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 recent years. At the sensing level, several novel tactile sensors have been developed, capable of providing accurate contact information. 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 policy rollouts. However, this integration has so far been underexplored, most notably in dynamic, contact-rich manipulation tasks where precision and reactivity are essential. This work therefore proposes a multimodal, visuotactile imitation learning framework that integrates a modular transformer architecture with a flow-based generative model, enabling efficient learning of 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. The experimental results highlight the effectiveness of our approach and show that adding tactile information improves policy performance, thereby underlining their combined potential for learning dynamic manipulation from few demonstrations. Project website: `https://sites.google.com/view/tactile-il`.**

I. INTRODUCTION

Robotic manipulation still remains far from matching human dexterity and efficiency [1], [2]. A promising direction toward closing this gap is leveraging human demonstration data for learning robotic manipulation through imitation [3], [4], [5], thereby actively exploiting humans' task understanding and advanced manipulation capabilities. Although numerous studies have shown that access to touch sensing benefits human manipulation performance [6], [7], [8], the majority of current works in imitation learning for manipulation are still missing out on this modality [4], [5], [9]. This raises the question of *whether learning robotic manipulation policies from human demonstrations could also benefit from incorporating tactile sensing*.

This work approaches this question by studying the impact of touch sensing for learning a dynamic task, namely, igniting matches. We argue that match lighting is a challenging and effective testbed because the task requires dynamic motion and compliance [10], which introduces additional complexity compared to standard, quasi-static tasks such as pick-and-place or insertion which have recently been addressed with multisensory learning approaches [11], [12], [13], [14], [15], [16]. Moreover, it is a task for which there is evidence that the availability of touch sensing impacts human performance [17]. Despite the task's relevance, to the best of our knowledge, it has only previously been investigated by Kronander et al. [10], who considered fixed match grasp poses in a precisely calibrated setup without including high-dimensional observations. Our work instead addresses more complicated scenarios, including varying grasp poses and striker paper orientations, while solely considering local embodied sensing, i.e., RGB wrist-camera images, the end effector velocity, and the information from an event-based optical tactile sensor (cf. Fig. 1).

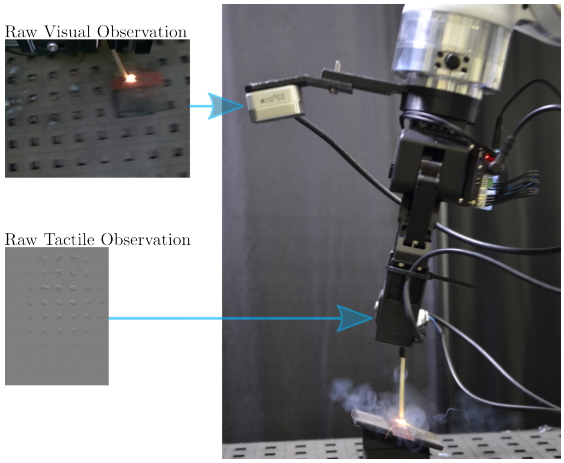

Fig. 1: Autonomous rollout of a policy that is conditioned on visual and tactile observations illustrated on the left. The policy controls the robot and, thereby, the contact configuration between the match and striker paper. As can be seen, the policy ensures sufficient force and velocity, resulting in successfully igniting the match. This work highlights the importance of tactile sensing for reliably solving the dynamic and delicate task of lighting up matches.

To address the challenges of this dynamic and intricate manipulation task, we also introduce a multi-modal learning from demonstrations framework. The proposed framework is built upon an expressive multi-modal flow matching policy [18] with a modular and efficient transformer-based policy architecture. This combination enables reactivity through fast real-time inference, and comparing different encoding and training strategies given the real-world observation data. To further restrict the human efforts for learning the task, we emphasise learning from a few demonstrations and consider only 20 available demonstrations. The experiments demonstrate the efficiency of the proposed framework and showcase 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 worse

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

throughout all evaluations. Additionally, we find that vision-only policies can benefit from employing a masked training procedure that exploits tactile observations during training. The results, therefore, underline that tactile information is a beneficial source of information for learning reliable robotic match lighting policies from few demonstrations.

Overall, we contribute a multi-modal framework for efficiently learning robust and reliable manipulation policies suitable for dynamic tasks such as lighting up matches. Moreover, we present a masked training procedure that exploits the tactile signals only during training and allows for increased success rates of vision-only policies. Lastly, we contribute an extensive evaluation conducted in our modular real-world match-lighting testing environment. The experiments across different policies and experiment configurations demonstrate the competitive performance of our proposed approach and its individual components. They also reveal that tactile observations are important for learning performant policies for dynamic tasks like match lighting and closely matching the human demonstration data.

II. RELATED WORK

Artificial tactile sensors are a promising technology to advance robotic manipulation as they enable the direct sensing of contacts [19]. Together with the emergence of commercial [20] and open-source tactile sensors [21], [22], the field of tactile robotic manipulation is gaining increased attention.

One approach to obtain tactile manipulation policies is through reinforcement learning [11], [12], [23]. Since reinforcement learning requires exploration, learning performant policies demands a vast amount of environment interactions. To account for this, previous works rely on simulation, allowing for fast sample generation while also mitigating the sim-to-real gap [12], [23], [24], [25]. Alternatively, [11] presented approaches for learning policies directly on real robots. This, however, requires a carefully designed experiment setup enabling autonomous exploration, as multiple hours of real-world interactions are needed for successful policy learning. Since our task of match lighting is challenging to simulate, and since safety considerations hinder realizing autonomous exploration on the real system, this work takes a different direction. We want to efficiently learn match lighting policies from few real-world expert demonstrations, thereby significantly reducing the data requirements.

