A3D: Adaptive Affordance Assembly with Dual-Arm Manipulation

Jiaqi Liang[1*], Yue Chen[1*], Qize Yu[1*], Yan Shen[1], Haipeng Zhang[1], Hao Dong[1], Ruihai Wu[1†]
[1]Peking University
jiaqiliang@stu.pku.edu.cn, yuechen@stu.pku.edu.cn, wuruihai@pku.edu.cn

Abstract

Furniture assembly is a crucial yet challenging task for robots, requiring precise dual-arm coordination where one arm manipulates parts while the other provides collaborative support and stabilization. To accomplish this task more effectively, robots need to actively adapt support strategies throughout the long-horizon assembly process, while also generalizing across diverse part geometries. We propose A3D, a framework which learns adaptive affordances to identify optimal support and stabilization locations on furniture parts. The method employs dense point-level geometric representations to model part interaction patterns, enabling generalization across varied geometries. To handle evolving assembly states, we introduce an adaptive module that uses interaction feedback to dynamically adjust support strategies during assembly based on previous interactions. We establish a simulation environment featuring 50 diverse parts across 8 furniture types, designed for dual-arm collaboration evaluation. Experiments demonstrate that our framework generalizes effectively to diverse part geometries and furniture categories in both simulation and real-world settings.

Introduction

Robotic furniture assembly (Funkhouser et al. 2011; Jones et al. 2021; Lee, Hu, and Lim 2021; Tian et al. 2022, 2025), the task of combining functional components such as chair base, legs, and arms into a fully constructed shape, with a focus on both the overall structure and functions of each part, is a critical capability for home-assistive robots.

Recent studies have addressed various aspects of robotic assembly, including motion planning (Suárez-Ruiz, Zhou, and Pham 2018; Sundaram, Remmler, and Amato 2001; Le, Cortés, and Siméon 2009; Zhang et al. 2020b), assembly pose estimation (Yu et al. 2021; Huang et al. 2020; Tie et al. 2025b; Jones et al. 2021; Shen et al. 2025), and RL-based combinatorial sequence search (Xu et al. 2023; Zhang, Tomizuka, and Li 2024; Funk et al. 2022; Ghasemipour et al. 2022). However, current robotic systems remain limited in their ability to assemble objects across diverse categories. Prior research has primarily focused on specific object types using a single robot arm (Heo et al. 2023; Lee, Hu, and Lim 2021). In contrast, generalizable furniture assembly requires vision understanding and bi-manual operation that frequently changes which part to hold to counter-balance the insertion force from the other hand. This presents new challenges to vision perception and precise manipulation. First,

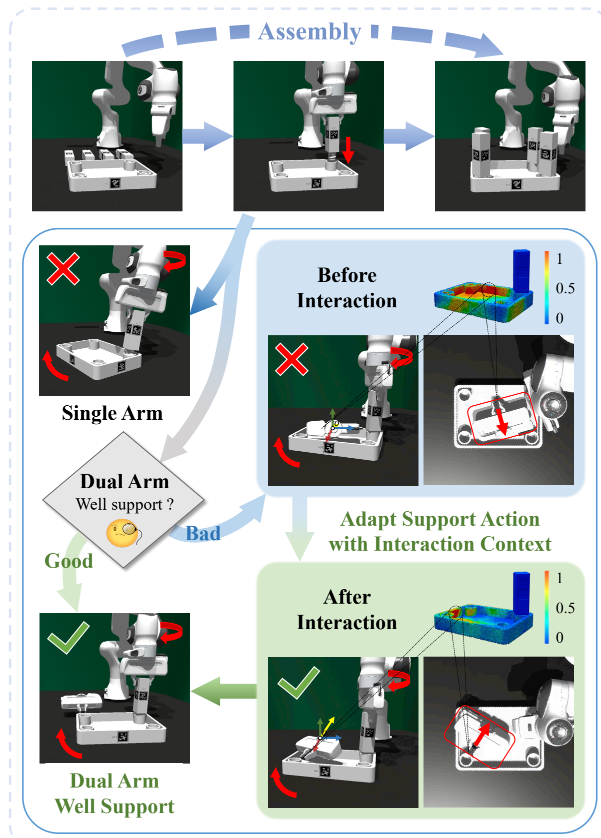

Figure 1: Procedure of assembling a furniture (Row 1). **Single Arm** may not stably assemble parts and a second robot is then introduced. **Before Interaction**, part kinematics and dynamics, indicated by affordance, are ambiguous, and the interaction may fail. **After Interaction**, the adapted affordance proposes actions for stable support during assembly.

[*]These authors contributed equally.
[†]Corresponding Authors.

assembling unseen furniture demands understanding functional affordances across various part geometries, requiring robots to identify viable support locations. Second, long-horizon assembly induces sequential state transitions where support strategies must dynamically refine based on part relations. Third, fine-grained assembly requires robust control skills to achieve precisly during contact-rich interactions.

To bridge these gaps, we introduce **A3D**: a framework that learns **A**daptive **A**ffordance **A**ssembly for collaborative **D**ual-arm manipulation. To enable geometric awareness, A3D leverages affordance as a representation of per-point actionability on objects for furniture assembly tasks. These per-point features are extracted hierarchically from local to global, effectively capturing detailed local geometry information for support and stabilization, as well as the contextual part relations that indicate whether the action would disturb other parts. This hierarchical structure enables A3D to localize stable support regions through fine-grained geometric cues while modeling higher-level part contexts to anticipate potential disturbances during manipulation.

However, static affordance derived solely from passive observations fails to account for critical kinematic (e.g., joint locations and limits) and dynamic uncertainties (e.g., contact direction and force), which might misdirect manipulation (Fig. 1). So we actively incorporate interaction feedback into affordance predictions, enabling dynamic adjustment of support strategies throughout assembly.

