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

Bi-Adapt: Few-shot Bimanual Adaptation for Novel Categories of 3D Objects via Semantic Correspondence

Jinxian Zhou[1,2*], Ruihai Wu[3*], Yiwei Liu[1], Yiwen Hou[1],
Xunzhe Zhou[1,4], Checheng Yu[1,4], Licheng Zhong[2], Lin Shao[1†]
[1]National University of Singapore [2]Shanghai Qi Zhi Institute [3]Peking University [4]The University of Hong Kong

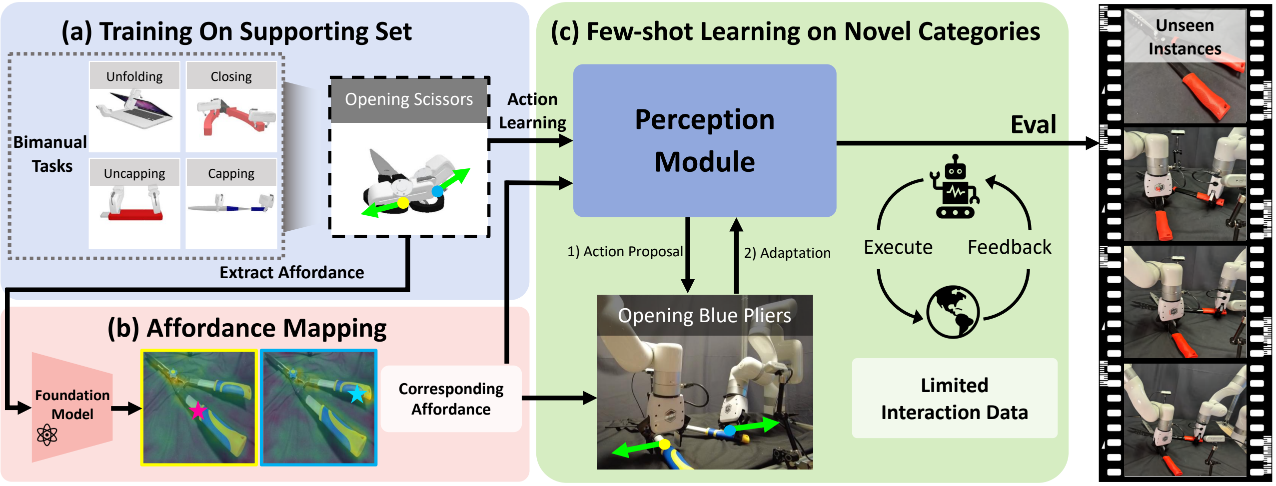


Fig. 1: We present **Bi-Adapt**, a novel framework designed for efficient learning of generalizable bimanual manipulation. It first learns point-level action on the supporting set for different bimanual tasks, then it predicts actions on novel categories based on the foundation-model-guided affordance, enabling cross-category generalization after few-shot adaptation.

***Abstract*— Bimanual manipulation is imperative yet challenging for robots to execute complex tasks, requiring coordinated collaboration between two arms. However, existing methods for bimanual manipulation often rely on costly data collection and training, struggling to generalize to unseen objects in novel categories efficiently. In this paper, we present Bi-Adapt, a novel framework designed for efficient generalization for bimanual manipulation via semantic correspondence. Bi-Adapt achieves cross-category affordance mapping by leveraging the strong capability of vision foundation models. Fine-tuning with restricted data on novel categories, Bi-Adapt exhibits notable generalization to out-of-category objects in a zero-shot manner. Extensive experiments conducted in both simulation and real-world environments validate the effectiveness of our approach and demonstrate its high efficiency, achieving a high success rate on different benchmark tasks across novel categories with limited data. Project website: `https://biadapt-project.github.io/`**

I. INTRODUCTION

Bimanual manipulation is common and crucial in human-centered environments, as some tasks inherently require the collaboration of both arms. However, obtaining high-quality and efficient manipulation strategies for complex bimanual manipulation tasks remains challenging for robots due to their high dependence on coordinated arm movements and the significant variation in 3D objects across different categories. Researchers have made great advances in bimanual manipulation [1]–[3]. However, some of these approaches focused only on simple tasks (e.g., ‘push’, ‘pull’) for a limited number of in-category objects. They performed poorly on unseen objects outside the training set. As a result, some methods aim to solve this generalization problem through training on large-scale datasets. These data-driven methods, such as ScrewMimic [2], based on imitation learning (IL), require the manual collection of extensive expert demonstrations. However, collecting real-world interaction data is labor-intensive, and training on a huge dataset is time-consuming. While these works have been extensively explored in the literature, efficiently generalizing manipulation strategies for complicated bimanual tasks, particularly leveraging prior knowledge on novel categories, remains a long-standing challenge.

Foundation models have demonstrated exceptional performance across various fields, particularly in computer vision and natural language processing [4]–[6]. For robotic manipulation, some works leverage the common-sense knowledge and generalization capabilities of current language and vision foundation models to improve robotic manipulation. For example, Shen et al. [7] combine accurate 3D geometry with rich semantics from 2D foundation models by leveraging distilled feature fields, enabling few-shot language-guided manipulation that generalizes across object poses, shapes,

---

[*] Equal contribution
[†] Corresponding author
Contact: isabella4444x@sjtu.edu.cn, linshao@nus.edu.sg.

