A Visuo-tactile Diffusion Policy Architecture for Multimodal Imitation Learning - WBCD Bimanual Robotics Competition Proposal

Team MARS, Mechanisms And Robotic Systems Lab, Purdue University

I. Introduction

In this proposal, our team from Purdue University presents a multimodal imitation learning approach to achieving fully autonomous bimanual robotic manipulation. By leveraging visuo-tactile sensing and supervised learning policy, our solution aims to enhance manipulation precision and generalizability across diverse specific tasks and different robotic platforms. It is designed to address the core challenges of the **Table Service Operation** and the **Packing Challenge in Logistics** tasks, in terms of operational speed, system reliability, and the learning curve for policy implementation.

In both tasks, one of the core challenges is the precise alignment of targeted objects. For instance, in the table service operation task, the alignment of the box and lid directly affects the success rate of closing the lid. However, relying solely on vision-based sensing is insufficient, for robotic components inherently occlude the target objects during and after grasping [1]. In tasks involving limited vision information and delicate physical interactions, such as reorienting and placing a wine glass, tactile sensing plays a crucial role, as vision alone cannot capture the fine-grained contact dynamics essential for robust manipulation [2].

This limitation motivates our approach, which introduces innovations in both hardware and algorithms to effectively integrate tactile sensing into vision-based policy learning. Our lab specializes in both hardware development, including tactile sensor design [1], [3], [4], imitation learning [5], [6], specifically visuo-tactile imitation learning [7], and visuo-tactile manipulation [8], [9]. Our expertise in combining vision and tactile sensing for bimanual robots positions us well to address the challenges presented in this competition.

II. Key Motivators for Our Approach

Pre-trained policies, such as π_0 [10], have demonstrated great potential in unlocking the full capabilities of flexible, general, and dexterous robotic systems. These hardware-agnostic learning models can be fine-tuned for diverse tasks and adapted to various robotic platforms. However, achieving the level of generality required for zero-shot real-world applications remains a challenge due to its reliability on fine-tuning with high-quality data.

As a stable, easy-to-train, and well-established visuomotor learning model, Diffusion Policy [11] stands out as an effective approach for dexterous manipulation. Its generalizability and transferability make it particularly well-suited for the bimanual competition tasks. Moreover, its effectiveness in policy learning has been demonstrated across real-world applications involving diverse objects manipulations and complex robot-object-environment interactions [12], [13].

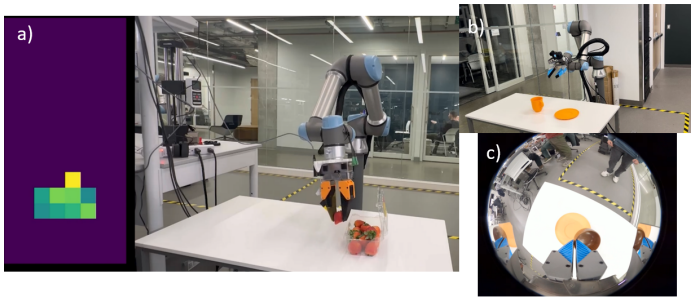

Fig. 1. (a): Previous tactile sensor implementation (b): UMI gripper on UR5e, (c): Camera view during first policy rollout

To enable portable, low-cost, and information-rich data collection for bimanual and dynamic manipulation demonstrations, the Universal Manipulation Interface (UMI) [14] facilitates direct skill transfer from in-the-wild human demonstrations to deployable robot policies. To enhance the UMI gripper with tactile sensing capabilities, we plan to integrate 3D-ViTac tactile sensors [15], as successfully demonstrated in previous experimental trials in our lab (Fig. 1a). This multimodal representation, combining visual and tactile data, can then be leveraged within diffusion policies for imitation learning, further improving precision and adaptability in manipulation tasks.

Since the competition sponsors and organizers support the interchangeability of grippers on the provided hardware (ARX X7, Galaxea R1), we are excited about integrating our custom grippers and algorithms into emerging products in the bimanual robotics industry. By enhancing precision and object identification in these tasks, we aim to develop a simple-to-train, plug-and-play solution that is both widely effective and adaptable for bimanual manipulation.

III. Technical Approach

A. Core Methodology: Multimodal Imitation Learning

Our approach employs a **visuo-tactile diffusion policy** trained through **supervised policy learning** to achieve high-precision manipulation.