Although existing simulation environments have facilitated progress in robotic manipulation, they remain limited in supporting the study of dual-arm furniture assembly. Previous works predominantly focus on single-arm manipulation or utilize limited furniture assets (Niekum et al. 2013; Suárez-Ruiz, Zhou, and Pham 2018; Kimble et al. 2020; Heo et al. 2023; Zhang, Tomizuka, and Li 2024; Jones et al. 2021; Li et al. 2020), failing to capture the unique physical and coordination challenges inherent in dual-arm assembly scenarios. To bridge this critical gap, we introduce a new evaluation environment extending FurnitureBench (Heo et al. 2023), featuring 4 assembly task categories with 50 geometrically diverse parts across 8 furniture types. Both qualitative and quantitative results from simulations and real-world experiments demonstrate the effectiveness of our framework. We also note that while our adaptation allows iterative refinement (max 3 rounds), most test cases succeeded after a single interaction (effective $k = 1$), demonstrating robustness and efficiency.

In conclusion, our contributions mainly include:

- We propose affordance learning framework for generalizable support and stabilization prediction in furniture assembly, enabling generalization across diverse parts.
- We further develop an adaptive module that uses interaction feedback to dynamically adjust support strategies during assembly based on previous interactions.
- We build a simulation environment for dual-arm collaborative assembly featuring 50+ geometrically diverse parts across 8 furniture types and 4 task categories.
- Extensive experiments in both simulation and real world demonstrate the effectiveness of our framework.

Related Work

Furniture Assembly

Furniture assembly is a prominent application in shape assembly, where individual components, each serving a distinct functional role (*e.g.*, chair arm, table leg), must be assembled following both geometric constraints and commonsense spatial and functional relations (e.g., a chair leg must be attached to the seat base with proper orientation and stability). The complexity arises from the need to reason about part functionality, structural dependencies, and physical constraints simultaneously. Previous research has mostly focused on assembly pose estimation (Li et al. 2020; Yu et al. 2021; Li et al. 2024; Huang et al. 2020). For instance, Li et al. (2020) learns to assemble 3D shapes from 2D images, while Huang et al. (2020) proposes image-free generative models for pose generation. However, these methods might neglect the challenges in dynamic robotic execution, particularly the need for precise dual-arm coordination, where one arm manipulates one part while the other actively provides collaborative support and stabilization throughout assembly. Addressing this challenge, especially adaptive support strategies and generalization across diverse geometries, is a core objective of our work.

Visual Affordance for Robotic Manipulation

Visual affordance (Gibson 1977) suggests possible ways for agents to interact with objects for various manipulation tasks. This approach has been widely used in grasping (Corona et al. 2020; Kokic, Kragic, and Bohg 2020; Zeng et al. 2018), articulated manipulation (Yuan et al. 2024; Tie et al. 2025a), and scene interaction (Nagarajan and Grauman 2020; Nagarajan et al. 2020). Point-level affordance, in particular, assigns an actionability score to each point, and thus enables fine-grained geometry understanding and improved cross-shape generalization in diverse tasks, such as articulated (Mo et al. 2021; Wang et al. 2022; Chen et al. 2024), and deformable (Wu et al. 2024; Wu, Ning, and Dong 2023; Wu et al. 2025; Wang et al. 2025) manipulation. For furniture assembly scenarios, where parts vary significantly in geometry and require precise dual-arm collaboration, we empower point-level affordance with the awareness of part geometry, and further leverage active interactions to efficiently query uncertain kinematic or dynamic factors for learning more accurate instance-adaptive visual affordance.

Method

Our goal is to enable effective dual-arm coordination for furniture assembly, where a tool arm executes assembly operations while a support arm provides adaptive stabilization to prevent part displacement and ensure task success. As shown in Fig. 2, our framework integrates two key components: (1) **Support Affordance Module** predicts initial affordance heatmaps and corresponding action directions from visual observations and operation points; (2) **Interaction Context Adaptation Module** leverages physical feedback from interaction history to adjust affordance predictions.

