Exploring the Role of Vision and Touch in Reinforcement Learning for Dexterous Insertion Tasks

Janis Lenz[1] **Inga Pfenning**[1] **Theo Gruner**[†,1,2] **Daniel Palenicek**[†,1,2] **Tim Schneider**[†,1]
Jan Peters[1,2,3,4]
[†] equal supervision [1] Institute for Intelligent Autonomous Systems, TU Darmstadt
[2] hessian.AI [3] German Research Center for AI (DFKI) [4] Centre for Cognitive Science

Abstract

Robotic insertion tasks pose significant challenges due to perceptual uncertainties and the need for precise control, especially in unstructured environments. Humans naturally integrate vision and touch to navigate such scenarios, yet replicating this seamless multimodal coordination in robotic systems remains an open research problem. This study investigates the roles of visual and tactile feedback in reinforcement learning for dexterous insertion tasks, aiming to understand their relative contributions. We analyze the interplay between these modalities by training and evaluating vision-only and vision-tactile policies on a range of insertion scenarios with varying levels of difficulty. Our findings reveal that tactile feedback significantly enhances task performance in precision-critical conditions, such as high-tolerance fits and challenging orientations, where visual input alone proves insufficient. Conversely, vision excels in initial alignment, highlighting the complementary strengths of both modalities. By systematically investigating the influence of the two modalities, we highlight the roles of each modality and their impact on designing effective reinforcement learning policies, contributing to the development of better multi-modal decision-making systems.

Keywords: dexterous insertion, tactile sensing, reinforcement learning

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the project Aristotle (ANR-21-FAI1-0009-01) and hessian.AI.

1 Introduction

Humans effortlessly solve daily insertion tasks, like plugging a plug into a socket, by leveraging vision, touch, and compliance. However, in robotics, such tasks remain challenging due to the inherent uncertainty in unstructured environments. Robotic perception systems often lack the precision to accurately determine the plug's and socket's positions, making success difficult given the small error margins typical of insertion tasks. While advances in tactile sensing technologies [1, 28, 20, 11] enable the incorporation of vision and touch to robot systems, effectively integrating these modalities into perception and control systems remains an open research problem.

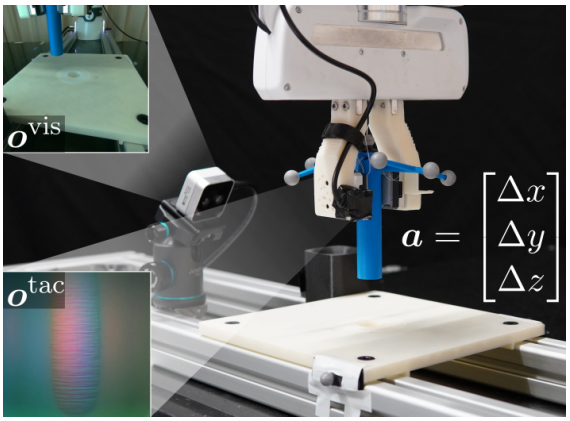

Figure 1: Dexterous insertion platform using vision and touch. o^{vis} shows the visual observation and o^{tac} the tactile observation.

Robotic dexterous manipulation has been an active field of research for a long time. Many approaches rely on vision as their sole sensing modality [5, 25, 22]. While these works highlight the importance of vision for robotic manipulation tasks, vision alone is often insufficient, as it provides no information on forces acting at the contact points and suffers from occlusion by the robot's end-effector.

Prior work has utilized tactile sensing for dexterous manipulation [19, 16, 17, 9]. Many recent approaches rely on vision-based tactile sensors [16, 17, 9], as they provide a rich, high-resolution representation of the contact patch while also allowing force estimation [10]. A challenge is that tactile data is often difficult to interpret, making designing control loops with tactile sensors non-trivial. To address this, many prior works choose a data-driven approach to extract interpretable features from tactile data [17] or directly map tactile data to actions via end-to-end learning [19].

Both tactile and vision sensors provide limited information about the environment. While tactile sensors can only provide information about objects the agent is in contact with, vision sensors suffer from occlusion and cannot measure contact forces. By combining both sources of information, one sensor can compensate for the other's shortcomings. The combination of vision and touch has been explored for 6D-pose estimation [8], grasping [15, 7], object manipulation, and insertion [21, 27]. Like our work, Lee et al. [21] also learns peg insertion with reinforcement learning from vision and touch. However, they first learn a representation of the multi-modal input with a Variational Auto Encoder (VAE) [18] and then optimize the policy on the representation. While such a scheme can simplify the policy learning algorithm, it also means that the representation is not task-specific and potentially contains unnecessary information.