The field of learning robotic manipulation policies from demonstration data [3] has lately received increasing attention [4], [5], [9], [26], [27]. Several works showed the effectiveness of training generative models based on expert demonstrations for obtaining advanced real-world manipulation [4], [5]. The field also benefits from efforts for building effective devices for collecting human demonstrations [28], [29]. However, the majority of works in imitation learning focus on quasi-static manipulation tasks and only incorporate RGB or RGBD cameras as external sensors without considering tactile information [4], [5], [9], [28]. This work follows the current efforts and proposes an efficient and modular multi-modal framework for learning from demonstrations by leveraging a generative model trained as a policy. Yet, it differs in that it considers tactile sensors as input modality and investigates the contact-rich and dynamic manipulation task of igniting matches. Only more recently, a few works investigated adding tactile sensing capabilities [13], [14], [30], [15] into imitation learning frameworks. Yet, these works also focused on quasi-static manipulation tasks and did not consider dynamic manipulation as we do herein. While [13] leverages diffusion policy for policy learning and [30], [15] use a standard mean-squared error behavioural cloning loss, none of the works investigated a flow-matching-based policy, which we find to be a key component for high success rates. In a concurrent effort, [31] introduces Reactive Diffusion Policy, which also achieves more reactive control, however, through hierarchical decomposition, requiring explicitly training two separate policies. Furthermore, this work introduces a masked training procedure, showcasing that considering tactile observations during training can enhance the inference performance of vision-only policies.

From a task-level perspective, [10] is closest related as it also investigates learning match lighting policies from human demonstrations. To achieve good task success rates, they propose employing a varying stiffness controller learned through information from a human-robot interface. Instead of learning a variable stiffness controller, this work directly learns a reactive policy capable of controlling the contact forces by varying the desired target poses. Moreover, this work extends upon [10] in that it considers a more realistic experimental setup, including varying match poses, striker paper orientations, and conditioning the policies onto high-dimensional image and tactile observations.

Overall, we contribute a framework for learning visuo-tactile robotic match lighting policies from human demonstrations and showcasing that tactile sensing is crucial for learning high performance policies on this dynamic task.

III. LEARNING MATCH LIGHTING POLICIES FROM DEMONSTRATIONS

This section describes our approach for learning the dynamic manipulation skill of lighting up matches from few real-world expert demonstrations and deploying the policies on the real system. In terms of sensing, this work exclusively considers local, embodied information, i.e., the image information from an Intel RealSense D405 camera mounted in the robot's wrist, an open source Evetac [22] tactile sensor mounted within the parallel gripper, and local velocity information (cf. Fig. 2). The following sections detail the learning framework, the policy architecture, the data collection, and the policy inference procedure.

A. Fast and Reactive Multi-modal Policies through Conditional Flow Matching

Our multimodal policy learning framework leverages a generative model as policy. Given the current observations, the model should output actions close to the demonstrations. Since the match lighting task is delicate and requires reactivity, we propose to learn a policy using flow matching [32].

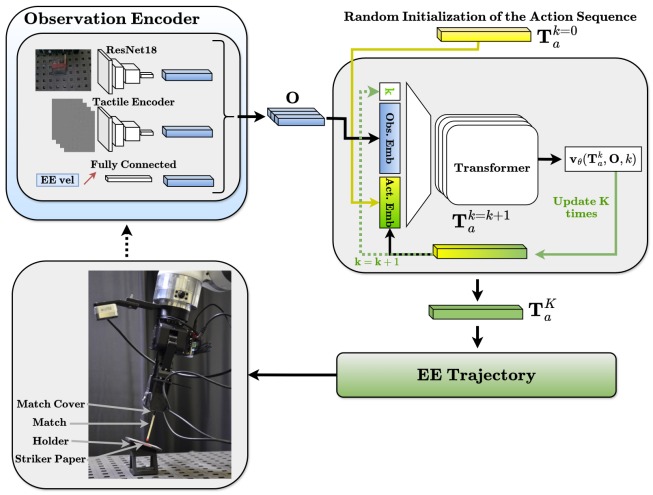

Fig. 2: Method Overview. Upon retrieving the current observations, they are first encoded individually inside the observation encoder and brought into a common shape, i.e., each modality contributes a latent vector of a fixed shape. These latent vectors, together with the current action sequence & time index, then serve as the input to the transformer architecture, which outputs velocities to iteratively refine the action sequence through flow matching. Upon retrieving the final desired end effector trajectory, it is sent to the robot and tracked through a Cartesian Impedance Controller. Note that we only apply the first action to maintain reactivity.

We learn an SE(3)-Rectified Linear flow model [9] that generates high-quality samples within low inference times. We impose a flow in SE(3), as the model should output the desired future trajectory of the robot end-effector, a sequence of $N = 16$ SE(3) poses, $\boldsymbol{T}_a = (T_a^1, \dots, T_a^N) \in SE(3)^N$.

