On the Importance of Tactile Sensing for Imitation Learning: A Case Study on Robotic Match Lighting

Niklas Funk[1], Changqi Chen[1], Tim Schneider[1], Georgia Chalvatzaki[1], Roberto Calandra[2], Jan Peters[1]

***Abstract*— The field of robotic manipulation has advanced significantly in the last years. At the sensing level, several novel tactile sensors have been developed. On a methodological level, learning from demonstrations has proven an efficient paradigm to obtain performant robotic manipulation policies. The combination of both holds the promise to extract crucial contact related information from the demonstration data and actively exploit it during the policy rollouts. However, despite its potential, it remains an underexplored direction. This work therefore proposes a multimodal, visuotactile imitation learning framework capable of efficiently learning fast and dexterous manipulation policies. We evaluate our framework on the dynamic, contact-rich task of robotic match lighting - a task in which tactile feedback influences human manipulation performance. Our experimental evaluations show that adding tactile information into the policies significantly improves performance and thereby underlines the importance of tactile sensing for contact-rich manipulation tasks.**

I. INTRODUCTION

Robotic manipulation remains far from matching the dexterity and efficiency of human hands [1], [2], [3]. In fact, the current trend of exploiting human demonstration data for learning robotic manipulation [4], [5], [6] actively exploits human task understanding and their advanced manipulation capabilities. Yet, while it is well-known that human manipulation heavily benefits from touch sensing [7], the majority of current works in imitation learning for manipulation is still missing out on this modality [5], [6], [8], [9]. Given the importance of touch for human manipulation, the question therefore arises whether robotic policies could also benefit from adding tactile sensing?

This work approaches the question through studying the impact of touch sensing for learning to ignite matches. We argue that match lighting is an effective testbed for examining the role of touch sensing in learning robotic manipulation from demonstrations. This is because the task requires dynamic motion and compliance [10], which introduces additional complexity compared to standard tasks such as pick-and-place or insertion [11], [12]. Moreover, it is a task for which there is evidence that the availability of touch sensing impacts human performance [13]. In fact, despite the task's relevance, to the best of our knowledge, it was previously only investigated in [10]. Yet, [10] considered fixed match grasp poses and a precisely calibrated setup without including high dimensional observations. Herein, we address more complicated scenarios including varying grasp poses and only considering the image of an RGB camera, the end effector velocity, and eventually the information from an event-based optical tactile sensor as observation (cf. Fig. 1).

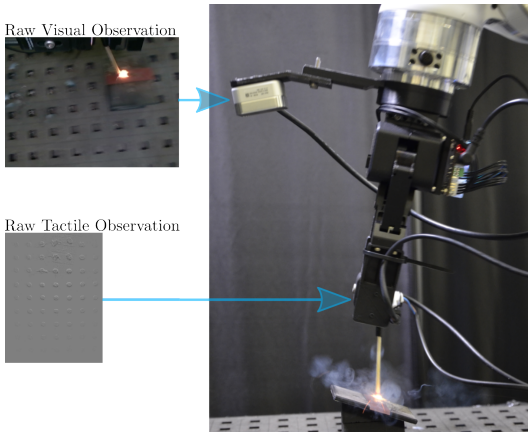

Fig. 1: Visualization of a visuotactile policy rollout in which the match is lit up successfully.

To solve this intricate manipulation task solely from local embodied sensing, we propose a multi-modal learning from demonstrations framework with an emphasis on learning from only 20 demonstrations. This data is then exploited to learn an expressive multi-modal flow matching policy [14], [15] suitable for reactivity and real-time inference. Given the few demonstrations, we employ a modular policy architecture which allows to compare different encoding strategies given the real world observation data.

Our experimental results demonstrate the efficiency of the proposed framework and showcases that the visuotactile policies can robustly light up matches across different scenarios and observation encoding strategies, despite learning from only 20 demonstrations. They also reveal that the vision-only policies perform considerably worse throughout all the evaluations, thereby underlining the importance of tactile sensing for obtaining reliable robotic match lighting policies.

