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

RAAP: Retrieval-Augmented Affordance Prediction with Cross-Image Action Alignment

Qiyuan Zhuang[1,2], He-Yang Xu[1], Yijun Wang[1], Xin-Yang Zhao[3], Yang-Yang Li[3], Xiu-Shen Wei[1†]

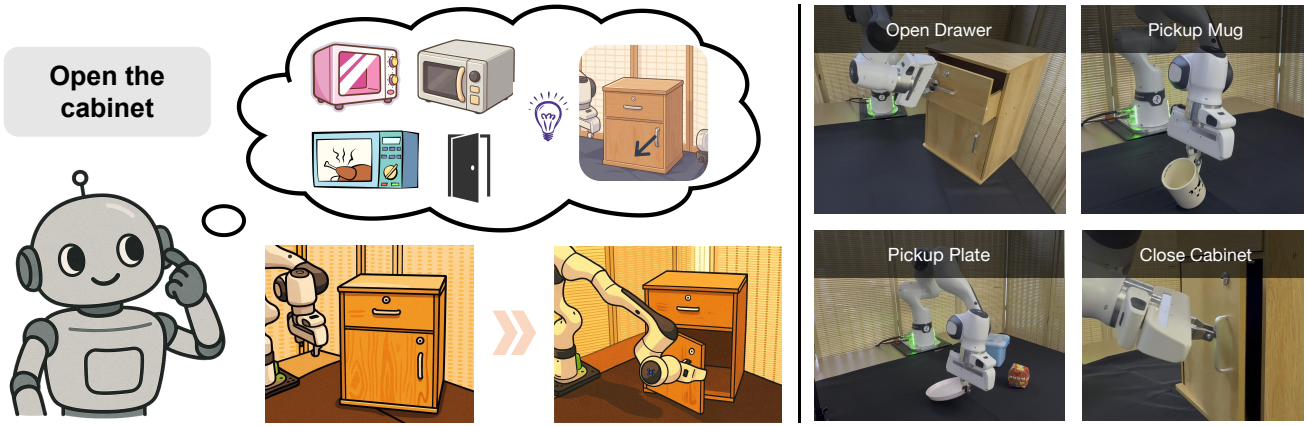

Fig. 1: When facing a novel task or unseen object category (e.g., "open the cabinet"), **Retrieval-Augmented Affordance Prediction (RAAP)** retrieves semantically related experiences (e.g., "opening a microwave") and transfers the corresponding affordances to guide execution.

***Abstract*— Understanding object affordances is essential for enabling robots to perform purposeful and fine-grained interactions in diverse and unstructured environments. However, existing approaches either rely on retrieval, which is fragile due to sparsity and coverage gaps, or on large-scale models, which frequently mislocalize contact points and mispredict post-contact actions when applied to unseen categories, thereby hindering robust generalization. We introduce Retrieval-Augmented Affordance Prediction (RAAP), a framework that unifies affordance retrieval with alignment-based learning. By decoupling static contact localization and dynamic action direction, RAAP transfers contact points via dense correspondence and predicts action directions through a retrieval-augmented alignment model that consolidates multiple references with dual-weighted attention. Trained on compact subsets of DROID and HOI4D with as few as tens of samples per task, RAAP achieves consistent performance across unseen objects and categories, and enables zero-shot robotic manipulation in both simulation and the real world. Project website: github.com/SEU-VIPGroup/RAAP.**

This work was supported by National Natural Science Foundation of China under Grant (62522602), Basic Research Program of Jiangsu under Grant (BK20250073), CIE-Tencent Robotics X Rhino-Bird Focused Research Program, and the Fundamental Research Funds for the Central Universities (4009002401, 2242025K30024). This work was also supported by the Big Data Computing Center of Southeast University.

† For Correspondence: weixs@seu.edu.cn
[1]School of Computer Science and Engineering, and Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Southeast University, Nanjing 211189, China
[2]Southeast University-Monash University Joint Graduate School, Southeast University, Suzhou 215123, China
[3]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

I. Introduction

Object affordances provide the perceptual basis for robotic fine manipulation in diverse and unstructured environments [1]–[4]. Given a manipulation task, a robot must reason about *where* to act (e.g., the graspable point on a drawer handle) and *how* to act (e.g., pulling direction to open it) from raw visual observations. Such reasoning should extend beyond object categories to fine-grained object parts and motion cues: for instance, identifying the exact rim of a bowl to pick it up, or the precise orientation to lift a mug by its handle. While visual affordance learning has been widely studied in robotics and computer vision [5]–[7], scaling affordance prediction to new object instances and categories remains a persistent challenge, especially under limited robot data.

A growing line of work addresses this challenge from two distinct paradigms. The first is the *retrieval-based paradigm*, where robots recall past demonstrations from a memory and transfer the associated affordances to new scenes [8]–[10]. This strategy has shown promising adaptability, as it leverages large collections of human and robot interaction data without requiring in-domain training, and can often generalize across tasks with minimal additional effort. However, retrieval methods face inherent limitations: the *sparsity problem*, where reliance on a single top-1 match makes predictions fragile; and the *coverage problem*, where the memory lacks semantically relevant instances, leading to failures on unseen categories.

The second is the *training-based paradigm*, which learns predictive affordance models directly from large-scale

data [11]–[15]. Such models can capture transferable visual patterns and achieve impressive generalization when abundant demonstrations are available. However, they also suffer from critical drawbacks. Many struggle to localize precise contact points or to predict reliable post-contact action directions [11]–[13], and some restrict affordances to static contacts only without modeling the dynamic component [14], [15]. As a result, training-based approaches alone remain insufficient for robust real-world generalization.