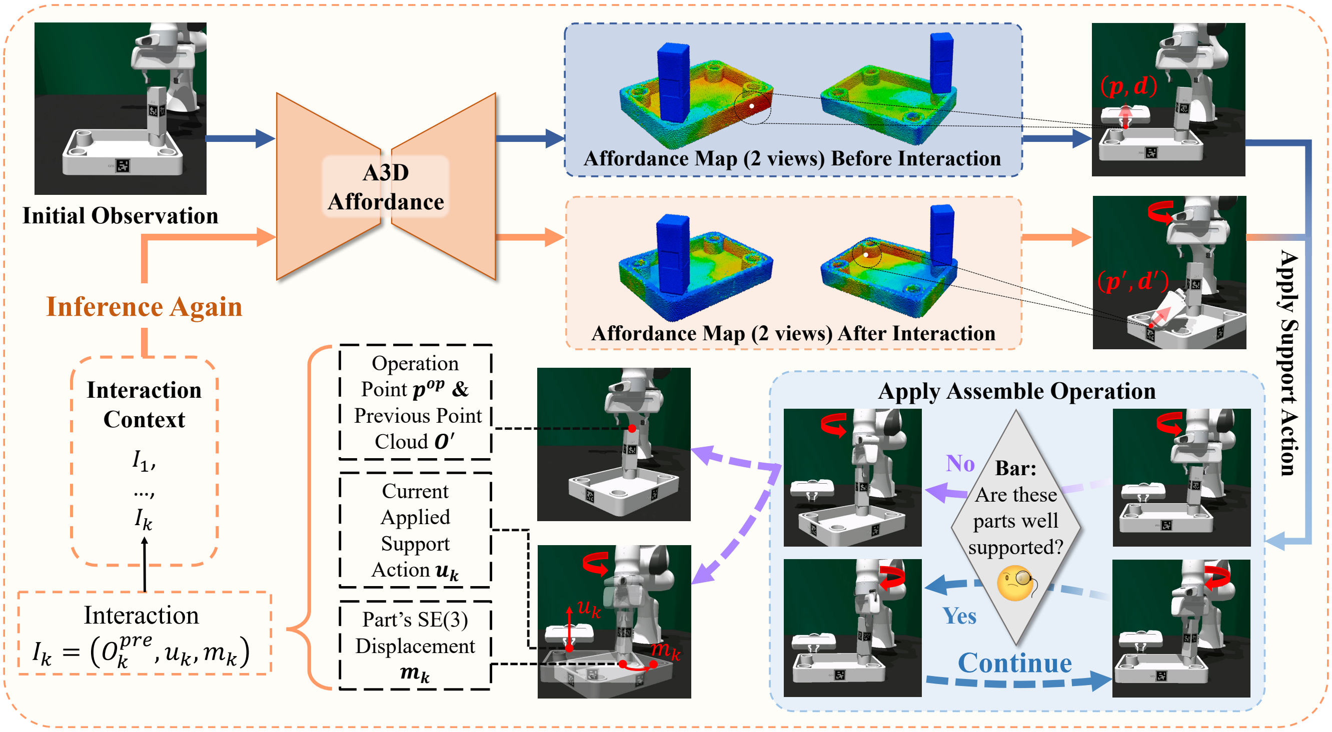

Figure 2: **Framework Overview.** At each operation stage, the policy takes the point cloud and the selected action point as inputs to predict the support action. The robot moves the gripper to the recommended pose to support the assembly. If part displacement occurs—indicating insufficient support—the system logs the pre-support point cloud, executed action, and displacement as interaction context, then re-predicts the support action using the updated point cloud and accumulated context.

Problem Formulation

We formulate this as learning a closed-loop adaptive policy $\pi(u_t|S_t, I_t)$. At each timestep t, the policy predicts the stabilization action u_t for the support arm, conditioned on the observed state S_t, and the interaction context I_t which records the history of previous assembly trials.

State: $S_t = (O_t, p_t^{op})$, where $O_t \in \mathbb{R}^{N\times6}$ represents a 3D partial point cloud of the furniture parts with surface normals, and p_t^{op} denotes the operation point where the tool gripper contacts the target part.

Action: $u_t = (p_t^{sp}, \mathbf{d_t})$, where $p_t^{sp} \in O_t$ is the support point and $\mathbf{d_t} \in SO(3)$ is the support gripper orientation.

Interaction Context: $I_t = \{(O_i, u_i, m_i)\}_{i=t-k}^{t-1}$ stores information from previous k interaction steps, where m_i denotes the base part displacement after step is.

Task Success: An episode succeeds if the primary operation reaches its geometric goal while maintaining base part displacement $m_i < \epsilon$ during execution.

Support Affordance Module

The Support Affordance Module employs an affordance–proposal–scoring architecture: the Affordance submodule predicts affordance maps and selects top-K candidate points; the Proposal submodule generates multiple candidate directions for each point; the Scoring submodule scores all point–direction pairs and selects the optimal support action (Steps 1–3, Fig. 3).

Visual Feature Extractor. PointNet++ (Qi et al. 2017) generates point-wise features $f_{p_i} \in \mathbb{R}^{128}$ from the point cloud O. Operation and support points are encoded via shared MLPs into $f_{op}, f_{sp} \in \mathbb{R}^{32}$, while gripper direction $\mathbf{d}$ and displacement m are encoded into $f_{\mathbf{d}}, f_m \in \mathbb{R}^{32}$.

Affordance Module. Module $\mathcal{A}$ predict an affordance score $a_p \in [0, 1]$ for each point p. It concatenates the operation-point feature $f_{p^{op}}$, the point feature f_{p_i} the operation-point encoding f_{op}, and interaction context f_I, feeds this into the MLP, and outputs a_{p_i}. The top-K points by score are then selected as support candidates.

Action Proposal Module. Action Proposal Module $\mathscr{P}$ implements a Conditional Variational Autoencoder (cVAE). The encoder processes the operation point feature $f_{p^{op}}$, candidate support point feature $f_{p^{sp}}$ in point cloud features, operation point embedding f_{op}, support point embedding f_{sp}, and interaction context feature f_I, to output latent vector $z \in \mathbb{R}^{128}$. The decoder then generates direction vector $\mathbf{d}$ from z.

Action Scoring Module. Action Scoring Module $\mathscr{S}$ predicts success scores $c \in [0, 1]$ for each action. An MLP takes concatenated features $f_{p^{op}}$, $f_{p^{sp}}$, f_{op}, f_{sp}, and $f_{\mathbf{d}}$ and outputs the success likelihood. A higher c suggests a greater chance for the support hand to collaborate effectively and complete the task.

