MAAP: Multi-Agent Active Perception for Collaborative Manipulation

Anonymous MEIS @ CVPR 2026 submission

Paper ID 2

Abstract

Multi-agent manipulation naturally produces multiple task-driven viewpoints, as each robotic arm carries a wrist-mounted camera and moves through the scene while acting. However, these distributed observations are typically under-utilized, and active perception in manipulation is still often treated as requiring a dedicated sensing agent. We introduce MAAP (Multi-Agent Active Perception), a framework that treats every arm as a dual-purpose agent: simultaneously a manipulator and a perception source. MAAP combines a VLM-based orchestrator for selecting feasible workspace-role configurations with low-level imitation controllers that aggregate multi-wrist observations for perception-aware execution. We evaluate on four collaborative manipulation tasks spanning no-occlusion to severe multi-phase occlusion. Results show that distributed wrist-camera perception consistently outperforms fixed global sensing and single-view active perception: the best MAAP variant achieves 73.0% average success vs. 62.5% (single active view) and 56.5% (global), with the largest gain on occlusion-heavy tasks (69% vs. 14%). We further show that frozen DINOv2 features substantially stabilize multi-view fusion, rescuing otherwise brittle channel fusion (24.0% → 70.5%).

1. Introduction

Collaborative multi-agent robotic manipulation holds great promise for complex real-world tasks that exceed the capabilities of a single arm [9, 23, 26, 34]. Tasks such as opening an appliance door while retrieving an object from inside, or coordinating multiple arms to assemble structures, inherently require multi-agent cooperation. As these tasks grow in complexity, so does the challenge of providing adequate visual perception of the workspace.

A fundamental barrier in many manipulation scenarios is *visual occlusion*: target objects may be hidden inside containers, obscured behind doors, or positioned beneath obstacles that no fixed camera can observe. Traditional approaches rely on a static overhead camera for a global view, but such cameras fundamentally cannot see inside enclosed spaces or around obstacles. This limitation becomes particularly acute in multi-agent settings, where the interaction of multiple arms with the environment creates dynamic viewpoint requirements.

Recent work has recognized the importance of *active perception*—the deliberate positioning of cameras to obtain task-relevant visual information [2, 3]. Classical approaches focused on next-best-view planning for 3D reconstruction [4, 21], while more recent work addresses active viewpoint selection and active vision strategies for manipulation [10, 17, 20, 28, 33, 36]. Active perception can also be induced from demonstration interfaces and teleoperation pipelines [30, 38]. Vision in Action (ViA) [35] is the most closely related work, introducing a dedicated 6-DoF robotic arm that serves exclusively as a camera platform for bimanual manipulation. However, this approach dedicates an entire robotic arm solely to camera positioning, eliminating its potential contribution to manipulation. Moreover, ViA reports that wrist cameras on execution arms are insufficient, as their motion is primarily dictated by manipulation needs.

We challenge this conclusion. In multi-agent settings where multiple arms are present, each equipped with a wrist-mounted camera, *every arm is already a potential perception agent*. Rather than designating one arm as a dedicated camera platform, we leverage wrist cameras on all manipulation arms to provide diverse, task-relevant viewpoints while these same arms simultaneously perform manipulation actions.

Crucially, since the role of "perception agent" is not fixed, the robots enjoy greater operational freedom, avoiding awkward kinematic configurations. A VLM-based task planner [1, 15, 19] distributes dual-purpose responsibilities—assigning both manipulation and perception goals—ensuring that agents are always positioned where they are most effective.

This insight motivates **Multi-Agent Active Perception (MAAP)**, which consists of: (1) a high-level VLM-based planner that decomposes tasks into role-specific sub-goals with complementary manipulation and perception responsibilities; and (2) imitation learning policies [11, 40] conditioned on proprioceptive states and wrist-camera observations that predict action sequences simultaneously accom-