In this paper, we present the **Retrieval-Augmented Affordance Prediction (RAAP)**, a framework that unifies the strengths of retrieval- and training-based approaches for task-aware affordance prediction. Our key insight is not merely that affordances can be represented as two components—a *static contact point*, indicating where to act, and a *dynamic action direction*, indicating how to act after contact—but that these components exhibit different uncertainty characteristics. While such representations have been explored in prior work [9], existing approaches typically treat them as jointly transferable attributes. In contrast, we observe that static contact localization is primarily governed by geometric correspondence and can be reliably transferred from a single well-matched reference, whereas post-contact action direction is inherently more ambiguous and requires aggregating evidence across multiple references.

RAAP resolves these issues through a complementary inference design. For static affordance, we adopt dense feature correspondence with the top-1 retrieved reference to localize contact points [8], [9]. For dynamic affordance, we introduce a retrieval-augmented alignment model that aggregates cues from multiple references [16]–[18]. By conditioning visual features on action vectors and integrating them via a dual-weighted attention mechanism, the model selectively emphasizes task-relevant priors while suppressing noisy or misaligned examples. Crucially, by consolidating information across diverse exemplars, RAAP significantly reduces directional prediction errors and improves robustness under visual and geometric variations. This design enables RAAP to achieve strong performance with data-scarce settings (as few as tens of samples per task). Moreover, our method enables zero-shot robotic manipulation in both simulation and real-world environments.

Our contributions are summarized as follows:

- We propose **RAAP**, a unified retrieval- and training-based paradigm that addresses the limitations of existing methods and enables generalization under data scarcity, achieving strong performance with only a handful of training samples per task.
- We design a novel **retrieval-augmented alignment model** that aggregates multiple references with dual-weighted attention, while treating static and dynamic affordances with complementary mechanisms.
- We conduct comprehensive evaluations on DROID, HOI4D, and real-world platforms, demonstrating that RAAP outperforms baselines (RAM, A0) in both *unseen-object* and *cross-category* generalization scenarios.

II. Related Work

A. Visual Affordance Learning for Robotics

Visual affordance learning [12], [19]–[23] aims to infer *where* and *how* an object can be interacted with from sensory inputs, providing crucial perceptual cues for downstream planning and control. Some prior works have employed pixel-level segmentation or detection to infer affordance regions [5], [24]. While intuitive, such heatmaps are often diffuse and lack the geometric precision required for complex manipulation. Subsequent work therefore explores richer representations, including dense visual correspondence [25] and per-point affordance prediction on 3D point clouds [26]. For articulated objects, keypoint-based affordances provide a compact abstraction of interaction [27], [28]. In parallel, UAD [14] distills affordance knowledge from foundation models to predict static contact maps from single images.

However, static affordances alone cannot capture the dynamic nature of interactions. Recent work therefore models dynamic affordances that encode both contact and post-contact motion, such as VRB [11] and A0 [13]. Instead of relying solely on large curated datasets, our work combines retrieval from a compact affordance memory with alignment-based learning to enable generalization under limited data.

B. Zero-Shot Robotic Manipulation

Zero-shot robotic manipulation [29]–[34] aims to enable robots to perform new tasks from high-level instructions without task-specific training. Existing approaches mainly follow two paradigms: large-scale imitation learning and retrieval-based methods. Large-scale imitation learning trains generalizable policies from diverse robot and human demonstrations [35], [36], often leveraging multimodal foundation models for semantic grounding [37]–[41], but remains heavily dependent on data scale and diversity.

Retrieval, as a general mechanism for leveraging prior instances across domains [42]–[46], has recently been adopted for zero-shot robotic manipulation. Some leverage semantic correspondence to transfer affordances across object categories [8], [47]. RAM [9] builds a cross-domain interaction memory for affordance transfer, while AffordDP [10] integrates retrieval with diffusion policies to transfer 3D affordances through semantic matching and geometric alignment. Our method follows this paradigm but introduces retrieval-augmented alignment that consolidates multiple references and explicitly decouples static and dynamic affordances, enabling robust zero-shot manipulation under limited data.

III. Method

In this section, we describe (A) how we construct an affordance memory from prior interactions, (B) how we retrieve task-relevant references and transfer static contact points, (C) how we predict dynamic action directions via retrieval-augmented alignment, and (D) how we lift 2D affordances into 3D for robotic execution. As illustrated in Fig. 2, RAAP decomposes affordance into static (contact point) and dynamic (action direction) components, which are predicted through complementary retrieval and alignment.