979-8-3315-8160-2/26/$31.00 ©2026 IEEE

appearances, and categories. Recent studies also suggest that the zero-shot generalization of distilled features from pre-trained foundation models holds strong power for cross-category affordance generalization via semantic correspondence, and thus guides the manipulation on novel object categories and unseen scenarios [8]–[10]. However, entirely relying on the off-the-shelf foundation models to extract affordance via semantic correspondence is not enough. The capability of obtaining positive affordance on novel categories is strictly upper-bounded by the capability of its upstream vision foundation model, the quality of images, and the accuracy of points selected. Besides affordance, the coordinated manipulation orientation is also imperative. The bimanual manipulation strategy without adaptation for affordance and manipulation orientation could easily lead to failure when directly evaluated on unseen categories.

Most prior work focuses on learning dual-arm stabilization policies through reinforcement learning (RL) [11]. But these methods are often costly and not efficient, as they typically require rolling out policies many times and struggle to achieve large-scale generalization. This process is time-consuming due to the complexity of bimanual manipulation.

To overcome the limitations above, in this work, we introduce **Bi-Adapt**, a foundation-model-based framework for efficient generalization for bimanual manipulation. Bi-Adapt accumulates prior knowledge about bimanual manipulation on a constrained supporting set, and then leverages semantic correspondence from the foundation model to transfer affordance to new categories, followed by few-shot learning on a small number of instances. Notably, with limited robot demonstrations on novel categories, our method can demonstrate a superior performance of the finetuned model evaluated zero-shot on more instances in novel categories.

To demonstrate the performance of our framework, we conduct experiments on diverse object categories over 5 challenging tasks. Extensive experiments, including both qualitative and quantitative results, validate the effectiveness of our framework and its individual components. In summary, our contributions are as follows:

- We propose a foundation-model-based framework for generalizable bimanual manipulation across object categories for different complex tasks.
- We introduce a few-shot adaptation strategy following contact point selection to enhance bi-manual collaboration efficiently.
- Our experiments provide strong evidence supporting the effectiveness of the proposed method.

II. RELATED WORK

A. Bimanual Manipulation with Affordance Learning

Bimanual manipulation offers substantial benefits over single-arm systems [12]. Initial efforts using traditional control methods [13]–[15] struggled with complex models and long computation time. As a result, learning-based approaches have been applied to dual-arm systems [16]–[18], though challenges like the sim-to-real gap and low sample efficiency persist. The sim-to-real gap also complicates policy transfer and limits real-world deployment. To address these challenges, some approaches pay attention to learning object-centric visual actionable affordance that use dense affordance maps to suggest action possibilities at every point on a 3D scan, to accelerate policy generation and benefit downstream robotic manipulation tasks [19]. Despite these advances, affordance-learning-based methods struggle to predict precise affordance maps on unseen objects. In our work, we reserve the point-level action direction learning, but utilize the pre-trained vision foundation model to obtain mapped affordances on novel categories to avoid time-consuming affordance learning for bimanual manipulation.

B. Foundation Models for Robotic Manipulation

Numerous works have sought to leverage the commonsense knowledge and generalization capabilities of foundation models for robotic manipulation, mainly focusing on open-world reasoning and goal specification [20]–[22]. In this work, we are interested in obtaining general affordance knowledge from existing foundation models to achieve affordance transfer. Recent advances in vision foundation models, such as DINOv2 [23] and DiFT [24], have demonstrated remarkable capabilities in finding semantic correspondences across objects. Particularly, the extracted features are versatile in mapping similar points across categories. Recent works have utilized features distilled from pre-trained foundation models for zero-shot or few-shot robotic manipulation on novel objects [25]. However, these works mainly focused on simple grasping-based tasks with a single arm, without considering adaptation for two-arm collaboration. As the affordance predictions are strictly subjected to the foundation model's capability, the viewpoint and image quality, leading to a certain proportion of mapped points may be negative for manipulation. In our work, we leverage the zero-shot features distilled from off-the-shelf foundation models to achieve affordance generalization via semantic correspondence across different categories. Compared to previous methods, we introduce a few-shot adaptation strategy after affordance mapping, to filter negative contact points and adjust action directions for better bimanual collaboration.

C. Efficient Adaptation for Robotic Manipulation

Various approaches have been explored for fast adaptation [26]–[28]. Never Stop Learning [29] was a pioneer in combining off-policy reinforcement learning with a simple fine-tuning procedure to adapt vision-based policies to changes in background, object shape, lighting, and robot morphology. Several studies have leveraged foundation models for policy transfer [30], though they primarily focus on high-level adaptation rather than fine-grained policy adaptation. AdaAfford [31] provides a way for affordance adaptation requiring only test-time interactions. Where2Explore [32] leverages geometric similarities across categories for few-shot learning. In our work, we introduce an efficient adaptation procedure following the contact-point selection. After the few-shot interactions on limited instances, the

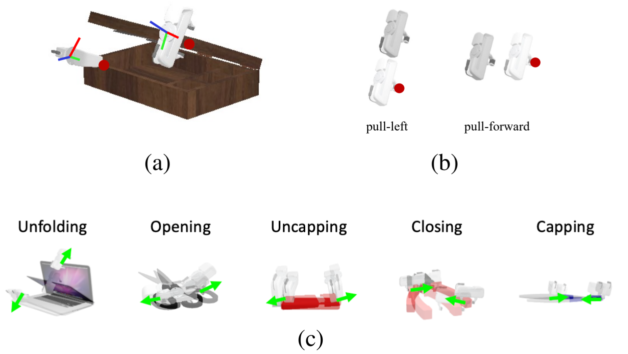


Fig. 2: (a) Simulation environment with gripper frames (red, green, blue axes for forward, leftward, upward); (b) Two $SE(3)$-parametrized action primitives, visualized by two key frames with red contact points and time steps from transparent to solid grippers; (c) Action definitions of different tasks.

fine-tuned model demonstrates promising generalization to unseen instances from novel categories.