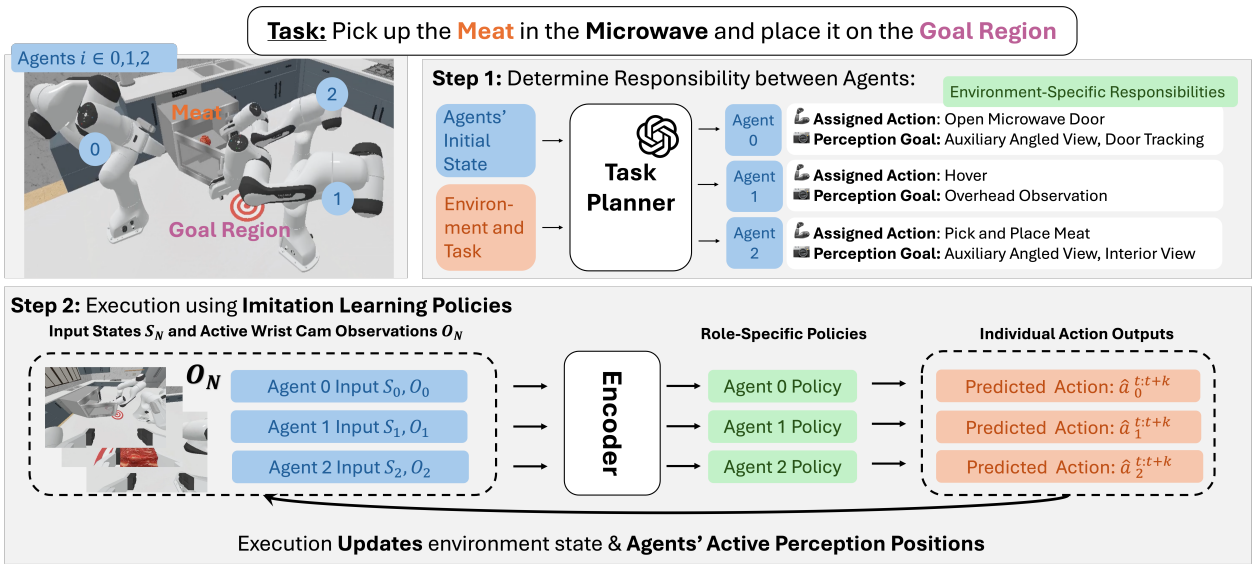

Figure 1. **System architecture of the MAAP framework. Step 1:** Given initial states and task specifications, a VLM-based task planner allocates dual-purpose responsibilities to each agent, assigning both manipulation action goals and task-relevant perception goals. **Step 2:** During execution, each agent uses a shared visual encoder and a role-specific policy to produce actions that both manipulate the environment and actively generate informative viewpoints.

plishing manipulation objectives and optimizing viewpoints. Fig. 1 illustrates the framework.

Our main contributions are: **(i)** We propose MAAP, where every manipulation arm serves as both an action executor and a visual sensor, eliminating dedicated perception hardware. **(ii)** We introduce a hierarchical framework combining VLM-based dual-purpose role assignment with perception-aware imitation learning policies. **(iii)** We demonstrate that distributed multi-view wrist-camera perception consistently outperforms both fixed global-view and single-view active perception (73.0% vs. 56.5% and 62.5% average success). **(iv)** We show that leveraging frozen DINOv2 [25] features stabilizes multi-camera fusion, improving channel fusion by 46.5 percentage points.

2. Related Work

Active Perception in Robotics. Active perception has a rich history [2, 3], from next-best-view planning [4, 21] to active viewpoint selection for manipulation [10, 17, 20, 28, 33, 36]. Recent work shows active perception can be induced from teleoperation [30, 38]. ViA [35] is most closely related, introducing a dedicated camera arm for bimanual manipulation. In contrast, MAAP distributes perception across *all* manipulation arms and supports dynamic role switching rather than concentrating perception in a single dedicated device.

Multi-Agent Manipulation. Multi-agent manipulation has gained attention with benchmarks such as RoboFactory [26] (built on ManiSkill3 [31]), featuring tasks requiring two to four agents. Related benchmarks further highlight coordinated multi-view sensing [9, 23, 34]. We adopt imitation learning using ACT [40] and DP3 [37] following the diffusion-policy formulation [11], and introduce multi-wrist observation fusion as a key design choice.