In this work, we explore the role of tactile sensing in improving robotic insertion tasks and analyze the interplay between visual and tactile inputs during complex manipulations. We adapt the platform proposed in [24] to include visual observations from an external camera and learn a policy using Dreamer-v3 [14]. Dreamer-v3 is an actor-critic algorithm that learns latent representations of sensory inputs to inform its policy. Each input modality — visual and tactile — is concatenated and encoded into an embedding, which is fed into a recurrent state-space model that captures temporal dynamics essential for continuous control. By jointly learning input representations and the policy, Dreamer-v3 effectively aligns the learning process of sensory inputs with action optimization. The latent representation is learned via a VAE, enabling the model to generalize across complex manipulative tasks and adapt to variations in input signals. This approach allows us to investigate how visual and tactile cues can complement each other in the context of precise and adaptive control.

2 Autonomously Learning Visual-Tactile Peg Insertion in the Real World

Preliminaries. We consider solving a challenging insertion task from visual and tactile feedback, which will be formulated as an infinite horizon finite-time partially observable Markov decision process (POMDP). We denote observations by $\boldsymbol{o} \in \mathcal{O}$, the (hidden) state by $\boldsymbol{h} \in \mathcal{H}$, actions by $\boldsymbol{a} \in \mathcal{A}$, the reward by $r \in \mathbb{R}$, and the discount factor by $\gamma \in [0, 1]$. To solve the decision-making problem, we leverage *Dreamer* [13], which learns a policy $\pi(\boldsymbol{a}_t|\boldsymbol{h}_t)$ conditioned on the latent state representation. To deal with the partial observability, Dreamer learns a recurrent state-space model $p(\boldsymbol{h}_t|\boldsymbol{h}_{t-1}, \boldsymbol{a}_{t-1}, \boldsymbol{o}_t)$ [12]. To see the direct influence of the different input modalities on the action predictions, we rewrite the conditional dependence of the policy as $\pi(\boldsymbol{a}_t|\boldsymbol{h}_{t-1}, \boldsymbol{a}_{t-1}, \boldsymbol{o}_t)$.

Hardware setup. Our setup is inspired by the work of [24] and extends the authors' setup. Figure 1 show our full task setup, which requires inserting a peg into a hole in the base plate. The base plate is modular and the holes can be swapped for different tolerances t with respect to the peg. The peg is being held by a parallel gripper equipped with Gelsight sensors [28] at the finger tips. We add an external Intel RealSense camera [4] for an external view of the scene. The observations are the downsampled RGB images of the scene camera $\boldsymbol{o}^{\text{vis}} \in \mathbb{R}^{64\times64\times3}$ and the tactile sensor

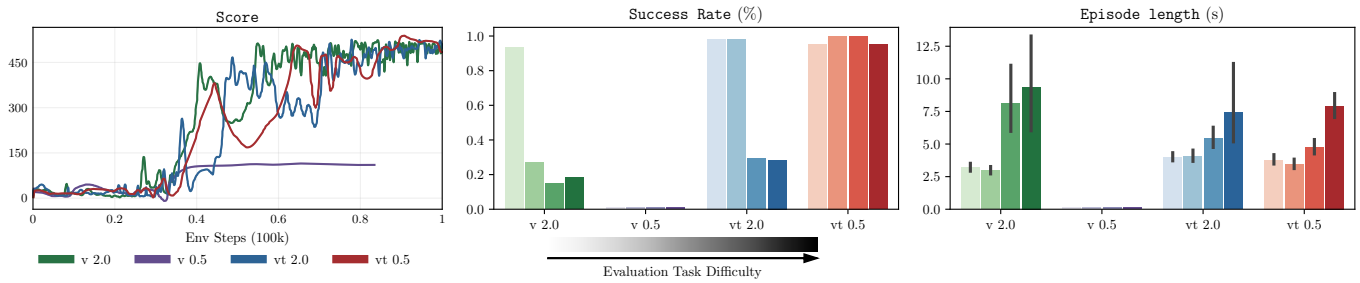

Figure 2: (left) Performances of the visual-only (v) and the visual-tactile (vt) policies trained on $2\,\mathrm{mm}$ tolerance insertion hole and $0.5\,\mathrm{mm}$ respectively. (middle) Final performance is reported as the success rate (middle) and the mean rollout length of successful insertions (right) over 20 trials across four varying insertion holes, with increasing evaluation task difficulty from left to right due to decreasing tolerances and increasing hole angles $(\mathrm{tol}, \alpha) \in \{(2, 0), (1, 4), (0.5, 0), (0.5, 4)\}[\mathrm{mm},\ ^\circ]$.