III. PROBLEM FORMULATION

A. General Setting

We place a 3D object on the ground, given its partially scanned point cloud observation $O \in \mathbb{R}^{N\times 3}$ and a task T, where N denotes the number of point clouds. The framework is required to propose two grippers' actions $u_1 = (p_1, R_1)$ and $u_2 = (p_2, R_2)$, in which $p_i \in O$ is a contact point and $R_i \in SO(3)$ is the gripper's orientation. Each action u_i includes contacting at point p_i and pulling to a specific direction without changing the orientation. Figure 2 shows our simulation environment, two action primitives, and presents the action definitions of different tasks.

B. Task Formulation

We formulate five benchmark tasks and set success judgments for them. For all tasks, success requires the movement distance of the target joint to be less than a set threshold while the object does not fall or flip. Here are the task-specific criteria:

- **Unfolding, Opening, Closing**: The rotational joint angle change must exceed 0.10 times the joint's range.
- **Uncapping, Capping**: Two parts of the prismatic joint are separated or drawn to each other by over 0.05 m.

IV. METHOD

Our goal is to achieve bimanual manipulation tasks on novel categories that efficiently learn from prior knowledge. **Method Overview.** Figure 3 shows our pipeline. First, we begin by learning bimanual point-level manipulation orientation on training categories to construct a supporting set (Section IV-A). Then, we transfer affordance from the training categories in the supporting set to novel categories by using the foundation model to map contact points (Section IV-B). After few-shot adaptation on novel categories (Section IV-C), the fine-tuned networks improve performance on unseen instances in novel categories. Section IV-D describes network architectures and the training strategy.

A. Action Learning for Building Supporting Set

To conduct cross-category few-shot adaptation tasks, we first need to build a supporting set that captures key affordances and action patterns from known categories. This supporting set serves as a knowledge prior, enabling the transfer of learned manipulation strategies to a broader range of novel object categories.

Visual manipulation affordance has fine-grained manipulation information and is capable of generalizing to unseen objects within the same category. Building upon previous works [1], [33], we design a Perception Module to learn collaborative visual actionable affordance and interaction policy for dual-gripper manipulation tasks over diverse objects in the supporting set. Given the vast combinatorial action space of two grippers, we mitigate the complexity by disentangling composite actions into two sequentially conditioned actions. Specifically, we design two coupled submodules in the Perception Module: the First Gripper Module $\mathcal{M}_1$(left), and the Second Gipper Module $\mathcal{M}_2$ (right), where $\mathcal{M}_1$ proposes u_1 and $\mathcal{M}_2$ proposes u_2 based on u_1. As indicated by the arrows in Figure 3, the training and inference procedures share the same architecture but follow opposite dataflow directions. For inference, the dataflow direction is intuitive: $\mathcal{M}_1$ proposes u_1, and $\mathcal{M}_2$ proposes u_2 conditioned on u_1. But this approach does not ensure that the u_1 is optimal for collaboration. To address this, we adopt a reversed dataflow strategy during training: $\mathcal{M}_2$ is trained first, followed by $\mathcal{M}_1$, which learns based on $\mathcal{M}_2$. Specifically, $\mathcal{M}_2$ is trained to generate suitable u_2 given the diverse u_1 samples from the training dataset. Once $\mathcal{M}_2$ is capable of generating u_2 that effectively collaborate with various u_1, $\mathcal{M}_1$ is trained to propose u_1 that facilitate successful collaboration with $\mathcal{M}_2$.

Each gripper module consists of two networks: the Action Proposal Network $\mathcal{A}$ and the Action Scoring Network $\mathcal{C}$. Given the contact point, $\mathcal{A}$ predicts the gripper orientation, while $\mathcal{C}$ evaluates whether the proposed action is suitable for manipulation. We detail the design of networks and training strategy in Section IV-D. However, this series of networks suffers significant drops in novel categories. To enable broader generalization, we leverage semantic correspondence to link different categories.

B. Affordance Transfer

When encountering an unfamiliar object, humans often recall experiences to determine how to interact with it. Inspired by this, we propose a similar approach to robots. Objects from different categories often share geometric similarities and semantic correspondence. For example, *Scissors* and *Pliers* have similar shapes and can be operated using similar actions. If a robot has mastered the skill of manipulating the Scissors, it can leverage this knowledge to handle the Pliers. The cross-category semantic correspondence of affordance can serve as a bridge to connect them. In our study, we use the foundation model to map contact points from objects in our supporting set to novel objects, guiding the manipulation.