Vision Foundation Models for Robotics. Self-supervised vision models such as DINO/DINOv2 [7, 25] (built on ViTs [12]) and robotics-oriented encoders such as R3M and MVP [24, 27] provide transferable features for downstream control [8, 18, 22, 39]. On the planning side, VLM-based task planners coordinate actions in long-horizon settings [1, 19], while vision-language-action models directly map observations to actions [5, 14, 16, 29]. For multi-agent cooperation, VIKI-R fine-tunes VLMs with chain-of-thought supervision [15]. Our work uses frozen DINOv2 features for stable multi-camera fusion and a VLM planner as an orchestrator for dual-purpose role dispatch.

3. Method

3.1. Problem Setting

We consider collaborative multi-agent manipulation tasks where N robotic arms share a workspace. Each arm $i \in \{0, \ldots, N-1\}$ is a 7-DoF Franka Panda with a parallel-jaw gripper and wrist camera $\mathcal{C}_i$. At each timestep t, agent i observes proprioceptive state $\mathbf{s}_i^t \in \mathbb{R}^8$ (7 joint positions + 1 gripper width) and wrist image $\mathbf{I}_i^t \in \mathbb{R}^{3\times H\times W}$. The policy outputs joint actions $\mathbf{a}^t \in \mathbb{R}^{8N}$. The central question: *how should the policy aggregate visual information from N wrist*

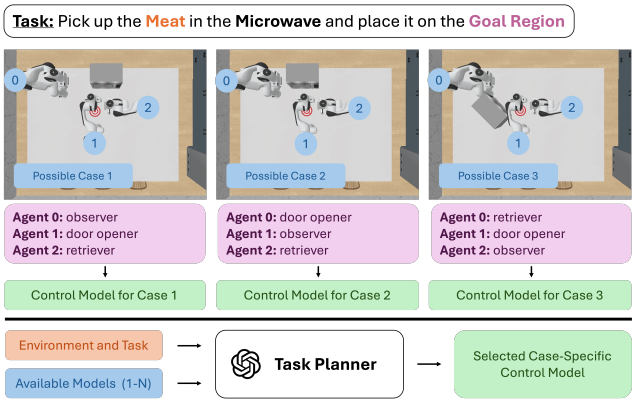

Figure 2. **Case-level orchestration with a policy bank (microwave task).** Feasible workspace cases induce different role allocations (observer/door opener/retriever); a VLM-based Task Planner selects the active case and dispatches the corresponding controller.

cameras to best support manipulation under occlusion?

3.2. Hierarchical Orchestration with a Policy Bank

MAAP adopts a two-level architecture separating *case selection* from low-level control (Fig. 2). At the high level, a VLM-based *Task Planner* selects which *workspace case* is feasible—i.e., which subset of arms can physically reach and interact with the target under kinematic constraints. At the low level, a *policy bank* contains a dedicated imitation controller trained for each case; once selected, the corresponding controller is invoked.

3.3. Dual-Purpose Active Perception

MAAP is built on *dual-purpose manipulation*: each arm is simultaneously a manipulator and a perception agent via its wrist camera. Role switching occurs at two levels:

Case-level switching. Scene geometry and reachability induce a small set of feasible workspace cases, each corresponding to different role allocations. The VLM orchestrator selects the active case and invokes the corresponding controller.

Temporal switching. Within a chosen case, the emphasis of manipulation vs. perception shifts over time. During demonstration collection, at each step one arm executes the primary manipulation while the other arm(s) reposition for informative wrist viewpoints; roles may swap as the task progresses. The imitation controller learns to reproduce this dual-purpose behavior. Fig. 3 visualizes this temporal role switching across three phases of the microwave task.

3.4. Perception Regimes and Fusion Strategies

We compare three perception regimes: **(1) Global View (GV):** a single fixed overhead camera; **(2) Single Active View (SAV):** one designated wrist camera; **(3) MAAP:** distributed multi-wrist perception with fusion.