The idea in conditional rectified linear flow matching is to impose a straight line path between samples from a noise distribution $\boldsymbol{a}_{t=0} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and samples from the dataset $\boldsymbol{a}_{t=1} \in \mathcal{D}$. The intermediate waypoints of the flow are thus defined by $\boldsymbol{a}_t = t\boldsymbol{a}_{t=1} + (1-t)\boldsymbol{a}_{t=0}$, $t \in [0, 1]$. The objective is then to learn the velocity field of this path $\mathrm{d}\boldsymbol{a}_t/\mathrm{d}t$, such that during inference, samples can be generated by starting from a random initial sample and iteratively refining the sample through Euler integration given a learned estimate of the velocity field. For our case of dealing with SE(3) action poses that should be generated $\boldsymbol{a}_t = (\boldsymbol{p}_t \in \mathbb{R}^3, \boldsymbol{r}_t \in SO(3))$, we decouple the translational and rotational component of the flow and obtain $\dot{\boldsymbol{p}}_t = (\boldsymbol{p}_1 - \boldsymbol{p}_t)/(1-t)$ & $\dot{\boldsymbol{r}}_t = (\mathrm{Log}(\boldsymbol{r}_t^{-1}\boldsymbol{r}_1))/(1-t)$ for the velocity field. Given the training data, we then train a parameterized Flow Matching model $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{p}_t, \boldsymbol{r}_t, \boldsymbol{O}, t)$, that, conditioned on the current observation $\boldsymbol{O}$ and "action" pose, outputs translation and rotation velocities ($\boldsymbol{v}_{\boldsymbol{\theta},\boldsymbol{p}} \in \mathbb{R}^3$ & $\boldsymbol{v}_{\boldsymbol{\theta},\boldsymbol{r}} \in \mathbb{R}^3$). The model is trained by minimizing $\mathcal{L} = ||\boldsymbol{v}_{\boldsymbol{\theta},\boldsymbol{p}} - \dot{\boldsymbol{p}}_t||^2 + ||\boldsymbol{v}_{\boldsymbol{\theta},\boldsymbol{r}} - \dot{\boldsymbol{r}}_t||^2$. During inference, we sample actions by iteratively refining random initial actions through $\boldsymbol{p}_{k+1} = \boldsymbol{p}_k + \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{p}_t, \boldsymbol{r}_t, \boldsymbol{O}, t)\Delta t$ & $\boldsymbol{r}_{k+1} = \boldsymbol{r}_k \mathrm{Exp}(\Delta t \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{p}_t, \boldsymbol{r}_t, \boldsymbol{O}, t))$. Note that we define the flow for the entire action sequence of 16 poses.

B. Policy Architecture

Following the previous section, our approach employs a parameterized SE(3)-Rectified Linear Flow matching model for obtaining the actions. At the core of this policy is a multimodal transformer architecture that receives observations from multiple sensors, including the RGB camera image, the current end-effector velocity, and, when available, observations from the Evetac tactile sensor. Transformers are particularly suitable for this task as they can seamlessly handle the multiple multimodal observations [33]. The resulting transformer-based policy architecture is illustrated in Fig. 2.

The observations are the crucial source of information for refining 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. The first 5 entries of this 64-dimensional vector are learnable weights that should inform the transformer about the type of observation modality. These latent vectors then serve as the input to a transformer for refining the action sequence. Importantly, the latent observations and entries of the action sequence enter the transformer as individual tokens. The modular policy architecture thus allows for seamlessly evaluating the policies' performance under different observation encoders. It also enables a masked training procedure that stochastically decides upon the modalities which are available in the transformer. The image observations are processed through a pre-trained ResNet 18 [34] or by training the ResNet from scratch. For the tactile observations, we consider the pre-trained model from [22], and training this architecture from scratch. These features (i.e., one per observation modality, one for each action in the action sequence, and one for the current time index) are the inputs to the transformer model, which consists of 4 layers with 4 attention heads. Inside the transformer, the inputs exchange information with each other and update their embeddings through multi-head attention [35]. In its standard implementation, all inputs exchange information with each other (including self-connections). Herein, we configure the transformer's attention mask to full connectivity between the observation tokens, while the action tokens solely cross-attend to the observation tokens. The value of the action tokens thus does not influence the update of the observation tokens. This choice is made because only the observations contain information on how to update the action sequence, while the action sequence only contains noise, especially at the beginning. Moreover, the self-attention within the action tokens is configured such that action poses in the sequence only attend to previous actions. In addition to this masking scheme regarding the actions, in the experiments, we will also investigate the effectiveness of employing stochastic masking at the observation level during training. In particular, we will train a single transformer model that is provided with tactile observation during training with a probability of 50%. Due to this stochasticity on the input level, the policy has to better align the latents of the vision and touch observations so that it can generate good outputs in both cases, i.e., when touch is available and when it is not.

The transformer's final output is the updated action features representing the velocity vectors for the iterative refinement, which is repeated $K = 5$ times. After obtaining the

final action sequence, it is sent to the controller and applied to the robot. Using this generative model as policy yields online action generation as illustrated in Fig. 2.

C. Data Collection

Similar to [10], we collect the demonstrations through kinesthetic teaching. This ensures that the human demonstrator directly feels the interaction forces between match and striker paper, and has been crucial for high task success rates during data collection. From a task-level perspective, to light up the match, the match tip must first be brought into contact with the striker paper. Subsequently, the match tip has to be moved along the striker paper while applying sufficient force with sufficient velocity.

