MAAP: Multi-Agent Active Perception for Collaborative Manipulation

Bruno N.Y. Chen[1], Heng Zhou[2], Li Kang[3], Xiufeng Song[3], Jiahua Ma[4], Zhemeng Zhang[3], Yiran Qin[5,*]

[1]Carnegie Mellon University [2]University of Science and Technology of China
[3]Shanghai Jiao Tong University [4]Sun Yat-sen University [5]The Chinese University of Hong Kong, Shenzhen

nybchen@cmu.edu *Corresponding: yiranqin@link.cuhk.edu.cn

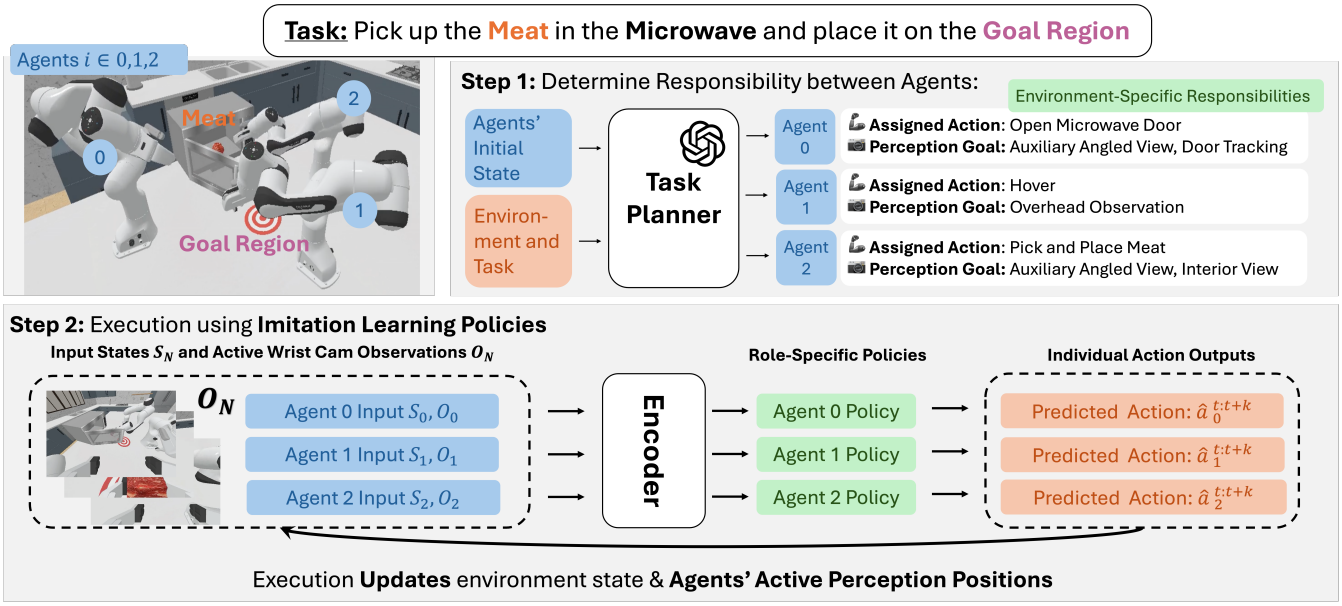

Fig. 1: **MAAP framework. Step 1:** A VLM-based task planner allocates dual-purpose responsibilities (manipulation + perception goals) to each agent. **Step 2:** Each agent uses a shared visual encoder and role-specific policy to produce actions that both manipulate and generate informative viewpoints.

***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 underutilized. 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 imitation controllers that aggregate multi-wrist observations. Across four collaborative tasks spanning no-occlusion to severe multi-phase occlusion, 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%). Frozen DINOv2 features substantially stabilize multi-view fusion, rescuing brittle channel fusion (24.0%→70.5%).**

I. Introduction

Collaborative multi-agent manipulation holds great promise for complex tasks that exceed single-arm capabilities [1]–[4]. A fundamental barrier is *visual occlusion*: targets may be hidden inside containers or obscured behind obstacles that no fixed camera can observe. Recent work on *active perception* [5], [6] addresses this by deliberately positioning cameras [7]–[14]. Recent work also shows that active perception can be induced from teleoperation [15], [16]. Vision in Action (ViA) [17] introduces a dedicated 6-DoF arm exclusively as a camera platform, but dedicates an entire arm solely to camera positioning. Moreover, ViA concludes that wrist cameras on execution arms are insufficient, as their motion is driven by manipulation needs.

We challenge this conclusion. In multi-agent settings with multiple arms, each equipped with a wrist camera, *every arm is already a potential perception agent*. Rather than designating one arm as a camera platform, we leverage wrist cameras on all manipulation arms for diverse viewpoints while these same arms simultaneously manipulate. Since perception roles are not fixed, robots enjoy greater operational freedom—avoiding awkward kinematic configurations. A VLM-based planner distributes dual-purpose responsibilities, ensuring agents are always positioned where most effective.