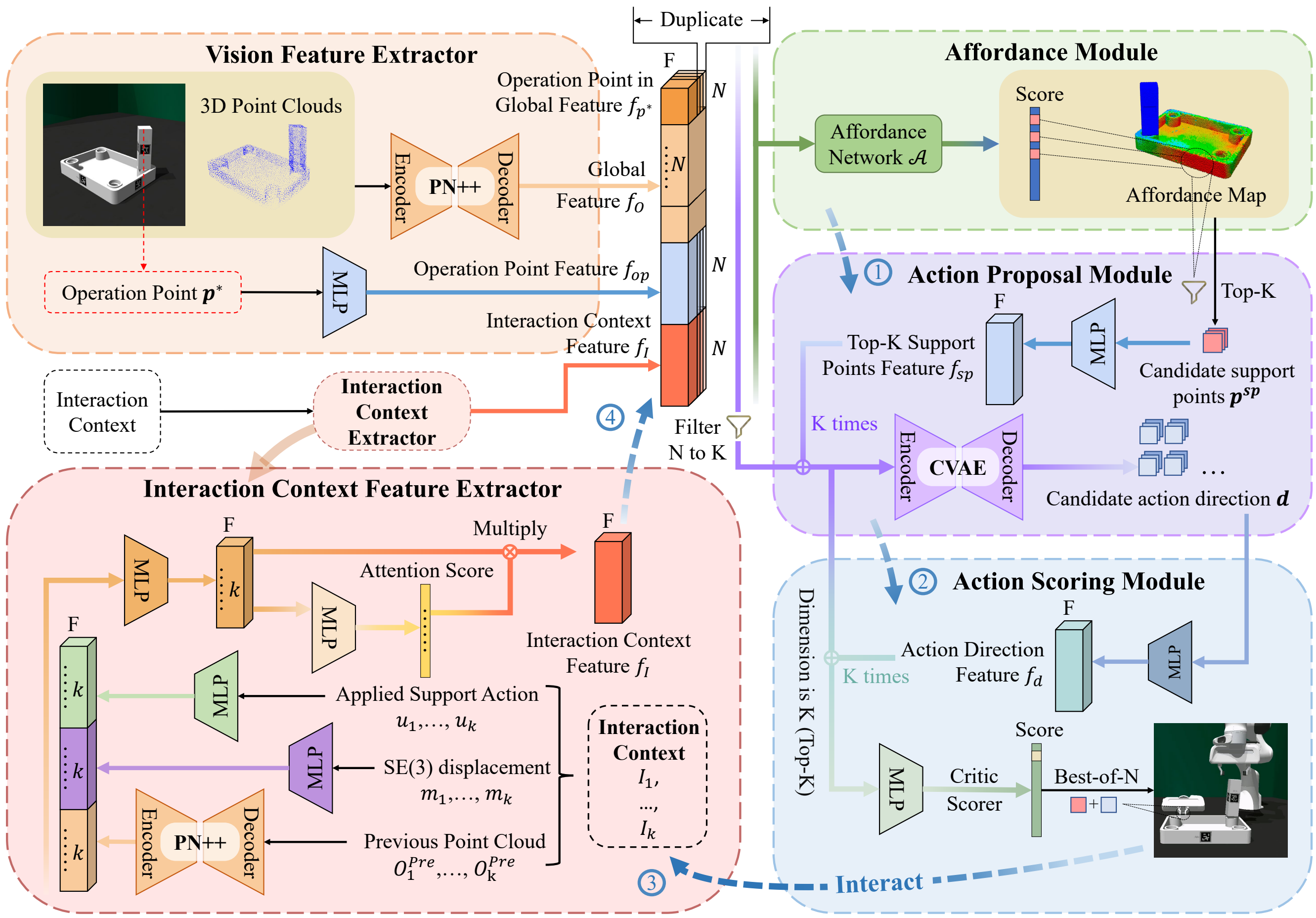

Figure 3: **Point-Level Adaptation Support Affordance Framework.** The model completes support decisions by extracting visual features, computing Top-K point-level affordances and generating candidate directions, scoring and selecting point–direction pairs, and extracting interaction context features.

Interaction Context Adaptation Module

If visual priors are insufficient (Fig. 3, step 4), the system records the support action and its feedback. The Context Extractor derives features from these records, concatenates them with current visual features, and feeds them back to the Affordance, Proposal, and Scoring submodules to refine predictions that adhere to physical dynamics.

Interaction Context Extractor Module. For interaction context $I_t = (O_i, u_i, m_i)_{i=t-k}^{t-1}$, we extract features for each historical step using the same encoders as above. Features are combined via

$$f_{I_i} = MLP(concat(f_{O_i}, f_{u_i}, f_{m_i})), i \in [t-k, t-1]. \quad (1)$$

To aggregate information from all previous interactions, we adopt a lightweight attention mechanism. Each previous interaction feature f_{I_i} is passed through an MLP to compute an attention weight w_i, and the final interaction context feature is obtained as a weighted average:

$$f_I = \frac{\sum_{i=t-k}^{t-1} f_{I_i} \times w_i}{\sum_{i=t-k}^{t-1} w_i}. \quad (2)$$

Train and Loss

Action Scoring Loss. The Action Scoring Module predicts a success score $\hat{r}$ and is trained with an MSE loss against a "real" score r . This real score combines the object's SE(3) movement distance g_d and a task-completion term g_c—the latter decreasing as completion improves—via weighted sum, then clamps the result to [0,1]:

$$r = \text{clamp}((1 - (\alpha \times g_d + \beta \times g_c)), 0, 1) \quad (3)$$

where α and β balance the distance and completion.

Action Proposal Loss. We evaluate the loss using cosine similarity and Kullback-Leibler (KL) divergence.

1. **Cosine Similarity Loss**: We use a cosine-similarity loss $\mathcal{L}_{\text{cosine}} = 1 - (\hat{\mathbf{d}} \cdot \mathbf{d})/(\|\hat{\mathbf{d}}\|\|\mathbf{d}\|)$ to align the predicted direction $\hat{\mathbf{d}}$ with the ground-truth $\mathbf{d}$.
2. **KL Divergence Loss**: We add a KL-divergence term to regularize the latent variable z inferred from $\hat{\mathbf{d}}$ and f^{in} towards a standard normal:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q(z|\hat{\mathbf{d}}, f^{in})||\mathcal{N}(0,1) \quad (4)$$