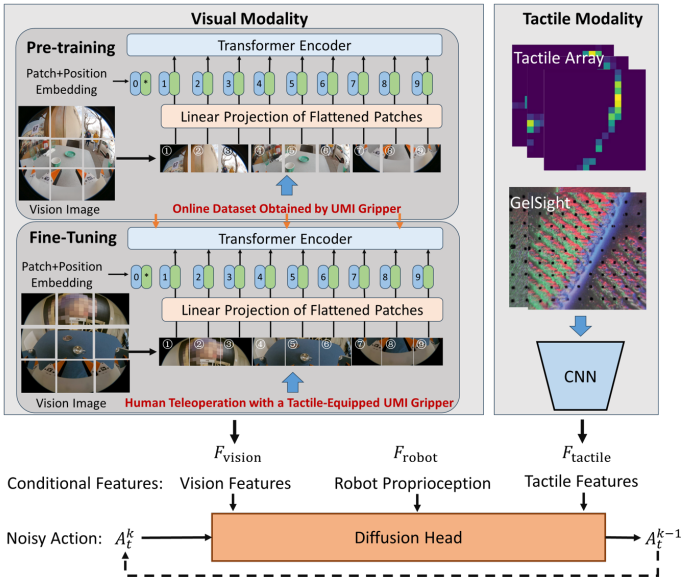

Fig. 2. Visuo-tactile Diffusion Policy Pipeline

On the hardware side, we propose a pair of integrated UMI grippers embedded with 3D-ViTac tactile sensors for both data collection and operation in bimanual tasks. Leveraging the open-source SDK and URDF from ARX X7/Galaxea R1, we anticipate that the integration of these new grippers will significantly enhance the hardware performance of the bimanual robots. However, a key challenge lies in embedding tactile sensing into data collection devices and robotic grippers while ensuring both efficient grasping and high-quality, scalable data collection.

After successfully replicating UMI grippers in our lab (Fig. 1b), we deployed the UMI gripper in real-world experiments and tested our first trained policy on a pick-and-place task as well as the cup arrangement policy [14] (Fig. 1c). Although the gripper experienced some jiggles and vibrations, we believe that its performance can be significantly improved with better training data and policy fine-tuning in the coming weeks.

On the algorithmic side, we introduce a two-stage adaptation of a pre-trained diffusion policy to enable the fine-tuning of large image-based models on tactile robot sensor modalities (Fig. 2). After data collection, we plan to use Rerun or Open X-Embodiment to visualize and analyze our dataset. The first stage involves vision-only pre-training, leveraging high-quality datasets. In the second stage, we implement tactile-aware fine-tuning, dynamically integrating and weighting features from tactile sensing. A critical aspect of this process is cross-modal synchronization, requiring hardware-triggered timestamp alignment between vision and tactile streams to ensure accurate data fusion.

To implement this approach at the policy layer, we need to explore **two key directions**. **First**, we aim to utilize tactile pre-training and representation learning to extract meaningful features from tactile input. **Second**, we focus on developing methods to integrate large-scale open-source visual data with a limited amount of tactile data to train a unified visual-tactile policy.

B. Innovation: Novel Sensor-Policy Co-Design

To not only address the challenges of standardized benchmarking tasks in the WBCD competition but also foster collaboration between researchers and industry professionals to bridge the gap between industry metrics and academic research, we propose three key novelties in our solution and their potential impacts.

The first novelty lies in the realization of high-quality tactile representation in an effective and cost-efficient manner. Converting raw tactile data into a high-quality representation for policy learning in a cost-effective way is crucial for industrial policy implementation.

The second novelty focuses on optimizing the balance between different multimodal data contributions, particularly leveraging varying amounts of visual and tactile data (e.g., 10,000 visual samples vs. 100 tactile samples) to achieve optimal learning efficiency. This directly relates to the effectiveness of our cross-modal data fusion strategy.

The third novelty addresses the challenge of fine-tuning a diffusion policy that is initially trained on visual data to seamlessly incorporate tactile sensing without disrupting prior visual training.

IV. Why Our Approach is Effective:

The primary challenge in table service operations is achieving precise box-lid alignment, which is difficult for vision-based systems due to occlusions. Vision data under occlusion struggles to accurately detect lid edges, limiting the system's ability to generate precise alignment commands. In contrast, a tactile-enhanced gripper provides a more effective solution by utilizing stress-strain feedback [16]. This tactile input enables fine-grained edge detection, allowing for better interpretation of contact dynamics and improving the success rate of lid/box alignment.