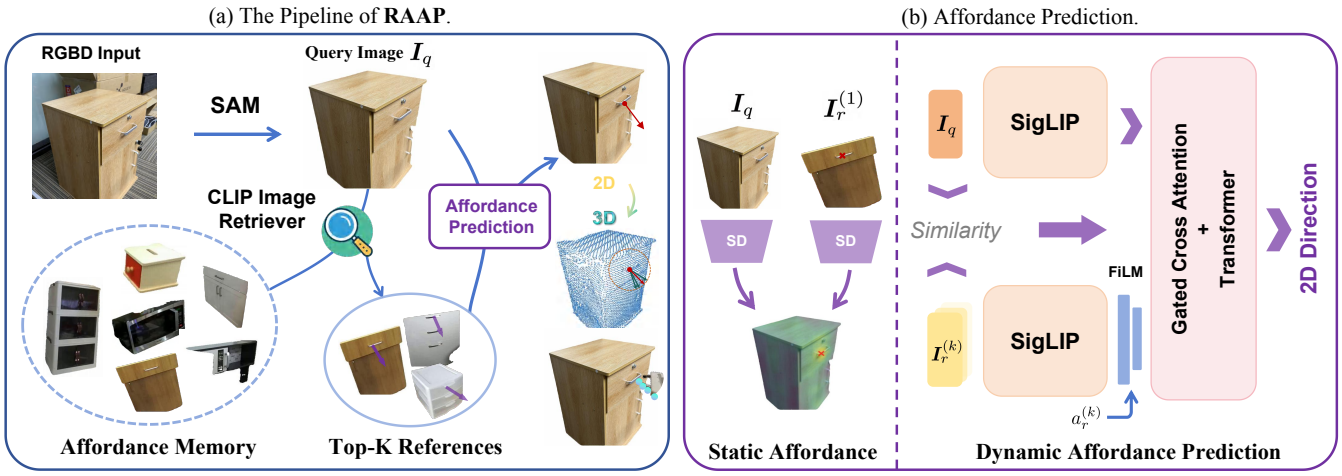

Fig. 2: **Overview of the Retrieval-Augmented Affordance Prediction (RAAP) framework.** (a) *Pipeline.* Given an RGB-D input and a task label, RAAP retrieves top-K references from an affordance memory using CLIP-based similarity. It then predicts a 2D affordance (contact point and action direction) and lifts it to 3D for execution. (b) *2D Affordance Prediction.* Static contact points are localized via dense correspondence using Stable Diffusion (SD) features, while dynamic action directions are inferred by a cross-image alignment module. Both query and reference images are encoded with a shared SigLIP-2 backbone; reference tokens are further modulated by their action vectors via FiLM, and fused with query tokens through gated cross-attention and a Transformer.

A. Affordance Memory

We represent a 2D affordance as a tuple $\mathcal{A}^{\text{2D}} = (\boldsymbol{c}^{\text{2D}}, \boldsymbol{a}^{\text{2D}})$, where $\boldsymbol{c}^{\text{2D}}, \boldsymbol{a}^{\text{2D}} \in \mathbb{R}^2$ denote the contact point (static affordance) and the post-contact motion direction (dynamic affordance), respectively. Both quantities are defined in the image coordinate frame. Given a target image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, the goal is to estimate the affordance that best facilitates task execution in the current context. At execution time, the 2D affordance can be lifted to 3D using object point clouds and camera intrinsics (see Sec. III-D).

To support retrieval-based generalization, we construct a visual affordance memory $\mathcal{R}$ that stores prior segmented object appearances and their associated interactions. We begin by applying Grounded-SAM [48] to segment the target object from the source image $\boldsymbol{I}$, producing a cropped image $\boldsymbol{I}_g$. Using a CLIP [49] image encoder, we extract an appearance embedding f from $\boldsymbol{I}_g$ for retrieval. Each memory entry stores a tuple of the form:

$$\mathcal{R} = \{(I_g, f, T, \mathcal{A}^{\text{2D}})\}, \tag{1}$$

where T is the task label, and $\mathcal{A}^{\text{2D}}$ denotes the 2D affordance annotation.

For dynamic affordance extraction, we convert annotated 2D trajectories into normalized motion directions. Specifically, each trajectory is reduced to its dominant orientation, which is then represented as a unit vector $\boldsymbol{a}^{\text{2D}}$. This ensures that the dynamic component encodes only the intended action direction, independent of trajectory length or scale.

The memory is populated from a composite dataset containing DROID and HOI4D. We filter out samples whose trajectories do not yield valid post-contact directions.

B. Retrieval and Static Affordance Transfer

Given a new target scene, we perform retrieval from the affordance memory in two stages. First, we use a CLIP text encoder [49] to filter memory entries by task relevance, identifying those whose task labels match or are semantically similar to the current task T. Within this task-constrained subset, we compute cosine similarity between the CLIP embeddings of the query image $\boldsymbol{I}_q$ and the stored exemplars $\{\boldsymbol{I}_r^{(k)}\}$, and retrieve the top-$K$ most similar entries. These retrieved examples will be used differently for static and dynamic affordance prediction.

To localize the static contact point $\boldsymbol{c}_q^{\text{2D}}$ in the query image, we adopt a dense feature matching strategy based on the top-1 retrieved example. Let $\boldsymbol{I}_r^{(1)}$ denote the top-1 retrieved reference image and $\boldsymbol{c}_r^{\text{2D}}$ its annotated contact point. We extract dense per-pixel visual features using a Stable Diffusion (SD) [50] encoder and upsample them to the original resolution. To estimate $\boldsymbol{c}_q^{\text{2D}}$, we compare the local feature around $\boldsymbol{c}_r^{\text{2D}}$ against all pixel features in $\boldsymbol{I}_q$ by identifying the location with maximal cosine similarity. This procedure enables direct transfer of static affordance through high-resolution correspondence in feature space, preserving fine spatial structures that are critical for accurate contact localization.

We emphasize that our framework treats static and dynamic affordance components separately: the contact point $\boldsymbol{c}_q^{\text{2D}}$ is estimated via one-shot dense matching from the most similar retrieved example (top-1), as SD-based features already provide reliable correspondences even under appearance variations, while the direction vector $\boldsymbol{a}_q^{\text{2D}}$ is predicted via a retrieval-augmented alignment model that consolidates directional priors across multiple references (see Sec. III-C).

C. Learning Dynamic Affordance via Retrieval-Augmented Alignment

Unlike static contact localization, dynamic action prediction is more abstract and often fails with a single reference. We therefore aggregate top-K exemplars to reduce directional ambiguity and mitigate errors from mis-retrieval. To predict the post-contact action direction $\boldsymbol{a}_q^{\text{2D}}$, we introduce a retrieval-augmented cross-image action alignment module that conditions on multiple prior interaction examples stored in the affordance memory. The model learns to align cross-scene visual cues and manipulation intents through token-level attention and dynamic weighting.

Given a query image $\boldsymbol{I}_q$ and a task label T, we first retrieve the top-K most visually similar entries $(\boldsymbol{I}_r^{(k)}, \mathcal{A}_r^{(k)}, s^{(k)})_{k=1}^{K}$ from the task-relevant subset of memory $\mathcal{R}$. Each entry contains a reference image $\boldsymbol{I}_r^{(k)}$, similarity score $s^{(k)}$, and annotated 2D affordance $\mathcal{A}_r^{(k)} = (\boldsymbol{c}_r^{(k)}, \boldsymbol{a}_r^{(k)})$. In this stage, only the directional component $\boldsymbol{a}_r^{(k)}$ is used, as the contact point $\boldsymbol{c}_q^{\text{2D}}$ has already been estimated.