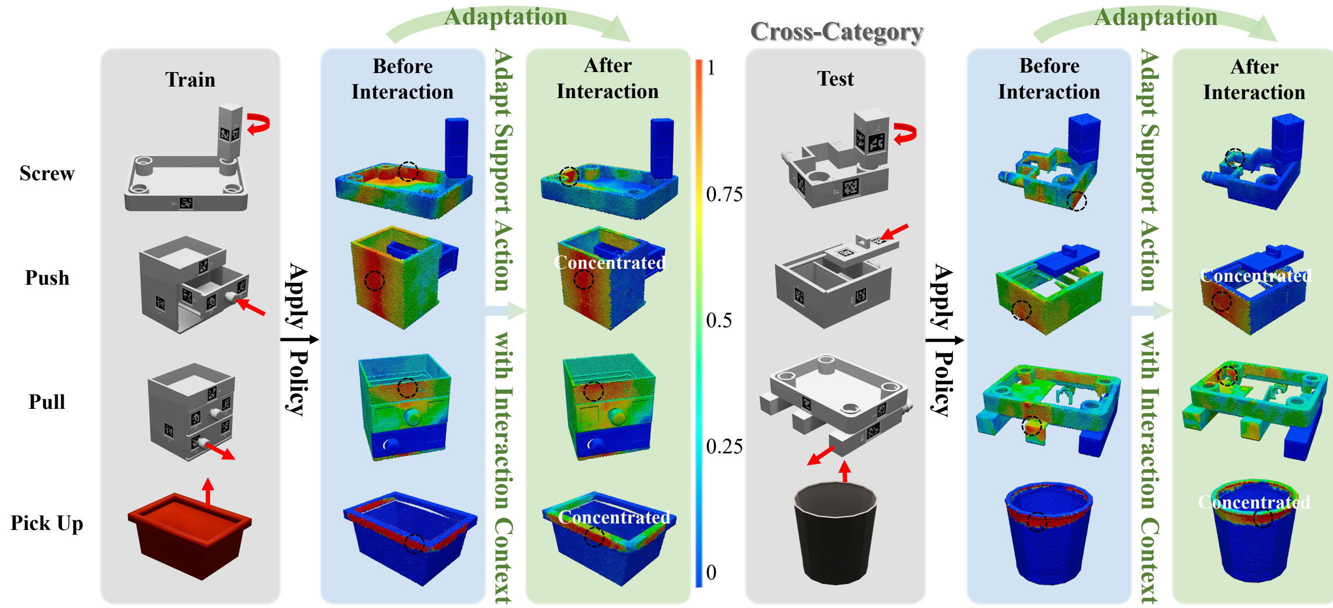

Figure 4: **Affordance Map.** The figure displays affordance heatmaps generated for various objects in simulation before and after interaction. Red arrows indicate the direction of part movement, and circled regions denote the highest-scoring areas. In the subplots labeled "Concentrated," it is evident that after interaction, high-scoring points converge more tightly at the correct locations; in the other subplots, the high-scoring points have shifted in accordance with the observed movement trends.

The overall Action Proposal loss then balances direction alignment and latent regularization:

$$\mathcal{L}_{proposal} = \lambda_{dir}\mathcal{L}_{cosine} + \lambda_{KL}\mathcal{L}_{KL} \tag{5}$$

where λ_{dir} and λ_{KL} weight the cosine similarity and KL terms, respectively.

Affordance Prediction Loss. Similar to Where2Act and DualAfford, we define each point's affordance score a as the predicted success probability of actions proposed by the Action Proposal Module, and evaluated by the Action Scoring Module. Concretely, for each point p_i we sample N support directions, score them via Action Scoring Network to obtain N action scores and average the top.

$$a_{p_i} = \frac{1}{K}\sum_{j=1}^{K} \mathscr{S}(f_{pi}^{in}, \mathscr{P}(f_{pi}^{in}, z_j)) \tag{6}$$

We then apply L1 loss to measure the difference between the predicted affordance score $\hat{a_{p_i}}$ and the ground-truth a_{p_i}:

$$\mathcal{L}_{\text{affordance}} = |\hat{a_{p_i}} - a_{p_i}| \tag{7}$$

Experiment

Setup

Environment. We build upon FurnitureBench in IsaacGym by extending it to support dual-arm coordination and modifying camera configurations, allowing us to study the collaborative support and stabilization using a second arm. To boost and evaluate policy generalization, we extend the assets by increasing object geometric diversity. For training, we collect 10k samples per task focusing on specific furniture types (e.g., desk, drawer, basket), each with multiple variants. Testing utilizes entirely unseen furniture types to validate cross-category generalization.

Tasks. We evaluate on four fundamental assembly operations: (1) **Screwing**: rotating components while the support arm provides counteracting force; (2) **Insertion**: pushing components along rails with support arm guidance; (3) **Extraction**: pulling components while the support arm stabilizes the base; (4) **Picking**: lifting and placing with dual-arm coordination.

Metrics. We use success rate as the evaluation metric. Success requires: target component reaching desired pose within tolerance, base structure remaining stable (displacement/rotation below thresholds), and secure grasping above specified height for picking tasks.

Baselines and Ablations

Our work targets adaptive support in dual-arm assembly, a novel setting not directly addressed by prior work. Thus, no existing method serves as a direct SOTA baseline. We compare against the following baselines and ablations:

- **Random**: Random selection of support points and directions.
- **Heuristic**: Support point selection by geometric rules.
- **3D Diffusion Policy (DP3)** (Ze et al. 2024): Imitation learning for support prediction with point cloud input.
- **LLM-Guided** (Comanici et al. 2025): Inferring support point and action using Gemini 2.5 Pro.

Method	Train Categories				Test Categories			
	Screw	Push	Pull	Pick Up	Screw	Push	Pull	Pick Up
Random	10.7%	11.1%	5.0%	13.9%	9.0%	7.2%	4.0%	6.5%
Heuristic	54.5%	70.9%	43.1%	37.5%	46.7%	52.8%	31.9%	31.4%
DP3	23.2%	41.5%	19.4%	22.9%	17.4%	22.1%	10.1%	11.5%
LLM-Guided	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
w/o Top-K	66.7%	73.5%	67.7%	**66.1%**	54.1%	53.7%	52.4%	41.1%
w/o Adaptation	54.9%	74.3%	76.3%	52.9%	34.2%	63.9%	55.6%	42.2%
Ours	**70.7%**	**80.6%**	**80.0%**	62.0%	**56.3%**	**67.9%**	**61.7%**	**47.1%**

Table 1: Comparison of baseline and ablation variants on the success rate metric.

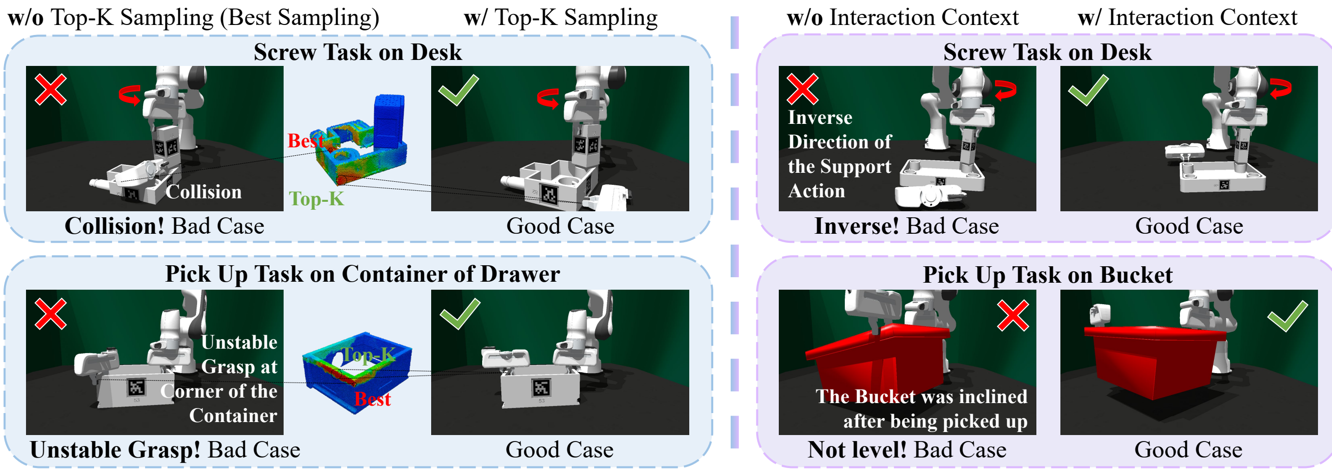

Figure 5: **Qualitative Analysis of Ablations.** (Left) Without Top-K sampling, the robot fails to find robust manipulation points. (Right) Without interaction context, the robot lacks physical awareness to adjust its actions.