1) Affordance Representation and Collection: We define the affordance as the contact points on the object at the first

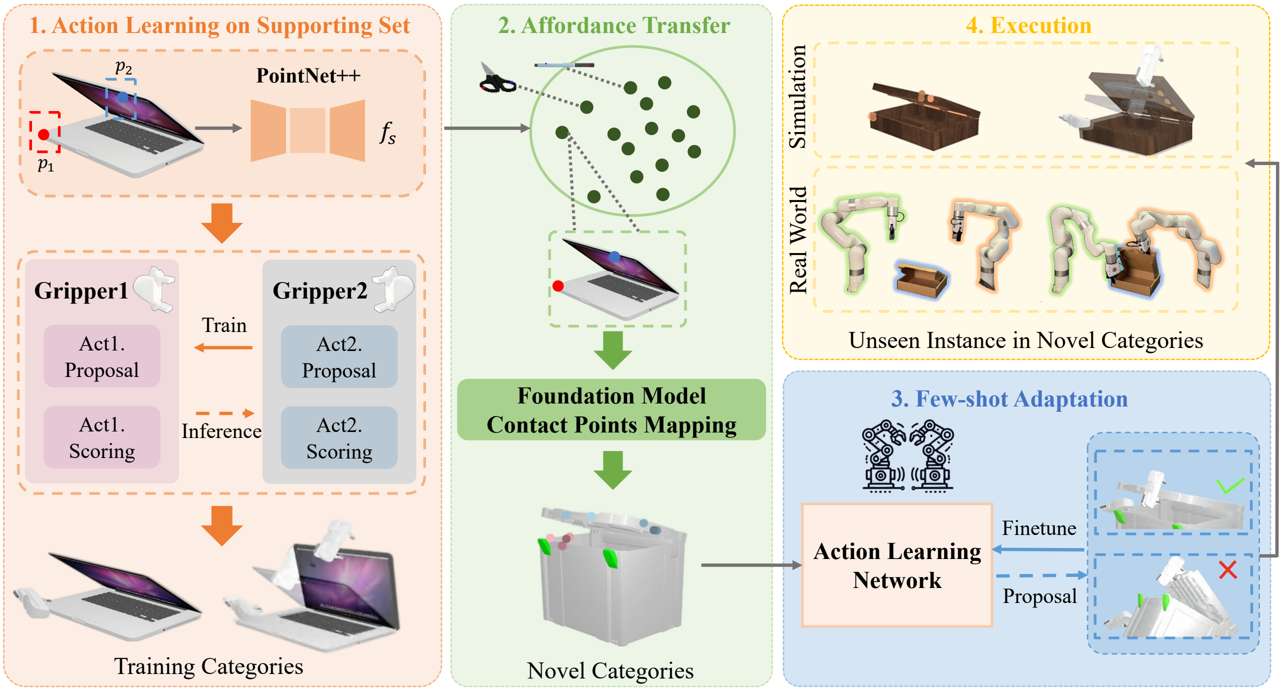


Fig. 3: **Pipeline. (Left)** We first train the Action Learning Network on the supporting set. The corresponding affordance and action distribution serve as our prior knowledge. **(Middle)** Then, we make contact points mapping from training categories to novel categories, leveraging the foundation model. **(Lower Right)** After that, the pre-trained network proposes actions based on mapped contact-point pairs on novel categories and fine-tuned with the interaction results. **(Upper Right)** Finally, the fine-tuned networks can facilitate manipulating unseen instances from novel categories with better performance.

frame when this contact takes place. For two grippers, we select one point for each, $p_1, p_2 \in O$.

When collecting data on training categories, we capture the RGBD image before the interaction and record the contact points. We only utilize the contact points with successful outcomes for affordance transferring.

2) *Contact Points Mapping via Semantic Correspondence:* When encountering an unseen object and a task, we first retrieve source objects within the same task in the supporting set. With the recorded contact points, we can sample multiple successful cases as our source data. Then we transfer the $2D$ affordance from the source objects to the target object. In our study, we utilize the emergent semantic correspondence ability from the foundation model to map the retrieved contact points to the new object.

Given a source image I_s with two contact points $p_{s_1}^{2D}$, $p_{s_2}^{2D}$, a target image I_t, we aim to find the corresponding points $p_{t_1}^{2D}$, $p_{t_2}^{2D}$ in I_t. With the foundation model, we extract the diffusion features (DIFT) of I_s and I_t as the steps shown in [24]. By adding noise to I_t first and denoising through the diffusion process next, the diffusion features are generated. As the diffusion features correspond to each pixel in I_t, with each source contact point p_s^{2D} and its diffusion feature f_s obtained, we can calculate the similarity in the form of:

$$similarity = \cos f_{p_s^{2D}} \cdot \cos f_{p_{tj}^{2D}} \tag{1}$$

for each pixel p_{tj}^{2D} in I_t. Then we can find the pixel $p_{t_i}^{2D}$ with the highest similarity to the source point $p_{s_i}^{2D}$ in the diffusion features of the target image I_t. Based on the partial point cloud of the target object generated from the depth map D, we back-project the pixel $p_{t_i}^{2D}$ to get the 3D contact point $p_{t_i}^{3D}$. Since the resulting contact points may not be suitable for manipulation on a specific task, we use multiple source images and their contact point pairs to obtain multiple contact-point pairs candidates, for further selection in Section IV-C.

C. Few-shot Adaptation

Not all contact point candidates are suitable for manipulation, as mentioned in Section IV-B.2. And the actions proposed by the previous Perception Module can easily lead to failure due to the significant variance in physical properties and geometries among different categories. To improve the performance on novel categories, we introduce the few-shot adaptation procedure, in which the networks are finetuned with the results obtained by executing proposed actions.

As shown in Figure 3, the pre-trained Perception Module predicts two grippers' orientations based on sampled contact-point pair candidates. The actions with the greatest likelihood of success are executed simultaneously, with the outcomes for updating the networks. Therefore, the Perception Module can better determine whether the proposed actions succeed and adjust them to fit the novel categories.