Action-Conditioned Reference Encoding To condition each reference image on its associated manipulation intent, we incorporate action vectors into the patch-level feature representation extracted by a shared SigLIP-2 encoder [51]. For each retrieved reference image $\boldsymbol{I}_r^{(k)}$, we obtain a sequence of patch tokens $\mathbf{F}_r^{(k)} \in \mathbb{R}^{N\times d}$, where N is the number of patches and d is the feature dimension.

The associated 2D action vector $\boldsymbol{a}^{(k)} \in \mathbb{R}^2$ is projected via a multi-layer perceptron (MLP) into modulation parameters $\gamma(\boldsymbol{a}^{(k)}), \beta(\boldsymbol{a}^{(k)}) \in \mathbb{R}^d$. We then apply FiLM-style modulation [52] to each patch feature independently:

$$\tilde{\mathbf{F}}_r^{(k)}[i] = \gamma(\boldsymbol{a}^{(k)}) \cdot \mathbf{F}_r^{(k)}[i] + \beta(\boldsymbol{a}^{(k)}), \quad \text{for } i = 1, \ldots, N. \tag{2}$$

This yields action-conditioned features $\tilde{\mathbf{F}}_r^{(k)} \in \mathbb{R}^{N\times d}$ that jointly encode visual context and motion intent. To distinguish different references, we add a learnable reference ID embedding $\mathbf{E}_{\text{ref}}^{(k)} \in \mathbb{R}^{1\times d}$:

$$\hat{\mathbf{F}}_r^{(k)} = \tilde{\mathbf{F}}_r^{(k)} + \mathbf{E}_{\text{ref}}^{(k)}. \tag{3}$$

The resulting tokens $\hat{\mathbf{F}}_r^{(k)}$ are used as memory inputs for attention-based alignment.

Cross-Attention Aggregation We apply a cross-attention module to fuse the action-conditioned reference tokens into the target representation, followed by a Transformer encoder [53] that refines the fused target representation after retrieval-conditioned alignment. The target image tokens $\mathbf{F}_q \in \mathbb{R}^{N_q\times d}$ serve as queries, while the action-conditioned reference tokens $\{\hat{\mathbf{F}}_r^{(k)}\}_{k=1}^{K}$ are concatenated along the sequence dimension, forming a unified key-value matrix $\mathbf{F}_{\text{mem}} \in \mathbb{R}^{K\cdot N\times d}$. This cross-attention mechanism enables the target representation to selectively incorporate directional cues from multiple retrieved exemplars.

To address the varying relevance of retrieved samples, we introduce a dual-weighting strategy that combines learned semantic relevance with pre-computed visual similarity. First, we compute global semantic features by averaging patch-level tokens:

$$\mathbf{z}_q = \frac{1}{N_q}\sum_{i=1}^{N_q} \mathbf{F}_q[i], \tag{4}$$

$$\mathbf{z}_r^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{F}}_r^{(k)}[i], \tag{5}$$

where $\mathbf{z}_q, \mathbf{z}_r^{(k)} \in \mathbb{R}^d$ represent the global semantic representations of the target and k-th reference image respectively. We compute semantic relevance weights using a lightweight gating network $w_k = \sigma(\text{MLP}([\mathbf{z}_q; \mathbf{z}_r^{(k)}]))$, where σ is the sigmoid activation and $[\cdot;\cdot]$ denotes feature concatenation.

The final attention weights $\{w_k^{\text{final}}\}_{k=1}^{K}$ are computed by combining the learned gating weights with the pre-computed similarity scores $\{s^{(k)}\}_{k=1}^{K}$:

$$w_k^{\text{final}} = \frac{\text{softmax}(s^{(k)}) \cdot w_k}{\sum_{j=1}^{K} \text{softmax}(s^{(j)}) \cdot w_j + \epsilon}, \tag{6}$$

where $\epsilon > 0$ is a small constant for numerical stability.

The fused target tokens are concatenated with a learnable [CLS] token and passed through a Transformer encoder. The final action direction $\boldsymbol{a}_q^{\text{2D}}$ is regressed from the [CLS] token via a lightweight MLP.

Learning Objective Ground-truth action directions $\hat{a}_q^{\text{2D}}$ are stored as unit vectors. We train the model by regressing the (x, y) components using a mean squared error (MSE) loss:

$$\mathcal{L} = \tfrac{1}{2}\,\|\boldsymbol{a}_q^{\text{2D}} - \hat{\boldsymbol{a}}_q^{\text{2D}}\|_2^2. \tag{7}$$

At inference time, the predicted vector $\boldsymbol{a}_q^{\text{2D}}$ is ℓ_2-normalized before evaluation.

D. Execution via Sampling-Based Affordance Lifting

To execute the predicted 2D affordance $(c_q^{\text{2D}}, a_q^{\text{2D}})$, we first lift it into the 3D workspace using camera intrinsics and the observed point cloud. The 2D contact point c_q^{2D} is projected to a set of candidate 3D points, from which we select the closest valid surface point $\boldsymbol{c}_q^{\text{3D}}$ as the target grasp region.

We then apply a dense grasp sampler [54] on the cropped point cloud around $\boldsymbol{c}_q^{\text{3D}}$ to generate a set of grasp proposals. The grasp pose closest to $\boldsymbol{c}_q^{\text{3D}}$ is selected for execution. To reach this selected grasp pose, we solve inverse kinematics (IK) to generate a collision-free joint-space trajectory, ensuring precise and feasible grasp approach planning.