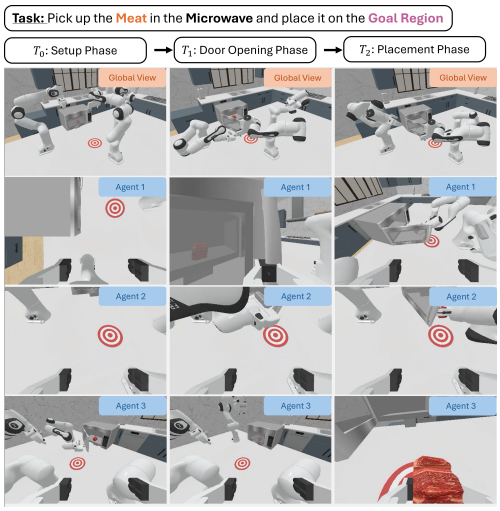

Figure 3. Temporal role switching on the microwave task: setup, door opening, and placement. At each stage, some agents perform primary manipulation while others reposition to provide informative wrist-camera viewpoints.

We use ACT [40] as the primary policy (CVAE + transformer [32], ResNet-18 [13] encoder, chunk size $k = 50$) and DP3 [37] as a complementary 3D policy family. Under MAAP, we evaluate four fusion strategies for ACT:

Multi-View (MV). Each camera encoded independently as separate tokens: $O_{\text{MV}} = [f_\theta(I_0); \ldots; f_\theta(I_{N-1})]$.

Spatial Fusion (SF). Images horizontally concatenated before encoding:

$$O_{\text{SF}} = f_\theta([I_0 \parallel \cdots \parallel I_{N-1}]_W), \ [\cdot]_W \in \mathbb{R}^{3\times H\times NW}. \quad (1)$$

Channel Fusion (CF). Images concatenated along the channel dimension:

$$O_{\text{CF}} = g_\phi([I_0 \parallel \cdots \parallel I_{N-1}]_C), \ [\cdot]_C \in \mathbb{R}^{3N\times H\times W}. \quad (2)$$

DINO Channel Fusion (DCF). Each image processed by a frozen DINOv2 [25] ViT-S/14 [12] backbone, then concatenated and projected via 1×1 convolution:

$$O_{\text{DCF}} = h_\psi([\text{DINOv2}(I_0) \parallel \cdots \parallel \text{DINOv2}(I_{N-1})]_C). \quad (3)$$

3.5. Task Suite

We develop four tasks on RoboFactory [26] (ManiSkill3 [31]), each with distinct occlusion challenges: **Stack Cube** (2 agents, no occlusion): coordinate to stack cubes. **Pot** (2 agents, localized occlusion): retrieve meat from inside a pot. **Microwave** (3 agents, heavy occlusion + coordination): open door and retrieve contents from the fully occluded interior. **Cart** (2 agents, structured occlusion): place a cube onto a shelf with partial geometric occlusion.

Table 1. ACT success rates (%) across perception regimes. **Bold**: best; underline: second best.

Task	GV	SAV	MAAP MV	MAAP SF	MAAP CF	MAAP DCF
Stack Cube	8	37	34	**51**	11	34
Pot	86	**99**	**99**	**99**	85	80
Microwave	32	14	47	42	0	**69**
Cart	**100**	**100**	**100**	**100**	0	99
Average	56.5	62.5	70.0	**73.0**	24.0	70.5

Table 2. DP3 success rates (%) under multi- vs. single-wrist inputs.

Task	DP3 (Multi)	DP3 (Single)
Stack Cube	5	0
Pot	85	80
Microwave	45	45
Cart	100	100
Average	58.8	56.3

4. Experiments

4.1. Setup

We generate expert trajectories using a motion-planning solver. We collect 50 RGB demonstrations for ACT and 100 point-cloud trajectories for DP3. ACT is trained with LeRobot [6] (batch 8, chunk $k = 50$, lr 10^{-5}) on a single RTX 4090. For CF, we expand the first convolution to $3N$ channels. For DCF, we use a frozen DINOv2 ViT-S/14 with a learnable 1×1 projection. We train for 50K steps on two-agent tasks and 200K on Microwave. Each policy is evaluated over 100 episodes with held-out seeds.

4.2. Main Results

Tables 1–2 report success rates across perception regimes.