Through just a few interactions on limited instances in the novel category, the Perception Module explores a novel category's semantic significant areas selected by the foun-

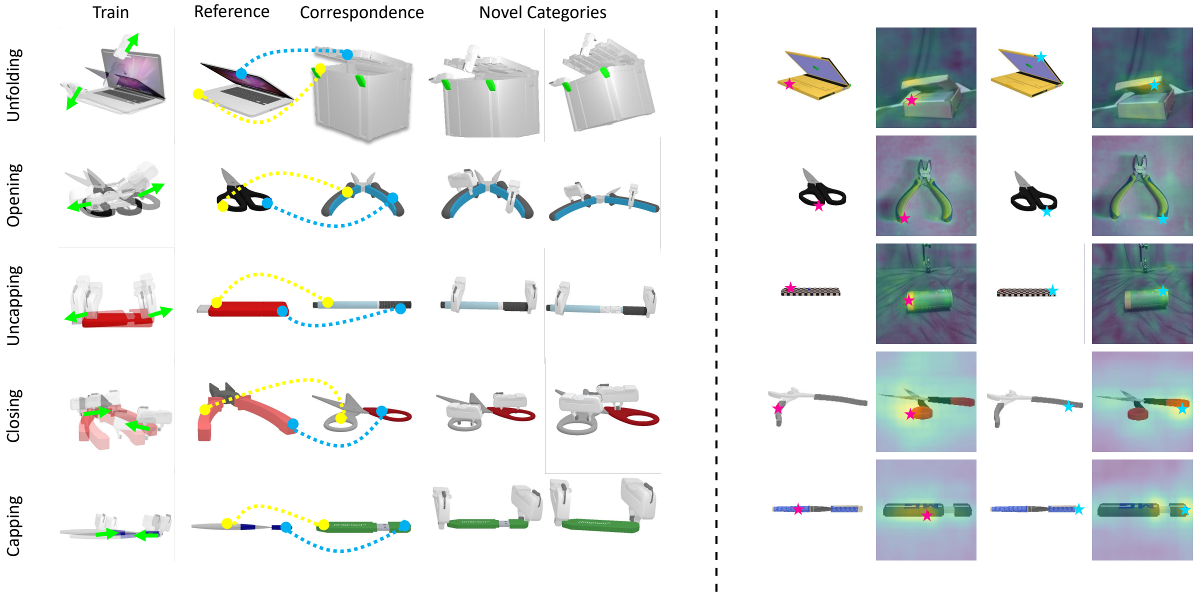


Fig. 4: (**Left**) Manipulation results across tasks. (**Right**) Cross-category affordance generalization. Each group shows a source image with contact points (★: first gripper, ★: second gripper) and a target image with inferred points. Highlight intensity indicates correspondence.

dation model, and better understands the geometrical and physical difference compared to learned categories. After the few-shot adaptation, the knowledge about the manipulation is transferred to novel categories. Thanks to the generalization of geometrical characteristics and physical properties within one category, our adapted module could manipulate unseen objects from this category without additional interactions.

D. Network Architecture and Training Strategy

Backbone Feature Extractors. The networks in the Perception Module receive three kinds of input entities or intermediate results: point cloud O, contact point p, and gripper orientation R. In different submodules, the backbone feature extractors share the same architectures. We use a segmentation-version PointNet++ (Qi et al., 2017 [34]) to extract per-point feature $f_s \in \mathbb{R}^{128}$ from O, and employ three MLP networks to respectively encode p and R into $f_p \in \mathbb{R}^{32}$ and $f_R \in \mathbb{R}^{32}$.

Action Scoring Loss. For $\mathcal{C}_2$, we use ground-truth results to supervise it. Given the interaction data with the ground-truth result r, where $r = 1$ means positive and $r = 0$ means negative, we can train $\mathcal{C}_2$ using the standard binary cross-entropy loss, in the form of:

$$\mathcal{L}_{\mathcal{C}_2} = r_i \log(\mathcal{C}_2(f^{in}_{\mathcal{C}_2}) + (1 - r_i) \log(1 - \mathcal{C}_2(f^{in}_{\mathcal{C}_2})). \tag{2}$$

We evaluate the first gripper's action by evaluating its potential for the second gripper to collaborate with.

Action Proposal Loss. $\mathcal{A}_1$ and $\mathcal{A}_2$ are implemented as cVAE [35]. For the j_{th} gripper, we adopt geodesic distance loss to measure the error between the reconstructed orientation $\hat{R}_j$ and ground-truth R_j, and KL Divergence to measure the difference between two distributions:

$$\mathcal{L}_{\mathcal{A}_j} = \mathcal{L}_{geo}(\hat{R}_j, R_j) + D_{KL}(q(z \mid \hat{R}_j, f^{in}_{\mathcal{A}_j}) \parallel \mathcal{N}(0, 1)). \tag{3}$$

TABLE I: Baseline comparisons of sample success rate (%).

Unseen instances in novel categories					
Method	Unfolding	Opening	Uncapping	Closing	Capping
Heuristic	24.70	19.82	31.33	33.74	43.20
M-Where2Act [33]	32.40	20.66	29.10	18.45	17.30
DualAfford [1]	33.50	21.90	19.60	25.50	35.00
Ours	**70.00**	**67.00**	**61.62**	**61.12**	**59.00**

For training Action Scoring Network $\mathcal{C}_1$ and $\mathcal{C}_2$, we balance the portion of successful cases and failed cases equally to ensure it can critically score. For training Action Proposal Network $\mathcal{A}_1$ and $\mathcal{A}_2$, we only use successful data for training. For training categories, we generate offline interaction data with multiple processes for training. For few-shot learning on novel categories, we sample affordances and propose actions to execute, collecting the online results to fine-tune networks.