After grasping, the 2D direction $\boldsymbol{a}_q^{\text{2D}}$ is transformed into a 3D displacement vector $\boldsymbol{\tau}_q^{\text{3D}}$ using the local surface orientation and camera geometry. The robot then executes the post-contact action by moving the end-effector along $\boldsymbol{\tau}_q^{\text{3D}}$. In simulation, we use position control for simplicity. In the real world, we adopt real-time Cartesian impedance control to modulate interaction forces during execution, allowing for compliant and safe manipulation.

IV. Experiments

A. Task-Aware Affordance Prediction

We first evaluate RAAP on task-aware dynamic affordance prediction, focusing on the post-contact action direction. Static affordance (i.e., contact point) is excluded here, as dense feature-based transfer has been extensively validated in prior work [8]–[10], and will instead be further assessed in our real-world experiments.

Datasets We experiment on subsets of the DROID [55] and HOI4D [56] datasets, adopting a 70/30 train–test split per task. To avoid potential data leakage, we remove samples with nearly identical object instances and viewpoints. For RAAP with $K > 0$, we further construct a task-aware affordance memory by merging semantically related tasks (e.g., *open drawer* and *open microwave*), which enables cross-task transfer of interaction cues during retrieval. On average, the DROID subsets provide about 18 samples per task, while HOI4D provides about 160. We use the same datasets as RAM [9], but apply additional cleaning to ensure fairer train–test separation.

Evaluation Metric Given a predicted unit vector $\boldsymbol{a}_q^{\text{2D}}$ and ground-truth $\hat{\boldsymbol{a}}_q^{\text{2D}}$, we report the Mean Angular Error (MAE):

$$\text{MAE} = \arccos\big(\langle \boldsymbol{a}_q^{\text{2D}}, \hat{\boldsymbol{a}}_q^{\text{2D}} \rangle\big) \cdot 180/\pi, \tag{8}$$

which directly measures angular discrepancy in degrees. Compared to vector MSE, MAE is invariant to magnitude and provides more interpretable directional accuracy.

Experimental Setting RAAP employs a 6-layer transformer with SigLIP-2 (base, patch16, 384 resolution) as the visual backbone, and all parameters are updated during training. For $K > 0$, each query retrieves 10–20 candidates from memory, from which K references are sampled per episode. To enhance training diversity, this sampling is repeated five times per query. We also apply data augmentation by horizontally flipping both the input images and the associated action directions. Models are trained for up to 50 epochs with early stopping if the training loss does not improve for 5 consecutive epochs. On a single NVIDIA RTX 4090 GPU, training one task of the DROID dataset requires approximately 85 seconds. All reported results are averaged over three runs with different random train–test splits to reduce variance.

Baselines We compare against two representative methods: RAM [9], a retrieval-and-transfer framework without multi-reference alignment, and A0 [13], a large-scale affordance-aware diffusion model trained on DROID, HOI4D, and ManiSkill (we report the A0-170M variant). Note that for A0, we only use the first two predicted 2D trajectory points to extract the contact location and action direction.

Main Results Figure 3 and Table I jointly present qualitative and quantitative comparisons. RAAP yields action directions that closely align with the ground truth across tasks, benefiting from the aggregation of multiple retrieved references rather than relying on a single exemplar. In contrast, RAM transfers features from only one reference, which often results in globally incorrect motion directions when the retrieved instance is not well aligned. A0 demonstrates reasonable generalization in *pickup* actions, but its predictions for *open/close* tasks deviate substantially from the ground truth in both direction and contact localization.

TABLE I: MAE ↓ for task-aware dynamic affordance prediction. The first six rows correspond to individual manipulation tasks from the DROID dataset, the seventh row reports the averaged performance on HOI4D pickup tasks (*bowl*, *bottle*, *mug*). The final row shows the overall average across all tasks.

Tasks	RAM	A0-170M	RAAP$_{K=0}$	RAAP$_{K=3}$
Open microwave	49.82	119.64	51.75	**37.00**
Close microwave	55.74	103.10	61.38	**30.88**
Open drawer	51.73	112.89	**15.45**	16.23
Close drawer	83.04	85.16	52.11	**48.25**
Pickup bowl	53.85	34.37	36.76	**31.45**
Pickup bottle	82.62	**33.86**	38.70	37.67
Pickup [obj] (HOI)	63.07	34.63	26.97	**26.36**
Overall avg.	62.84	74.81	40.45	**32.55**

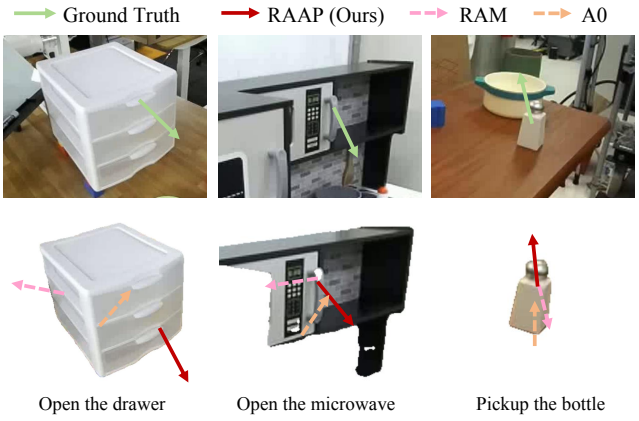

Fig. 3: Qualitative comparison of 2D affordance predictions on the DROID dataset. The first row shows the input RGB image with ground-truth affordances (contact point and action direction), and the second row visualizes predictions from RAM, A0, and RAAP ($K = 3$).

Table I further compares RAAP with RAM, A0, and our ablations using different numbers of retrieved references ($K = 0$ vs. $K = 3$). RAAP with retrieval ($K = 3$) achieves the lowest average error of $32.55°$, representing a reduction of nearly 50% relative to RAM ($62.84°$) and over 50% relative to A0 ($74.8°$), while also surpassing the no-retrieval variant ($K = 0$). The benefit of retrieval is particularly evident on the *open/close* tasks in DROID, where RAM and A0 often fail to predict meaningful directions. On HOI4D pickup tasks, RAAP also maintains the best overall accuracy. These results demonstrate that retrieval-augmented alignment substantially enhances dynamic affordance prediction, especially for semantically complex manipulations.