II. LEARNING MATCH LIGHTING POLICIES FROM DEMONSTRATIONS

Fig. 1 depicts the components of our robotic match lighting environment. In terms of sensing, this work exclusively considers local, embodied information i.e., the image infor-

[1]Technical University of Darmstadt. [2]LASR Lab, TU Dresden.
Corresponding author: Niklas Funk. Email: niklas@robot-learning.de

This work has received funding from the German Research Foundation (DFG) Emmy Noether Programme (CH 2676/1-1), the EU's Horizon Europe project ARISE (Grant no.: 101135959). This work was also partly supported by DFG as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden, and by Bundesministerium für Bildung und Forschung (BMBF) and German Academic Exchange Service (DAAD) in project 57616814.

mation from an Intel RealSense D405 camera mounted in the robot's wrist, an open source Evetac [16] tactile sensor mounted within the Franka Panda's parallel gripper, and local velocity information. While Evetac naturally returns asynchronous event information, for compatibility with the other sensors, we integrate its events for $40\,\mathrm{ms}$, thereby converting the event information into image form. In line with this choice, we also collect all the other sensor information at $25\,\mathrm{Hz}$. Since the task is delicate, image (or tactile image) resolution might be crucial. Thus, we maintain a high resolution of 320×240 pixels for all image modalities.

Similar to [10], we collect the demonstrations through kinesthetic teaching. This procedure ensures that the human demonstrator directly feels the interaction forces between the match and the striker paper. This feedback has been crucial for achieving high task success rates during data collection.

Our multimodal policy learning framework follows the recent trend of leveraging generative models as policies for robotic manipulation. Since the match lighting task is delicate and requires reactivity, we propose to employ a model based on flow matching [17], [8]. In particular, we follow [8] and learn an SE(3)-Rectified Linear flow model. We impose a flow in SE(3), as the model's output should be the desired future trajectory of the robot end-effector. In other words, the resulting policies' action space is a sequence of $N = 16$ SE(3) poses, $\boldsymbol{T}_a = (T_a^1, \ldots, T_a^N) \in SE(3)^N$. However, unlike [8], the core of our policy is a multimodal transformer architecture [18], that receives as inputs observations from multiple sensors, including the RGB camera image, the current end-effector velocity, and, when available, observations from the Evetac tactile sensor. The policy internally then outputs velocity update vectors to refine the action sequence. This process is repeated for 5 iterations and returns the complete refined action sequence. The observations are the crucial source of information to refine the actions. Since we later want to compare different sensor combinations, we ensure modularity, i.e., the individual observation modalities are first encoded individually into latent vectors of dimension 64. These latent vectors then serve as the input to a transformer for refining the action sequence. Since we only consider 20 demonstrations for each task, we evaluate the policies' performance under different observation encoders. For the image observation we consider a pre-trained ResNet 18 [19] and training the ResNet from scratch. For the tactile observations, we consider the pre-trained model from [16], and training this architecture from scratch. If not stated differently, we use the pre-trained observation encoders but also optimize them while optimizing the policy for action generation.

III. EXPERIMENTAL RESULTS

This section evaluates our trained policies on the match lighting task. We consider two task versions. One in which the match is always grasped with the same pose, and a more complicated one, where the grasping location is varied within translational offsets of $\pm 1\,\mathrm{cm}$ & rotational perturbations of $\pm 10\,°$. For both tasks we collected 20 successful demonstrations within 1 hour. We then trained our models for 500 epochs. This evaluation reports the mean performance across task and model configurations. For each combination we trained 3 seeds, and evaluated the last checkpoint through performing 5 rollouts on the real system, respectively.

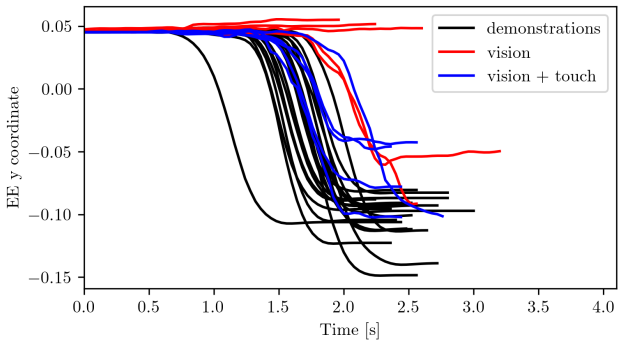

Fig. 2: Comparing Policy Rollouts with the Demonstration Data. Qualitatively, the visuotactile (vision + touch) policies better match the demonstration data than the vision-only policies. This plot considers the y-coordinate of the end effector which corresponds to the direction in which the robot needs to accelerate in order to light up the matches.