To demonstrate the necessity of the proposed module, we compare with the following ablated versions:

- **w/o Top-K**: Selecting only the single highest-scoring point instead of Top-K candidates.
- **w/o Adaptation**: Removing adaptation with interaction.

Results and Analysis

Fig. 4 demonstrates the predicted affordance before and after the interactions, for different tasks, over training and novel object categories. Before the interaction, the learned affordance might be ambiguous (indicating a larger number of points that are plausible for manipulation) due to the uncertainty of object kinematics and dynamics. After a support action executed by the second robot, on the point selected by the proposed affordance, the affordance will be adapted by the interaction feedback. Eventually, the manipulation regions indicated by the adapted affordance will be more concentrated on plausible support points.

Moreover, the learned and adapted affordance, and the corresponding policy can generalize to novel geometries and categories, as point-level affordance aggregates both the low geometry (indicating where can be manipulated) and overall structure (indicating where to support).

Tab. 1 shows the quantitative results, and our proposed framework outperforms all baseline and ablation methods. **Heuristic** method, though more effective than **Random** actions, requires manual rule design for each task and even object. **DP3** lacks the understanding of diverse shapes and categories. **LLM-Guided** approaches lack essential 3D geometry and a low-level fine-grained action understanding for precise manipulation.

For the analysis of **ablations**, Tab. 1 and Fig. 5 together showcase the effectiveness of the proposed components.

For **Top-K Sampling**, generates a wider set of high-quality candidate actions for the support goal, for the following Action Scoring Module to further select the best actions. On the contrary, if the framework only selects the best point indicated by the learned affordance, chances are that on this selected point the best action direction is worse than the action directions sampled on other points with high affordance scores. The left side of Fig. 5 illustrates two failure cases when Top-K sampling is omitted. In the top-left case, although the affordance module provides the highest-scoring contact point, its combination with the direction proposed by the action module fails to provide optimal support due to the collision problem. In the bottom-left case, the highest-scoring contact point is located on a corner of the box that yields an unstable grasp, making successful execution highly improbable. In our scenarios with complex geometries and

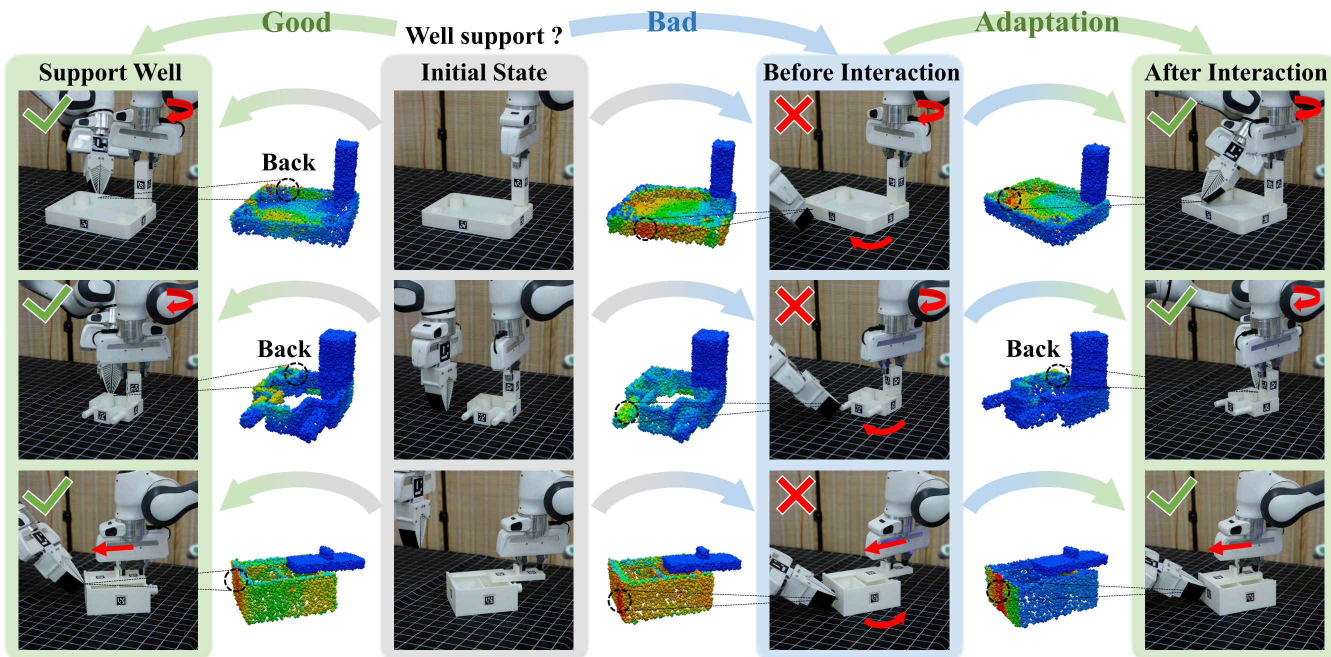

Figure 6: **Real-World Experiments.** We validate our framework in real-world conditions. The experiments include three scenarios:screw a desk leg, screw chair leg, and push a cabinet door. The left path ("Good") shows our policy directly finding a stable support. The right path ("Bad" to "Adaptation") show the ability to adapt its support strategy to ultimately succeed.

diverse tasks, **Top-K sampling** effectively expands the high-quality search space, markedly improving the probability of selecting the best action.