4.3. Analysis

Overall trends. The MAAP perception regime delivers the strongest overall performance: across fusion methods, policies with coordinated wrist-camera observations consistently outperform the static-camera baseline without requiring a dedicated perception arm. SF performs best on average (73.0%), followed closely by DCF (70.5%) and MV (70.0%).

Task-level interpretation. The benefit of MAAP is most pronounced on *Microwave* (severe occlusion): DCF reaches 69% vs. 14% SAV and 32% GV—while one arm executes manipulation, others reposition to supply informative wrist viewpoints of critical regions. On near-saturated tasks (*Pot*, *Cart*), MAAP remains competitive (99–100%). Interestingly, MAAP also improves on *Stack Cube* (no explicit occlusion): SF achieves 51% vs. 37% SAV and 8% GV, suggesting wrist cameras capture complementary alignment cues beyond occlusion removal.

Fusion strategy analysis. No single fusion dominates: SF excels on Stack Cube (geometric correspondence), DCF on Microwave (semantic stability). The CF-vs-DCF contrast is striking: CF collapses to 24.0% average (0% on Microwave and Cart), while DCF achieves 70.5%—a **46.5-point** improvement. Raw pixel concatenation disrupts pretrained convolutional structure; frozen DINOv2 features provide camera-agnostic representations that stabilize multi-view learning under limited demonstrations.

DP3. DP3 shows smaller sensitivity to view constraints (58.8% multi vs. 56.3% single), consistent with point-cloud conditioning being more view-robust than RGB-only policies.

5. Conclusion

We presented Multi-Agent Active Perception (MAAP), a framework for collaborative manipulation that treats each arm as a dual-purpose agent—simultaneously acting on the environment and providing task-relevant wrist-camera observations. Across four multi-agent tasks (24 control-layer settings, 100 episodes each), distributed multi-wrist perception consistently improves success over non-active global perception and single-view active perception. On average, MAAP with Spatial Fusion achieves 73.0% success versus 62.5% for SAV and 56.5% for the global baseline, with the most pronounced gains on occlusion-heavy tasks (Microwave: 69% vs. 14% SAV and 32% GV), while remaining competitive on near-saturated tasks. MAAP also improves precision without explicit occlusion (Stack Cube: 51% vs. 37% SAV), suggesting benefits beyond occlusion removal.

A particularly strong signal is the stability gap between CF and DCF: CF collapses to 24.0% while DCF reaches 70.5%, a **46.5-point** improvement, demonstrating that camera-agnostic pretrained features substantially stabilize multi-view learning. These results suggest that collaborative manipulation itself can serve as an effective active perception mechanism.

Limitations. Our evaluation is simulation-only with limited demonstrations per task. Active perception behaviors arise from expert demonstrations rather than end-to-end optimization, and experiments focus on the control layer without quantifying gains from the full hierarchical stack.

Future Work. Promising directions include adaptive fusion, joint manipulation-perception optimization, VLA integration [5, 14, 16, 29], and sim-to-real transfer.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1, 2

[2] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. 1, 2

[3] Ruzena Bajcsy, Yiannis Aloimonos, and John K. Tsotsos. Revisiting active perception. *Autonomous Robots*, 42(2):177–196, 2018. 1, 2

[4] Michel Breyer, Lionel Ott, Jen Jen Chung, Roland Siegwart, and Juan Nieto. Closed-loop next-best-view planning for target-driven grasping. *arXiv preprint arXiv:2207.10543*, 2022. 1, 2

[5] Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 2, 4

[6] Remi Cadene, Simon Alibert, Alexander Soare, et al. LeRobot: State-of-the-art machine learning for real-world robotics. https://github.com/huggingface/lerobot, 2024. 4

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[8] Tianxing Chen, Yao Mu, Zhixuan Liang, Zanxin Chen, Shijia Peng, Qiangyu Chen, Mingkun Xu, Ruizhen Hu, Hongyuan Zhang, Xuelong Li, and Ping Luo. G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation. *arXiv preprint arXiv:2411.18369*, 2024. 2

[9] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 1, 2

[10] Ricson Cheng, Arpit Agarwal, and Katerina Fragkiadaki. Reinforcement learning of active vision for manipulating objects under occlusions. In *Proceedings of The 2nd Conference on Robot Learning (CoRL)*, 2018. 1, 2