RAAP requires approximately 4 seconds per inference on an RTX 4090 GPU, which is substantially faster than RAM (50 seconds per inference) but slower than A0 (0.2 seconds).

Ablation Studies Table II reports an ablation on the proposed dual-weighted attention module, comparing the full model against several variants. Removing either gating ("w/o Gat-

TABLE II: Ablation on dual-weighted attention for dynamic affordance prediction (MAE ↓, degrees). Results are reported on the DROID dataset for two task categories: *open/close* (e.g., microwave) and *pickup* (e.g., bowl, bottle).

Variant	Open/Close	Pickup
RAAP (Full)	**33.09**	**34.56**
w/o Gating	36.14	42.92
w/o Similarity	39.47	38.96
w/ Uniform Weights	37.63	36.54

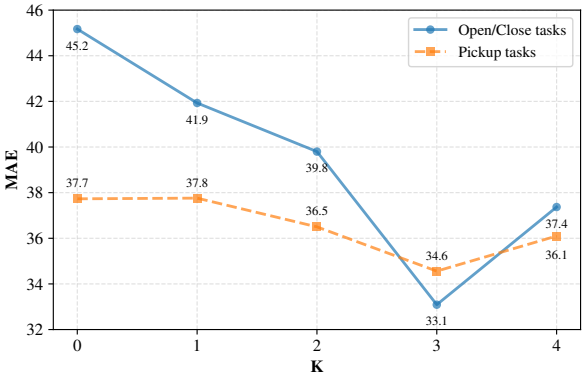

Fig. 4: MAE ↓ of RAAP as the number of retrieved references K varies from 0 to 4 on the DROID dataset.

ing") or similarity weights ("w/o Similarity") leads to marked performance drops, confirming that the two weighting signals are complementary: similarity provides a strong prior on appearance, while gating suppresses misaligned references. Uniform weighting (averaging all retrieved references with equal weights) performs better than either single-weight variant, but remains inferior to the full model. RAAP (Full) achieves the lowest errors, highlighting the importance of jointly leveraging both gating and similarity cues for robust dynamic affordance prediction.

We further examine the effect of the number of retrieved references K on dynamic affordance prediction using the DROID dataset. As shown in Fig. 4, on open/close tasks, RAAP achieves a substantial gain when increasing K from 0 to 3, reducing the error from 45.2° to 33.1°. A similar but milder trend is observed on pickup tasks. Notably, $K = 3$ consistently yields the best performance across both task types, whereas too few references (e.g., $K = 0$ or $K = 1$) leave the model vulnerable to mis-retrieval, and too many (e.g., $K = 4$) introduce noisy exemplars that slightly degrade accuracy. These results validate our design choice of aggregating a moderate number of references for robust dynamic prediction.

B. Simulation and Real World Experiments

Experimental Setting Real-world experiments were conducted on a Franka Emika Research 3 robot equipped with its native parallel-jaw gripper. Perception was provided by RGB-D input from either an Intel RealSense D455 (front view) or an Intel RealSense L515 (side view), with only one camera used per task and no multi-view fusion. All models were trained exclusively on subsets of the DROID and HOI4D datasets without using real-world demonstrations. Unless otherwise specified, RAAP and the baseline methods (RAM and A0) follow the same training and inference configurations as in Sec. IV-A, with RAAP using retrieval with $K = 3$ by default. In both real-world and simulation studies, control follows the execution protocol described in Sec. III-D.

TABLE III: Real-world manipulation success rates (20 trials per task). The top block shows *unseen-object* generalization, and the bottom block shows *cross-category* generalization.

Settings	Tasks	RAM	A0	RAAP (Ours)
Unseen Objects	Open drawer	70%	0%	**85%**
	Close drawer	60%	5%	**80%**
	Pickup bowl	85%	65%	**90%**
	Pickup bottle	50%	35%	**75%**
	Pickup mug	**70%**	25%	**70%**
Unseen Categories	Open cabinet	70%	0%	**80%**
	Close cabinet	75%	15%	**100%**
	Pickup plate	80%	20%	**85%**
	Pickup watering can	50%	10%	**60%**

We evaluate two generalization scenarios: (i) *unseen-object* generalization, where the task (e.g., *open drawer*) remains the same but the manipulated object instance differs from those observed during training; and (ii) *cross-category* generalization, where affordances are transferred to semantically related object categories (e.g., training only on *open microwave* but testing on *open cabinet*). Performance is measured by the success rate (SR), averaged over 20 trials per task with randomized object placements.

Real-World Results The results in Table III reveal clear differences across methods. A0 consistently fails across both *pickup* and *open/close* tasks due to poor contact point prediction, which prevents stable grasping and reliable execution. RAM achieves moderate success, but its predicted action directions are often misaligned with the intended motion, leading to unstable or incomplete executions.

RAAP demonstrates the most reliable performance across both generalization settings. In *unseen-object* tasks, it improves over RAM by 15–25 points on *open/close drawer* and achieves the highest success rates in all pickup tasks. For *cross-category* generalization, RAAP attains 100% success on *close cabinet*. This task is relatively less sensitive to precise contact localization since the gripper can push on the cabinet door without needing to grasp the handle, and approximate directional accuracy is sufficient. In contrast, *close drawer* requires accurate directional prediction to overcome the drawer's resistance; errors in action direction lead to frequent failures for RAM and A0. RAAP, by leveraging retrieval-augmented cues, maintains high success despite these physical challenges. We observe occasional failures in pickup tasks (e.g., *pickup mug*, 70%), primarily due to the difficulty of grasping small object parts such as handles, even when the predicted 2D affordances are correct. As shown in Fig. 1, RAAP successfully handles both unseen-object and cross-category scenarios.