$\boldsymbol{o}^{\mathrm{tac}} \in \mathbb{R}^{64\times64\times3}$ at $25\,\mathrm{Hz}$. The policy outputs the relative new positions $\boldsymbol{a} = [\Delta x, \Delta y, \Delta z]^\intercal$ of the end-effector at $10\,\mathrm{Hz}$. The target positions are passed to the *franky* control library [2], which integrates *ruckig* [6] for smooth motion planning and leverages Franka's internal Cartesian impedance controller for execution. To ensure safe exploration, the workspace $\mathcal{W}$ is restricted to the area around the hole. The peg's pose is continuously tracked using OptiTrack [3], but this data is used exclusively for evaluation and not provided as input to the policy. The reward function

$$r = \underbrace{5 \cdot (0.1 - |\boldsymbol{p}_g - \boldsymbol{p}_e|)}_{r_d\text{: proximity to the goal}} \quad \underbrace{+500 \cdot \mathbb{1}_{\{\mathcal{G}\}}(\boldsymbol{p}_g)}_{r_g\text{: terminal reward upon reaching goal}} \quad \underbrace{-50 \cdot \mathbb{1}_{\{\mathcal{R}\}}(\boldsymbol{p}_r)}_{r_r\text{: peg rotational penalty}} \quad \underbrace{+10^{-3} \cdot |\boldsymbol{a}|}_{r_a\text{: action penalty}}$$

is comprised of four components. (i) r_d proximity to the goal, (ii) r_g a terminal reward upon reaching the goal $\mathcal{G} = \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{p}_g - \mathbf{x}| < (5, 5, 5)[\mathrm{mm}]\}$, (iii) r_r a penalty to prohibit large rotational deviations $\mathcal{R} = \{\boldsymbol{\theta}\,|\,|\boldsymbol{\theta} - \boldsymbol{p}_r| > 10^\circ\}$ of the end-effector rotation $\boldsymbol{p}_r$, and (iv) r_a an action penalty to encourage smooth motions. For more details of the base-setup, we refer to the previous work [24].

3 Experimental Results

We present two studies, which are designed to showcase the importance of touch for dexterous insertion tasks. The first involves training vision-based policies with and without the added sense of touch. The second analyzes the relative importance of each modality for action prediction.

The impact of tactile feedback during learning. We train four policies – two based solely on vision and the other two combining vision with tactile feedback – for insertion tasks involving different tolerances t between the insertion hole and the peg. The first set of experiments is performed for $\mathrm{tol} = 2\,\mathrm{mm}$ over three seeds, where minimal jamming of the peg is expected. The second configuration uses a hole with $\mathrm{tol} = 0.5\,\mathrm{mm}$, which is anticipated to result in considerable jamming during insertion and increased interaction forces between the gripper and the peg. We report the training curves and the final performance of the respective models solving separate evaluation tasks in Figure 2. In the training phase, all policies, except the vision-only policy trained on the $0.5\,\mathrm{mm}$ tolerance, successfully ($r > 500$) completed the peg insertion after 100k environment steps. During the evaluation, the $2\,\mathrm{mm}$ and $1\,\mathrm{mm}$ configurations could be solved reliably by all policies within 2.5 to 4 s. In the more challenging $0.5\,\mathrm{mm}$ tolerance tasks, both the vision-only and vision-tactile policies exhibit a significant drop in success rates, except for the vision-tactile policy trained on the $0.5\,\mathrm{mm}$ configuration, which maintains strong performance. The increased episode length for successful insertions is directly related to more frequent jamming and tilting, highlighting the `vt 0.5` policy's ability to effectively handle these challenges.