V. EXPERIMENTS

To evaluate our framework's ability, we seek to answer the following research questions:

Q1: How successful are our proposed actions on unseen objects in novel categories, and how does it compare to other baselines for bimanual manipulation tasks? (Section V-C)

Q2: How's the cross-category affordance generalization ability via semantic correspondence, and how does it compare to the prior method without using the foundation model? (Section V-D)

Q3: How's the effectiveness of the few-shot adaptation strategy, and does it improve the performance compared with no adaptation? (Section V-E)

Q4: How efficient is few-shot adaptation, and how much data is required for fine-tuning? (Section V-F)

A. Environment Settings and Data

Environment Settings. Following Dualafford [1], we use SAPIEN [36] with NVIDIA PhysX for simulation, and we

TABLE II: Ablation studies on sample success rate (%). Training categories and novel categories (seen/unseen instances).

Method	Unfolding			Opening			Uncapping			Closing			Capping		
	Train	Novel		Train	Novel		Train	Novel		Train	Novel		Train	Novel	
		Seen	Unseen		Seen	Unseen		Seen	Unseen		Seen	Unseen		Seen	Unseen
Ours w/o AT	75.50	47.95	37.76	63.50	32.61	31.23	43.22	41.30	21.71	76.00	39.29	38.70	75.00	39.13	36.00
Ours w/o FA		62.50	62.96		61.90	47.83		53.19	52.94		45.65	48.72		59.52	57.58
Ours		**68.00**	**70.00**		**72.00**	**67.00**		**63.00**	**61.62**		**56.00**	**61.12**		**64.00**	**59.00**

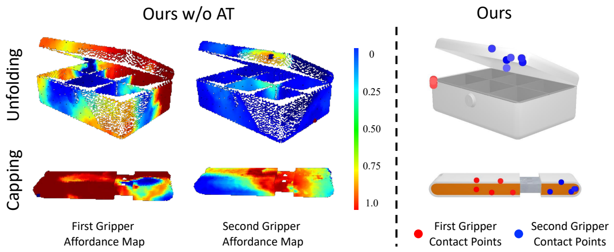


Fig. 5: Visualization of affordance comparison.

use two Franka Panda Flying grippers as robot actuators. **Data.** We conduct experiments with the large-scale PartNet-Mobility [37] and ShapeNet [38] dataset. We use the terms **training categories** and **novel categories** to denote whether objects of the same category are presented in the supporting set. In the **novel categories**, a smaller proportion (less than 30%) of instances that are exclusively used for few-shot learning are called **seen instances**, while many more unseen instances are only used for evaluation and are called **unseen instances**. In summary, we use 61 articulated objects spanning 6 categories.

B. Evaluation Metrics

Sample success rate. We use the sample success rate to assess the method's ability to propose successful actions. Notably, we compare our method using one budget of interaction data (50 demonstrations on novel categories) to compare with other baselines and ablation versions, shown in Table I and Table II.

C. Overall Performance

Baselines. To answer Q1, we compare our framework with several baselines, as shown in Table I. This comparison includes diverse methods to address the challenge of cross-category generalization for bimanual manipulation. They were evaluated on unseen instances in novel categories over 5 complex bimanual manipulation tasks.

Heuristic is a manually-scripted approach, where we acquire the ground-truth object poses and hand-engineer a set of rules for different tasks. For example, in the uncapping task, two contact points are sampled on each part of the prismatic joint, and the grippers are oriented to pull in opposite directions. **M-Where2Act** is a dual-gripper version of Where2Act [33] approach. While Where2Act originally considers interactions for a single gripper, we first train a separate model for each gripper and then combine them without considering collaboration. **DualAfford [1]** is a dual-gripper affordance learning method considering two arms' actions as a combination. Unlike **M-Where2Act**, it models the action of one gripper conditioned on the action of the other gripper, enabling coordinated dual-arm manipulation. This approach serves as our baseline for affordance transferring and few-shot adaptation. Experiments demonstrate that our method outperforms all baselines regarding success rate. Section IV-B.2 illustrates the manipulation results generated from our method. The **Heuristic** baseline shows a low success rate since the manually engineered rules are difficult to fit all objects with highly varied geometries of shapes. Moreover, this approach is time-consuming and labor-intensive, and impractical to generalize. The **M-Where2Act** approach struggles with tasks strongly demanding two arms' coordination. Although **DualAfford** achieves relatively high performance on training categories, its performance drops dramatically when encountering unseen objects from novel categories. The long training procedure with a large amount of data is also costly. Our work employs a similar bimanual-conditioned action learning method as prior knowledge.

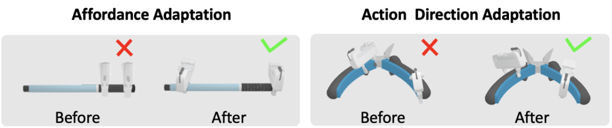


Fig. 6: Visualization of the adaptation results.

To answer Q2 and Q3, we compare to ablated versions of our framework to verify each component's effectiveness: **Ours w/o AT** is our method without transferring affordance based on semantic correspondence, where the Perception Module, after finetuning with 50 demonstrations of novel categories(the same as our full version's few-shot adaptation strategy), and then evaluated on unseen instances. **Ours w/o FA** is our method without a few-shot adaptation strategy, where the pre-trained Perception Module is directly evaluated on unseen instances with all mapped contact points via semantic correspondence. Table II clearly shows that each component improves our framework's performance. We detail the analysis in Section V-D and Section V-E.