The **Adaptation Mechanism** based on the interaction context enhances the perception of real-time physical properties, endowing the model with 'physical awareness'. The right side of Fig. 5 presents failure examples when this module is removed. In the top-right of the "Screw" task, although providing a seemingly correct support action, without interaction contexts, the model is unaware of the complex thread direction. This leads to an inverse action that cannot support the tightening operation well. In the bottom-right of the "Pick-up" task, the initial grasp causes the bucket to incline; by incorporating this tilt as interaction feedback, the model can automatically adjust the contact point and successfully lift the target object. As shown in Fig. 4, after incorporating interaction context, the model not only highlights contactable regions more accurately, but also reveals differences in physical interaction properties such as force direction and stability, significantly enhancing its perception and understanding of interaction states.

Real-Word Experiments

We set up two Franka Panda with the furniture positioned between them. Three RealSense cameras capturing 3D point cloud are mounted around the scene. Robot control is managed through ROS (Quigley et al. 2009) and the frankapy library (Zhang et al. 2020a). Fig. 6 demonstrates the complete pipeline from scene perception and adaptive affordance prediction based on interaction feedback.

We evaluate each task over 15 trials with varying furniture configurations on 3 tasks. As shown in Tab. 2, our method significantly outperforms baselines and achieves high success rates in real-world assembly tasks. Fig. 6 shows real-world observations, affordance and adaptation. Additional videos are provided in the supplementary material.

Method	Screw	Push	Pick Up
Random	0 / 15	0 / 15	0 / 15
Heuristic	8 / 15	10 / 15	5 / 15
DP3	4 / 15	6 / 15	5 / 15
Ours	**11 / 15**	**12 / 15**	**9 / 15**

Table 2: Real world experimental results.

The primary failure mode in real-world stems from motion planning limitations.The RRTConnect algorithm cannot find feasible trajectories due to robotic arm or environmental constraints. In the future work, we plan to develop a policy for motion refinement to improve real-world robustness.

Conclusion

We propose A3D, a framework that learns adaptive affordances for dual-arm furniture assembly by identifying optimal support and stabilization locations. Our approach combines dense geometric representations for cross-geometry generalization with an adaptive module that leverages interaction feedback to dynamically adjust strategies. Experiments demonstrate superior performance in both simulation and real-world settings.

Acknowledgments

The authors gratefully acknowledge the hard work and close collaboration of the team members. In particular, Jiaqi contributed to the code implementation, experimental work, and real-robot experiments; Yue provided mentoring and contributed to experimental design, real-robot experiments, and manuscript preparation; Qize contributed to portions of the experiments, including real-robot experiments, as well as figure preparation and website development. Yan assisted with manuscript polishing, and Ruihai contributed to idea generation and overall research advising.

The funding for this work was unexpectedly withdrawn after the camera-ready submission, and Ruihai covered all associated costs.

References

Chen, Y.; Tie, C.; Wu, R.; and Dong, H. 2024. EqvAfford: SE(3) Equivariance for Point-Level Affordance Learning. arXiv:2408.01953.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Corona, E.; Pumarola, A.; Alenya, G.; Moreno-Noguer, F.; and Rogez, G. 2020. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5031–5041.

Funk, N.; Chalvatzaki, G.; Belousov, B.; and Peters, J. 2022. Learn2assemble with structured representations and search for robotic architectural construction. In *Conference on Robot Learning*, 1401–1411. PMLR.

Funkhouser, T.; Shin, H.; Toler-Franklin, C.; Castañeda, A. G.; Brown, B.; Dobkin, D.; Rusinkiewicz, S.; and Weyrich, T. 2011. Learning how to match fresco fragments. *Journal on Computing and Cultural Heritage (JOCCH)*, 4(2): 1–13.

Ghasemipour, S. K. S.; Kataoka, S.; David, B.; Freeman, D.; Gu, S. S.; and Mordatch, I. 2022. Blocks assemble! learning to assemble with large-scale structured reinforcement learning. In *International Conference on Machine Learning*, 7435–7469. PMLR.

Gibson, J. J. 1977. The theory of affordances. *Hilldale, USA*, 1(2): 67–82.

Heo, M.; Lee, Y.; Lee, D.; and Lim, J. J. 2023. FurnitureBench: Reproducible Real-World Benchmark for Long-Horizon Complex Manipulation. In *Robotics: Science and Systems*.

Huang, J.; Zhan, G.; Fan, Q.; Mo, K.; Shao, L.; Chen, B.; Guibas, L.; and Dong, H. 2020. Generative 3D Part Assembly via Dynamic Graph Learning. arXiv:2006.07793.

Jones, B.; Hildreth, D.; Chen, D.; Baran, I.; Kim, V. G.; and Schulz, A. 2021. Automate: A dataset and learning approach for automatic mating of cad assemblies. *ACM Transactions on Graphics (TOG)*, 40(6): 1–18.

Kimble, K.; Van Wyk, K.; Falco, J.; Messina, E.; Sun, Y.; Shibata, M.; Uemura, W.; and Yokokohji, Y. 2020. Benchmarking protocols for evaluating small parts robotic assembly systems. *IEEE robotics and automation letters*, 5(2): 883–889.

Kokic, M.; Kragic, D.; and Bohg, J. 2020. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters*, 5(2): 3352–3359.

Le, D. T.; Cortés, J.; and Siméon, T. 2009. A path planning approach to (dis) assembly sequencing. In *2009 IEEE International Conference on Automation Science and Engineering*, 286–291. IEEE.

Lee, Y.; Hu, E. S.; and Lim, J. J. 2021. IKEA furniture assembly environment for long-horizon complex manipulation tasks. In *2021 ieee international conference on robotics and automation (icra)*, 6343–6349. IEEE.

Li, Y.; Mo, K.; Duan, Y.; Wang, H.; Zhang, J.; and Shao, L. 2024. Category-level multi-part multi-joint 3d shape assembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3281–3291.

Li, Y.; Mo, K.; Shao, L.; Sung, M.; and Guibas, L. 2020. Learning 3d part assembly from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 664–682. Springer.