Figs. 1 & 2 depict the components of our real-world match lighting environment. Throughout the demonstrations, we record all sensor data, i.e., the image from the wrist-mounted Intel RealSense D405 camera, an open-source Evetac [22] tactile sensor mounted within a Robotis RH-P12-RN gripper attached to the end effector of a 7-DoF Franka Panda, and the local end-effector velocity information. We also record the end-effector poses that the robot moves through. They contain the trajectory information that the robot should follow. Yet, we want to emphasize that the policy framework only operates on the level of local poses expressed in the current end effector frame. While Evetac naturally returns asynchronous event information, for compatibility with the other sensors, we convert the events into image form by integrating them for every pixel for a duration of $40\,\mathrm{ms}$. 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. As shown in Figs. 1 & 2, for the images of the wrist-mounted camera, we ensure that the match and the tip of the match is fully observable during the trajectories. Moreover, we found that using the striking surfaces of regular paper matchboxes resulted in short durability after a few experiments. We, therefore, decided to 3D-print a thin rectangular plate to hold the striker paper. In its standard configuration, the plate is raised and placed with an angle of $20°$ relative to the table (cf. Fig. 1). We used long standard matches with dimensions of $(100\,\mathrm{mm} \pm 5\,\mathrm{mm}) \times (4\,\mathrm{mm} \pm 1\,\mathrm{mm}) \times (4\,\mathrm{mm} \pm 1\,\mathrm{mm})$ to keep the fire at a sufficient distance from the silicone surfaces of the tactile sensors mounted inside the gripper. Lastly, we also 3D printed hollow cylindrical cones to cover the upper $45\,\mathrm{mm}$ of the matches. This was necessary to significantly increase the longevity of the silicone gels that cover the tactile sensor, which could rip easily when in direct contact with the matches.

D. Policy Inference and Robot Control

We use the Cartesian Impedance Controller from [36] to move the robot during the autonomous policy rollouts. We tuned the controller's stiffness and damping values on a few of the collected demonstration trajectories. The gains have been chosen such that replaying the trajectories obtained during kinesthetic teaching yields task success when tracked using this control strategy. We rely on the Robotic Operating System (ROS) to gather the sensor observations. Policy inference is run asynchronously, and only the first action of the action sequence is applied by the controller before updating the action sequence based on the most recent model inference. The policies run online in real-time as action generation, i.e., policy inference, only takes $0.028\,\mathrm{s}$ for our largest vision+touch policies on an NVIDIA 3090 GPU.

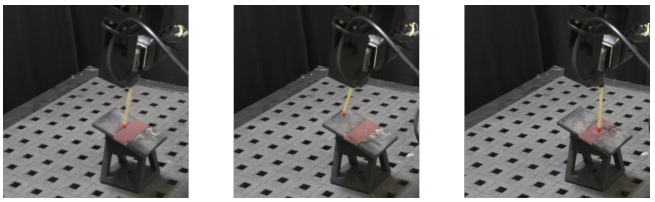

Fig. 3: Visualizing the versatility of the initial configurations during the experiments. Left: Fixed grasp pose strategy. Middle & Right: Two examples of the variable grasp initialization. Note how the initializations yield different configurations w.r.t. distance and angle between match and striker paper that the policies have to handle for solving the task.

IV. EXPERIMENTAL RESULTS

This section evaluates our proposed approach. It is structured along four main questions to investigate the importance of tactile sensing and the effectiveness of our approach for the dynamic manipulation task of lighting up matches: **A:** How important is tactile feedback for obtaining performant match lighting policies? **B:** Can the vision-only policies benefit from leveraging the tactile information during training? **C:** How does our proposed approach perform compared to baselines? and **D:** Are the policies robust w.r.t. generalizing to novel scenarios?

The following evaluation considers 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 $\pm 10\,°$ (cf. Fig. 3). For both tasks, we collected 20 successful demonstrations within 1 hour. We then trained our models for 500 epochs. The evaluations report the mean performance together with the standard deviation across task and model configurations. We trained 3 seeds per combination and evaluated the last checkpoint through 10 rollouts on the real system.

A. How important is tactile feedback for obtaining performant match lighting policies?

Fixed Grasp Pose. In the fixed grasp pose scenario (cf. Fig. 3, left), the vision+touch policies outperform the vision-only policies, achieving a mean success rate of 87% compared to 33%, with both having a standard deviation of 12%. Apart from the differences in success rate, Fig. 4 reveals that the rollouts of the vision+touch (also referred to as visuotactile) policies better match the demonstration data. 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

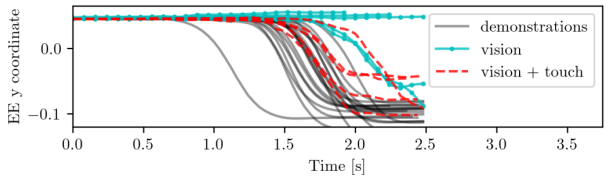

Fig. 4: Comparing the demonstrated trajectories with trajectories from rolling out different policies, considering the y-coordinate of the end effector. The y-coordinate is the direction along the striker paper in which the robot needs to accelerate to light up the matches. Qualitatively, the vision+touch policies generate rollouts that better match the demonstrations compared to the vision-only policies, indicating that the tactile observations contain important information for explaining and matching the demonstrations.