In logistics packing tasks, our multimodal system enhances performance in several ways. It enables faster adaptation to novel objects by leveraging the generalization capability of pre-trained policies, allowing the system to handle a wide range of items without extensive retraining [17]. Additionally, it improves grasp stability through contact-aware force sensing, ensuring more secure and reliable object handling [18], [19]. Furthermore, the system enhances object identification and sorting by utilizing a multimodoal algorithm, which refines classification and placement based on tactile feedback [20].

V. Conclusion

With this multimodal imitation learning solution, our framework leverages diverse data sources to enhance the performance of bimanual robots. Its ability to generalize across different platforms is essential for providing robust and transferable solution for bimanual robotics in both research and industry settings. By integrating visuo-tactile sensing, policy learning, and bimanual hardware, we aim to establish a new benchmark in WBCD's manipulation tasks.

REFERENCES

[1] S. Athar, G. Patel, Z. Xu, Q. Qiu, and Y. She, "Vistac toward a unified multimodal sensing finger for robotic manipulation," *IEEE Sensors Journal*, vol. 23, no. 20, pp. 25 440–25 450, 2023.

[2] S. Q. Liu and E. H. Adelson, "Gelsight fin ray: Incorporating tactile sensing into a soft compliant robotic gripper," in *2022 IEEE 5th International Conference on Soft Robotics (RoboSoft)*, 2022, pp. 925–931.

[3] O. Yu and Y. She, "Feelit: Combining compliant shape displays with vision-based tactile sensors for real-time teletaction," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 13 853–13 860.

[4] J. Liu, W. Li, S. Yu, S. Blanchard, and S. Lin, "Fatigue-resistant mechanoresponsive color-changing hydrogels for vision-based tactile robots," *Advanced Materials*, vol. TBD, Sep 2024.

[5] Z. Xu and Y. She, "LeTac-MPC: Learning model predictive control for tactile-reactive grasping," *IEEE Transactions on Robotics*, 2024.

[6] ——, "Leto: Learning constrained visuomotor policy with differentiable trajectory optimization," *IEEE Transactions on Automation Science and Engineering*, pp. 1–12, 2024.

[7] Z. Xu, R. Uppuluri, X. Zhang, C. Fitch, P. G. Crandall, W. Shou, D. Wang, and Y. She, "UniT: Unified tactile representation for robot learning," 2024. [Online]. Available: https://arxiv.org/abs/2408.06481

[8] Y. Zhou, P. Zhou, S. Wang, and Y. She, "In-hand singulation and scooping manipulation with a 5 dof tactile gripper," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 5238–5243.

[9] Y. Du, P. Zhou, M. Y. Wang, W. Lian, and Y. She, "Stick roller: Precise in-hand stick rolling with a sample-efficient tactile model," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 2312–2318.

[10] Physical Intelligence Company, "Openpi blog," 2024, accessed: 2024-02-07. [Online]. Available: https://www.physicalintelligence.company/blog/openpi

[11] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2024.

[12] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "UMI on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers," in *Proceedings of the 2024 Conference on Robot Learning*, 2024.

[13] M. Seo, H. A. Park, S. Yuan, Y. Zhu, and L. Sentis, "Legato: Cross-embodiment imitation using a grasping tool," *IEEE Robotics and Automation Letters*, vol. 10, no. 3, pp. 2854–2861, 2025.

[14] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[15] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3d vitac:learning fine-grained manipulation with visuo-tactile sensing."

[16] J. Zhao, Y. Ma, L. Wang, and E. H. Adelson, "Transferable tactile transformers for representation learning across diverse sensors and tasks," 2024.

[17] M. Bauza, A. Bronars, Y. Hou, I. Taylor, N. Chavan-Dafle, and A. Rodriguez, "simple: A visuotactile method learned in simulation to precisely pick, localize, regrasp, and place objects," *Science Robotics*, 2024.

[18] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.

[19] A. Rodriguez, "The unstable queen: Uncertainty, mechanics, and tactile feedback," *Science Robotics*, vol. 6, no. 54, p. eabi4667, May 2021, pMID: 34043542.

[20] A. R. M. Bauza, A. Bronars, "Tac2pose: Tactile object pose estimation from the first touch," *IJRR*, 2023.