D. Cross-category Affordance Generalization

Section IV-B.2 visualizes the affordance mapping results via semantic correspondence, illustrating the foundation model's strong power in cross-category affordance generalization. As shown in Figure 5, the affordance map generated from **Ours w/o AT** for novel categories is much less accurate than our method's contact-point-pair candidates.

Compared with **Ours w/o AT**, we observe that the semantic correspondence could act as an essential guidance for the cross-category affordance generalization, and thus

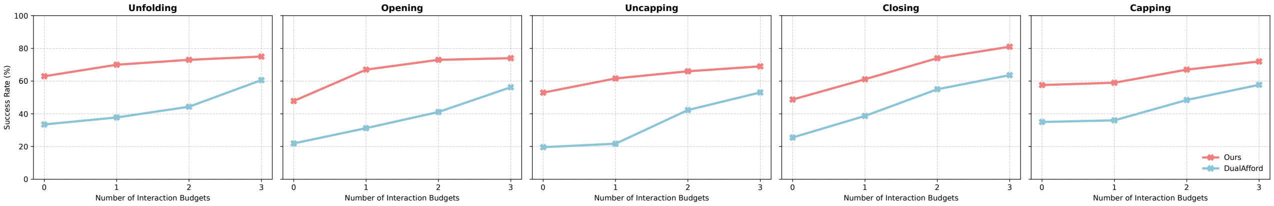
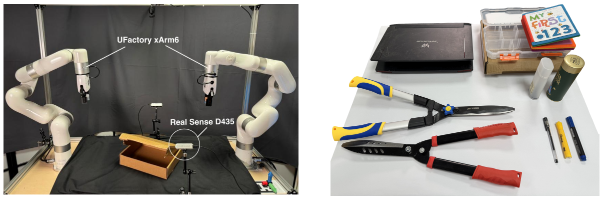
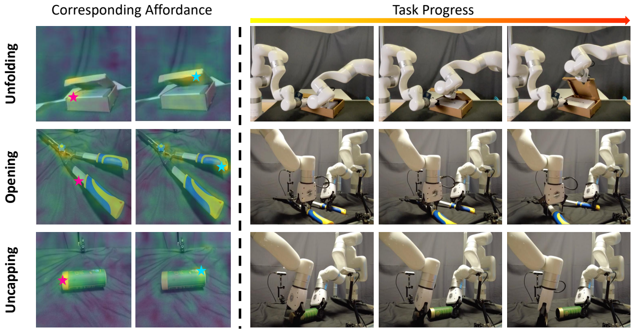


Fig. 7: Efficiency comparison of Ours and DualAfford with varying novel-category interaction data on unseen instances.



(a) Robot Setup (b) Objects



(c) Real-world tasks. Each row represents a task. From left to right, we respectively show the corresponding affordance on novel categories (★: first gripper, ★: second gripper) and the manipulation progress.

Fig. 8: Real-world Experiments.

significantly improves the overall performance (Q2).

E. Effectiveness of Adaptation

Compared with **Ours w/o FA**, which directly uses all contact points mapped by the foundation model and directions predicted by pre-trained networks, we find that this strategy can help filter negative contact points and adjust action directions to suit new categories.

As shown in Figure 6, after adaptation, the contact points selected and each gripper's orientation both tend to be more suitable for manipulation on novel categories, addressing Q3.

As discussed in Section IV-B, we try to find multiple contact point pairs through semantic correspondence approaches when encountering an unseen object, and these multiple pairs serve as candidates for sampling and selection. Since actions proposed by the pre-trained networks can lead to failure on unseen instances, the adaptation procedure adaptation eliminates negative candidates and inaccurate action directions.

F. Effeciency

To answer Q4, we evaluate the efficiency of our method on few-shot learning for novel categories with varying numbers of interaction data. Figure 7 shows the comparison with Dualafford using the same number of budgets of interactions(a budget of 50 interactions in total). Notably, our method's success rate at budget 0 is obtained via the semantic correspondence. We can see that, even with a small number of interactions, our method consistently outperforms DualAfford, demonstrating superior initial performance obtained from the foundation model. And the success rate of our approach rises faster and higher as the interaction budget increases. Particularly on the task **Closing**, 3 budgets of demonstrations lead to a dramatic increase (more than 30 %), compared to the performance without finetuning. These results highlight the advantage of our approach in achieving higher efficiency and better performance on unseen instances.

Fig. 9: Cross-category generalization for **unfolding** task.

G. Real-World Experiments

Figure 8 shows our real-world setup and objects used. We conducted real-world experiments using two UFactory xArm6 equipped with a UFactory xArm Gripper, and a front-view RealSense D435 camera to capture RGB-D observations. We use SAM [5] to segment a mask of the target object from the scene and project the segmented image. We obtain the object's mesh using AR code [39] and estimate its pose using FoundationPose [40]. Subsequently, we sample the object's point cloud based on the estimated pose. We first obtain the corresponding affordance on the 2D image, and then back-project the pixel-level points to get the 3D contact points based on depth information, as described in Section IV-B. Given the task, object point cloud, and the resulting points, our framework can propose bimanual action. Figure 8c shows our real experiments on different tasks. And Figure 9 illustrates our framework's superior generalization that it can successfully manipulate varied unseen instances across categories for the **unfolding** task.

VI. CONCLUSIONS

We propose Bi-Adapt, a foundation-model-based framework for bimanual manipulation generalization with adaptation. Bi-Adapt transfers affordance across categories via semantic correspondence. We introduce a few-shot learning strategy to adapt the pre-trained model's action direction prediction on novel categories with limited interactions. Experiments evaluate our framework's improved performance on adaptation for unseen instances in novel categories.