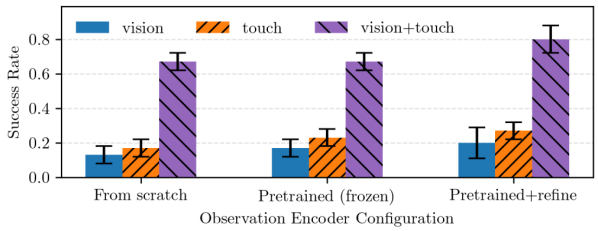

Fig. 5: Comparing the success rates (mean and std deviation) of different policies on the variable grasp pose task. Across different observation encoding strategies, the vision+touch policies consistently outperform the vision-only policies by at least 50%, thereby highlighting the importance of tactile sensing for obtaining reliable match lighting policies. The vision+touch policies also outperform a touch-only baseline underlining that touch alone is insufficient for high success rates.

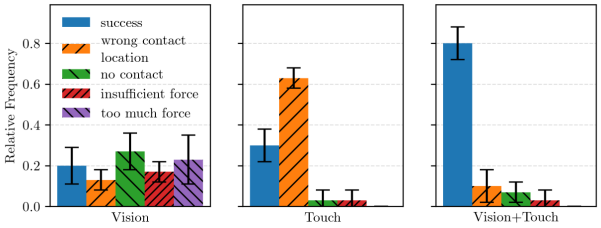

Fig. 6: Comparing success rates and different failure modes (mean and std deviation) for the vision-only, touch-only, and vision+touch policies in the variable grasp pose evaluation, for the pretrained+refine encoding strategy (cf. Fig. 5). The vision+touch policies reduce the failure rate by over 40%, substantially decreasing contact-related failures — not applying force, applying insufficient force, or applying excessive force — compared to vision-only policies, and also reducing wrong-contact-location failures compared to touch-only policies.

that 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 procedure for the variable grasping poses (cf. Fig. 3, middle & right), yet considering a wider class of observation encoders. In particular, we train policies with the pre-trained encoders and either freeze or optimize them during policy training. We also investigate training the observation encoders from scratch. As presented in Fig. 5, in this new, more complicated scenario, there remains a significant difference between the vision-only and vision+touch policies in terms of success rate. Importantly, the superior performance of the visuotactile policies holds across the observation encoding strategies, and adding the tactile observations improves the task success rates by at least 50%. While the best visuotactile policies achieve an average success rate of 80%, the best performing vision-only policies only reach success rates of up to 20%. The visuotactile policies also reliably outperform the touch-only baseline, demonstrating that only the combination of visual perception from the RGB wrist-camera images and tactile perception leads to robust policy performance. Given the variability of the task and the selected observation modalities, visual input is essential for reliably guiding the match tip toward the striker paper—especially since the policy operates solely on local end-effector velocities without access to the global end-effector poses. However, it is the integration of vision and touch that enables high success rates, as tactile perception is critical for minimizing contact-related failures and significantly improving performance.

Fig. 6 provides a more detailed comparison in that it differentiates between different failure modes of the policies. We consider four types of failures: 1) making contact in the wrong location, i.e., the tip of the match not making contact with the striker paper, 2) not making contact at all during the policy rollout, i.e., the policy accelerating along the striker paper but without making contact, 3) insufficient contact force, i.e., making contact in the right location but without sufficient force resulting in the match not lighting up, and 4) applying too much force during the rollout, i.e., the policy pressing the tip of the match with too much force against the striker paper which results in the match sliding through the fingers. The last failure case is mainly related to the policy missing the transition between the approaching phase of the task and the phase of accelerating along the striker paper. As shown in the comparison, both, the touch-only and the vision-only policies exhibit significantly increased failure rates. Most failures of the touch-only policies stem from making contact in the wrong location, confirming that the task also needs spatial understanding. The vision-only policies' failures related to resolving the current contact state (no contact, insufficient force & too much force) are the most prominent ones with 27%, 17%, & 23%, respectively. In contrast, adding the tactile observations yields significantly reduced failure rates. The few failure cases of the vision+touch policies are mainly related to not making contact in the right location (10%), while none of the vision+touch policies apply too much contact force.

Lastly, Fig. 7 illustrates the evolution of the attention weights of the individual inputs of the transformer w.r.t. the update of the fifth action of the action sequence for a visuotactile policy (trained using pre-trained weights but further refining the encoder during training). In other words, it shows how the inputs contribute to updating the action. As can be seen, initially, the vision observation from the RealSense is the most important modality. This is expected, as the camera information is crucial to moving the robot closer to the striker paper. The event-based tactile sensor does not provide any information during this phase, as there

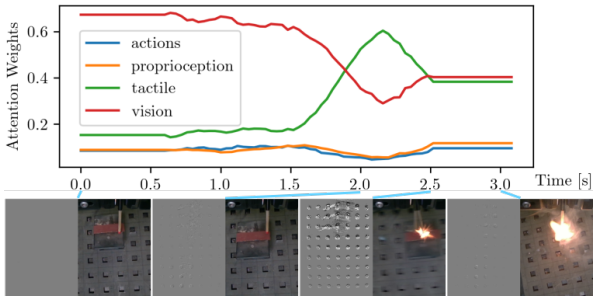

Fig. 7: Visualizing the evolution of the attention weights over time for one exemplary trajectory. The bottom images show the task progression. The plot shows the weights that are attributed to the individual inputs of the transformer: 1) the actions, 2) the proprioception observation (end effector velocity), 3) the tactile observation from Evetac, and 4) the vision observation from the Realsense camera. The weights are w.r.t. to updating the fifth action of the desired end-effector trajectory, which is computed for every observation along the rollout. At the beginning and end of the trajectory (when there are no tactile signals), vision is the most important modality. Once there are changes in contact configuration, touch is the most important modality for action generation, therefore highlighting that touch provides important feedback for controlling the contact configuration.