Simulation Results We conduct a controlled study in MuJoCo [57] with a UR5e manipulator, following the same control protocol as in real-world experiments. To test cross-

Fig. 5: Qualitative results on *pickup kettle* in MuJoCo with a UR5e manipulator. RAAP successfully transfers handle-oriented affordances to kettles.

category generalization, RAAP is trained only on the *pickup mug* task of HOI4D and evaluated on *pickup kettle*, with the kettle placed at different positions within the workspace.

Over 20 trials, RAAP ($K = 3$) achieves an SR of **85%**, outperforming RAM (70%) and A0 (10%). Qualitatively (Fig. 5), RAAP transfers handle-oriented action cues from mugs to kettles, enabling stable approach-and-lift behaviors. RAM, limited to single-exemplar transfer, frequently misaligns the predicted direction when the retrieved instance is not geometrically well matched, while A0 fails to localize valid graspable regions. This simulation study complements the real-world results, showing that RAAP provides robust cross-category transfer under geometric variations.

V. Conclusion

We presented the **Retrieval-Augmented Affordance Prediction (RAAP)**, a framework that unifies retrieval- and training-based approaches for task-aware affordance prediction. By decoupling static contact localization and dynamic action direction, RAAP leverages dense correspondence for pixel-level transfer while employing a retrieval-augmented alignment model to aggregate multiple references with dual-weighted attention. Our experiments on DROID, HOI4D, and real-world robotic platforms show that RAAP substantially improves generalization to unseen objects and novel categories, achieves competitive performance with only a handful of training samples per task, and enables zero-shot robotic manipulation in both simulation and the real world. These results indicate that RAAP provides a data-efficient and generalizable solution for fine-grained robotic manipulation.

Despite these promising results, several limitations remain. First, RAAP currently focuses on short-horizon tasks; extending it to multi-object and long-horizon sequential interactions would further test its robustness. Second, RAAP operates in an open-loop setting, and coupling it with closed-loop control policies could further improve robustness in dynamic and unstructured environments.

References

[1] L. Jamone, E. Ugur, A. Cangelosi, L. Fadiga, A. Bernardino, J. Piater, and J. Santos-Victor, "Affordances in psychology, neuroscience, and robotics: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 1, pp. 4–25, 2018.

[2] N. Yamanobe, W. Wan, I. G. Ramirez-Alpizar, D. Petit, T. Tsuji, S. Akizuki, M. Hashimoto, K. Nagata, and K. Harada, "A brief review of affordance in robotic manipulation research," *Advanced Robotics*, vol. 31, no. 19-20, pp. 1086–1101, 2017.

[3] M. Hassanin, S. Khan, and M. Tahtali, "Visual affordance and function understanding: A survey," *ACM Computing Surveys*, vol. 54, no. 3, 2021.

[4] S. Yike, C. Xiaogang, W. Yijun, L. Yang, Z. Yuewei, W. Qinlei, G. Shangkai, and G. Xiaorong, "Connecting minds and devices: A fifty-year review of brain-computer interfaces," *Chinese Journal of Electronics*, vol. 34, no. 5, pp. 1464–1474, 2025.

[5] T.-T. Do, A. Nguyen, and I. Reid, "AffordanceNet: An end-to-end deep learning approach for object affordance detection," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 5882–5889.

[6] H. Wu, Z. Zhang, H. Cheng, K. Yang, J. Liu, and Z. Guo, "Learning affordance space in physical world for vision-based robotic object manipulation," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2020, pp. 4652–4658.

[7] P. Mandikal and K. Grauman, "Learning dexterous grasping with object-centric visual affordances," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2021, pp. 6169–6176.

[8] 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 *Proceedings of the European Conference on Computer Vision*, 2024, pp. 222–239.

[9] 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 *Proceedings of the Conference on Robot Learning*, 2025, pp. 547–565.

[10] S. Wu, Y. Zhu, Y. Huang, K. Zhu, J. Gu, J. Yu, Y. Shi, and J. Wang, "AffordDP: Generalizable diffusion policy with transferable affordance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 6971–6980.

[11] 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, pp. 13 778–13 790.

[12] K. Mo, L. 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, pp. 6813–6823.

[13] R. Xu, J. Zhang, M. Guo, Y. Wen, H. Yang, M. Lin, J. Huang, Z. Li, K. Zhang, L. Wang, Y. Kuang, M. Cao, F. Zheng, and X. Liang, "A0: An affordance-aware hierarchical model for general robotic manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 13 491–13 501.

[14] 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 *Proceedings of the IEEE International Conference on Robotics and Automation*, 2025, pp. 3822–3831.

[15] 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, pp. 2252–2261.

[16] Y. Yu, W. Ping, Z. Liu, B. Wang, J. You, C. Zhang, M. Shoeybi, and B. Catanzaro, "RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs," in *Advances in Neural Information Processing Systems*, 2024, pp. 121 156–121 184.

[17] Z. Haohui, L. Bo, H. Jingyi, S. Chao, H. Pingkuan, and N. Evgeny, "A parallel multi-demonstrations generative adversarial imitation learning approach on uav target tracking decision," *Chinese Journal of Electronics*, vol. 34, no. 4, pp. 1185–1198, 2025.

[18] S. Han, S. Lee, M. Cha, S. O. Arik, and J. Yoon, "Retrieval augmented time series forecasting," in *Proceedings of the International Conference on Machine Learning*, 2025, pp. 21 774–21 797.

[19] S. Liu, S. Tripathi, S. Majumdar, and X. Wang, "Joint hand motion and interaction hotspots prediction from egocentric videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3282–3292.

[20] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," in *Proceedings of the Conference on Robot Learning*, 2025, pp. 1541–1566.