Explaining the role of touch during an episode. To evaluate the influence of the two different observation modalities, we leverage Shapley value [26] analysis. Shapley values provide a systematic way to assess the influence of different input components of $\boldsymbol{x}$ on the output of a model $f(\boldsymbol{x})$. In Dreamer, the input $\boldsymbol{x}$ for the action prediction $\boldsymbol{a}_t = f(\boldsymbol{x})$ consists of the previous hidden state $\boldsymbol{h}_{t-1}$, the visual $\boldsymbol{o}^{\mathrm{vis}}_{t-1}$ and tactile information $\boldsymbol{o}^{\mathrm{tac}}_{t-1}$, and the previous action $\boldsymbol{a}_{t-1}$. To evaluate the individual impact of each feature, we mask them by substituting their values with a default placeholder. This results in a total of $2^N, N = 4$ distinct features, each identified by an integer label, allowing us to systematically analyze their respective contributions to the model's performance. Formally, we denote $\mathcal{F}$ as the set of the N different features, then the Shapley value of feature $i \in \mathcal{F}$ is the attribution that the feature has on the outcome of the model

$$\phi_i(\boldsymbol{x}) = |\mathcal{F}|^{-1} \sum_{\mathcal{S}\subset\mathcal{F}\setminus\{i\}} \binom{|\mathcal{F}|-1}{|\mathcal{S}|} \left(f(\boldsymbol{x}_{\mathcal{S}\cup\{i\}} - f(\boldsymbol{x}))\right). \tag{1}$$

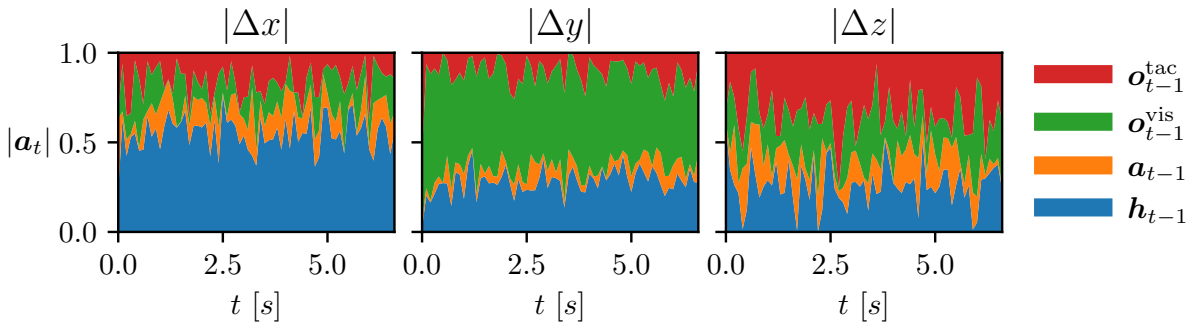

Figure 3: Shapley value analysis over a single exemplary trajectory of the `vt=0.5` experiment on the most challenging hole ($\text{tol} = 0.5\,\text{mm}, \alpha = 4°$). We report the individual contributions of the four input modalities on the action prediction in x-, y-, and z-direction.

Calculating Shapley values often becomes intractable due to the exponential scaling with 2^N [23]. However, in our case, with only 16 possibilities, we compute the Shapley values of the individual contributions exactly shown in Figure 3. Along the x-axis, the model predominantly relies on its hidden state, with minimal influence from vision and tactile inputs. This may be attributed to the camera's orientation, which is primarily aligned along the x-direction, making depth perception more challenging. In contrast, along the y-axis, visual observations contribute most to action prediction, suggesting that visual feedback plays a key role in aligning the peg with the insertion hole. Finally, along the z-axis, tactile feedback has the greatest influence on action prediction, likely because vertical movements of the end-effector lead to increased tilts and jams, resulting in the highest impact on the contact forces applied between the peg and the hole.

4 Conclusion

In this work, we presented a comprehensive analysis of the interplay between visual and tactile feedback for robotic insertion tasks. Our results demonstrate that incorporating tactile sensing can significantly improve success rates on challenging insertions with tight tolerances (0.5mm) and varied orientations (up to 4° tilted insertion hole) that vision alone struggles to solve. Through Shapley value analysis, we revealed that different input modalities dominate action prediction along different axes – vision primarily guides alignment in the camera plane, while tactile feedback is crucial for controlling vertical movements where contact forces are highest. However, it is essential to acknowledge that this analysis is confined to single trajectories and could benefit from additional explainability studies, such as gradient-based analyses. Furthermore, categorizing trajectories into distinct insertion stages may enhance the study's depth and facilitate statistical evaluation.