is no change in contact configuration. However, once contact is made, the tactile inputs gain importance and become the most important entity. This holds true until the match ignites, which signals successful task execution. The other inputs, i.e., attention to the other actions and to the proprioception observation, stay low throughout the trajectory.

Overall, based on the findings from these experiments, we conclude that touch is a crucial sensing modality for learning performant match lighting policies from few demonstrations, and that it is particularly effective for reducing contact-related failure modes of vision-only policies.

B. Can the vision-only policies benefit from leveraging tactile information during training?

While the previous section showed the importance of conditioning the policies onto tactile signals, this section investigates whether vision-only policies can benefit from leveraging tactile information during policy training. In particular, we exploit the transformer architecture's natural capability to handle input sequences of different lengths and investigate the effectiveness of the masked training procedure (cf. Sec. III-B). During the masked training, the policy either receives all of the input modalities or all of the input modalities except for the tactile signals. The masking probability is set to 50%. Since the policy uses the same transformer independent of the masking, it has to align the latent spaces to generate meaningful outputs given the different input combinations. This experiment now investigates whether the masked training procedure can improve the performance of the vision-only policies in the variable grasp pose scenario. We start with the pre-trained encoders and optimize them during the training. This choice is made because the pre-trained encoders already provide a meaningful embedding when the respective modality is

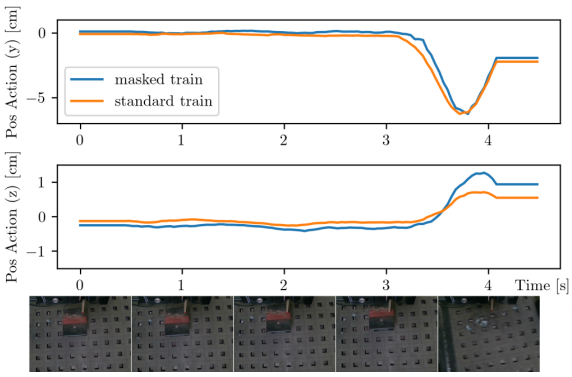

Fig. 8: Comparing the policy predictions of two vision-only policies (one trained with the standard procedure, the other one with the masked one). We visualize the policy predictions for the y- and z-component of the 5th action on a trajectory that was obtained by rolling out the standard policy and on which it fails to establish sufficient contact between the match and the striker paper. As shown, the policy that underwent the masked training procedure proposes different actions, i.e., moving closer to the striker paper before accelerating along the striker paper (as shown for the z-predictions when $T < 3.5\,\mathrm{s}$). Additionally, it proposes to accelerate at a later point in time along the striker paper, as shown for the y-axis predictions.

TABLE I: Success Rate of different vision-only policies in the variable grasp scenario. The policies differ regarding the training procedure, i.e., whether they are trained with the standard procedure or with masked training that considers the tactile signals during training. The masked training procedure, i.e., leveraging touch during training, is effective and yields increased success rates.

	Training Configuration	
	Standard Training	Masked Training
Success Rate	20% (8%)	**40% (8%)**

important. This is particularly important for the tactile representation, as the masked training procedure indirectly forces the optimization of the vision encoder to account for the missing tactile information.

As shown in Tab. I, the vision-only policies that have undergone the masked training procedure achieve significantly higher success rates, increasing the number of successful rollouts by a factor of 2 and achieving an overall success rate of 40%. In particular, while the policies trained with the standard procedure often fail to establish contact between the match and the striker paper (46% in this experiment), the policies that underwent the masked training procedure exhibit a significantly decreased probability of this failure mode (25%). To underline this finding quantitatively, Fig. 8 compares the differently trained policies regarding action

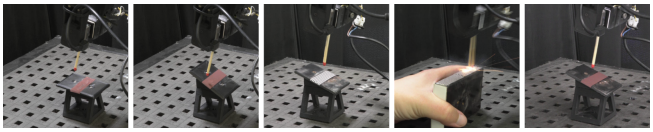

Fig. 9: Visualizing the experiment setups considered in the generalization experiments. From left to right: Mounting angle of 5°; Mounting angle of 30°; Nominal mounting angle of 20° but different striker paper; Using an actual handheld matchbox; Varying lighting conditions (decreased illumination). Note how the different mounting angles alter the angle & distance between match & striker paper, and how the other changes affect the visual appearance.

TABLE II: Success Rate of different visuotactile policies in the variable grasp scenario. The policies differ w.r.t. architectures and training objectives. Our proposed architecture together with the flow-matching objective yields best performance.