TABLE I: Success Rate on the Variable Grasp Pose Task.

Policies	Observation Encoders		
	Pretrained + optimize	Pretrained (frozen)	Train from Scratch
Vision-only	13%	6%	20%
Vision + Touch	**80%**	73%	73%

Fixed Grasp Pose. The visuotactile policies outperform the vision-only policies, achieving a success rate of 86% compared to 33%. Fig. 2 reveals that apart from the differences in success rate, the rollouts of the visuotactile policies better match the demonstrations. In particular, the visuotactile policy evaluations better align in terms of the timing of accelerating along the striker paper, which corresponds to the end-effectors y-axis. This finding hints at the fact that the vision-only policies struggle to precisely detect the point in time of making contact since this indicates that the acceleration phase along the striker paper should follow.

Variable Grasp Pose. We repeat the same procedure for the scenario of considering variable grasping poses. Yet, in this scenario we consider a wider class of observation encoders. In particular, we train policies with the pre-trained encoders and either freeze or optimize them when training the policies to fit the dataset. We also consider training the observation encoders from scratch. As presented in Tab. I, in this new, more complicated scenario, there remains a significant difference between the vision-only and visuotactile (vision+touch) policies. The superior performance of the visuotactile policies also holds across the different observation encoding strategies. We, therefore, conclude that touch is a crucial sensing modality to learn performant match lighting policies from few demonstrations. In the future, we want to investigate whether these findings transfer to different tasks and look into further improving the overall performance of the visuotactile policies.

REFERENCES

[1] S. K. Sampath, N. Wang, H. Wu, and C. Yang, "Review on human-like robot manipulation using dexterous hands." *Cogn. Comput. Syst.*, vol. 5, no. 1, pp. 14–29, 2023.

[2] Y. Huang, D. Fan, H. Duan, D. Yan, W. Qi, J. Sun, Q. Liu, and P. Wang, "Human-like dexterous manipulation for anthropomorphic five-fingered hands: A review," *Biomimetic Intelligence and Robotics*, p. 100212, 2025.

[3] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *Journal of machine learning research*, vol. 22, no. 30, pp. 1–82, 2021.

[4] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, no. 1, pp. 297–330, 2020.

[5] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid, "Aloha unleashed: A simple recipe for robot dexterity," in *8th Annual Conference on Robot Learning*, 2024.

[6] 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*, p. 02783649241273668, 2023.

[7] M. R. Cutkosky and J. M. Hyde, "Manipulation control with dynamic tactile sensing," in *6th international symposium on robotics research, Hidden Valley, Pennsylvania*, 1993.

[8] N. Funk, J. Urain, J. Carvalho, V. Prasad, G. Chalvatzaki, and J. Peters, "Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching," *arXiv preprint arXiv:2409.04576*, 2024.

[9] A. Prasad, K. Lin, J. Wu, L. Zhou, and J. Bohg, "Consistency policy: Accelerated visuomotor policies via consistency distillation," *arXiv preprint arXiv:2405.07503*, 2024.

[10] K. Kronander and A. Billard, "Learning compliant manipulation through kinesthetic and tactile human-robot interaction," *IEEE transactions on haptics*, vol. 7, no. 3, pp. 367–380, 2013.

[11] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing," *arXiv preprint arXiv:2410.24091*, 2024.

[12] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao, "Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation," *arXiv preprint arXiv:2310.16917*, 2023.

[13] R. S. Johansson, "The effects of anesthesia on motor skills," [Online] - https://www.youtube.com/watch?v=0LfJ3M3Kn80, [Accessed 15-12-2024].

[14] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[15] R. T. Chen and Y. Lipman, "Riemannian flow matching on general geometries," *arXiv preprint arXiv:2302.03660*, 2023.

[16] N. Funk, E. Helmut, G. Chalvatzaki, R. Calandra, and J. Peters, "Evetac: An event-based optical tactile sensor for robotic manipulation," *IEEE Transactions on Robotics*, 2024.

[17] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[18] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.