[11] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025. 1, 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[14] Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilve Wang, Longzan Luo, Xiaoqi Li, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, and Shanghang Zhang. Lift3d policy: Lifting 2d foundation models for robust 3d robotic manipulation, 2025. 2, 4

[15] Li Kang, Xiufeng Song, Heng Zhou, Yiran Qin, Jie Yang, Xiaohong Liu, Philip Torr, Lei Bai, and Zhenfei Yin. Viki-r: Coordinating embodied multi-agent cooperation via reinforcement learning. *arXiv preprint arXiv:2506.09049*, 2025. 1, 2

[16] M. J. Kim et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2, 4

[17] Quoc V. Le, Ashutosh Saxena, and Andrew Y. Ng. Active perception: Interactive manipulation for improving object detection. Technical report, Stanford University, 2010. Technical Report. 1, 2

[18] Y. Li et al. Theia: Distilling diverse vision foundation models for robot learning. *arXiv preprint arXiv:2407.20179*, 2024. 2

[19] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022. 1, 2

[20] Y. Liu, S. Mu, X. Chao, Z. Li, Y. Mu, T. Chen, S. Li, C. Lyu, X. Zhang, and W. Ding. Avr: Active vision-driven robotic precision manipulation with viewpoint and focal length optimization. *arXiv preprint arXiv:2503.01439*, 2025. 1, 2

[21] Haoran Ma et al. Active perception for grasp detection via neural graspness field. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2

[22] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? *arXiv preprint arXiv:2303.18240*, 2023. Introduces CortexBench and VC-1. 2

[23] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins. In *CVPR*, 2025. Also available as arXiv:2409.02920. 1, 2

[24] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2023. 2

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel

Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2, 3

[26] Yiran Qin, Li Kang, Xiufeng Song, Zhenfei Yin, Xiaohong Liu, Xihui Liu, Ruimao Zhang, and Lei Bai. Robofactory: Exploring embodied agent collaboration with compositional constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10075–10085, 2025. 1, 2, 3

[27] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning (CoRL)*, 2022. 2

[28] Jinghuan Shang and Michael S. Ryoo. Active vision reinforcement learning under limited visual observability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

[29] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning (CoRL)*, 2022. Also available as arXiv:2109.12098. 2, 4

[30] S. A. Sontakke et al. Roboclip: One demonstration is enough to learn robot policies. *arXiv preprint arXiv:2310.07899*, 2023. 1, 2

[31] Stone Tao, Fanbo Xiang, Arth Shukla, et al. ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI. *arXiv preprint arXiv:2410.00425*, 2024. 2, 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 3

[33] Guokang Wang, Hang Li, Shuyuan Zhang, Di Guo, Yanhong Liu, and Huaping Liu. Observe then act: Asynchronous active vision-action model for robotic manipulation. *arXiv preprint arXiv:2409.14891*, 2024. 1, 2

[34] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifan Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A SimulAted part-based interactive ENvironment. In *CVPR*, pages 11097–11107, 2020. 1, 2

[35] Haoyu Xiong, Xiaomeng Xu, Jimmy Wu, Yifan Hou, Jeannette Bohg, and Shuran Song. Vision in action: Learning active perception from human demonstrations. In *Proceedings of The 9th Conference on Robot Learning*, pages 5450–5463. PMLR, 2025. 1, 2

[36] Y. Zaky, G. Paruthi, B. Tripp, and J. Bergstra. Active perception and representation for robotic manipulation. *arXiv preprint arXiv:2003.06734*, 2020. 1, 2

[37] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 2, 3

[38] Q. Zeng, C. Li, J. S. John, Z. Zhou, J. Wen, G. Feng, Y. Zhu, and Y. Xu. Activeumi: Robotic manipulation with active perception from robot-free human demonstrations. *arXiv preprint arXiv:2510.01607*, 2025. 1, 2

[39] Yi Zhang et al. Dinobot: Robot manipulation via retrieval and alignment using dino features. *arXiv preprint arXiv:2402.13181*, 2024. 2

[40] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023. 1, 2, 3