Policy	Success Rate
See, Hear, and Feel [15]	50% (8%)
Diffusion Policy (DP architecture + DDIM) [5]	30% (14%)
Our architecture (III-B) + DDIM	53% (5%)
Ours (Our architecture (III-B) + Flow (III-A))	**80% (8%)**

generation. It visualizes the translational outputs for the 5th action in the sequence along the y- (direction of acceleration along the striker paper) and the z-direction (controlling the height of the match tip). For the comparison, we consider a trajectory that has been obtained by rolling out the policy trained with the standard procedure. During this trajectory, the policy failed to establish contact between the match and the striker paper. Considering the z-component of the predicted action, before the start of the sideways motion, the policy that was trained using the masked procedure outputs lower values, thereby indicating that it wants to move the end effector lower, increasing the probability of making contact with the striker paper. Considering the y-direction, it is also evident that the policy trained using the standard procedure aims to move along the striker paper earlier. This behaviour again increases the probability of accelerating too early without making proper contact with the striker paper. We conclude that the masked training procedure increases the success rates of vision-only policies. Therefore, the availability of tactile observations can improve policy performance, even when tactile feedback is only provided during training.

C. How does our proposed approach perform compared to baselines?

This section compares the performance of visuotactile policies trained with our proposed approach against baseline methods on the variable grasp pose task. As baselines, we consider See, Hear, Feel [15], which similarly employs a multi-head attention-based policy architecture but differs in its training strategy: it relies on an explicit behaviour cloning loss, directly regressing to the action prediction. We also compare against the vanilla implementation of Diffusion Policy (DP) [5] that uses the DDPM sampler during training and DDIM for sampling during inference. To disentangle the effect of the policy architecture and the sampler, we create a third baseline that consists of our proposed policy architecture (cf. Sec. III-B) trained with DDPM, and using DDIM during inference. To ensure a fair comparison, all policies are trained on the same vision+touch data, equipped with identical observation encoder architectures, and configured to maintain a comparable number of parameters. The previously mentioned choices ensure that all the policies achieve real-time inference given our control frequency of $25\,\mathrm{Hz}$.

The results in Tab. II showcase that our proposed approach outperforms the baselines. The lower success rates of the See, Hear, and Feel [15] baseline stem from difficulties in reliably reaching the striker paper, which we attribute to the behavioral cloning loss struggling with multi-modal data during the approach phase. While the DP baseline employs a more expressive generative model, the vanilla version performs worse in this comparison, with only 30% successes on average. Notably, exchanging the DP architecture with our proposed model architecture (our architecture + DDIM) improves success rates to 53%. Compared to our proposed architecture, the DP architecture concatenates the encodings of the individual observation modalities without enforcing equal dimensionality within the policy network. We hypothesise that this design choice accounts for the observed performance gap, which may be particularly impactful given the low-demonstration regime. Indeed, introducing a bottleneck that aligns the dimensionality of the encoded observations before feeding them into the DP architecture improves performance to 47% (5%), supporting this hypothesis. Nevertheless, it still performs slightly worse compared to employing our architecture + DDIM. Lastly, the results showcase another improvement in performance to 80% upon combining our architecture with the proposed flow-matching generative model. Given the budget of only 5 inference iterations, the straight-line rectified flow produces more precise and less noisy actions compared to DDIM, which is crucial for task success in the match lighting task that requires both precision and reactivity. To further support this choice of limiting inference to $K{=}5$ iterations, we conducted an ablation with $K{=}10$ for our proposed approach. While the choice of $K{=}10$ increased the inference time to $55\,\mathrm{ms}$, performance dropped to a 50% (10%) success rate, due to the less regular action updates, which increase the likelihood of the policies accelerating before establishing sufficient contact. Overall, these results underline the importance of our approaches' individual components and their competitive performance.

TABLE III: Success Rate of our vision+touch policies when evaluating the policies in novel, previously unseen scenarios.

Evaluation Configuration	Success Rate
Different Mounting Angle (5°)	77% (5%)
Different Mounting Angle (30°)	67% (5%)
Nominal Mounting Angle (20°), Different Striker Paper	77% (9%)
Actual Matchbox (handheld) & Different Striker Paper	70% (8%)
Varying Lighting Conditions	67% (5%)

D. Are the policies robust w.r.t. generalizing to novel scenarios?

This last experiment evaluates whether the visuotactile policies can generalize to novel, previously unseen scenarios (cf. Fig. 9). We evaluate the following variations: (1) altering the angle between the match and the striker paper by mounting the paper at previously unseen angles of 5° and 30°; (2) using a different, dotted striker paper; (3) replacing the 3D-printed mount with an actual matchbox that is handheld; and (4) varying lighting conditions by increasing or decreasing the intensity of the external light source. This evaluation considers the visuotactile policy with the pretrained+refine training procedure and the variable grasp initialization.

In addition to providing rollout videos in the supplementary material, quantitatively, as shown in Tab. III, the policies generalize well to the 5° mounting angle and to the different striker paper, with only a 3% drop in mean success rate compared to the policies' mean success rate

of 80% in the nominal scenario. In contrast, performance decreases by 10%, 13%, and 13% for the handheld matchbox, 30° mounting angle, and varying lighting conditions, respectively. The slightly lower successes at 30° can be attributed to the match starting closer to the 3D-printed holder, leaving less room for adjusting the angle relative to the striker paper. We further hypothesize that the handheld matchbox and lighting variations introduce the most substantial visual perturbations, leading to reduced successes as the policies more often fail to align the match tip with the striker paper in the initial phase. Despite being trained on only 20 demonstrations, the results suggest that the learned visuotactile policies exhibit robustness for variations beyond the training scenario, generalizing reasonably well to new conditions, with the overall performance remaining above the performance of the previously investigated baselines in the nominal scenario. Future improvements could be achieved by adding demonstrations for the more challenging scenarios and by refining visual data augmentation strategies.