This motivates **Multi-Agent Active Perception (MAAP)**: (1) a VLM-based planner [18]–[20] assigning complementary manipulation and perception responsibilities; and (2) imitation learning policies [21], [22] that predict actions simultaneously accomplishing manipulation and optimizing viewpoints (Fig. 1). Our contributions: **(i)** MAAP treats every arm as both action executor and visual sensor, eliminating dedicated perception hardware. **(ii)** A hierarchical framework

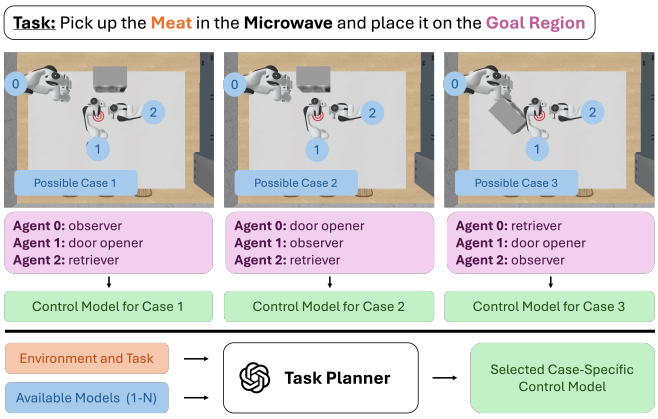

Fig. 2: Case-level orchestration (microwave task). Feasible workspace cases induce different role allocations; a VLM-based planner selects the active case and dispatches the corresponding controller.

combines VLM-based role assignment with perception-aware imitation learning. **(iii)** Distributed multi-view perception outperforms global and single-view active perception (73.0% vs. 56.5%/62.5%). **(iv)** Frozen DINOv2 features stabilize multi-camera fusion (+46.5 points).

II. Method

Problem setting. N 7-DoF Franka Pandas share a workspace, each with a wrist camera $\mathscr{C}_i$. At each timestep, agent i observes proprioceptive state $\mathbf{s}_i^t \in \mathbb{R}^8$ and wrist image $\mathbf{I}_i^t \in \mathbb{R}^{3\times H\times W}$. The policy outputs joint actions $\mathbf{a}^t \in \mathbb{R}^{8N}$. Core question: *how should the policy aggregate N wrist-camera observations under occlusion?*

Hierarchical orchestration. A VLM-based Task Planner selects the feasible *workspace case*—which arm subset can reach the target—then dispatches a case-specific imitation controller from a pre-trained policy bank. We pre-enumerate feasible cases, collect demonstrations under each, and train one policy per case (Fig. 2).

Dual-purpose active perception. Each arm is simultaneously manipulator and perception agent. Role switching occurs at two levels: (1) *case-level*—the VLM selects workspace configurations with different role allocations; (2) *temporal*—within a case, manipulation vs. perception emphasis shifts over time. During data collection, one arm executes primary manipulation while others reposition for informative viewpoints; roles swap as the task progresses.

Perception regimes and fusion. We compare: **Global View (GV)**—fixed overhead camera; **Single Active View (SAV)**—one wrist camera; **MAAP**—distributed multi-wrist fusion. We use ACT [21] (CVAE + transformer [23], ResNet-18 [24], chunk $k = 50$) as primary policy and DP3 [25] (diffusion on point clouds [22]) as complementary 3D family. Under MAAP with ACT, we evaluate four fusion strategies: *Multi-View (MV)*: each camera encoded independently as separate tokens. *Spatial Fusion (SF)*: images horizontally concatenated before encoding. *Channel Fusion (CF)*: images concatenated along channels; first convolution expanded to $3N$ channels. *DINO Channel Fusion (DCF)*: frozen DINOv2 [26] ViT-S/14 [27] features concatenated and projected via 1×1 convolution. Self-supervised vision models [26], [28] and robotics-oriented encoders [29]–[31] provide transferable features for downstream control [32]–[34].

Tasks. Four tasks on RoboFactory [1]/ManiSkill3 [35]: *Stack Cube* (2 agents, no occlusion), *Pot* (2 agents, localized occlusion), *Microwave* (3 agents, severe occlusion + coordination), *Cart* (2 agents, structured occlusion).

III. Experiments

Setup. We collect 50 RGB demonstrations for ACT and 100 point-cloud trajectories for DP3 using a motion-planning solver. ACT is trained with LeRobot [36] (batch 8, lr 10^{-5}) on a single RTX 4090 for 50K steps (2-agent tasks) or 200K steps (Microwave). Each policy is evaluated over 100 episodes with held-out seeds.

Main results. Tables I–II report success rates across perception regimes.