Mo, K.; Guibas, L. J.; Mukadam, M.; Gupta, A.; and Tulsiani, S. 2021. Where2act: From pixels to actions for articulated 3d objects. In *CVPR*.

Nagarajan, T.; and Grauman, K. 2020. Learning Affordance Landscapes for Interaction Exploration in 3D Environments. In *NeurIPS*.

Nagarajan, T.; Li, Y.; Feichtenhofer, C.; and Grauman, K. 2020. EGO-TOPO: Environment Affordances from Egocentric Video. In *CVPR*.

Niekum, S.; Chitta, S.; Barto, A. G.; Marthi, B.; and Osentoski, S. 2013. Incremental Semantically Grounded Learning from Demonstration. In *Robotics: Science and Systems*, volume 9, 10–15607. Berlin, Germany.

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

Quigley, M.; Conley, K.; Gerkey, B.; Faust, J.; Foote, T.; Leibs, J.; Wheeler, R.; and Ng, A. Y. 2009. ROS: an open-source Robot Operating System. In *ICRA Workshop on Open Source Software*, 5. Kobe, Japan.

Shen, Y.; Wu, R.; Ke, Y.; Song, X.; Li, Z.; Li, X.; Fan, H.; Lu, H.; and dong, H. 2025. BiAssemble: Learning Collaborative Affordance for Bimanual Geometric Assembly. arXiv:2506.06221.

Suárez-Ruiz, F.; Zhou, X.; and Pham, Q.-C. 2018. Can robots assemble an IKEA chair? *Science Robotics*, 3(17): eaat6385.

Sundaram, S.; Remmler, I.; and Amato, N. M. 2001. Disassembly sequencing using a motion planning approach. In *Proceedings 2001 ICRA. IEEE International Conference on*

Robotics and Automation (Cat. No. 01CH37164), volume 2, 1475–1480. IEEE.

Tian, Y.; Jacob, J.; Huang, Y.; Zhao, J.; Gu, E.; Ma, P.; Zhang, A.; Javid, F.; Romero, B.; Chitta, S.; Sueda, S.; Li, H.; and Matusik, W. 2025. Fabrica: Dual-Arm Assembly of General Multi-Part Objects via Integrated Planning and Learning. arXiv:2506.05168.

Tian, Y.; Xu, J.; Li, Y.; Luo, J.; Sueda, S.; Li, H.; Willis, K. D.; and Matusik, W. 2022. Assemble them all: Physics-based planning for generalizable assembly by disassembly. *ACM Transactions on Graphics (TOG)*, 41(6): 1–11.

Tie, C.; Chen, Y.; Wu, R.; Dong, B.; Li, Z.; Gao, C.; and Dong, H. 2025a. ET-SEED: Efficient Trajectory-Level SE(3) Equivariant Diffusion Policy. arXiv:2411.03990.

Tie, C.; Sun, S.; Zhu, J.; Liu, Y.; Guo, J.; Hu, Y.; Chen, H.; Chen, J.; Wu, R.; and Shao, L. 2025b. Manual2Skill: Learning to Read Manuals and Acquire Robotic Skills for Furniture Assembly Using Vision-Language Models. arXiv:2502.10090.

Wang, Y.; Wu, R.; Chen, Y.; Wang, J.; Liang, J.; Zhu, Z.; Geng, H.; Malik, J.; Abbeel, P.; and Dong, H. 2025. DexGarmentLab: Dexterous Garment Manipulation Environment with Generalizable Policy. *arXiv preprint arXiv:2505.11032*.

Wang, Y.; Wu, R.; Mo, K.; Ke, J.; Fan, Q.; Guibas, L.; and Dong, H. 2022. AdaAfford: Learning to Adapt Manipulation Affordance for 3D Articulated Objects via Few-shot Interactions. *European conference on computer vision (ECCV 2022)*.

Wu, R.; Lu, H.; Wang, Y.; Wang, Y.; and Dong, H. 2024. UniGarmentManip: A Unified Framework for Category-Level Garment Manipulation via Dense Visual Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, R.; Ning, C.; and Dong, H. 2023. Learning Foresightful Dense Visual Affordance for Deformable Object Manipulation. In *IEEE International Conference on Computer Vision (ICCV)*.

Wu, R.; Zhu, Z.; Wang, Y.; Chen, Y.; Wang, J.; and Dong, H. 2025. GarmentPile: Point-Level Visual Affordance Guided Retrieval and Adaptation for Cluttered Garments Manipulation. arXiv:2503.09243.

Xu, J.; Kim, S.; Chen, T.; Garcia, A. R.; Agrawal, P.; Matusik, W.; and Sueda, S. 2023. Efficient tactile simulation with differentiability for robotic manipulation. In *Conference on Robot Learning*, 1488–1498. PMLR.

Yu, M.; Shao, L.; Chen, Z.; Wu, T.; Fan, Q.; Mo, K.; and Dong, H. 2021. Roboassembly: Learning generalizable furniture assembly policy in a novel multi-robot contact-rich simulation environment. *arXiv preprint arXiv:2112.10143*.

Yuan, C.; Wen, C.; Zhang, T.; and Gao, Y. 2024. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*.

Ze, Y.; Zhang, G.; Zhang, K.; Hu, C.; Wang, M.; and Xu, H. 2024. 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations. arXiv:2403.03954.

Zeng, A.; Song, S.; Yu, K.-T.; Donlon, E.; Hogan, F. R.; Bauza, M.; Ma, D.; Taylor, O.; Liu, M.; Romo, E.; et al. 2018. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *ICRA*.

Zhang, K.; Sharma, M.; Liang, J.; and Kroemer, O. 2020a. A modular robotic arm control stack for research: Franka-interface and frankapy. *arXiv preprint arXiv:2011.02398*.

Zhang, X.; Belfer, R.; Kry, P. G.; and Vouga, E. 2020b. C-space tunnel discovery for puzzle path planning. *ACM Transactions on Graphics (TOG)*, 39(4): 104–1.

Zhang, X.; Tomizuka, M.; and Li, H. 2024. Bridging the sim-to-real gap with dynamic compliance tuning for industrial insertion. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 4356–4363. IEEE.