V. CONCLUSION

This work investigated the importance of tactile sensing for performing the dynamic manipulation task of match lighting. The policies were learned from demonstration data obtained through kinesthetic teaching. We also introduced a policy learning framework combining a flow matching generative model for fast and efficient action generation and an expressive modular multi-modal transformer architecture. The experimental results showcase the effectiveness of our approach compared to competitive baselines, and that tactile feedback is important for learning performant match lighting policies from few demonstrations. Across all task variations, our proposed vision+touch policies outperformed vision-only policies, increasing the number of successful policy rollouts almost by a factor of 3. By analysing the visuotactile policies' attention weights, we confirmed that tactile observations gain importance during the contact-rich interactions between the match and the striker paper. Moreover, we also showed that exploiting the tactile signals during training and employing a masked training procedure can benefit vision-only policies and yield increased success rates. Yet, the improved vision-only policies still cannot reach the performance of the visuotactile policies. Lastly, we showed that the individual components of our approach are essential for obtaining policies with high success rates, and that the visuotactile policies are robust and can generalize to novel task variations. Taken together, these findings highlight the synergistic potential of integrating tactile sensing with suitable policy architectures to learn performant policies for dynamic manipulation tasks like match lighting. Future work should investigate transferring these findings to other manipulation tasks and further improving the policy performance by, e.g., learning from unsuccessful policy rollouts.

ACKNOWLEDGMENT

We thank Erik Helmut and Rickmer Krohn for helping with the 3D printing and attention analysis. 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), the AICO grant by the Nexplore/Hochtief Collaboration with TU Darmstadt. This work was partly supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) 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 (SECAI, School of Embedded and Composite AI).

REFERENCES

[1] S. K. Sampath, N. Wang, H. Wu, and C. Yang, "Review on human-like robot manipulation using dexterous hands." *Cogn. Comput. Syst.*, 2023.

[2] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *JMLR*, 2021.

[3] 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*, 2020.

[4] 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 *CoRL*, 2024.

[5] 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," *IJRR*, 2023.

[6] A. B. Vallbo, R. S. Johansson *et al.*, "Properties of cutaneous mechanoreceptors in the human hand related to touch sensation," *Hum neurobiol*, 1984.

[7] B. B. Edin, G. Westling, and R. S. Johansson, "Independent control of human finger-tip forces at individual digits during precision lifting." *The Journal of physiology*, 1992.

[8] E. Pavlova, Å. Hedberg, E. Ponten, S. Gantelius *et al.*, "Activity in the brain network for dynamic manipulation of unstable objects is robust to acute tactile nerve block: an fmri study," *Brain research*, 2015.

[9] 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.

[10] K. Kronander and A. Billard, "Learning compliant manipulation through kinesthetic and tactile human-robot interaction," *ToH*, 2013.

[11] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-rl for insertion: Generalization to objects of unknown geometry," in *ICRA*, 2021.

[12] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, "Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning," in *ICRA*, 2022.

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

[14] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao, "Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation," in *CoRL*, 2024.

[15] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *CoRL*, 2023.

[16] R. Feng, D. Hu, W. Ma, and X. Li, "Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation," in *CoRL*, 2025.

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

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

[19] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE T-RO*, 2020.

[20] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[21] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone *et al.*, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, 2018.

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

[23] A. Church, J. Lloyd, N. F. Lepora *et al.*, "Tactile sim-to-real policy transfer via real-to-sim image translation," in *CoRL*, 2022.

[24] T. Bi, C. Sferrazza, and R. D'Andrea, "Zero-shot sim-to-real transfer of tactile control policies for aggressive swing-up manipulation," *IEEE RA-L*, 2021.

[25] P. Ojaghi, R. Mir, A. Marjaninejad, A. Erwin, M. Wehner, and F. J. Valero-Cuevas, "Curriculum is more influential than haptic feedback when learning object manipulation," *Science Advances*, 2025.

[26] T. Ablett, Y. Zhai, and J. Kelly, "Seeing all the angles: Learning multiview manipulation policies for contact-rich tasks from demonstrations," in *IROS*, 2021.

[27] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *CoRL*, 2021.

[28] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in *RSS*, 2023.

[29] D. Mukashev, S. Seitzhan, J. Chumakov *et al.*, "E-bts: Event-based tactile sensor for haptic teleoperation in augmented reality," *IEEE T-RO*, 2024.

[30] T. Ablett, O. Limoyo, A. Sigal, A. Jilani, J. Kelly, K. Siddiqi, F. Hogan, and G. Dudek, "Multimodal and force-matched imitation learning with a see-through visuotactile sensor," *IEEE T-RO*, 2024.

[31] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu, "Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation," *arXiv preprint arXiv:2503.02881*, 2025.

[32] K. Black, N. Brown, D. Driess *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[33] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE PAMI*, 2023.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[35] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[36] "Franka Interactive Controllers," https://github.com/nbfigueroa/franka_interactive_controllers, [Accessed 02-09-2024].