**36**	92	**40**	**96**	**64.8**
w/ AnyGrasp	60	16	92	8	72	49.6
w/o masked features	12	0	**96**	28	80	43.2

performance in capturing task-relevant object regions. As shown in the last row of Tab. IV, the success rate decreases markedly, especially for tasks requiring precise contact point predictions, such as OpenDrawer, which underscores the importance of incorporating mask-enhanced features for accurate pose-centered affordance prediction.

Effectiveness of joint pose-centered learning. To further evaluate the effectiveness of the pose-centered affordance learning paradigm, which jointly predicts contact points and contact poses, we compare our model with and without AnyGrasp. For the model with AnyGrasp, we replace the output of 6-DoF pose with the grasp pose obtained by filtering the grasp proposals generated by AnyGrasp based on the predicted contact point from our model. As shown in the third row of Tab. IV, the performance of our model outperforms the model with AnyGrasp, which demonstrates the advantage of jointly learning contact points and poses for producing reliable and consistent manipulation actions.

Compatibility with robot data. Beyond training only on human demonstrations, we also investigate whether RoboPCA can leverage robot demonstrations to further improve the performance of pose-centered affordance prediction. We collect 2K data samples based on DROID [14] for evaluation, each comprising an RGB-D frame, the interacting object mask, the contact point, and the corresponding contact pose. As shown in the second row of Tab. IV, incorporating robot data improves performance among most of the tasks, demonstrating that RoboPCA is compatible with robot demonstrations and can effectively benefit from additional robotic experience.

V. CONCLUSION AND FUTURE WORK

In this work, we propose **RoboPCA**, a pose-centered affordance prediction model that jointly predicts the contact points and poses given task instructions. To reduce collection costs, we devise **Human2Afford**, which automatically recovers scene-level 3D information and extracts pose-centered affordances from unlabeled human demonstrations. Extensive experiments show that RoboPCA achieves higher accuracy in contact point prediction, stronger consistency between contact points and manipulation poses, and generalization across tasks and categories. Ablation studies further validate the effectiveness of mask-enhanced features, the pose-centered affordance learning paradigm, and compatibility with robot data. Future work will extend our approach to cross-embodiment and larger datasets, enabling more versatile, robust manipulation across diverse objects and scenarios.

REFERENCES

[1] J. J. Gibson, *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[2] J. J. Gibson, "The senses considered as perceptual systems." 1966.

[3] S. Qian, W. Chen, M. Bai, X. Zhou, Z. Tu, and L. E. Li, "Affordancellm: Grounding affordance from vision language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[4] Y. Tang, W. Huang, Y. Wang, C. Li, R. Yuan, R. Zhang, J. Wu, and L. Fei-Fei, "Uad: Unsupervised affordance distillation for generalization in robotic manipulation," in *CoRL 2024 Workshop on Mastering Robot Manipulation in a World of Abundant Data*.

[5] T. Ma, Z. Wang, J. Zhou, M. Wang, and J. Liang, "Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping," *arXiv preprint arXiv:2411.12286*, 2024.

[6] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[7] Y. Ju, K. Hu, G. Zhang, G. Zhang, M. Jiang, and H. Xu, "Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation," in *European Conference on Computer Vision*, 2024.

[8] Y. Kuang, J. Ye, H. Geng, J. Mao, C. Deng, L. Guibas, H. Wang, and Y. Wang, "Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation," in *Conference on Robot Learning*, 2025.

[9] F. Liu, K. Fang, P. Abbeel, and S. Levine, "Moka: Open-vocabulary robotic manipulation through mark-based visual prompting," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[10] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox, "Robopoint: A vision-language model for spatial affordance prediction in robotics," in *Conference on Robot Learning*. PMLR, 2025.

[11] T. Ma, J. Zheng, Z. Wang, Z. Gao, J. Zhou, and J. Liang, "Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation," *arXiv preprint arXiv:2505.11865*, 2025.

[12] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2023.

[13] C. Tang, D. Huang, W. Dong, R. Xu, and H. Zhang, "Foundationgrasp: Generalizable task-oriented grasping with foundation models," *IEEE Transactions on Automation Science and Engineering*, 2025.

[14] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[15] B.-W. Yin, J.-L. Cao, M.-M. Cheng, and Q. Hou, "Dformerv2: Geometry self-attention for rgbd semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[16] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning affordance grounding from exocentric images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[17] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[18] A. Delitzas, A. Takmaz, F. Tombari, R. Sumner, M. Pollefeys, and F. Engelmann, "Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[19] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *2018 IEEE international conference on robotics and automation (ICRA)*, 2018.

[20] H. Geng, Z. Li, Y. Geng, J. Chen, H. Dong, and H. Wang, "Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[21] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[22] H. Chen, B. Sun, A. Zhang, M. Pollefeys, and S. Leutenegger, "Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[23] R. Xu, J. Zhang, M. Guo, Y. Wen, H. Yang, M. Lin, J. Huang, Z. Li, K. Zhang, L. Wang, *et al.*, "A0: An affordance-aware hierarchical model for general robotic manipulation," *arXiv preprint arXiv:2504.12636*, 2025.

[24] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[25] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[26] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "Egodex: Learning dexterous manipulation from large-scale egocentric video," *arXiv preprint arXiv:2505.11709*, 2025.

[27] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[28] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the cross-domain manipulation interface," *arXiv preprint arXiv:2407.15208*, 2024.

[29] H. Bharadhwaj, A. Gupta, V. Kumar, and S. Tulsiani, "Towards generalizable zero-shot manipulation via translating human interaction plans," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[30] M. K. Srirama, S. Dasari, S. Bahl, and A. Gupta, "Hrp: Human affordances for robotic pre-training," *arXiv preprint arXiv:2407.18911*, 2024.

[31] A. S. Chen, S. Nair, and C. Finn, "Learning generalizable robotic reward functions from" in-the-wild" human videos," *arXiv preprint arXiv:2103.16817*, 2021.

[32] M. Chang, A. Prakash, and S. Gupta, "Look ma, no hands! agent-environment factorization of egocentric videos," *Advances in Neural Information Processing Systems*, 2023.

[33] K. Shaw, S. Bahl, and D. Pathak, "Videodex: Learning dexterity from internet videos," in *Conference on Robot Learning*. PMLR, 2023.

[34] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

[35] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, D. Shah, F. Xia, D. Sadigh, and S. Kirmani, "Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation," *arXiv preprint arXiv:2409.16283*, 2024.

[36] S. Haldar and L. Pinto, "Point policy: Unifying observations and actions with key points for robot manipulation," *arXiv preprint arXiv:2502.20391*, 2025.

[37] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," *arXiv preprint arXiv:2205.09991*, 2022.

[38] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[39] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.

[40] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *arXiv preprint arXiv:2402.10885*, 2024.

[41] Z. Liang, Y. Mu, H. Ma, M. Tomizuka, M. Ding, and P. Luo, "Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[42] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[43] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*, 2024.

[44] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[45] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[46] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at internet scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[47] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[48] Y. Xiao, Q. Wang, S. Zhang, N. Xue, S. Peng, Y. Shen, and X. Zhou, "Spatialtracker: Tracking any 2d pixels in 3d space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[49] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[50] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.