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

Task	GV	SAV	MAAP			
			MV	SF	CF	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 II: 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

Analysis. MAAP delivers the strongest overall performance: SF achieves 73.0% average, followed by DCF (70.5%) and MV (70.0%). The gain is most pronounced on *Microwave* (severe occlusion): DCF 69% vs. SAV 14% and GV 32%. On near-saturated tasks (*Pot*, *Cart*), MAAP remains competitive (99–100%). Interestingly, MAAP also helps on *Stack Cube* (no occlusion): SF 51% vs. SAV 37%, suggesting complementary alignment cues from multiple wrist views. No single fusion dominates: SF excels on Stack Cube (geometric correspondence), DCF on Microwave (semantic stability). CF collapses to 24.0% avg while DCF achieves 70.5% (**+46.5 points**), showing frozen DINOv2 features stabilize multi-view learning. DP3 shows smaller view sensitivity (58.8% vs. 56.3%), consistent with 3D conditioning being more view-robust.

Conclusion. Collaborative manipulation itself is an effective active perception mechanism. Frozen DINOv2 features are critical for stable multi-view fusion. Future work includes adaptive fusion, integration with vision-language-action models [37]–[40], and real-hardware transfer.

REFERENCES

[1] Y. Qin, L. Kang, X. Song, Z. Yin, X. Liu, X. Liu, R. Zhang, and L. Bai, "Robofactory: Exploring embodied agent collaboration with compositional constraints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10075–10085, October 2025.

[2] 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," in *CVPR*, 2025. Also available as arXiv:2409.02920.

[3] T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Z. Li, Q. Liang, X. Lin, Y. Ge, Z. 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.

[4] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su, "SAPIEN: A SimulAted part-based interactive ENvironment," in *CVPR*, pp. 11097–11107, 2020.

[5] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[6] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, no. 2, pp. 177–196, 2018.

[7] M. Breyer, L. Ott, J. J. Chung, R. Siegwart, and J. Nieto, "Closed-loop next-best-view planning for target-driven grasping," *arXiv preprint arXiv:2207.10543*, 2022.

[8] H. Ma *et al.*, "Active perception for grasp detection via neural graspness field," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

[10] R. Cheng, A. Agarwal, and K. Fragkiadaki, "Reinforcement learning of active vision for manipulating objects under occlusions," in *Proceedings of The 2nd Conference on Robot Learning (CoRL)*, 2018.

[11] J. Shang and M. S. Ryoo, "Active vision reinforcement learning under limited visual observability," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[12] 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.

[13] G. Wang, H. Li, S. Zhang, D. Guo, Y. Liu, and H. Liu, "Observe then act: Asynchronous active vision-action model for robotic manipulation," *arXiv preprint arXiv:2409.14891*, 2024.

[14] Q. V. Le, A. Saxena, and A. Y. Ng, "Active perception: Interactive manipulation for improving object detection," tech. rep., Stanford University, 2010. Technical Report.

[15] 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.

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

[17] H. Xiong, X. Xu, J. Wu, Y. Hou, J. Bohg, and S. Song, "Vision in action: Learning active perception from human demonstrations," in *Proceedings of The 9th Conference on Robot Learning*, vol. 305 of *Proceedings of Machine Learning Research*, pp. 5450–5463, PMLR, 2025.

[18] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. Jauregui Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[19] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," *arXiv preprint arXiv:2209.07753*, 2022.

[20] L. Kang, X. Song, H. Zhou, Y. Qin, J. Yang, X. Liu, P. Torr, L. Bai, and Z. Yin, "Viki-r: Coordinating embodied multi-agent cooperation via reinforcement learning," *arXiv preprint arXiv:2506.09049*, 2025.

[21] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[22] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

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

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.

[25] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[26] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[28] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, pp. 9650–9660, 2021.

[29] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[30] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning (CoRL)*, 2022.

[31] A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, P. Abbeel, J. Malik, D. Batra, Y. Lin, O. Maksymets, A. Rajeswaran, and F. 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.

[32] Y. Zhang *et al.*, "Dinobot: Robot manipulation via retrieval and alignment using dino features," *arXiv preprint arXiv:2402.13181*, 2024.

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

[34] T. Chen, Y. Mu, Z. Liang, Z. Chen, S. Peng, Q. Chen, M. Xu, R. Hu, H. Zhang, X. Li, and P. Luo, "G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation," *arXiv preprint arXiv:2411.18369*, 2024.

[35] S. Tao, F. Xiang, A. Shukla, *et al.*, "ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI," *arXiv preprint arXiv:2410.00425*, 2024.

[36] R. Cadene, S. Alibert, A. Soare, *et al.*, "LeRobot: State-of-the-art machine learning for real-world robotics." `https://github.com/huggingface/lerobot`, 2024.

[37] A. Brohan, N. Brown, J. Carbajal, *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

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

[39] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2022. Also available as arXiv:2109.12098.

[40] Y. Jia, J. Liu, S. Chen, C. Gu, Z. Wang, L. Luo, X. Li, P. Wang, Z. Wang, R. Zhang, and S. Zhang, "Lift3d policy: Lifting 2d foundation models for robust 3d robotic manipulation," June 2025.