[21] R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, Q. Fan, X. Chen, L. Guibas, and H. Dong, "VAT-Mart: Learning visual action trajectory proposals for manipulating 3D ARTiculated objects," in *Proceedings of the International Conference on Learning Representations*, 2022, pp. 1–23.

[22] Y. Wang, R. Wu, K. Mo, J. Ke, Q. Fan, L. J. Guibas, and H. Dong, "AdaAfford: Learning to adapt manipulation affordance for 3D articulated objects via few-shot interactions," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 90–107.

[23] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, and H. Dong, "RLAfford: End-to-end affordance learning for robotic manipulation," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2023, pp. 5880–5886.

[24] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2015, pp. 1374–1381.

[25] D. Hadjivelichkov, S. Zwane, L. Agapito, M. P. Deisenroth, and D. Kanoulas, "One-shot transfer of affordance regions? affcorrs!" in *Proceedings of the Conference on Robot Learning*, 2022, pp. 550–560.

[26] Y. Qin, B. Huang, Z.-H. Yin, H. Su, and X. Wang, "DexPoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation," in *Proceedings of the Conference on Robot Learning*, 2022, pp. 594–605.

[27] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "kPAM: Keypoint affordances for category-level robotic manipulation," in *Proceedings of the International Symposium of Robotics Research*, 2019, pp. 132–157.

[28] R. Xu, F.-J. Chu, C. Tang, W. Liu, and P. A. Vela, "An affordance keypoint detection network for robot manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2870–2877, 2021.

[29] S. Sontakke, J. Zhang, S. Arnold, K. Pertsch, E. Bıyık, D. Sadigh, C. Finn, and L. Itti, "RoboCLIP: One demonstration is enough to learn robot policies," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 55 681–55 693.

[30] H. Huang, F. Lin, Y. Hu, S. Wang, and Y. Gao, "CoPa: General robotic manipulation through spatial constraints of parts with foundation models," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024, pp. 9488–9495.

[31] Y. Ding, H. Geng, C. Xu, X. Fang, J. Zhang, S. Wei, Q. Dai, Z. Zhang, and H. Wang, "Open6DOR: Benchmarking open-instruction 6-dof object rearrangement and a vlm-based approach," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024, pp. 7359–7366.

[32] P. Li, T. Liu, Y. Li, M. Han, H. Geng, S. Wang, Y. Zhu, S.-C. Zhu, and S. Huang, "Ag2Manip: Learning novel manipulation skills with agent-agnostic visual and action representations," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024, pp. 573–580.

[33] K. Fang, F. Liu, P. Abbeel, and S. Levine, "MOKA: Open-world robotic manipulation through mark-based visual prompting," in *Proceedings of Robotics: Science and Systems*, 2024.

[34] H. Bharadhwaj, A. Gupta, V. Kumar, and S. Tulsiani, "Towards generalizable zero-shot manipulation via translating human interaction plans," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2024, pp. 6904–6911.

[35] K. Black, N. Brown, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," in *Proceedings of Robotics: Science and Systems*, 2025.

[36] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An open-source vision-language-action model," in *Proceedings of The Conference on Robot Learning*, 2025, pp. 2679–2713.

[37] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, 2024.

[38] P. Yuxin, W. Zishuo, L. Geng, Z. Xiangtian, Y. Sibo, and H. Hulingxiao, "A survey on fine-grained multimodal large language models," *Chinese Journal of Electronics*, vol. 35, no. 2, pp. 1–33, 2026.

[39] F. Tianhao, Y. Zehua, Y. Zhisheng, M. Chenxiang, H. Yang, L. Yingwei, W. Xiaolin, and W. Zhenlin, "A survey on the scheduling of DL and LLM training jobs in gpu clusters," *Chinese Journal of Electronics*, vol. 34, no. 3, pp. 881–905, 2025.

[40] Y. Shen, X.-S. Wei, Y. Sun, Y. Song, T. Yuan, J. Jin, H. Xu, Y. Yao, and E. Ding, "Explanatory Instructions: Towards unified vision tasks understanding and zero-shot generalization," in *Proceedings of the International Conference on Machine Learning*, 2025.

[41] H.-T. Yu, X.-S. Wei, Y. Peng, and S. Belongie, "Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation," *arXiv preprint arXiv:2504.14988*, 2025.

[42] X.-S. Wei, Y.-Z. Song, O. M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-Grained image analysis with deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8927–8948, 2022.

[43] Y. Suo, F. Ma, L. Zhu, and Y. Yang, "Knowledge-enhanced dual-stream zero-shot composed image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 951–26 962.

[44] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 241–257.

[45] W. Hao, G. Junqi, and B. Rongfang, "Multi-use learning instance for optimized image retrieval," *Chinese Journal of Electronics*, vol. 34, no. 3, pp. 1002–1005, 2025.

[46] X.-S. Wei, Y. Shen, X. Sun, P. Wang, and Y. Peng, "Attribute-Aware deep hashing with self-consistency for large-scale fine-grained image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 904–13 920, 2023.

[47] J. Zhu, Y. Ju, J. Zhang, M. Wang, Z. Yuan, K. Hu, and H. Xu, "DenseMatcher: Learning 3D semantic correspondence for category-level manipulation from a single demo," in *Proceedings of the International Conference on Learning Representations*, 2025, pp. 1–22.

[48] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, and F. Yan, "Grounded SAM: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

[50] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[51] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[52] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 3942–3951.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6000–6010.

[54] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 964–15 973.

[55] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, V. Guizilini, D. A. Herrera, M. Heo, K. Hsu, J. Hu, M. Z. Irshad, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn, "DROID: A large-scale in-the-wild robot manipulation dataset," in *Proceedings of Robotics: Science and Systems*, 2024.

[56] 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, pp. 21 013–21 022.

[57] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.