Limitation and Future Work. Currently, this work is unable to complete more complex long-horizon tasks. A unified framework could be considered for further research.

ACKNOWLEDGMENT

This research/project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2024).

REFERENCES

[1] Y. Zhao, R. Wu, Z. Chen, Y. Zhang, Q. Fan, K. Mo, and H. Dong, "Dualafford: Learning collaborative visual affordance for dual-gripper manipulation," 2023. [Online]. Available: https://arxiv.org/abs/2207.01971

[2] A. Bahety, P. Mandikal, B. Abbatematteo, and R. Martín-Martín, "Screwmimic: Bimanual imitation from human videos with screw space projection," 2024. [Online]. Available: https://arxiv.org/abs/2405.03666

[3] Y. Mu, T. Chen, S. Peng, Z. Chen, Z. Gao, Y. Zou, L. Lin, Z. Xie, and P. Luo, "Robotwin: Dual-arm robot benchmark with generative digital twins (early version)," *arXiv preprint arXiv:2409.02920*, 2024.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[7] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, "Distilled feature fields enable few-shot language-guided manipulation," *arXiv preprint arXiv:2308.07931*, 2023.

[8] Q. Wang, H. Zhang, C. Deng, Y. You, H. Dong, Y. Zhu, and L. Guibas, "Sparsedff: Sparse-view feature distillation for one-shot dexterous manipulation," *arXiv preprint arXiv:2310.16838*, 2023.

[9] 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," *arXiv preprint arXiv:2401.07487*, 2024.

[10] Y. Wang, G. Yin, B. Huang, T. Kelestemur, J. Wang, and Y. Li, "Gendp: 3d semantic fields for category-level generalizable diffusion policy," *arXiv preprint arXiv:2410.17488*, 2024.

[11] Y. Chen, Y. Geng, F. Zhong, J. Ji, J. Jiang, Z. Lu, H. Dong, and Y. Yang, "Bi-dexhands: Towards human-level bimanual dexterous manipulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 2804–2818, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265801841

[12] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations," 2021. [Online]. Available: https://arxiv.org/abs/2107.14483

[13] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, *et al.*, "Dual arm manipulation—a survey," *Robotics and Autonomous systems*, vol. 60, no. 10, pp. 1340–1353, 2012.

[14] S. S. Mirrazavi Salehian *et al.*, "A unified framework for coordinated multi-arm motion planning," *The International Journal of Robotics Research*, vol. 37, no. 10, pp. 1205–1232, 2018.

[15] A. Suarez, A. E. Jimenez-Cano, V. M. Vega, G. Heredia, A. Rodriguez-Castaño, and A. Ollero, "Design of a lightweight dual arm system for aerial manipulation," *Mechatronics*, vol. 50, pp. 30–44, 2018.

[16] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang, "Towards human-level bimanual dexterous manipulation with reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5150–5163, 2022.

[17] Y. Lin, A. Church, M. Yang, H. Li, J. Lloyd, *et al.*, "Bi-touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning," *IEEE Robotics and Automation Letters*, 2023.

[18] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Goal-conditioned dual-action imitation learning for dexterous dual-arm robot manipulation," *IEEE Transactions on Robotics*, 2024.

[19] R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, *et al.*, "Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects," *arXiv preprint arXiv:2106.14440*, 2021.

[20] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu, *et al.*, "Rt-trajectory: Robotic task generalization via hindsight trajectory sketches," *arXiv preprint arXiv:2311.01977*, 2023.

[21] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[22] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[23] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[24] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, "Emergent correspondence from image diffusion," *Advances in Neural Information Processing Systems*, vol. 36, pp. 1363–1389, 2023.

[25] Y. Wang, M. Zhang, Z. Li, T. Kelestemur, K. Driggs-Campbell, J. Wu, L. Fei-Fei, and Y. Li, "D^{3}fields: Dynamic 3d descriptor fields for zero-shot generalizable rearrangement," 2024. [Online]. Available: https://arxiv.org/abs/2309.16118

[26] K. Farid and N. Sakr, "Few-shot system identification for reinforcement learning," in *2021 6th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*. IEEE, 2021, pp. 1–7.

[27] K. Rakelly, A. Zhou, C. Finn, *et al.*, "Efficient off-policy meta-reinforcement learning via probabilistic context variables," in *International conference on machine learning*. PMLR, 2019, pp. 5331–5340.

[28] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[29] R. C. Julian, B. Swanson, G. S. Sukhatme, S. Levine, C. Finn, and K. Hausman, "Efficient adaptation for end-to-end vision-based robotic manipulation," *ArXiv*, vol. abs/2004.10190, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216036080

[30] T. Kagaya, Y. Lou, T. J. Yuan, S. Lakshmi, J. Karlekar, S. Pranata, N. Murakami, A. Kinose, K. Oguri, F. Wick, and Y. You, "Envbridge: Bridging diverse environments with cross-environment knowledge transfer for embodied ai," 2024. [Online]. Available: https://arxiv.org/abs/2410.16919

[31] 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 *European conference on computer vision*. Springer, 2022, pp. 90–107.

[32] C. Ning, R. Wu, H. Lu, *et al.*, "Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[33] K. Mo, L. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," 2021. [Online]. Available: https://arxiv.org/abs/2101.02692

[34] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[35] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[36] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 097–11 107.

[37] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[39] AR Code, "Ar code," 2022, accessed: 2024-09-28. [Online]. Available: https://ar-code.com/

[40] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 868–17 879.