Our results suggest that the information the agent extracts from vision is effectively complemented by tactile sensor data. In particular, when the hole tolerances are smaller and vision alone is insufficient for robust insertion, the agent learns autonomously to rely more heavily on tactile sensing. Hence, our work demonstrates that RL is a powerful tool for learning robust policies from multi-modal sensor inputs.

As the vision-tactile policies manage to insert the peg successfully even for tight tolerances and varying inclination angles, increasing the task difficulty to different objects is a logical next step. In particular, screw and light-bulb insertion is a challenging task with many real-world applications, which might significantly benefit from tactile feedback. A crucial future direction is optimizing the representation learning part during policy optimization. Leveraging pre-trained vision and visual-tactile models could be beneficial to reduce the number of real-world interactions required.

References

[1] Robotics & Prosthetics | SynTouch, Jan. 2023. URL https://syntouchinc.com/robotics. [Online; accessed 4. Oct. 2024].

[2] franky, 2024. URL https://github.com/TimSchneider42/franky.

[3] Motion Capture Systems, 2024. URL optitrack.com.

[4] RealSense d405, 2024. URL intelrealsense.com/depth-camera-d405.

[5] M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba. Learning dexterous in-hand manipulation. *International Journal of Robotics Research*, 39(1), 2020.

[6] L. Berscheid and T. Kröger. Jerk-limited real-time trajectory generation with arbitrary target states. *arXiv:2105.04830*, 2021.

[7] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[8] S. Dikhale, K. Patel, D. Dhingra, I. Naramura, A. Hayashi, S. Iba, and N. Jamali. Visuotactile 6d pose estimation of an in-hand object using vision and tactile sensor data. *IEEE Robotics and Automation Letters*, 7(2):2148–2155, 2022.

[9] 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 *International Conference on Robotics and Automation*, 2021.

[10] N. Funk, P. O. M"uller, B. Belousov, A. Savchenko, R. Findeisen, and J. Peters. High-resolution pixelwise contact area and normal force estimation for the gelsight mini visuotactile sensor using neural networks. In *Embracing Contacts - Workshop at ICRA 2023*, 2023.

[11] 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*, 40:3812–3832, 2024.

[12] D. Hafner, T. P. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

[13] D. Hafner, T. P. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[14] D. Hafner, J. Pasukonis, J. Ba, and T. P. Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[15] Y. Han, K. Yu, R. Batra, N. Boyd, C. Mehta, T. Zhao, Y. She, S. Hutchinson, and Y. Zhao. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. *IEEE/ASME Transactions on Mechatronics*, 2024.

[16] F. R. Hogan, J. Ballester, S. Dong, and A. Rodriguez. Tactile Dexterity: Manipulation Primitives with Tactile Feedback. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2020–31. IEEE. doi: 10.1109/ICRA40945.2020.9196976.

[17] T. Kelestemur, R. Platt, and T. Padir. Tactile Pose Estimation and Policy Learning for Unknown Object Manipulation. *arXiv*, Mar 2022. doi: 10.48550/arXiv.2203.10685.

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[19] L. Lach, N. Funk, R. Haschke, S. Lemaignan, H. J. Ritter, J. Peters, and G. Chalvatzaki. Placing by touching: An empirical study on the importance of tactile sensing for precise object placing. In *International Conference on Intelligent Robots and Systems*, 2023.

[20] M. Lambeta, P. Chou, S. Tian, B. H. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[21] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3): 582–596, 2020.

[22] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *International Journal of Robotics Research*, 37(4-5):421–436, 2018.

[23] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[24] D. Palenicek, T. Gruner, T. Schneider, A. Böhm, J. Lenz, I. Pfenning, E. Krämer, and J. Peters. Learning tactile insertion in the real world. *40th Anniversary of the IEEE International Conference on Robotics and Automation*, 2024.

[25] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In D. Kragic, A. Bicchi, and A. D. Luca, editors, *International Conference on Robotics and Automation*, pages 3406–3413. IEEE, 2016.

[26] L. S. Shapley. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

[27] 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 *8th Annual Conference on Robot Learning*, 2024.

[28] W. Yuan, S. Dong, and E. H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.